DATA SHARING AND DATA REUSE: AN INVESTIGATION OF DESCRIPTIVE INFORMATION FACILITATORS AND INHIBITORS

Angela Patricia Murillo

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2016

Approved by:

Jane Greenberg

Mohammed Hossein Jarrahi

Robert Losee

William Michener

Reagan Moore

Arcot Rajasekar

ABSTRACT

Angela Patricia Murillo: Data Sharing And Data Reuse: An Investigation Of Descriptive
Information Facilitators And Inhibitors
(Under the direction of Jane Greenberg)

This dissertation examines how descriptive information inhibits or facilitates data sharing and reuse. DataONE serves as the test environment. The objective is to identify descriptive information made discoverable through DataONE and subsequently determine what of this descriptive information is helpful for scientists to determine data reusability. This study uses a mixed method approach, which includes a data profiling assessment in the form of a quantitative and qualitative content analysis and a quasi-experiment think-aloud. A quantitative and qualitative content analysis was conducted on a stratified sample of data extracted from DataONE to examine types of descriptive information made available through the shared data. Participants searched a quasi-experiment interface and thought-aloud about what information inhibited or facilitated them to determine data reusability. Additionally, participants completed a post result usefulness survey, post search rank order survey, and a post search factors survey. The quantitative and qualitative content analysis shows that the shared data contains 30 unique pieces of descriptive information found in the records. The quasi-experiment think-aloud indicates that scientists found pieces of descriptive information particularly useful for their ability to determine data reusability. These include: (a) the data description, (b) the attribute table, and (c) the research methods. In conclusion, metadata schema, member node standards, and community standards, impact what types of descriptive information are provided through the shared data. Attribute and unit lists, research methods information, and succinctly written abstracts facilitate data reuse. However long abstracts and having the same information in multiple places, and the exclusion of data descriptions inhibit data reuse. The findings and recommendations assist funding agencies and scientific organizations in understanding the current state of data being shared and prioritizing how to meet the needs of scientists regarding data reuse. This dissertation provides guidance to developers of current and future data sharing environments and infrastructures, research data management and scientific communities, scientific data managers, creators of data management plans, and funding agencies; and has implications beyond DataONE.

Dedicated To My Parents: Eddie William Murillo And Bertha Murillo

ACKNOWLEDGEMENTS

I am truly thankful and grateful to my committee: (Drs.) Jane Greenberg, Mohammed Hossein Jarrahi, Robert Losee, William Michener, Reagan Moore, and Acrot Rajasekar. I deeply appreciate all of your time, thoughts, ideas, and encouragement. Thank you to the rest of the SILS professors particularly Dr. Diane Kelly and Dr. Barbara Wildemuth for your helpful advice and encouragement along the way; thank you for your generosity, time, and guidance.

Thank you to the community for the many research opportunities and researcher communities that I have had the opportunity to be a part of. Thank you to the SILS community and to the many opportunities for research, teaching, and service you've provided me. Thank you to DataONE for all of the research opportunities, the opportunity to explore your data, and for your generous funding. Thank you to the to the Metadata Research Center, CODATA, the National Consortium for Data Science, and the Earth Science Information Partners for all of your generous research and funding opportunities. A special thank you to the DigCCurr Fellowship for funding so many years of my doctorate studies, thank you for the wonderful opportunity to work with you. And lastly, thank you UNC-Writing Center!

Thank you to all the many great friends and colleagues that have travelled this long and wonderful path with me particularly, my doctoral cohort friends. I miss seeing all of your smiling encouraging faces and I'm so glad we are still able to connect from time to time. It has been so wonderful to have you by my side during this long journey and to see where this journey has taken all of you too. Hugs and love ☺

Thank you to my writing buddies especially Rachael Clemens, Ericka Patillo, and Leslie Thomson, I will miss coffee shopping with you all. Thank you Dr. Ashlee Edwards, Sami Kaplan, Debbie Maron, Sarah Ramdeen, and Jewel Ward, I definitely couldn't have finished this without each of your support. To those I've forgotten to acknowledge, sorry and thanks to you too! To triangle area coffee shops and libraries, I definitely couldn't have done this without you. To all past friends and colleagues, I wish you all well too!

To my outside of doctoral studies friends: Greg, Trish, Chris, Megan, Katelyn, Rachel, and Mindy, thank you all who listened to me endlessly talk about my research. Special thanks to Kjersti Kyle and Emily Zaentz; love you both! To my best girl, Xan, I could not have done this without you; thanks for putting up with endless hours of watching me work. To my family, especially my mom and sister, thank you for everything, I love you all always!

Lastly again, thank you to my wonderful committee for helping me through this process. And lastly, to Jane Greenberg, you are an amazing mentor, advisor, and friend and I will forever be in your debt.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANDS	Australian National Data Service
AGU	American Geophysical Union
CEISIN	Center for International Earth Science Information Network
CCSDS	The Consultative Committee for Space Data Systems
CDSN	Collaborative data sharing network
CIB	Collaborative information behavior
CODATA	Committee on Data for Science and Technology
DANS	The Netherlands SURF Foundation and Data Archiving and Networked Services
DataONE	Data Observation Network for Earth
DCC	Digital Curation Centre
DFC	DataNet Federation Consortium
DFG	Germany’s Deutsche Forschungsgemeinschaft
EML	Ecological Metadata Language
ESA	Ecological Society of America
ESIP	Earth Science Information Partners
EU	European Union
FGDC CDSGM	Federal Geographic Data Committee Content Standard for Digital Geospatial metadata
GEON	Geosciences Network
GPD	Genomic and proteomic databases

ICPSR	Inter-University Consortium for Political and Social Research
ISO	International Organization for Standardization
KNB	Knowledge Network for Biocomplexity
LIS	Library and Information Science
LTER	Long Term Ecological Research
OAIS	Reference Model for an Open Archival Information System
ORNL	Oak Ridge National Laboratory
NASA	National Aeronautics and Space Administrator
NSF	National Science Foundation
NIH	National Institutes of Health
SEAD	Sustainable Environment Actionable Data
TerraPop	Terra Populus

CHAPTER I: INTRODUCTION

Data sharing and reuse is a topic of growing significance due to technological developments and changes in scientific practices and processes. A variety of factors have influenced these developments and changes. These factors include the data deluge that describe the immensity of data availability (Bell, Hey, & Szalay, 2009; Hey & Trefethen, 2003) and changes in journal and grant agency policies that provide data sharing guidelines for scientists (National Institutes of Health, 2007; National Science Foundation, 2010b; Stodden, Guo, & Ma, 2013). These factors are also driving and shaping the fourth paradigm, Jim Grey's description of today's data-intensive science (Hey, Tansley, & Tolle, 2009).

As part of this change, the benefits of scientific data sharing and reuse are increasingly being documented. These include the ability to extract additional value from data, enable reproducible research, enable researchers to ask new questions of existing data, and advance the state of science in general (Borgman, 2012). The potential benefits of data sharing have placed pressure on the scientific community and funding agencies to provide infrastructure solutions for data sharing and reuse.

To advance infrastructure developments for data sharing and reuse, the U.S. National Science Foundation supports the investigation of solutions via the Sustainable Digital Data Preservation and Access Network Partners (DataNet) program that launched in 2007. The Data Observation Network for Earth (DataONE), a DataNet, provides cyberinfrastructure for "open, persistent, robust, and secure access to … earth science observational data" (DataONE, 2013d). Scientists participating in DataONE are able to deposit, search, and reuse data available through

DataONE tools; this provides scientists the opportunity to reap the benefits of data sharing and reuse.

For the last five years I have actively participated in DataONE: as an intern during the summer of 2012, as a member of the Preservation and Metadata Working Group (2012-2014), participating in DataONE User Group (DUG) meetings, and now through my dissertation activities. Through my DataONE work and my research activities, I have made the following observations:

1. Advancing cyberinfrastructure is crucial due to changes in the sciences,
2. Examining current cyberinfrastructure advancement is required to ensure success, and
3. Developing a better understanding of data sharing and reuse will further ensure that cyberinfrastructure development succeeds.

These observations highlight an important research and literature gap in current research; that is the examination of data sharing and reuse while considering the impact of new infrastructures.

The goal of this dissertation research is to address this research gap through a thorough examination of data sharing and reuse within the newly developed cyberinfrastructure of DataONE. The overarching goal and research question is: "how does descriptive information influence scientists' ability to share and reuse data within DataONE".

This dissertation is organized as follows: Chapter 2 provides a review of the relevant literature; Chapter 3 examines research methods and theoretical frameworks that informed this dissertation; Chapter 4 describes the research preparations and rationale for this study; Chapter 5 describes the research questions, context, and terminology; Chapter 6 describes the research methods and study procedures; Chapter 7 describes the results of this study; Chapter 8 provides a discussion of the results; and lastly, Chapter 9 contains a conclusions, study limitations, and potential future work.

CHAPTER II: LITERATURE REVIEW

This literature review provides the necessary background for understanding the research problem described in Chapter 1. This chapter includes a review of literature of (a) the DataNet and DataONE, (b) data sharing and reuse in the sciences, (c) scientific data management, and (d) selected infrastructure and interoperability factors. This chapter provides an understanding of the current research conducted on these topics, as well as research gaps.

DataNet and DataONE

Jim Gray's notion of the Fourth Paradigm predicted and identified the tremendous changes in the scientific process over the last decade (Hey et al., 2009). Many researchers have described and examined these changes in the scientific process, specifically investigating how new cyberinfrastructures are affecting and changing scientific research (Berman, Fox, & Hey, 2003; Borgman, 2000a; Edwards, 2011). Changes in how scientific data are created, managed, and analyzed introduce new questions regarding how to organize data and ensure long-term use. Furthermore, these questions form a growing research area for scientists, academics, and government. Granting agencies such as the United States National Science Foundation (NSF) and National Institutes of Health (NIH) have initiated programs to advance the development of long-term sustainable data infrastructures, interoperable data preservation and access services, and cyberinfrastructure capabilities (National Science Foundation, 2006), in order to support data utilization within this changing environment. In 2007, the NSF launched the Sustainable Digital Data Preservation and Access Network Partners (DataNet), commonly referred to as the DataNet program, via request for proposals, 07-601. The DataNet program seeks to address the challenges in science by "creating a set of exemplar national and global data research infrastructure organizations that provide unique opportunities to communities of research to advance science and/or engineering research and learning" (National Science Foundation, 2006, para. 1). As both a geoscientist and an information scientist, I am keenly interested in assisting scientists in their efforts; and my personal experience inspired and shaped my dissertation research. The work presented in this dissertation asks questions about whether the DataNets assist scientists the way they were intended to.

This section has two purposes: (a) to provide an overview of the DataNet program and (b) to extensively explore DataONE. DataONE is the chief focus in this chapter since it is the test environment for this dissertation. First, an overview of the DataNet program and a brief description of all projects funded by the DataNet program are provided. Next, DataONE cyberinfrastructure is described including the: (a) Coordinating Nodes, (b) Member Nodes, (c) Investigator Toolkit, and (d) Education and Outreach Program. Lastly, research studies specific to DataONE are examined including: (a) general investigations, (b) research and development, (c) data sharing investigations, and (d) frameworks, models, and theoretical studies.

The DataNet

In 2007, the National Science Foundation Office of Cyberinfrastructure announced a request for proposals for the Sustainable Digital Data Preservation and Access Network Partners (DataNet). The purpose of the program is to address the challenges of: "how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams" and to "address that challenge by creating a set of exemplar national and global data research infrastructure organizations" (National Science

Foundation, 2006, para. 1) . These organizations are tasked with providing opportunities for the advancement of science and/or engineering research and education.

These new organizations called the "DataNet Partners" were envisioned to combine library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise. Each area of expertise is needed to accomplish the goals of the DataNet partners. The goals of the DataNet partners are to:

- Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;
- Continuously anticipate and adapt to changes in technologies and in user needs and expectations;
- Engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and
- Serve as component elements of an interoperable data preservation and access network (National Science Foundation, 2006, para. 2).

The DataNet Partners were asked to provide: (a) a vision and rationale that meets critical data needs, (b) an organizational structure that provides expertise and cyberinfrastructure capabilities, (c) activities to provide for the full data management lifecycle, and (d) feasible and viable models for long-term technical and economic sustainability. Furthermore, the call for proposals stated that the program was not intended to support narrowly-defined, discipline-specific repositories (National Science Foundation, 2006, para. 4).

In the initial program solicitation, the estimated number of awards was five in total for two review cycles, one for fiscal year 2008 and the other for fiscal year 2009. The limited amount of direct plus indirect cost was to be a total of $20,000,000 for up to 5 years, with a potentially renewal for another 5 years based on performance and funds. The proposals were limited to academic institutions and non-profit, non-academic organizations associated with educational or research activities in the United States (National Science Foundation, 2006). A second call was originally to be released in 2009, however it was delayed for several years.

Funded Projects

To date, the National Science Foundation has funded five DataNet Partners. For the first-round, two DataNet Partners were funded: Data Observation Network for Earth (DataONE)[1] and Data Conservancy[2]. During the second round of funding, three additional projects were added: the Sustainable Environment Actionable Data (SEAD)[3], the DataNet Federation Consortium (DFC)[4], and Terra Populus (TerraPop)[5]. Each of these projects is described in the following section.

Data Observation Network for Earth (DataONE)

DataONE aims to ensure the “preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data” (“What is DataONE?” DataONE, 2013m). DataONE’s mission is to “enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it” and the vision is to be “commonly used by researchers, educators, and the public to better understand and conserve life on earth and the environment that sustains it” (“What is DataONE?” DataONE, 2013m).

[1] http://www.dataone.org/

[2] http://dataconservancy.org/

[3] http://sead-data.net/

[4] http://datafed.org/

[5] http://www.terrapop.org/

DataONE provides cyberinfrastructure through three primary elements: Coordinating Nodes, Member Nodes, and the Investigator Toolkit. DataONE has an extensive education and outreach program, and is headquartered at the University of New Mexico. DataONE goals, cyberinfrastructure, outreach and educational programs, and research studies specific to DataONE are further described in section 2.1.2.

Data Conservancy

Data Conservancy aims to create solutions for digital research data collections, curation, and preservation challenges. The Data Conservancy is a first-round DataNet Partner, alongside DataONE. It is headquartered at the Sheridan Libraries at Johns Hopkins University. The Data Conservancy builds tools and services with the hope that these tools will incentivize scientists and researchers to participate in data curation efforts. The Data Conversancy has made advances in four areas: (a) a research program that examines research practices across multiple disciplines to understand data curation tools and services needed to support the research community, (b) an infrastructure development program that develops cyberinfrastructure based on data management and curation services, (c) a data curation education and professional development program, and (d) development of sustainable models for long-term data curation. The Data Conservancy addresses their goals through three interdependent teams: (a) the infrastructure research and development team, (b) the broader impacts team, and (c) the sustainability team (The Sheridan Libraries at Johns Hopkins University, 2013).

The Data Conservancy promotes data preservation and reuse across disciplines with tools and services. Users can set up their own instances of Data Conservancy and use the software for advanced queries of interoperable data objects. The Data Conservancy attempts to be discipline agnostic so that users can search across disciplines and have access to all the data they need. The partners include Johns Hopkins Sheridan Libraries, National Snow and Ice Data Center, Marine Biological Laboratory, Cornell University, Earth Science Information Partners (ESIP), University of Illinois Center for Informatics Research in Science and Scholarship, National Center for Atmospheric Research, Sustainable Environment Actionable Data (SEAD), Tessella, and UCLA's Center for Embedded Networked Sensing. The partners work together to achieve the overall goal of the Data Conservancy, which is to assist with data curation efforts through the building of tools and services (The Sheridan Libraries at Johns Hopkins University, 2013).

Sustainable Environment Actionable Data (SEAD)

The primary goal of SEAD, the Sustainable Environment Actionable Data project is to enable management of heterogeneous data for long-term community use at a low cost. SEAD is a second-round DataNet Partner headquartered at the University of Michigan – School of Information. Its partner institutions include the University of Illinois at Urbana-Champaign and Indiana University. SEAD develops lightweight data services designed to meet sustainability projects. Key services include the ability to build a data collection as you work, publish and preserve data, and find future collaborators. The software is open source and made up of three components: (a) an Active Content Repository which provides secure project spaces where data can be collected, shared, annotated, analyzed, used, and published, (b) a Community Research Profile and Analytic Service which tracks information about real-world entities (e.g. people, projects, centers), and (c) a Virtual Archives which provides the discovery and preservation layer of the SEAD services suite (SEAD, 2013).

The DataNet Federation Consortium (DFC)

The DataNet Federation Consortium (DFC) supports science and engineering collaborations by providing a policy-driven national data management infrastructure through a community-based approach. The DataNet Federation Consortium is a second-round DataNet Partner that is headquartered at the University of North Carolina – Chapel Hill. The goals are to create a national data infrastructure to enable collaborative research on shared data collections through managing the collection lifecycle. This collection lifecycle includes project collection, support, collection sharing with other projects, collection publications, processing pipelines, and preservation. This national data infrastructure is organized through a federation of independent data grids. This infrastructure uses a bottom-up federation approach to integrate existing data management systems through use of the iRODS policy-based data grid technology. The shared data collection is created from datasets located at remote sites and the collection is managed by policies that control data ingest, data access, collection properties, and administrative tasks (DataNet Federation Consortium, 2014).

The DFC is organized through six communities of practice and each community is responsible for specific deliverables in their defined areas. The communities include: (a) Science and Engineering, (b) Facilities and Operations, (c) Technology and Research, (d) Policies and Standards, (e) Education and Outreach, and (f) Sustainability. These communities define the governance policies that are needed to build national infrastructure. Some of these policies include: policies for collaboration and reuse, metadata extraction, automating data management, retention, disposition, analysis, workflow, provenance, and sustainability (DataNet Federation Consortium, 2014).

Terra Populus (TerraPop)

The TerraPop is developing infrastructure to make it easier for researchers to use data describing people alongside other data. These data include: population censuses and surveys, land cover information from remote sensing, climate records from weather stations, and statistical land use records from federal agencies. This infrastructure permits data interoperability across time, space, and scientific domains, which allows for the study of interactions between population and environment. Terra Populus is a second-round DataNet Partner based primarily out of University of Minnesota. The partners include the Minnesota Population Center, the Institute on the Environment – University of Minnesota, the University of Minnesota Libraries, CEISIN (Center for International Earth Science Information Network) at Columbia University, and the Inter-University Consortium for Political and Social Research (ICPSR) (University of Minnesota, 2013).

The TerraPop's main goals are to: (a) collect, preserve, integrate and describe datasets that measure changes in the world's population and environment over the past two centuries, (b) develop tools and procedures to manage and disseminate the data collections, (c) carry out education and outreach to engage the scientific community and the public, and (d) establish an organizational structure to ensure the long-term sustainability of the project (University of Minnesota, 2013).

Data Observation Network for Earth (DataONE)

Overview of DataONE

DataONE is a leading exemplar of an international science initiative and a first-round DataNet Program. DataONE provides a "distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered earth observational data" ("What is DataONE?" DataONE, 2013m). DataONE provides this framework through three primary cyberinfrastructure elements that support the user community. These elements are: (a)

Coordinating Nodes, (b) Member Nodes, and (c) the Investigator Toolkit. DataONE also has an extensive education and outreach program (DataONE, 2013d).

DataONE Cyberinfrastructure and Organization

Member Nodes are either existing or new data repositories that install the DataONE Member Node application programming interfaces (APIs) and expose their data or metadata to DataONE. Coordinating Nodes are responsible for cataloging content, managing replication of content, and providing search and discovery mechanisms. Lastly, the Investigator Toolkit is a "modular set of software and plug-ins that enables interactions with the DataONE infrastructure through commonly used analysis and data management tools" (Michener, Allard, et al., 2011, p. 6). DataONE also has an education and outreach program. Each element enables important functions of DataONE cyberinfrastructure by providing support to scientists (DataONE, 2013d).

Figure 1 provides an overview of DataONE cyberinfrastructure, showing the locations of the Member Nodes, Coordinating Nodes, and provides some examples of items in the Investigator Toolkit.

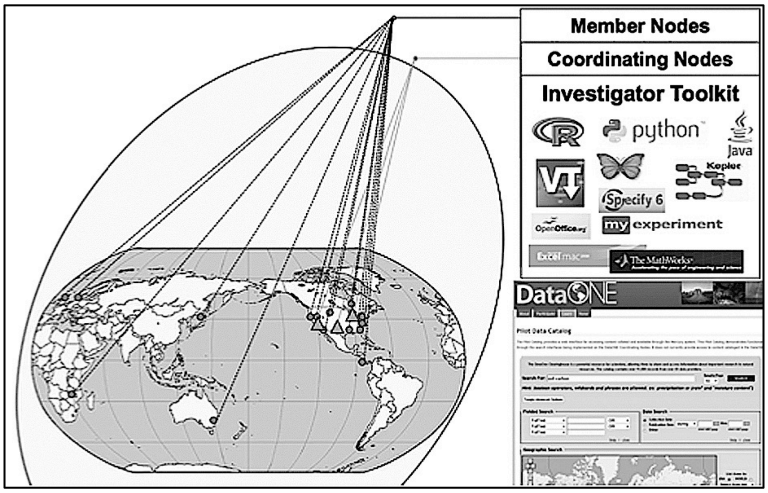


Figure 1. DataONE Cyberinfrastructure (Michener, Vieglais, et al., 2011).

Coordinating Nodes

Coordinating Nodes maintain a complete catalog of all data and provide core DataONE services, including search and discovery. DataONE Core Cyberinfrastructure Team members who are responsible for the technical infrastructure of DataONE manage the Coordinating Nodes. Currently, there are three coordinating nodes; these are located at the University of New Mexico, the University of California at Santa Barbara, and the University of Tennessee in collaboration with the Oak Ridge National Laboratory (ORNL). Each of these three Coordinating Nodes provides equivalent services to the DataONE users and Member Nodes. Each Coordinating Node has provisioned several hundred terabytes of disk space, which assist with data replication. It is very likely that more coordinating nodes will need to be added as DataONE grows ("Coordinating Nodes" DataONE, 2013c).

Coordinating Nodes provide services include harvesting metadata and network-wide indexing and replication of Member Node holdings. The Coordinating Nodes make it easier for scientists to discover data where they reside and also enable data repositories to make their data and services more broadly available. The Coordinating Nodes also determine which Member Nodes hold data when users are using tools in the Investigator Toolkit. Additionally the Coordinating Nodes manage user identities ("Coordinating Nodes" DataONE, 2013c).

Member Nodes

Member Nodes expose their data and metadata through a common set of interfaces and services, the Member Node service interface. Member Nodes are selected based on "evaluation criteria that include factors such as diversity of data holdings, readiness to participate, community leadership, and resources available"(Michener, Vieglais, et al., 2011, p. 6). Member Nodes are a distributed network of data centers, science networks, and organizations. Current

Member Nodes include: the Knowledge Network for Biocomplexity (KNB), the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), the South Africa National Parks (SanParks), the Ecological Society of America (ESA) Data Registry, the USGS Core Sciences Clearinghouse, the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), the University of California Curation Center (UC3) Merritt Repository, the Long Term Ecological Research Network (LTER), the Cornell Lab of Ornithology Avian Knowledge Network, ONEShare, the Taiwan Forestry Research Institute (TFRI), the USA National Phenology Network, the Sustainable Environment Actionable Data (SEAD) Virtual Archive, Dryad, the Earth Data Analysis Center (EDAC), and the Gulf of Alaska Data Portal (DataONE, 2013a, 2013m). Feedback from current and potential Member Node operators indicated a preference for a staged implementation of the DataONE services ("Member Nodes" DataONE, 2013a; Michener, Allard, et al., 2011). Appendix 1 contains the list of all Member Nodes

Investigator Toolkit

The Investigator Toolkit gives scientists access to customizable tools that support all aspects of the data lifecycle. These tools allow users to find, use, and contribute data into DataONE. Some of these tools were custom written for DataONE and others are existing tools that have been modified to use DataONE API. Tools may have their own interfaces that can be called by DataONE tools. Investigator Toolkit tools include:

- **ONEMercury**: A web-based tool for searching data held by DataONE Member Nodes.
- **DMP*Tool***: A data management planning tool that provides researchers a way to develop practical data management plans consistent with agency requirements.

- **Zotero, Mendeley, and Context Objects in Spans (COinS)**: A bibliographic tool. DataONE includes COinS tags in search results in order to simplify data citation information into bibliographic tools.
- **DataUp**: Assist scientists in creating metadata, checking for best practices, obtaining a unique identifier, and depositing their data into a repository. It is available as either an open source add-in for Microsoft Excel or a web-based application. DataUp links to ONE*Share* so that users can directly deposit their data.
- **ONE*R***: Provides users the ability to access data from DataONE network of repositories and save the data within R.
- **Morpho**: A metadata editor for Ecological Metadata Language (EML). This can be used to submit data and metadata to a Member Node, which runs Metacat, using the Metacat interface in Morpho.
- **Workflow Tools**: DataONE is collaborating with Kepler and VisTrails to provide workflow tools to users.

There are other tools in development to further assistance to scientists using DataONE ("Investigator Toolkit" DataONE, 2013j).

Figure 2 and 3 provide a detailed overview of how DataONE cyberinfrastructure works. Figure 2 gives an overview of how this cyberinfrastructure interacts with the Dryad Member Node. Figure 3 provides a graphical representation of DataONE cyberinfrastructure, including the purposes and objectives of each of the three-cyberinfrastructure elements.

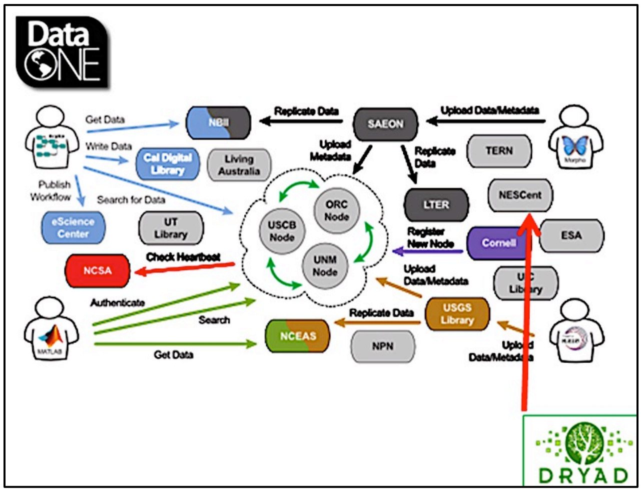


Figure 2. Cyberinfrastructure of DataONE, example of Dryad.

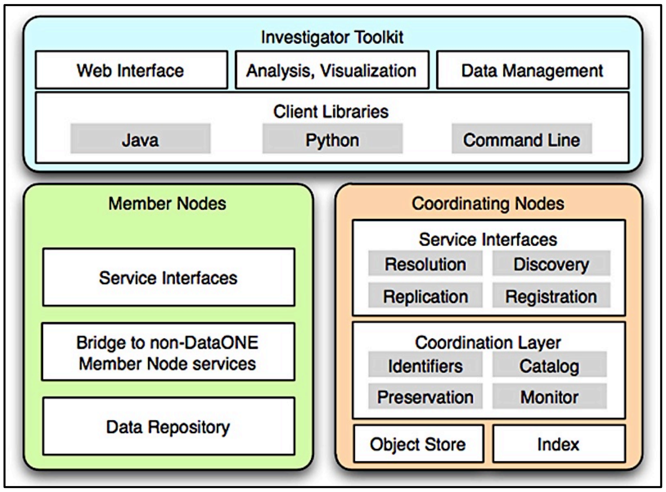


Figure 3. Reference Architecture (DataONE, 2013e)

<u>Education and Outreach Program</u>

Through a working group structure DataONE engages a community of partners that focus on identifying, describing, and implementing DataONE cyberinfrastructure, governance, and sustainability models. Working Groups are composed teams of experts that collaborate to achieve DataONE goals. During the preliminary research for this dissertation Working Groups were:

- Community Education and Engagement

- Data Integration and Semantics
- Data Preservation and Metadata
- Distributed Storage
- Federated Security
- Public Participation in Science and Research
- Scientific Exploration, Visualization, and Analysis
- Scientific Workflows and Provenance
- Sociocultural Issues
- Sustainability and Governance
- Usability and Assessment

Working groups as of July 2015 are: Cyberinfrastructure, Usability and Assessment, Sustainability and Governance, and Community Engagement and Outreach ("Working Groups" DataONE, 2013n).

DataONE Users Group (DUG) is another way that DataONE engages with the user community. DUG members are a community of data authors, users, and stakeholders. The primary function of DUG is to represent the needs and interests of DataONE users by providing guidance to DataONE in achieving its vision and mission. Members of DUG include representatives from Member Nodes, Coordinating Nodes, and research networks, libraries, data centers, scientists, educators, and policy makers. DUG holds an annual meeting to identify the evolving technical challenges and opportunities of DataONE ("DataONE Users Group" DataONE, 2013f).

DataONE educational programs include training activities, education modules, graduate courses, and research experience via DataONE summer internship programs.

- **Training**: DataONE provides training to scientists, students, and data managers via presentations, workshops, and online resources on topics including: management, preservation, analysis, and visualization of data. These include training activities such as Member Node implementation workshops, data management workshops, and metadata workshops. These workshops have been conducted at professional conferences such as the International Digital Curation Conference, Ecological Society of America (ESA) annual conference, Earth Science Information Partners (ESIP) annual conference, and American Geophysical Union (AGU) annual conference ("Training Activities" DataONE, 2013l).
- **Educational Modules**: DataONE produces educational modules in PowerPoint that can be downloaded and included in teaching materials. These modules include: (a) Why Data Management, (b) Data Sharing, (c) Data Management Planning, (d) Data Entry and Manipulation, (e) Data Quality Control and Assurance, (f) Data Protection and Backups, (g) Metadata, (h) How to Write Good Quality Metadata, (i) Data Citation, and (j) Analysis and Workflows. These are available under a CCO-No rights reserved license and can be enhanced and reused for any purpose. DataONE does ask for appropriate citation and attribution. DataONE also provides educational webinars ("Education Modules" DataONE, 2013g).
- **Graduate Education**: DataONE provides graduates courses hosted by the University of New Mexico at the Walter E. Dean Environmental Information Management Institute. The courses are geared towards Masters and PhD students and professionals with bachelor's degrees in biology, geology, ecology, environmental sciences, environmental engineering, geography, and science librarianship. This institute provides conceptual and

practical hands-on training for managing and preserving data and information. The institutes are three-week summer institutes where participants gain valuable experience on all aspects of the data lifecycle. Courses include Environmental Information Management, Environmental Data Analysis and Visualization, and Spatial Data Management in Environmental Science ("Graduate Courses" DataONE, 2013h; University of Minnesota, 2014).

- **Summer Internship Research Programs**: Since 2009, DataONE has funded and mentored four to eight students through a summer internship program. Student interns are chosen each summer to work on projects related to DataONE. Students are paired with mentors to assist them through the internship program. Interns present their work at the fall DataONE All Hands Meeting. Topics have included metadata registration, evaluation of ontology coverage, provenance models, and data policies. Eligible students include undergraduates, graduate students, and postgraduates who have received their degree in the last five years ("Internships" DataONE, 2013i).

Research Studies Specific to DataONE

Research studies have investigated DataONE's cyberinfrastructure and organization, research and development, data sharing and reuse, and frameworks, models, and theory.

General Investigations

Lee, Zhang, Zimmerman, and Lucia (2009) provided an overview and analysis of the DataONE cyberinfrastructure. This research analyzed DataONE cyberinfrastructure through examining: (a) universal difficulties with data, (b) domain coverage and science drivers, and (c) anticipated impacts of DataONE. The researchers discussed how scientists and engineers do not have a working knowledge of basic data management concepts such as metadata and ontologies and that DataONE infrastructure assist in the ability for scientists and engineers to discover data, share, and reuse data. Additionally, the authors discussed how the DataONE outreach programs promote best practices and educational opportunities at all levels (k-12, undergraduate, graduate, professional and public) (Lee et al., 2009).

In Michener, Allard et al. (2011), the authors described how the participatory design of DataONE enables cyberinfrastructure development. The authors examined how stakeholder communities were identified and influenced the design of DataONE. DataONE uses four approaches to identify and understand community perceptions through (a) a baseline assessment of environmental scientists, (b) creation of personas and user scenarios, (c) usability testing, and (d) engaging scientists. The authors analyzed the information received through the four approaches and generate information needed to assist in developing and implementing DataONE (Michener, Allard, et al., 2011).

Michener, Vieglais et al. (2011) provided an overview of DataONE architecture and community engagement activities and discuss the role of identifiers in DataONE. This research examined five activities that are central to DataONE including: (a) discovery and access, (b) data integration and synthesis, (c) education and training, (d) community building, and (e) data sharing; and described how identifiers are key to ensuring these activities are accomplished successfully. The researchers introduced how EZID facilitates acquisition and management of persistent identifiers for data and other data objects. EZID allows users to create and mange globally unique identifiers for data (University of California - California Digital Library, 2016). The authors described how persistent identifiers provide (a) uniqueness, (b) authority, (c) opacity, (d) immutability, (e) resolvability, and (f) granularity of data. This research suggested that EZID makes it simple for digital object producers to obtain and maintain long-term identifiers for their digital content (Michener, Vieglais, et al., 2011).

In Allard's (2012a) paper, the author explored eScience and eScience librarianship, and used DataONE to demonstrated how cyberinfrastructure can be developed, particularly in reference to librarians. The author suggested that librarians could assist in particular demands of eScience including: information across disparate vocabularies, heterogeneous information artifacts, and diverse paradigms. Furthermore, this research suggested that DataONE is an exemplar for how to build infrastructure with recent scientific changes (Allard, 2012a).

Science research practices and processes are significantly changing. Programs such as DataONE are responding to these changes. In Allards's (2012b) book chapter, the author introduced a selection of global data management initiatives responding these changes. Allard discussed specific programs including the Australian National Data Service (ANDS), Germany's Deutsche Forschungsgemeinschaft (DFG), the Netherlands SURF Foundation and Data Archiving and Networked Services (DANS), and the UK's Digital Curation Centre (DCC). From the United States, Allard described DataONE and the context in which DataONE is situated by providing background information of the Coalition for Networked Information and the National Science Foundation (NSF). Much of this book chapter discussed DataONE's organizational structure. Additionally, this research included eight cyberinfrastructure challenges of DataONE. These eight challenges include: (a) inconsistent service interface specifications, (b) lack of reliable unique identifier production and resolution, (c) data longevity and availability is dependent on repository lifespan, (d) inconsistent search semantics and effectiveness, (e) varying service interactions and data models, (f) access to quality metadata limits reuse of data, (g) lack of shared identity and access control policies, and (h) difficulties in placing data near analysis, visualization, and other computational services. The author discussed each of these cyberinfrastructure challenges and shows specifically how DataONE addresses them (Allard, 2012b).

Research and Development

A growing body of literature documents technical research and development for DataONE. Dexter, Cobb, Vieglais, Jones and Lowe (2011) described a pilot collaboration to deploy and operate DataONE Member Node software stack on TeraGrid infrastructure. This research described how this collaboration opens up the possibility to add large scale computing with DataONE data, metadata, and workflow tools. Additionally, this collaboration allowed for use of the Investigator Toolkit to retrieve datasets relevant to TeraGrid. This research showed how testing implementations assist in the understanding of DataONE infrastructure; in this case the testing of large scale computing infrastructure, the TeraGrid (Dexter et al., 2011).

Studies from DataONE summer internships have also addressed specific technical aspects of the DataONE. For example, Enriquez et al. (2010) investigated the policies, practices, and implications of current data attribution behavior in the environmental sciences. This study reviewed 500 papers published between 2000 and 2010 across six journals. The findings suggested that journal policies are specific and clearly stated. For example, very few repositories, journals, and funding agencies suggested best practice for data citations. This research also examined the practices of researchers citing data in their papers and found that data citations were most positively correlated with certain repositories, such as GenBank. This research also found that tracking DOIs for reuse was difficult with the standard retrieval resources (Enriquez et al., 2010). Additionally, Piwowar and Vision (2013) conducted a study to investigate patterns of data reuse with open data citations. This study analyzed 10,555 studies that created microarray datasets to investigate open data citation and reuse patterns. Results indicated that third-party reuse continued to increase over six years, citation benefits were most clear for the years 2004 and 2005, and a steady increase of dataset reuse since 2003 for microarray data (Piwowar & Vision, 2013).

I conducted research related to metadata registries in relation to my DataONE summer internship. This research examined metadata registration use in the sciences, and provided DataONE the knowledge to consider and create a community crowd sourced metadata online dictionary (http://www.yamz.net/). The study used a multi-method approach to examine metadata registration within the sciences through an evaluation of the literature, evaluation of current registries, and a survey of scientists. A metric that consisted of (a) services/tools, (b) education/training, (c) resources/documentation, (d) technical support, and (g) access/barriers was created and used to examine metadata registries. The criteria that emerged as most important were access/barriers and services/tools. Additionally, a survey of scientists provided information about registry use in the scientific community. Feedback included that there were too many registries and that these were too time consuming and complicated, although most scientists did conclude that standardization was important. This study provided the necessary feedback for the development of Yamz (Murillo, Greenberg, Kunze, & Boone, 2012).

Data Sharing and Reuse Investigations

Several studies examined major technical challenges that affect data sharing and reuse, and how DataONE is addressing these challenges. In Reichman, Jones, and Schildhauer's (2011) study, the authors analyzed data sharing and reuse in the ecological sciences and discussed the challenges of data sharing in reference to DataONE. This study stated that reviews of ecological informatics suggested "three major technological challenges regarding data sharing including: data dispersion, heterogeneity, and provenance" (Reichman et al., 2011, p. 703). This study indicated that DataONE provides assistance with these technological issues through well-curated, federated data repositories and provenance tracking systems. Furthermore, this study indicated that traditionally ecologists have little incentive to share information due to these challenges. Additionally, the authors suggested that the most effective strategy to encourage data sharing would be to alter the reward system. Lastly, the authors suggested DataONE provide the opportunity to realize the potential of fully reproducible science (Reichman et al., 2011).

Researchers Sayogo and Pardo have conducted two studies that examine data sharing and DataONE. In Sayogo and Pardo's (2011a) study the authors explored the critical challenges facing researchers involved in data sharing and how these challenges influence researchers willingness to openly share data. The researchers' conducted a survey of 1,320 researchers from 2009-2010. The results indicated that the two main determinants of publishing research datasets were data management and attribution of the data ownership. Other factors considered significant included data management skills and organizational support for data management. This study looked specifically at scientists making their data available for sharing, but did not address scientists reusing data (Sayogo & Pardo, 2011a).

Frameworks, Models, and Theoretical Investigations

Several studies explore DataONE from the perspectives theoretical frameworks. Allard and Allard's (2009) study explored DataONE as a transdisciplinary organization. The authors addressed the role of transdisciplinary research in earth and environmental sciences and indicated that a transdisciplinary approach makes it possible for global questions to be answered through emphasizing collaborations between scientists and information professionals. Additionally, this study explored how DataONE fits within the transdisciplinary index. The transdisciplinary index contains seven categories: (a) leadership/governance, (b) communication, (c) engagement, (d) purpose, (e) integration, (f) adaptability, and (g) context. This article showed how DataONE contains elements in each category and concluded that DataONE could be identified as a transdisciplinary organization (Allard & Allard, 2009).

The uncertainty framework presents another theoretical perspective that scholars have used to explore DataONE. Lagoze and Patzke (2011) examined the first two DataNet Partners, DataONE and Data Conservancy through the lens of an uncertainty framework to understand the sociocultural issues that influence cyberinfrastructure projects. This study investigated the framework for technological, organizational, scientific, and institutional uncertainty. The findings indicated that two issues need to be addressed for sustainability of these projects including ways to ensure economic stability and the creation of an administrative structure for effective long-lived virtual organization (Lagoze & Patzke, 2011).

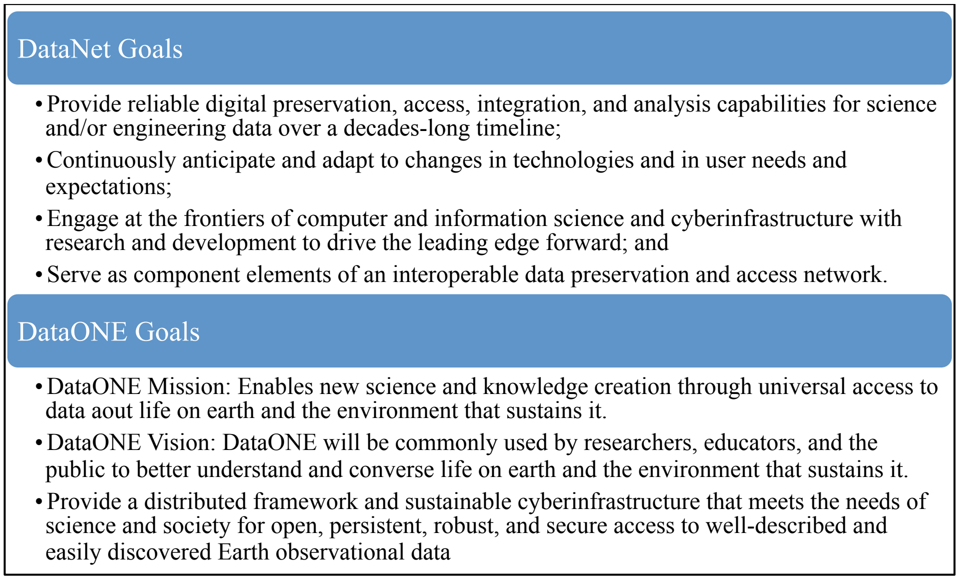
Sayogo and Pardo conducted a second study (2011b) on data sharing specifically examining how DataONE could be considered a collaborative data sharing network. In this study the authors analyzed DataONE through research from the cross-boundary information sharing and integration framework (CBIS), the transnational knowledge networks (TKNs), and the collaborative networks (CNs) literature. The authors suggested that DataONE is comprised of a multidimensional combination of interorganizational and cross-boundary sharing to achieve its collective purpose. The authors described how DataONE contains factors that are characteristics of CBIS's and TKN's. Ultimately, the authors suggested that there are five main characteristics of CDSN's (a) collaborative social actors, (b) shared common goals, (c) one-way or bi-directional information flow, (d) mediated and dynamic collaboration structure within a trusted network, and (e) collaboration supported with an interoperable infrastructure, and that DataONE embodies all of these. Furthermore, the authors discussed how the characteristics of CDSN's are common and compatible to the goals of sharing data and information (Sayogo & Pardo, 2011b).

Lastly, Aydinoglu (2011) applied complexity framework his dissertation to investigate how DataONE behaves like a complex adaptive system. The study specifically applied the complex adaptive systems perspective to a virtual scientific collaboration, which had not been applied in the past. The main research question was "How can the emergence of DataONE – a multidisciplinary, multinational, and multi-institutional scientific collaboration – be explored from a complex adaptive systems perspective" (Aydinoglu, 2011, p. 7). This study used a multi-method case study approach with semi-structured interviews, naturalistic observations, and a survey to answer the research question. Emergence is particularly important for studying complex adaptive systems; therefore this study was conducted during year two of DataONE. The contributions of this study included: (a) the building of a complexity framework for virtual scientific collaborations and (b) the complexity framework was applied to the real case of DataONE as a virtual scientific collaboration. Overall, it was found that DataONE does operate as a complex system which makes this virtual scientific collaboration resilient, adaptive, and successful (Aydinoglu, 2011).

Conclusion

This section provided an overview of the DataNet program and provided a more in depth account of DataONE. To summarize, the DataNets and DataONE provide the infrastructure for scientists to respond to the changes occurring within the scientific research process. DataONE has built specific cyberinfrastructure by establishing Coordinating Nodes and connecting

Member Nodes. Additionally, DataONE has created cyberinfrastructure through the development of the Investigators Toolkit, as well as established an Education and Outreach program in order to address the specific goals of the DataNets and DataONE. The goals of the DataNet and DataONE are summarized in Figure 4.



DataNet Goals

- Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;
- Continuously anticipate and adapt to changes in technologies and in user needs and expectations;
- Engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and
- Serve as component elements of an interoperable data preservation and access network.

DataONE Goals

- DataONE Mission: Enables new science and knowledge creation through universal access to data aout life on earth and the environment that sustains it.
- DataONE Vision: DataONE will be commonly used by researchers, educators, and the public to better understand and converse life on earth and the environment that sustains it.
- Provide a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data

Figure 4. DataNet and DataONE Goals (DataONE, 2013d; National Science Foundation, 2006)

Scholars have conducted research investigations on this infrastructure. These studies include investigations that provided a general overview of DataONE (Allard, 2012a, 2012b; Lee et al., 2009; Michener, Allard, et al., 2011; Michener, Vieglais, et al., 2011), studies that discussed specific technical aspects of DataONE (Dexter et al., 2011; Enriquez et al., 2010), studies that investigated data sharing within DataONE (Reichman et al., 2011; Sayogo & Pardo, 2011a), and studies that provided a model, framework, or theoretical lens in reference to DataONE (Allard & Allard, 2009; Aydinoglu, 2011; Lagoze & Patzke, 2011; Sayogo & Pardo, 2011b).

Data Sharing and Reuse in the Sciences

There is an extensive and growing body of literature investigating data sharing and reuse across a number of disciplines. This review addresses the major themes found in this literature. This attention is driven partially by concerns over the data deluge, which is discussed in section 2.3; as well as granting agency requirements for data management and data sharing plans along with grant proposals (Hey & Trefethen, 2003; National Institutes of Health, 2007; National Science Foundation, 2010a; Tenopir et al., 2011). Additionally, data sharing is important for the scientific community. As noted, the sharing of data provides the ability to extract additional value from data, enable reproducible research, enable others to ask new questions of existing data, and advances the state of science in general (Borgman, 2012; Lord & Macdonald, 2003). Furthermore, there has been discussion that data gathered and generated from public funding should be made available so that other scientists can reuse this data and so that the public has access to data that they financially supported (Borgman, 2012; Lord & Macdonald, 2003).

Themes and Factors Associated with Data Sharing

Table 1 provides an overview themes and factors found throughout the literature.

Table 1

Themes and Factors Associated with Data Sharing

THEMES	**Scientific Practice**	**Journal Requirements and Policies**	**Granting Agencies**	**Additional Factors**
FACTORS	Avoid duplication	Data Deposition	Data Management Plans	Financial concerns
	Scientific Reputation	Journal Mandates	Data Sharing Policies	Informational/organizational ownership
	Extract Additional Value	Data Withholding	Making publically funded data available to the public	Coworker helpfulness
	Encourages diversity	Journal Impact Factor		Secondary use of data
	Assist with reproducibility			Intended users

	New methods to test existing data			Work Experience
				Data Value

Several reports have demonstrated the many reasons why scientists want to share data. For example, in Lord and McDonald's (2003) report on e-Science, the authors described how sharing data can extract additional value and avoid duplication of existing work. The OECD's (2002) report, indicated that sharing data reinforces scientific inquiry, encourages diversity in analysis, and promotes new ways to test hypotheses or methods of analyzing data (Lord & Macdonald, 2003). Other reasons that sharing data are important include: (a) making results of publicly funded data available, (b) enabling others to ask new questions, (c) advancing the state of science, and (d) reproducing research (Borgman, 2010).

Funding agencies have put pressure on scientists to make their data available, since data are so expensive and time consuming to create, by asking for data management plans and creation of data sharing policies (Ceci, 1988; Cohen, 1995; National Institutes of Health, 2007; National Science Foundation, 2010a; Sayogo & Pardo, 2011a). Many journals have policies that they will not publish articles if scientists do not make their data publically available and most journals have additional policies associated with data sharing (Blumenthal, Campbell, Anderson, Causino, & Louis, 1997; Cohen, 1995; McCain, 1995; Sayogo & Pardo, 2011a; Sieber, 1988).

Research addressing data sharing began as early as the late 1980's. For example, Sieber (1988) stated that scientists have increased pressure to share data since many funding agencies and journals require data management as a condition for funding and publication. Scientific reputation also provides motivation for scientists to share data (Brown, 2003; Ceci, 1988; Cohen, 1995; Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences, 2009; Sayogo & Pardo, 2013). As described in Sieber (1988) scientists who are unwilling to share data could seem fraudulent. Therefore willingness to share data assists in creating a positive reputation with fellow scientists. However, even though many researchers postulate that reputation is a factor for data sharing, it is still not pervasive in the sciences (Sieber, 1988).

Data Sharing Research Studies

Data Sharing Studies

Researchers have examined data sharing practices through a range of methods including direct report from scientists, journal policy studies, and bibliometric studies. Several direct report studies have investigated scientists' attitude towards data sharing to gain an understanding of motivations (Baru, 2007; Blumenthal et al., 2006, 1997; Ceci, 1988; Constant, Kiesler, & Sproull, 1994; Sayogo & Pardo, 2011a; Tenopir et al., 2011; Zimmerman, 2003). Ceci's (1988) survey study suggested that scientists generally want data sharing as a norm in science. However, scientists also stated that their reason for not sharing was mostly financial (Ceci, 1988). Constant, Kiesler, and Sproull's (1994) study suggested that attitudes towards information ownership and work experience were factors highly correlated with sharing and organizational ownership moderately correlated. Other factors included if a coworker had been previously unhelpful, the researcher was less likely to share. This study used exchange and expressive theory of information sharing to investigate attitudes towards sharing through hypothetical situations in experimental conditions rather than through direct observation (Constant et al., 1994).

Data withholding is another factor that researchers have investigated. Blumenthal et al. (1997) indicated that while data withholding was not a widespread phenomenon, it had affected researchers enough to need further investigation. Blumenthal et al. (2006) conducted a second study, surveying geneticists and life scientists. Two types of data withholding were considered, verbal withholding (withholding in verbal exchange about unpublished research) and publishing withholding (withholding in and around the publishing process). The findings indicated that geneticists and life scientists participated in some form of data withholding, and that publishing withholding was the most common. Overall the findings suggested that data sharing is perceived as more beneficial than data withholding (Blumenthal et al., 2006).

Zimmerman's (2003) dissertation focused on secondary use of data, meaning the combining or comparing of data sources to answer new research questions. The findings indicated that tacit and informal knowledge are key in the process of data sharing. Furthermore this study indicated that it is difficult to anticipate all intended users, that there is little incentive to document datasets with complex metadata standards, and that scientists used publications to locate publically available data. Zimmerman analyzed the findings from the point of view of communities of practice to show how members share knowledge within their community or group (Zimmerman, 2003).

Time is also a key data sharing factor, both when it comes to data creation and data management. Birnthotlz and Bietz's (2003) study compared the data sharing practices of three different scientific disciplines: earthquake engineering, HIV/AIDS research, and space physics. The findings indicated time and effort to create data were a factor in how possessive scientists were about their data and that those working in collaborative environments negotiated more to form data sharing relationships (Birnholtz & Bietz, 2003). Tenopir et al. (2011) surveyed scientists to explore data sharing practices and perceptions. Scientists indicated that insufficient time and lack of funding were key reasons why they did not share data. Scientists also indicated that lack of organizational support was a barrier for them to share data (Tenopir et al., 2011).

Journal Policy Studies

Journal policies are another data sharing factor. As discussed, journals have put pressure on scientists to deposit their data. McCain's (1995) study investigated approximately 850 natural science journals and examined data deposition policy statements. Between 1992-1994 many journals began to contain policies to make research data available to readers. Approximately two-thirds of the journal's policy statements addressed some form of data sharing. This study revealed that enforcement including refusal to publish without evidence of deposition and that editor-author negotiation included data sharing. This research also demonstrated some of the early efforts of journals to enforce data sharing (McCain, 1995).

Bibliometric Studies

Research-related information (RRI) refers to data, software, electronic documents, and images that are associated with the research process. McCain's (2000) study focused on sharing of research-related information. This study analyzed bibliographic records from 1988-1998 downloaded from SCISEARCH to explore if RRI were also available. The results indicated molecular biology, general biology, and medicine had the greatest concentration of accessible RRI. Certain types of RRI, particularly data computations, software, and e-documents were found concentrated in biology, chemistry, and astronomy. Overall, this study indicated a general increase in scientists making their RRI available (McCain, 2000).

Data deposition is another way that researchers have explored data sharing. Brown (2003) explored the use of genomic and proteomic databases (GPD), such as GenBank and Protein Data Bank, and journal mandates of data deposition for publication accepted to investigate how these policies were reshaping the way molecular biologists published their work. Brown used a two-step approach and qualitative analysis was gathered using surveys and case studies to determine database use and acceptance. Quantitative analysis was collected using citation and bibliographic analysis, and was combined with an examination of instructions to authors' sections of molecular biology journals. The results indicated that over half of the scientists access GPDs weekly for the following reasons: to deposit data, to examine newly deposited data, and to examine previously data deposited. The results also indicated that the scientists believed that sharing of genomic and proteomic data are fundamental to the advancement of science. The results of the analysis of instructions showed that journals began requesting deposition in 1983 and by 1990 the majority of the journals requested data deposition before publication. This citation and bibliographic study was conducted to demonstrate the penetration of GPD into the molecular biology field. The conclusion of this study was that the data explosion in molecular biology is paralleled by the growth in usage and acceptance of GPD (Brown, 2003).

Data withholding has been discussed earlier through Blumenthal's survey work that investigated scientists' perceptions of withholding. Noor, Zimmerman, and Tetter (2006) examined the frequency in which published studies failed to submit their data to GenBank. The authors examined six journals that had explicit policies requiring data deposition. The conclusions were that no journals had complete compliance with requirements for data deposition to GenBank. However, the majority of the articles were submitted along with their data (Noor et al., 2006). Ochsner and colleagues (2008) reported opposite results from Noor. The authors surveyed articles from twenty journals in Medline. The findings from this study were that the rate of deposition of datasets was less than 50%, which indicates that many researchers did not deposit datasets and many journals were not positioned to enforce their own policies (Ochsner et al., 2008).

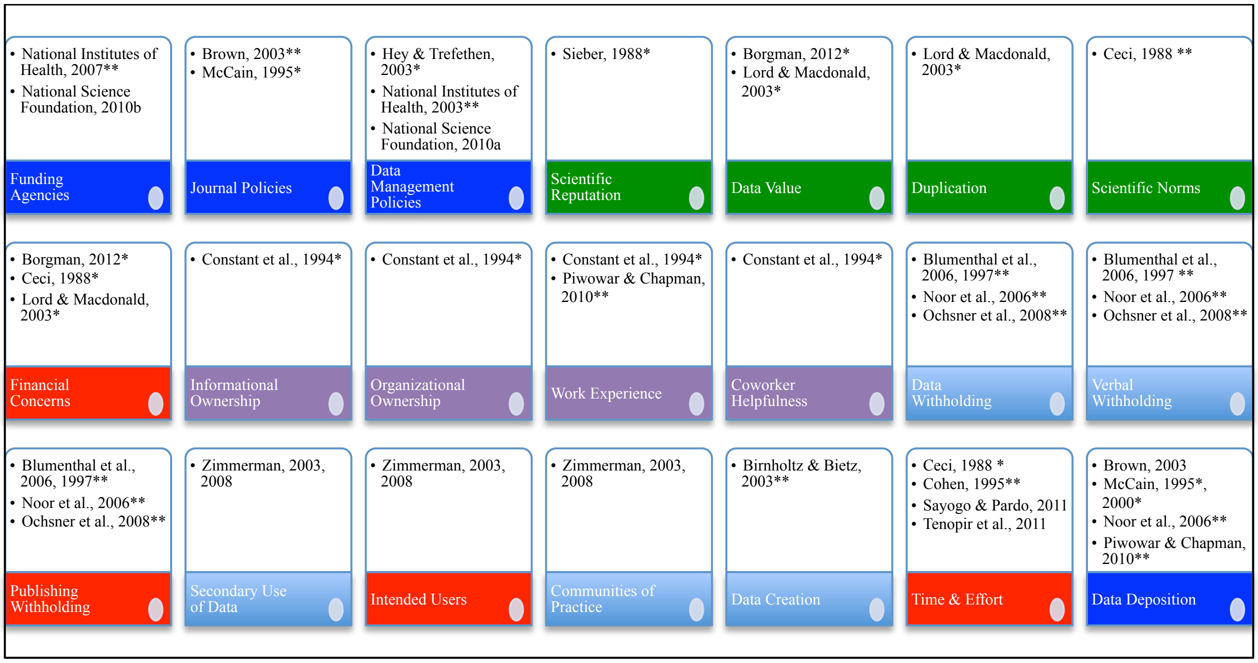
Lastly, Piwowar examined specific factors associated with data sharing. In Piwowar and Chapman (2010) the authors analyzed whether data sharing was associated with funder or publisher requirements, journal impact factor, and studies that were led by more experienced primary investigators. The results indicated that more than half of the studies made their raw datasets available. Factors that had increased data sharing included: author experience and studies published in high-impact journals. The National Institutes of Health (NIH) data sharing plan requirement was not found to increase data sharing behavior (Piwowar & Chapman, 2010).

Piwowar (2011) used full-text queries to identify a set of studies that generated data and then determined if these data were deposited in data repositories. The results indicated that data sharing had increased significantly since 2001. Studies from highly cited institutions had a higher rate of data sharing. Human and cancer studies had the lowest rate of sharing than any other discipline studied. Authors that had shared in the past were also more likely to share again. Authors in their early years of their profession were less likely to share than older scholars. Male authors were more likely to share. Some of the limitations of the study were that it did not consider direct peer-to-peer sharing or collaborative sharing (Piwowar, 2011).

Conclusion

This chapter described studies that investigated data sharing and reuse. While all of the studies discussed above provide insight they have some limitations. Studies that rely on interview or survey results that rely on self-report, could yield possible inflated or biased results (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). It is possible that scientists may exaggerate their participating in data sharing since most scientists believe it is important and vital to the sciences. Bibliometric studies are limited because they only examine data deposition. A scientist could deposit their data into a scientific data repository, however no one may download the data for reuse.

Figure 5 provides an overview of the themes found in this review.



• National Institutes of Health, 2007** • National Science Foundation, 2010b Funding Agencies	• Brown, 2003** • McCain, 1995* Journal Policies	• Hey & Trefethen, 2003* • National Institutes of Health, 2003** • National Science Foundation, 2010a Data Management Policies	• Sieber, 1988* Scientific Reputation	• Borgman, 2012* • Lord & Macdonald, 2003* Data Value	• Lord & Macdonald, 2003* Duplication	• Ceci, 1988 ** Scientific Norms
• Borgman, 2012* • Ceci, 1988* • Lord & Macdonald, 2003* Financial Concerns	• Constant et al., 1994* Informational Ownership	• Constant et al., 1994* Organizational Ownership	• Constant et al., 1994* • Piwowar & Chapman, 2010** Work Experience	• Constant et al., 1994* Coworker Helpfulness	• Blumenthal et al., 2006, 1997** • Noor et al., 2006** • Ochsner et al., 2008** Data Withholding	• Blumenthal et al., 2006, 1997 ** • Noor et al., 2006** • Ochsner et al., 2008** Verbal Withholding
• Blumenthal et al., 2006, 1997** • Noor et al., 2006** • Ochsner et al., 2008** Publishing Withholding	• Zimmerman, 2003, 2008 Secondary Use of Data	• Zimmerman, 2003, 2008 Intended Users	• Zimmerman, 2003, 2008 Communities of Practice	• Birnholtz & Bietz, 2003** Data Creation	• Ceci, 1988 * • Cohen, 1995** • Sayogo & Pardo, 2011 • Tenopir et al., 2011 Time & Effort	• Brown, 2003 • McCain, 1995*, 2000* • Noor et al., 2006** • Piwowar & Chapman, 2010** Data Deposition

Key	
🟦	Policy based
🟩	Motivator
🟥	Inhibitors
🟪	Both Motivator and Inhibitor
**	Life/Biological Sciences
*	Mixed sciences (not earth science focused)

Figure 5. Key Data Sharing Themes in Academic Research Studies

Additional limitations to the current studies include many of these studies focus on motivators, inhibitors, and journal and grant agency policies. Additionally, many of these studies were conducted within the life and biological sciences and not in the earth sciences, leaving a gap in the current knowledge of scientific data sharing. While the investigations do provide an insight into data sharing, there are many other factors to be considered such as technical factors, descriptive information, and reuse requirements.

Data Management in the Sciences

Factors important to scientific data management include the data deluge (Hey & Trefethen, 2003), fourth paradigm (Hey et al., 2009), and technical considerations of the data lifecycle (Consultative Committee for Space Data Systems, 2002; Higgins, 2008; Lord, Macdonald, Lyon, & Giaretta, 2004). Challenges and opportunities have come to the forefront as scientific data management has changed dramatically. There have been a significant increase in the amount of data being produced (Bell et al., 2009; Borgman, Wallis, & Enyedy, 2007), significant changes in scientific practice (Ailamaki, Kantere, & Dash, 2010; Beran, van Ingen, & Fatland, 2010), and new terminology associated with these changes (Atkins, Hey, & Hedstrom, 2011; Higgins, 2008). This review examines the recent past and current state of scientific data management by examining the data deluge and fourth paradigm. Additionally this review investigates key scientific data management factors including: (a) the data lifecycle, (b) data types, and (c) scientific metadata.

The Data Deluge

Concerns stemming from the growing data deluge are increasingly gaining attention from the scientific, computer science, and information science communities. As new technologies have been created, the research process has changed and the data that are created have also changed. Hey and Trefethen, first described and discussed the term data deluge (Hey & Trefethen, 2003). In general, the term data deluge can be defined as the concern that the amount of data being produced exceeds the current ability to structure the data in an organized fashion (CODATA, 2002; Consultative Committee for Space Data Systems, 2002; Hey & Trefethen, 2003; Lord & Macdonald, 2003; Lord et al., 2004).

Hey and Trefethen (2003) discussed how technologies are increasing the amount of data being produced. The authors presented summaries of scientific research groups that are significantly increasing the amount of data they are generating. VISTA projected that by 2014 the archive would contain multiple petabytes of data, while in 2004 it only produced 10 Terabytes per year. For NASA, it was predicted that "data volumes of more than 10 fold in the 5 year period 2000 to 2005" (Hey & Trefethen, 2003, p. 7). In the domain of particle physics there is a similar trend and it was projected that "by 2015, particle physicists will be using exabytes" (Hey & Trefethen, 2003, p. 8). Each of these examples suggests the same trend; the amount of data being created is increasing at an alarming rate.

Lord and McDonald (2003) reinforce the previously described perspectives and offer a European Union point of view. As stated on page 5, "science is being transformed by accelerating change in information technology, with huge increases in computing power and network bandwidth, accompanied by an explosion in data volumes and information" (Lord & Macdonald, 2003). This report examined the "curation of digital primary research data generated in academic and scientific research in the United Kingdom, with a focus on data enabled by e-Science … data-intensive, computing intensive, collaborative, dispersed science" (Lord & Macdonald, 2003, p. 9). The data for this study were gathered through JCSR call for proposals, questionnaires to researchers, and by conducting face-to-face interviews with experts and researchers. This study indicated that not only the volume of data "are increasingly exponentially, but the number of digital objects is increasing, additionally the objects themselves are heterogeneous and increasing in complexity" (Lord & Macdonald, 2003, pp. 9–10). Beyond discussing the data deluge and the heterogeneity of scientific data, this report also examined the importance of archiving and curation of scientific data, funding challenges, repositories, and standards (Lord & Macdonald, 2003).

Historical View of the Data Deluge

Although data deluge is a fairly new term, this problem has existed for quite some time. Lide (1981) described a similar concern that "the proliferation of scientific literature and data files has taxed our ability to store, retrieve, and assimilate information" (Lide, 1981, p. 1343). Lide discussed how the amount of data tends to discourage a logical approach to decision-making, which in essence makes it difficult for scientists to know how to make sense of the data. Lide states that "automation of the measurement process has led to vast increases in the amount of data being generated in every science discipline…the instrumentation advances of the past two decades have resulted in the production of far more data than can possibly be handled" (Lide, 1981, p. 1343).

In 1990, the Committee on Data for Science and Technology (CODATA) released a series of papers from their 11th annual conference in a book titled "Scientific and Technical Data in a New Era" (Glaeser, 1990). This book presented data deluge concerns for multiple sciences including in materials sciences and engineering, environmental sciences, geosciences, and space sciences and demonstrated how this concern has been relevant for many years (Glaeser, 1990).

How the Data Deluge is Changing the Scientific Data Management

Many studies describe the data deluge and how this is changing the scientific research process. In 2002, the Consultative Committee for Space Data Systems (CCSDS) released their Blue Book, which discussed concerns about the data deluge particularly in reference to how computers are changing the scientific process. "A tremendous growth in computational power, and in networked bandwidth and connectivity, has resulted in an explosion in the number of organizations making digital information available" (Consultative Committee for Space Data Systems, 2002, pp. 1–3). Organizations such as the DataONE assist in making these scientific data available to users.

Baru (2007) described that as the amount of scientific data increases, there is a general need to examine how these data are being produced, organized, accessed, and shared in order to discover best practices in organizing data. Baru suggested "there is a need to examine long-term preservation strategies and archiving, and a need to evaluate the future value of the current data" (Baru, 2007, p. 113). Additionally, Baru suggested that advances in technology have changed how scientists are conducting research and these changes have influenced research practices, contributing to the data deluge. As Baru described, a steady decrease in costs, increase in storage capacity and reliability in hardware, in addition to cheaper and faster processors have provided great benefits for computational and e-Science disciplines (Baru, 2007). Technology has given scientists the ability to create and store vast amounts of data and essentially is changing how the scientific process is occurring.

Borgman (2007) examined these issues in further detail by specifically looking at how the data deluge is affecting scientists in the small science communities. Prior to this work, much attention was placed on big data science fields. Borgman investigated habitat ecology as an example of a small science community and how the use of embedded sensors is changing the field. The author concluded that the data deluge is affecting all scientific communities, including small science (Borgman et al., 2007).

Additional studies examined how the data deluge is changing scientific practice. Bell, Hey, and Szalay (2009) discussed how the data deluge is leading to data-intensive science. This article explored how the nature of the scientific process is changing in that simulations and experiments are yielding more data than in the past. This work also suggested that Moore's Law, inexpensive bandwidth, and faster computers have led to this data explosion. Lastly this work described aspects of the fourth paradigm, which will be discussed in greater detail in the following section (Bell et al., 2009).

There have been investigations of systems and technological aspects of the data deluge. Ailamaki, Kentere, and Dash (2010) discussed how the data deluge will only worsen with time with improvements in instruments and simulation models. The authors suggested that the current tools are incapable of supporting the data needs of scientists: "unfortunately, today's commercial data-management tools are incapable of supporting the unprecedented scale, rate, and complexity of scientific data collection and processing" (Ailamaki et al., 2010, p. 68). This concern has led to an explosion of data management tools.

There has been research regarding the data deluge from specific discipline perspectives. Beran, van Ingen, and Fatland (2010) discussed the data deluge from the geoscience perspective. This study explored how the data deluge is affecting the way geoscientist's collect their data, particularly how low-cost in situ sensors are changing earth science research. These sensors make an unprecedented amount of data available to scientists not only in the geosciences, but other fields (Beran et al., 2010).

The National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization prepared a report in 2011 that provided a synthesis of the issues in scientific data management and the scientific process. This report described the cultural and social changes, changes in roles and responsibilities, economic considerations, and other issues regarding how the data deluge. Overall this report suggested that "a policy of retaining all scientific data is impractical and therefore recommends that the NSF support researchers and research communities undertaking the effective triaging of data for retention, archiving, and deletion" (Atkins et al., 2011, p. 7).

Changes in Scientific Process and the Fourth Paradigm

Jim Gray (1996) described six distinct phases in data management. He described these phases chronologically throughout history as (a) manual record managers, (b) punched-card record managers, (c) programmed record managers, (d) on-line network databases, (e) relational databases and client-server computing, and (f) multimedia databases. As stated,

> this easy access to information will transform the way we do science, the way we manage our businesses, the way we learn, and the way we play. It will both enrich and empower us and future generations (Gray, 1996, p. 46).

This article influences his later work on scientific paradigms throughout history, specifically the concept of the fourth paradigm.

Jim Gray's Fourth Paradigm

During Jim Gray's final presentation (Hey et al., 2009), he introduced the concept of the fourth paradigm and summarized scientific paradigms throughout history. As described, thousands of years ago science was empirical, over the last few hundred years the theoretical branch was developed, over the last few decades computational science has been developed, and today science is in a data exploration phase, which is considered eScience. Scientific research that is conducted today is through unified theory, experiments, and simulations. As stated, "the techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science" (Hey et al., 2009, p. xix). This differentiation is described with "data is captured by instruments or generated by simulators, processed by software, the information/knowledge is stored in computer, and scientists analyze databases or files using data management and statistics" (Hey et al., 2009, p. xix), which is quite different than the way science has been conducted in the past.

These changes in the scientific process have led to concerns regarding the management of scientific data. Hey described how "the world of science has changed, and since scientific practice is changing, these new data practices need to be analyzed" (Hey et al., 2009, p. xix). Since data are being created and captured differently than in previous scientific paradigms, the way we work with scientific data also needs to change. Furthermore, the tools needed to work within this new paradigm need to be analyzed.

How the Fourth Paradigm is Changing Data Management in the Sciences

Studies have described the fourth paradigm and how this is changing the scientific research process. Beran, van Ingen, and Fatland (2010) discussed how new technologies are changing geoscience research. They described how the data itself are changing and becoming more diverse. They also described how scientists have new challenges with how to discover relevant data and assemble that data for analysis or modeling. The authors suggested that this problem needs to be addressed by information technology in order to ensure that scientists are able to conduct this new type of research (Beran et al., 2010).

Researchers have discussed concerns that are occurring with this new scientific paradigm. Ailamaki, Kantere, and Dash (2010) described how one of the problems is the current technical solutions. The authors suggested that there is a lack of complete solutions in commercial database system, which has led scientists to develop or adopt application-specific solutions. Some of these application-specific solutions are software such as scientific collaboration software and scientific workflow software. Another concern that the authors discussed is the need for general-purpose data-management systems that can handle specific technical aspects.

Lastly, the authors discussed the key aspects to scientific data management which include: “workflow management, management of metadata, data integration, data archiving, and data processing” (Ailamaki et al., 2010, p. 73). Organizations such as DataONE are striving to create environments to promote the type of data management described in this work.

Another key concern is the change in large scale and geographically distributed studies that are becoming more common. As described by Agarwal et al. (2010), large-scale synthesis studies conducted by geographically distributed science teams are becoming more common. New technologies that are being used in the scientific process are enabling collaboration. Atkins, Hey, and Hedstrom (2011) described how “data volumes, computing power, software, and network capacities are all on exponential growth paths, and research collaborations are expanding dramatically” (Atkins et al., 2011, p. 3). Examples of these large-scale collaborative projects include CERN, the National Virtual Observatory, and the Geoscience Network or GEON. As described, “crucial data collections in the social, biological, and physical sciences are now online and remotely accessible” …providing the ability for …“groups collaborating across institutes, time zones, sharing data, and complementing expertise” (Atkins et al., 2003, p. 9). Additionally, these data and technology are no longer restricted to just a few research groups (Atkins et al., 2003). The DataONE and other DataNet programs are trying to address the large scale geographically distributed programs that are becoming more common in the sciences.

As described above, the data deluge and the fourth paradigm are the two primary drivers of the current changes within the scientific process. The following section specifically examines key elements regarding scientific data management including new terminology and data lifecycle models, data types, and scientific metadata.

Emerging Concerns of Data Management in the Sciences

Terminology and Data Lifecycle Models

As changes in the scientific process have occurred, new terminology has emerged. Lord and MacDonald (2003) introduced working definitions of curation, archiving, and preservation in relation to scientific data. This report defines these terms: curation - managing from the point of creation; archiving - a curation activity to ensure data are properly selected, stored, and can be accessed; and preservation - an activity within archiving where data are maintained over time. These terminologies are described throughout the literature and models for data management. Overall, these terms and the various data lifecycle models that have been created are "concerned with managing change over time" (Lord & Macdonald, 2003, p. 12). The following section discusses the new terminology and lifecycle models that are being created as ways of understanding and organizing scientific data.

The Consultative Committee for Space Data Systems (CCSDS) is an organization of space agencies including the British National Space Centre (BNSC), the Canadian Space Agency (CSA), the European Space Agency (ESA), the National Aeronautics and Space Administration (NASA), and many others. These organizations worked together to publish the 2002 Blue Book to address issues of data management in the sciences and to define an International Organization for Standardization (ISO) Reference Model for an Open Archival Information System (OAIS). The purpose of the OAIS model is to facilitate an "understanding of what is required to preserve and access information for the long term" (Consultative Committee for Space Data Systems, 2002, pp. 3–1). An OAIS "is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a designated community" (Consultative Committee for Space Data Systems, 2002, pp. 1–1). The model accommodates both physical and digital data, and addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination. Figure 6 shows the environment surrounding an OAIS. The key components to this environment are producers, consumers, and management. The overall lifecycle model is shown in Figure 7 and contains the entire lifecycle including ingest, data management, access, and archival storage. The model also shows how descriptive information fits into the cycle and how each community (producers, management, and consumers) are interacting with the data management lifecycle (Consultative Committee for Space Data Systems, 2002). This model continues to be used as a reference for data management in many organizations. The CCSDS has written updated versions of the original Blue Book, the 2011 Magenta Book that focuses on recommended best practices (Consultative Committee for Space Data Systems, 2011).

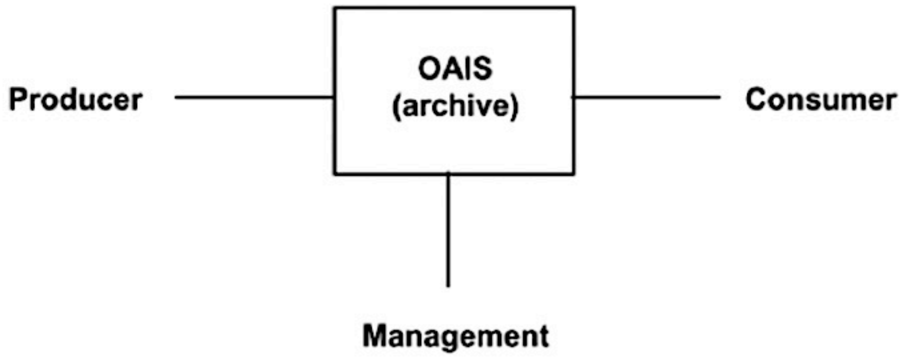


Figure 2-1: Environment Model of an OAIS

Figure 6. Environment Model of the OAIS (Consultative Committee for Space Data Systems, 2002)

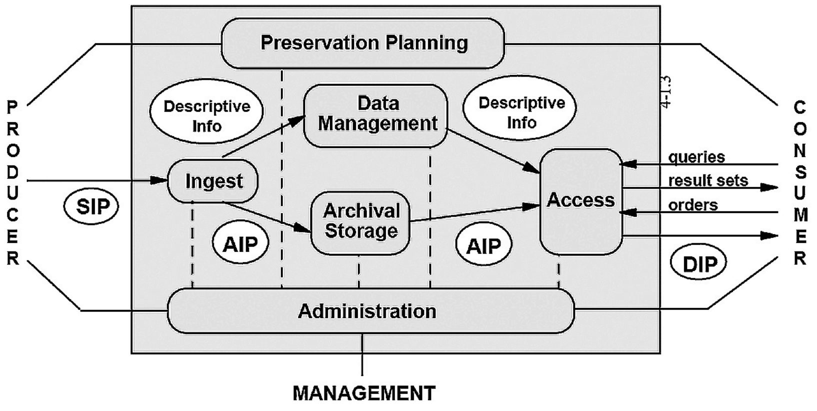


Figure 4-1: OAIS Functional Entities

Figure 7. Simple OAIS Model (Consultative Committee for Space Data Systems, 2002)

The National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization Report, 2011, discussed another data lifecycle model developed by the Interagency Working Group on Digital Data in 2008 (shown below in Figure 8). The vision for this model is that,

> data creation, collection, documentation, analysis, preservation, and dissemination could be appropriately, reliably, and readily managed, thereby enhancing the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society (Atkins et al., 2011, p. 10).

This model focuses on the planning, creation, keeping, and deposition of the data. This model also illustrates the importance of technical requirements, human resources and professional skills, organizations and entities, and policy (Atkins et al., 2011).

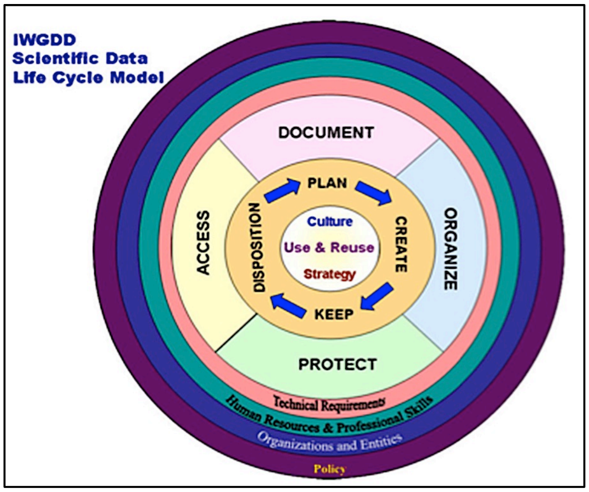


Figure 8. Interagency Working Group on Digital Data (Atkins et al., 2011)

The Digital Curation Centre also developed a data curation lifecycle model shown below in Figure 9. This model provides a "high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle" (Digital Curation Centre, 2004, para. 1). As discussed in Higgins, 2008, the purpose of the DCC data lifecycle model is to "provide a high-level view, it can be used in conjunction with relevant reference models, frameworks, and standards to help plan activities" (Higgins, 2008, p. 135). The model is meant to complement standards such as the OAIS reference model and is used as a training tool for DCC projects (Digital Curation Centre, 2004).

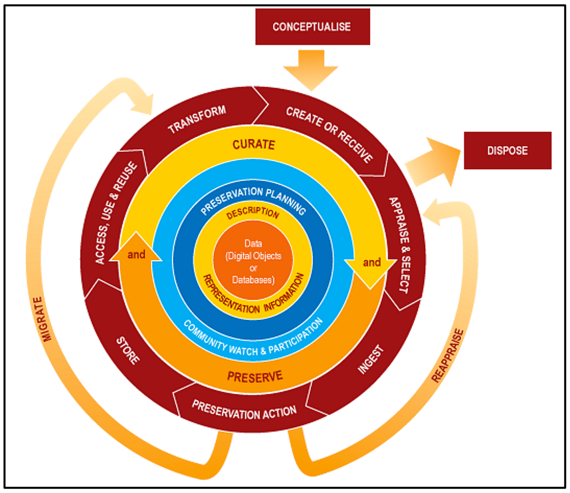


Figure 9. DCC Lifecycle Model (Digital Curation Centre, 2004)

There have been multiple other models created, all with the intent to assist with the management of scientific data. Several of these models were described in Bill LeFurgy's blog, "Life Cycle Models for Digital Stewardship" and included a model from JISC, DigitalNZ, and CASPAR. As described in the blog these models illustrated basic stages that content moves through from creation, preservation, and access over time (LeFurgy, 2012). The amount of new models and terminology continues to grow and are used by researchers, scientists, and practitioners in a variety of fields.

Data Types

Scholars have created classification schemes for scientific data. For example, CODATA created a detailed scheme for classifying scientific and technical data which can be described with three broad classes of data: (a) Class A – repeatable measurements on well-defined systems, (b) Class B – observational data, and (c) Class C – statistical data. While these three broad classes provide definitions for the general data types, the distinctions between the classes are not always clear. For example most biological data are both Class A and Class B, and sometimes can be considered Class C (Lide, 1981).

Scientific data are incredibly heterogeneous and diverse. Anderson (2004) examined how scientists use many types of data; observational, experimental, computer generated, and measurement (Anderson, 2004). Similarly, Borgman, et al. (2007) categorized scientific data into "observations, computations, experiments, and record-keeping" (Borgman et al., 2007, p. 19). The authors defined observational data as 'data associated with specific place and time, which can be used in cross-sectional or longitudinal studies.' They defined computational data as 'data resulting from executing a computer model or simulation, and experimental data as data resulting from laboratory studies.' Lastly, the authors defined recording keeping data as 'data which include government, business, public, and private life data' (Borgman et al., 2007). These are the most common data types created through the scientific process.

Ailamaki et al. (2010) described observational data as data that are "collected through detectors; input is digitized, and output is raw observational data," and differentiate it from simulation data by defining simulation data as data that are "produced through simulators that take as input the values of simulation parameters" (Ailamaki et al., 2010, p. 69). However, data can be described beyond observational and simulation data and include heterogeneous information-intensive data. Heterogeneous data are common in sociology, biology, and psychology (Ailamaki et al., 2010). Data gathering practices should also be considered. As described by Borgman and colleagues, "data on the same variables are gathered by multiple means, that data exist in many states and in many places, and that publication practices often drive data collection practices" (Borgman et al., 2007, p. 17).

Not only do scientists need to access a variety of data they need to access it from various locations. As described by Agarwal and colleagues,

> carbon-climate measurement site data encompasses sensor measurements field observations, laboratory analysis of field samples, as well as anecdotal descriptions…other ancillary data includes relatively infrequent measurements of variables such as soil carbon…soil characteristics, and site disturbances…each measurement is often annotated (Agarwal et al., 2010, p. 2324).

From the above, the heterogeneity and diversity of scientific data are made evident showing the complexity of scientific data.

Due to the complexity and cost of generating scientific data there has been a great deal of attention given to data sharing and reuse. Major funding agencies have asked for mandatory data management and data sharing plans to be included in grant applications. As described by Baru, "there is interest in ensuring that existing data are used to the fullest extent possible" (2007, p. 114). For example, the GEON cyberinfrastructure addresses "the need in the geosciences to interlink and share multi-disciplinary dataset…heterogeneity across sub-disciplines. … therefore, a data sharing culture was built in to the project from its very inception" (Baru, 2007, p. 114). This need for data sharing and reuse has fueled the creation of scientific data tools that assist scientists in the organization of their data. However sharing, reuse, and use in general is nearly impossible without proper metadata.

Metadata

The current literature has described the importance of metadata in the scientific process. Lide described that "it is ironic that the scientific and technical information community has failed to develop a precise vocabulary for many key concepts" (Lide, 1981, p. 1343) and indicated the need for a controlled vocabulary. Atkins and colleagues described how important the metadata is to understanding the data, "without explicit schema and metadata, the interpretation is only implicit and depends strongly on the particular programs used to analyze it" (Atkins et al., 2011, p. 11). Without the metadata the actual data becomes unreadable and often is lost. As described by Atkins and colleagues "ultimately, such uncurated data are as good as lost, even if the bits are stored forever, because they cannot be interpreted correctly" (Atkins et al., 2011, p. 11).

Due to the importance of metadata, scientists and information scientists have worked to create standards for scientific data. For example, in 1988, the Data Bank on Enzymes and Metabolic Pathways created a format that included 253 subject fields divided into 16 groups and allowed for representation of information in text and digital form. The structure was created very broad in scope and made it possible to present symbolic, numeric, and graphical information (Selkov, Goryanin, Kaimachnikov, Shevelev, & Yunus, 1990).

Jones, Berkley, Bojilova, and Schildhauer (2001) examined ecological and biological science metadata standards. The authors examined how the modular Metacat framework helped researchers customize and revise their metadata. This article examined the Ecological Metadata Language (EML) and the National Biological Information Infrastructure's Biological Data Profile used by ecological scientists. The authors described other metadata standards that are "deep or broad enough for effectively documenting biological data" (Jones et al., 2001, p. 67) and included Dublin Core and the Global Information Locator Service. Another metadata standard discussed in this article was the U.S. Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (CSDGM), a relevant standard for geospatial data. However, it still has "notable omissions in taxonomic coverage and other biological relevant information" (Jones et al., 2001, p. 67). The article concluded that for the ecological community the EML core concepts incorporated into a biological profile of the CSDGM made the most useful standard for ecological researchers (Jones et al., 2001).

Edwards et al. (2011) discussed the Long Term Ecological Research (LTER) program which has established its own practices of metadata based on the community needs. As described in the article, the LTER program consists of 26 sites and each site is responsible for managing their own research data. One of the goals of the project is to improve data exchange among the sites and the major challenges to this project have been (a) the heterogeneity of the data, (b) the wide dispersal of LTER sites, and (c) the multiple metadata schemas. In order to standardize metadata practices the community adopted EML as their metadata standard (Edwards et al., 2011).

A group of geoscientists created GEO-SEED to assist with their discipline's metadata concerns. Brazier and colleagues (2009) discussed metadata needs in the geosciences and described the metadata discovery web-based service, GEO-SEED. GEO-SEED serves as a metadata repository for the geosciences and assists geoscientists describe their data. GEO-SEED provides a "standard vocabulary encoded in standard Semantic Web language OWL and RDF for metadata acquisition" (Brazier et al., 2009, p. 356). This system allows descriptions to be interpreted by machines and humans (Brazier et al., 2009).

The above examples demonstrate how the scientific community is integrating metadata into their systems. Metadata standards are often designated for specific communities or data types. Riley (2009) provided a comprehensive list of metadata standards. Darwin Core is a standard for biological objects or data. The Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC/CSDGM) is a metadata standard for geospatial information. The OpenGIS Geographic Markup Language (GML) describes spatial and temporal features, topologies, and observation methods. The ISO 19115 is an international geospatial metadata standard. MathML is W3C's recommendation for encoding of mathematical information (Riley, 2009).

Scientists also have to consider auxiliary information that is important to their scientific process. As described in Lide, "it is also important to record and preserve all the auxiliary information required to use the data with confidence, such as calibration standards and ambient temperature and pressure" (1981, p. 1346). The notes regarding calibrations and any external factors that could affect the outcome of the data and needs to be recorded in order to ensure data reproducibility. These are considered part of the auxiliary information important to the scientific process, and just as with metadata, if this information is not recorded the data itself could become unusable for scientists.

Conclusion

Examination of data deluge literature indicates that the amount of data being produced by the sciences has increased at an extraordinary rate and this increase in data has led to a thorough examination of how to manage the amount of data (Ailamaki et al., 2010; Baru, 2007; Bell et al., 2009; Beran et al., 2010; CODATA, 2002; Consultative Committee for Space Data Systems, 2002; Hey & Trefethen, 2003; Lord & Macdonald, 2003). Examination of the Fourth Paradigm literature indicates that scientific processes have become more data and computation intensive (Agarwal et al., 2010; Ailamaki et al., 2010; Beran et al., 2010; Gray, 1996; Hey et al., 2009).

Section two discusses (a) new terminology and data lifecycle models, (b) data types, and (c) metadata. New terminology and data lifecycle models have been created in order to assist in understanding scientific data management (Atkins et al., 2011; Consultative Committee for Space Data Systems, 2002; Higgins, 2008; Lord & Macdonald, 2003), as well as data types (Agarwal et al., 2010; Ailamaki et al., 2010; Anderson, 2004; Baru, 2007; Borgman et al., 2007; Lide, 1981). Lastly, this review discusses the importance of metadata and describes metadata types (Atkins et al., 2011; Brazier et al., 2009; Edwards et al., 2011; Jones et al., 2001; Lide, 1981; Riley, 2009; Selkov et al., 1990).

While this review covers a fair amount of the growing body of literature and information regarding scientific data management there are several limitations. A more detailed discussion of data types and metadata, and a discussion of provenance and other aspects of managing scientific data need to be further examined. These topics are covered in the following section.

Selected Infrastructure and Interoperability Factors

One of the most pressing challenges facing researchers in data management is understanding the full range of factors that impact data sharing and reuse. The previous sections discussed many of these factors; however additional factors need to be addressed. The following section addresses interoperability and infrastructure considerations for data management and data sharing and reuse.

Taylor (2004) defined interoperability as "the compatibility of two or more systems such that they can exchange information and data and can use the exchanged information and data without special manipulation" (Taylor, 2004, p. 369). The National Research Council (NRC) (1994) defined an information infrastructure as:

> a framework in which communications networks support higher-level services for human communication and access to information. Such as infrastructure has an architectural aspect – a structure and design – that is manifested in standard interfaces (Borgman, 2000a, p. 26).

Together these factors related to interoperability and infrastructure can heavily influence scientists' ability to share and reuse data.

This chapter addresses selected interoperability and infrastructure factors. The selection of factors was informed by two key considerations: (a) the dissertation research being pursued in connection with this literature review and (b) background research to conceptualize and define potential factors as a baseline for this dissertation research. The factors are organized as follows: data tools, provenance, metadata, ontologies, and lastly data and data models.

Data Tools and Applications

Informatics researchers and domain scientists have created a range of tools to aid data management during data activities along the full data lifecycle from capture to manipulation. These tools reflect the combined effort of the scientific, computer science, and information science communities to deal with the changes in the scientific process described throughout this literature review. The range and availability of tools has grown a great deal. While it is impractical to review all tools, the remainder of this section examines a selected set of scientific data management tools.

Scientific Data Repositories

The growth of scientific data centers and data repositories provides the scientific community tools to manage and organize data. This section describes selected scientific data repositories that are available, many of which have their own specializations and core audience and are fundamental tools for scientists to deposit data for sharing and to locate data for reuse.

Lide (1981) described some of the earliest data centers such as the World Data Center. The World Data Center was established during the first International Geophysical Year in 1957, and collects geosciences and astronomical data (Lide, 1981, p. 1346). Another online data service that Lide described is the Chemical Information System (CIS). The CIS included a central hub known as the Structure and Nomenclature Search Systems (SANSS) and contained nearly 200,000 chemical compounds. These systems were some of the earliest data repositories. Some of the key concerns of these systems include high cost, standardization, and quality assurance (Lide, 1981). These complications still exist today, and projects such as the DataONE are striving to alleviate these complications.

The biomedical field also created data repositories. As described by Bussard (1990), by the late 1980's to early 1990s, the production of DNA sequence data were larger than the capacity of the two major data banks for DNA sequences, GenBank and EMBL. The author described reasons why data repositories were important which included: intellectual effort, financial effort, duplication of research, and new discoveries. Furthermore, the author described concerns regarding the data management which included: issues about data ownership, ethics, dissemination of information, and national interests (Bussard, 1990). These concerns still exist today.

Another field that created data banks relatively early is materials science. For example, the Institute of Inorganic Chemistry, Academy of Sciences in Novosibirsk developed a data bank for electronic materials (Db EMAP). This data bank contained three types of databases: physiochemical properties, structural characteristics, and selected physical properties (Kuznetsov, Titov, Borisov, & Vetroprakhov, 1990). These are just a few examples of the diversity of early scientific data repositories. As described in Fredje and Meinard, the number of data banks in the biological sciences increased dramatically, from 1980 to 1988 the number of databases increased from 400 to 3,457 (Fredje & Meinard, 1990).

Since the 1990s, many new scientific data repositories have been created. Most of these are available online for scientists to download data for their research needs. For example, http://www.ncdc.noaa.gov/cdo-web/search, is a website where visitors can download data in relation to climate data. This website is part of www.data.gov, which contains multiple portals for various types of scientific data (United States of America, 2016). GenBank provides users with annotated collections of DNA sequences and has over 126 billion bases in 135 million sequence records. The purpose of GenBank is to "provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information" ("GenBank Overview" National Center for Biotechnology Information, 2013).

Marcial and Hemminger (2010) examined 100 web-based scientific data repositories to identify the major characteristics. The findings indicated that the majority of these data repositories were in the geosciences, medicine, biology, and astronomy. There were fewer data repositories in the marine sciences, mathematics, chemistry, social sciences, physics, ecology, and multidisciplinary disciplines. The findings also provided a summary of the common file types found in these data repositories. These file types included archives, statistical analysis, GIS, extensible markup, flat files, images, movie/multimedia, word processor files, spreadsheets, presentations, proprietary, and web pages. The study also discussed the business type of each repository including: federal centers, university centers, partnerships, institutes, non-profits organizations, publishers, state government agencies, world data centers, and societies (Marcial & Hemminger, 2010).

Scientific Workflow Software

Over the last several years, the scientific and related informatics communities have created scientific workflow software to assist scientists. Scientific workflow software "allow scientists to specify large computational experiments involving a range of different activities, such as data integration, modeling and analysis, and visualization, to name a few" (Abramson, Enticott, & Peachey, 2008, p. 392). Additionally, these systems support the reproducibility of experiments and allow the reuse of produced artifacts (Lifschitz, Gomes, & Rehen, 2011). Some examples of scientific workflow software include: VisTrails, Kepler, Triana, and Taverna.

DataONE Investigators Toolkit includes VisTrails and Kepler. VisTrails is an open source scientific workflow software and provenance management tool that supports simulations, data exploration, and visualization. This workflow software assists with exploratory tasks, and was designed to manage rapidly evolving and iterative workflows. Something that distinguishes VisTrails from other software is that it maintains provenance of data products, the workflows that derive these products and their executions (VisTrails, 2014).

The Kepler Project is a National Science Foundation project led by a team from UC Davis, UC Santa Barbara, and UC San Diego. The software is an open source scientific workflow application which is designed to help scientists, analysts, and computer programmers create, execute, and share models and analysis across a broad range of scientific and engineering disciplines. The software also helps users share and reuse data, workflows, and components (Kelper/Core, n.d.).

Scientific Collaboration Tools

Scientists incorporate collaboration tools into their work environment. Agarwal (2010) describes that "a broad array of collaboration and data analytics tools are now available" to support scientific teams. Examples of scientific collaboration software include fluxdata.org and HUBzero. Fluxdata.org is a scientific collaboration portal that serves the community of scientists who analyze the FLUXNET carbon-flux synthesis dataset. The FLUXNET dataset consists of flux and meteorological data that have been collected worldwide and submitted to the central database fluxdata.org. These measurements are collected from a global network of 400 carbon-flux measuring sites, which provides carbon, water, and energy flux measurements (Agarwal et al., 2010; "Home - Fluxdata.org," n.d.).

HUBzero allows the creation of active websites that support scientific collaboration, scientific discovery, learning, and educational activities. This software provides researchers the ability to create groups, delegate responsibilities, and manage roles. Scientists are able to upload files, tools, and data; as well as use wiki and blog services. Additionally, there is support for social networking such as tagging, rating, commenting, and citations. HUBzero supports the ability for users to build projects, publish datasets, and use computational tools including simulation and modeling tools (“HUBzero | DataONE,” n.d., “HUBzero - Platform for Scientific Collaboration,” n.d.).

These are just two examples of the many scientific collaboration tools that have been recently created to assist scientists in their collaboration efforts. Although many collaboration tools exist there is still much work that needs to be done. Agarwal (2010) described that “although many collaboration portals have been designed to support science, many of them are not adopted by the intended users or are quite limited in functionality” (Agarwal et al., 2010, p. 2332).

Outside of data tools, there are many other infrastructure and interoperability factors that need to be considered including: provenance, metadata, ontologies, data and data models.

Provenance

Taylor defined provenance as “the origin of an archival document or collection…the origin may be an organization, officer, or person that created, received, or accumulated and used the item or the record in the collection” (Taylor, 2004, p. 375). More specifically, provenance is “information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness” (W3C, 2013). Cheah and Plale (2012) stated that data provenance is a key piece of metadata that describes the lifecycle of a data product and is crucial in a scientists’ ability to reproduce and reuse results.

Researchers have explored ways to ensure quality provenance. For example, Cheah and Plale (2012) examined provenance information for correctness, completeness, and relevance. This work examined correctness through contextual provenance and completeness though a structural analysis. Cheah and Plale suggested that researchers use data provenance to assess data quality, but none have taken this approach to evaluate the quality of the provenance itself. Cheah and Plale captured provenance for one month of data from different scientific data products and analyzed the provenance records. The findings indicated that the researchers were able to establish quality dimensions of correctness and completeness as a measure of provenance quality. Furthermore, the analysis suggested that provenance quality should be partitioned into a contextual and a structural problem (Cheah & Plale, 2012).

Barkstrom (2010) summarized provenance issues related to earth science data. The author described three distinct data production paradigms for earth science data, each of which have their own versioning structure. These include climate data record production, operational dataset production, and exploratory production. The author then described a mathematical framework developed for three provenance-tracking activities. These activities included: tracing the history of data production; tracing the history of the custody; and tracing the history of Intellectual Property Rights transfers. The author stated that much of the work on data provenance has emerged from the bioinformatics, chemistry, and nuclear physics community. Additionally the author described some key differences between earth science data and data from other fields, mainly that earth science data are produced and distributed as files that are often too voluminous to store in databases. The author also analyzed how these differences affect developments such as the Open Provenance Model. While this is an excellent resource, the

Open Provenance Model does not include support for Earth science error assessments, custodial provenance, or intellectual property rights provenance (Barkstrom, 2010).

DataONE provides a list of software tools to assist scientists with provenance concerns for scientific data. These include Bitbucket, CrowdLabs, Kepler, Pegasus, Taverna, and VisTrails.

Metadata

Taylor (2004) described metadata as an encoded description of an information package (Taylor, 2004, p. 371). More specifically as defined by Greenberg, metadata is "structured data about an object that supports functions associated with the designated object" (Greenberg, 2003, p. 1876). There are many types of metadata associated with scientific data, as well as many registries and other tools for scientific metadata. There are also various classifications of metadata including descriptive, structural, and administrative. Furthermore, most metadata for research datasets are "individual, personal, and non-interoperable" (Jeffery, Asserson, & Houssos, 2013, p. 2).

Scientific communities have investigated metadata for their community needs. For example, Michener et al. (1997) presented a study of generic metadata descriptors for nongeospatial ecologic data. The authors examined potential:

> benefits and costs associated with developing and implementing metadata for nongeospatial data; proposed a set of generic metadata descriptors; and presented alternative strategies for metadata implementations that meet different organizational or investigator-specific objectives (Michener et al., 1997, p. 331).

Additionally, researchers have investigated the need for effective standards of data and metadata exchange, specifically through creating systems for this process. Devarakonda and colleagues

(2010) described a number of data repositories serving field ecologists and how these repositories need solutions for distributed search. Therefore, scientists have turned to a harvest and index approach to enable integration of metadata from multiple sources. Mercury is a toolset that allows for metadata authoring, harvesting, indexing, and searching across a range of metadata standards including Federal Geographic Data Committee Content Standard for Digital Geospatial metadata (FGDC CDSGM), Ecological Metadata Language (EML), Global Change Master Directory's Director Interchange Format (GCMD DIF), Dublin Core, and ISO 19115. Mercury's purpose is to make it easy for users to find and use relevant data, as well as the associated metadata (Devarakonda et al., 2010).

Other developers have focused on the creation of metadata registries. As described in King et al. (2010), a registry is a "location in an organization where definitions are stored and maintained … metadata registries contains definitions for data models and resource registries contain definitions of an accessible resource" (King et al., 2010, p. 127). The authors described how these metadata registries and resource registries are fundamentally required for establishing interoperability and the ability to exchange data (King et al., 2010).

Ontologies

As defined in Taylor, an ontology is "a formal representation of language that identifies specific terms, usually from a defined subject area, and lays out the relationships that exist between terms" (Taylor, 2004, p. 373). Gruber described the ontology of a program by defining a set of representational terms, where definitions are associated with entities in the universe of discourse (e.g. classes, relations, functions, etc.) with human-readable text describing what the names mean and formal axioms that constrain the interpretation, and well-formed use of these terms (Gruber, 1995, p. 908). Ma and Fox (2013) provided a simple definition of ontology, “a shared conceptualization of domain knowledge” (2013, p. 31).

Ma and Fox (2013) conducted a study to investigate progress on geologic time ontologies. Projects such as OneGeology Europe and the United Stated Geoscience Information Network have applied ontologies to harmonize distributed geologic data and improve functionalities of web portals. The authors analyzed the structures and characteristics of several geologic time ontologies. The authors suggested that collaborative work should be conducted to share best practices, discuss design patterns, and build ontologies for the broader community. Furthermore, the authors suggested that geologic ontology work should address the challenges of balancing expressiveness, implementability, and maintainability (Ma & Fox, 2013).

Goranova, Shishedjiev, and Georgieva (2011) explored the development of a scientific data ontology as a way to formalize the management and integration of information, services, and processes. This work explored service-oriented data processing, searching, storing and visualization of scientific data from experiments and numerical simulations in a distributed and heterogeneous environment. The authors suggested that scientists need semantic approaches to formal description of experiments for analysis, annotation, and sharing of results. The authors suggested that ontologies would help achieve these goals and they propose the use of the base scientific data ontology (SDO). Additionally, the researchers examined other similar ontologies related to scientific data including: Earth System Grid ontology (ESG), Extensible Observation ontology (OBOE), Virtual Solar-Terrestrial Observatory ontology (VSTO), and the common ontology for scientific experiments (EXPO). The researchers suggested that scientists need an ontology to process heterogeneous data for modeling, analysis, and visualization, and proposed that the scientific data ontology (SDO) would be able to fulfill that need through generic scientific relationships (Goranova et al., 2011).

DataONE suggested Protégé, an open source ontology editor that allows users to create ontologies both in the Frames and Web Ontology Language (OWL) framework. Users can import ontologies, edit ontologies, create new ontologies, save ontologies in several formats including XML expressions of RDF and OWL, and visualize ontologies in graphical forms ("Protege" DataONE, 2013k).

Data and Data Models

Scientific data and scientific data modeling can be quite complicated. Heterogeneous data produced by different communities with varying practices and assumptions, and who organize data differently with different representation schemes, encodings, file formats create a difficult environment to implement data reuse and interdisciplinary science (Wickett, Sacchi, Dubin, & Renear, 2013). The following provides a summary some of the research being conducted regarding scientific data and scientific data models.

Patel, Okamoto, Dascalu, and Harris (2012) proposed a web-enabled software application to address some of the challenges related to data interoperability in scientific modeling and simulation. The proposed software toolkit dynamically generated data processing tools that are capable of performing conversions on data, based on the user's specifications. The authors described some of the issues related to data processing of geospatial data including: data conversions, data filtering, merging, sorting, grouping, and data scaling. Furthermore, the authors suggested that the development of web-enabled tools with specific subsystems include: user management, file formats, data structures, data structure operations, workflows, and dynamic code generators would considerably alleviate the current problems (Patel et al., 2012).

Additionally, researchers have explored the creation of conceptual models for data representation. Wicket, Sacchi, Dubin, and Renear (2013) presented two complementary conceptual models, the Basic Representation Model and the Systematic Assertion Model and demonstrated how these models together can provide an analytical account of scientific data. The authors discussed how there are two major classes of models: digital preservation models and scientific data models. Additionally, the authors discussed how scientific data models support retrieval and meaningful use and reuse of data, while digital preservation models identify the broader socio-technical environment. Examples of scientific data models include OBOE and the Semantic Sensor Network Ontology. Examples of digital preservation models include the Open Archival Information System (OAIS) and the Long-term Access through Networked Services (PLANETS). The authors suggested that both types of models are essential in the development of systems to support effective data curation; however since neither address the basic nature of data and datasets a gap still remains. Therefore, the authors suggested that the Basic Representation Model which accounts for key entities and relationships involved in the representation of digital objects, and the Systematic Assertion Model which accounts for how these relationships come to be established for scientific data, bridges this gap. Lastly, the authors provided an example from biodiversity data to demonstrate how these models together provided the missing account of entities and relationships involved in the creation and representation of scientific data (Wickett et al., 2013).

Some researchers have focused on models pertaining to types of study. For example, Parsons (2011) conducted a study to investigate how modelers in the Arctic research community access and use data for creation of models. Parsons described: (a) that the diversity of models makes understanding their data requirements complex, (b) data discovery is increasingly complicated, and (c) aspects of data handling for models (assessment, acquisition, preparation) have not been investigated. The author suggested that data should be independently understandable by a designated user community according to the OAIS model; however, user communities for a dataset can change over time. Data quality and peer-review of data are becoming increasingly important. For example, the new journal Earth System Science Data has been established for publication of high-quality data. Modelers must address issues of scale, coverage, and any techniques. The author proposed that there are consistencies in how modelers assess, acquire, and prepare their data, and these consistencies should inform how to design a data system that understands the scientists' needs. The author compared and contrasted the needs of different modelers in three case studies: The Community Sea Ice Model, Snow Model, and Multiple Element Model and provided recommendations. Recommendations included: (a) there should be coherence with the CMIP (Coupled Model Intercomparison Project) data formats, grids and naming conventions, (b) data centers should indicate which products are suitable for which applications, (c) data managers should provide multiple formats, and (d) data centers should have guidelines on how their data should be citation to give fair acknowledgment. The author also provided nine propositions for how to make improvements for the data modeling community, and indicated that one simple principle emerges which is that the data are more important than the system (Parsons, 2011).

Literature Review Conclusions

This literature review provided the background needed to understand topics important to this dissertation and the appropriate knowledge needed to recognize thematic research gaps within current literature. While there remains some literature that was not included, the literature review were analyzed and summarized to exhaustion in order to provide the background needed for this dissertation study. The following chapter provides an analysis of the research methods and theoretical frameworks utilized throughout this literature in order to analyze methodological research gaps and to provide insight into the research methods utilized for this dissertation.

CHAPTER III: RELEVANT RESEARCH METHODS AND THEORETICAL FRAMEWORKS

In preparation for this dissertation a broader contextual scope of the topics and themes pertinent to this research were examined and provided in Chapter II: Literature Review. Additionally, it was crucial to thoroughly investigate the methods and theoretical frameworks utilized throughout this research. The following chapter provides an analysis of these methods and theoretical frameworks, and as such certain points from Chapter Two are reemphasized as needed throughout the chapter. This section provides an overview of selected theoretical frameworks and research methods that have been used to study DataONE and data sharing and reuse and informs the study design of this dissertation.

Theoretical Research Specific to the DataONE

A number of studies have examined DataONE's organization and infrastructure from a theoretical framework including collaborative data sharing networks, complex adaptive systems, transdisciplinary organization, and uncertainty framework. The section below briefly describes these studies, as they were more thoroughly discussed in the previous chapter.

Allard and Allard (2009) investigated DataONE through the perspective of a transdisciplinary organization and concluded that DataONE does fit the criteria as indicated by the transdisciplinary index. A transdisciplinary organization exist when "researchers are from different disciplines and they work jointly to create a shared conceptual framework" (Allard & Allard, 2009, p. 2). Sayogo and Pardo (2011b) analyzed DataONE from several theoretical perspectives including: cross-boundary information sharing, integration frameworks, transnational knowledge networks, and collaborative networks. Ultimately this study suggested that DataONE is a collaborative data-sharing network (CDSN) and demonstrated this by examining components of DataONE's organization and infrastructure based on the five main characteristics of CDSNs.

Lagoze and Partzke's (2011) study explored DataONE through the uncertainty framework to determine factors that could impact long-term stability. The uncertainty framework provides "a conceptual model for understanding the dimensions and interdependencies of a problem space" (Lagoze & Patzke, 2011, p. 374). The findings indicated that two structural issues needed to be addressed for sustainability: (a) economic stability and (b) the creation of an administrative structure (Lagoze & Patzke, 2011).

Lastly, Aydinoglu's (2011) analyzed DataONE organization from the perspective of complex adaptive systems. The study was conducted during year two of DataONE, when the collaboration was still in the emerging stage. According to Aydinoglu, "if scientific collaborations behave like complex adaptive systems, they should demonstrate basic features of such systems" (Aydinoglu, 2011, p. 7). While there is not a unified complex adaptive system theory; there are concepts including agents, interactions, co-evolution, and emergence that can be used to examine systems. The author employed these tools and concepts to analyze DataONE (Aydinoglu, 2011).

Studies Specific to Data Sharing and Reuse

A growing body of research has employed a variety of research methods and theoretical frameworks to examine data sharing and reuse have employed methods including: literature reviews, surveys, interviews, observations, data deposition, bibliometric, and experimental studies. While some of these studies were described previously in Chapter Two, this section focuses on the methods or theories used in these studies.

Literature Review Research in Data Sharing and Reuse

Researchers have produced literature reviews and commentaries in order to gain an understanding of key topics, themes, and concerns for data sharing and reuse. For example, Sieber's (1988) research described a variety of data-sharing situations that occur within the science, for example the differences in sharing of quantitative and qualitative data. Additionally, this research provided tools for researchers regarding data sharing including equitable sharing agreements and described how to formulate data sharing agreements. Sieber described how scientists were very naive regarding data sharing, and suggested that the lack of guidelines is part of the problem contributing to this naivety. The author concluded that scientists are obligated to demonstrate the validity of their results and that ethical scientists should seek ways to formulate data sharing agreements (Sieber, 1988).

Researchers have described how sharing impacts the reputation of scientists. For example, Cohen's (1995) article focused on materials sharing, specifically knockout mice (genetically modified mice). This study stated that most researchers share knockout mice freely and willingly, but some have developed a reputation to be less willing. Additionally, the author suggested that the increase in cost and competition have challenged what has traditionally been a sharing culture in biological research. This article described some reasons for unwillingness to share, which included: (a) if the requester was in direct competition and (b) to protect postdoctoral researchers who were still using the mice for projects. Lastly, the article suggested that in these small communities, pressure from within the community to share are more likely to increase sharing rather than grant or journal policies (Cohen, 1995).

Baru's (2007) study examined the Geosciences Network (GEON) as a "useful case study to consider issues related to data discovery, integration, and provenance" (Baru, 2007, p. 114). Baru discussed the importance of metadata in sharing and reuse by describing the need for "what" (descriptive metadata), "how" (lineage, provenance), and the "why" (contextual information), in order to reuse data (Baru, 2007). This research promoted the GEON project as a way for scientists to share data and describes the types of metadata needed for reuse. Additionally, this was one of the few studies that reports specifically on the nuances of metadata, describes the intricacies of metadata, and addresses how metadata affects data sharing and reuse.

Lastly, researchers provided arguments for the benefits of sharing data. Borgman (2010) stated that data sharing is vital: "(a) to make the results of publically funded data available to the public, (b) to enable others to ask new questions of extant data, (c) to advance the state of science, and (d) to reproduce research" (Borgman, 2010, p. 2). Borgman further analyzes the four arguments for data sharing in her 2012 article by examining these arguments from a science, social science, and humanities perspective. Additionally, the author explores the motivations and incentives of stakeholders (Borgman, 2012).

Survey Research in Data Sharing and Reuse

Studies have employed survey methodology to investigate data sharing and reuse. Ceci's (1988) conducted two surveys to gain an understanding of scientists' opinion of data sharing including if scientists felt that it was their professional responsibility to share data and what would lead them to reject a colleague's request for sharing. The analysis indicated that scientists in all fields desire data sharing to be the norm (Ceci, 1988).

Researchers have been interested in scientists' reasons for refusing to share data. Blumenthal et al. (2006, 1997) surveyed scientists regarding data withholding. The 1997 study asked about delays in reporting of research results in order to: (a) allow for patent application, (b) protect scientific lead, (c) slow dissemination of underlying results, (d) allow time for negotiation of a patent, or (e) resolve disputes over ownership. Blumenthal's 2006 study surveyed scientists regarding their data withholding, and specifically measured personal characteristics and research characteristic (Blumenthal et al., 2006, 1997).

Researchers have also explored what influences scientists to publish their datasets. Sayogo and Pardo's (2013, 2011a) mixed methods study included visualization and an online survey. The survey results were analyzed using descriptive statistics and ordered logistic regressions. The dependent variable was the propensity of researchers to publish their datasets online or in a research network. The independent variables included: (a) organization involvement, (b) data management skills, (c) misinterpretation of data, (d) legal and condition on use, (e) appreciation and acknowledgement of data reuse, (f) economic motives, and (g) institutional requirements. Through data visualization, the researchers showed that most scientists only shared data within one network and very few researchers connected to several networks. These studies also showed how items such as data management skills and organizational involvement affected data sharing (Sayogo & Pardo, 2013, 2011a).

Tenopir et al. (2011) found similar factors associated with data sharing and reuse. This study described scientists' practices and perceptions of the barriers and enablers of data sharing through a two-part survey; the first portion focused on demographics and the second scientists' relationships with data. The findings included time, lack of funding, satisfaction with current practice, and lack of support for data management were all considered factors (Tenopir et al., 2011).

Interview Research in Data Sharing and Reuse

Researchers have conducted interviews in order to investigate data sharing and reuse. While many of the studies listed above also incorporated semi-structured interviews along with surveys; the research described below employed interviews as their main source of data collection.

Dr. Ann Zimmerman's (2003) dissertation and subsequent (2008) study employed semi-structured interviews in order to investigate secondary use of data among ecologists. Her main research question was: "What are the experiences of ecologists who use shared data" (Zimmerman, 2003, p. 110). Zimmerman conducted 20 face-to-face and telephone interviews with ecologists and conducted inductive open coding qualitative analysis in order to investigate which topics emerged in the data. Zimmerman described how qualitative analysis is suited well for exploratory investigations where little is known about the topic. Zimmerman stated that there is a lack of direct research and theory regarding sharing and reuse of data.

Zimmerman created her interview guide and questions based on a conceptual framework that drew from theories developed in the fields of history, philosophy, and sociology of scientific knowledge. These concepts included the theory of measurement as a social technology, from Theodore Porter, which is similar to Bruno Latour's concept of circulating reference. Furthermore, Zimmerman incorporated communities of practice and the notion of inscriptions and boundary objects from the sociology of scientific knowledge from Latour and Woolgar (1986); Star and Griesemer (1989), and Wenger (1998), (Zimmerman, 2003, p. 99). These theories provided Zimmerman a conceptual framework to investigate data sharing and reuse among ecologists (Zimmerman, 2003, 2008).

Data Deposition and Bibliometric Research in Data Sharing and Reuse

Several studies have analyzed data sharing by focusing on whether scientists deposited data alongside their research articles. These studies have examined scientists depositing data into repositories, examined journal policies, and in some cases have examined both data deposition and journal policies. The studies employed a variety of statistical methods, as well as content analysis to analyze data sharing.

McCain's (1995) study examined journal policies of natural science, medical, and engineering journals as well as compliance enforcement of these policies. The researcher reviewed 850 journals to investigate whether policies were included in these journals. The researcher found that out of the 850 journals, 132 contained some reference to research related information (RRI) in their policy statements. The author then analyzed the journal policies through content analysis and discussed the types of journals that included sharing policies in relation to sponsorships by learned societies, discipline, and types of data (McCain, 1995).

McCain conducted a follow-up study in 2000 which included electronic research related material and RRI. McCain conducted bibliographic analysis and determined if there was RRI associated with the article. McCain specifically explored the types of data made available by researchers such as data compilations, software, and documents; and furthermore cross-referenced these data types with disciplines. McCain was looking for data that researchers made available online and placed a URL in their article in order to provide readers access to their data (McCain, 2000).

Brown (2003) was interested in investigating both journal instructions and data deposition. The researcher gathered qualitative data through surveys and case studies to determine the use of data deposition into genomic and proteomic databases (GPD).

Additionally, the researcher collected bibliographic data to examine journal instructions regarding data deposition. The surveys were conducted to learn how scientists perceive GPD and additionally case studies explored how scientists used GPD in their everyday activities (Brown, 2003).

Noor, Zimmerman, and Teeter (2006) investigated how journal policies affected scientists data deposition practices, particularly scientists depositing DNA sequences into GenBank. The authors examined articles from six journals to see if data were deposited alongside the articles. The authors concluded that "none of the journals had complete compliance with the requirements for DNA sequences to be submitted to GenBank" (Noor et al., 2006, p. 1113). Ochsner et al. (2008) conducted a similar study, focusing on microarray datasets. This study explored articles from 20 journals and analyzed articles to see if the microarray data were deposited. The findings indicated that less than 50% of the data were deposited and noted that the amount of effort it took to deposit microarray data are particularly difficult due to their highly contextual nature and complicated metadata structure (Ochsner et al., 2008).

Piwowar and Chapman (Piwowar, 2011; Piwowar & Chapman, 2010) conducted studies that investigated data deposition of microarray data. In the first study, Piwowar and Chapman (Piwowar & Chapman, 2010) examined 397 biomedical microarray studies to investigate what factors influenced data deposition of the microarray datasets. The authors employed multivariate logistic regression to evaluate whether factors, such as impact factor, journal policies, number of authors, and authors career length influenced authors decisions to deposit data alongside their research article (Piwowar & Chapman, 2010). In a second study, Piwowar first identified 11,603 articles that were described to possibly have an associated microarray dataset deposited into

Gene Expression Omnibus (GEO) or ArrayExpress. The author then conducted first order factor analysis on 124 bibliometric attributes that revealed 15 factors that described authorship, funding, institution, publication, and domain environments. This analysis indicated that authors were most likely to deposit data if they had prior experience in sharing, had published in an open access journal, if the study was published in a journal with a strong data sharing policy, or if the study was funded by a large NIH grant (Piwowar, 2011; Piwowar & Chapman, 2010).

Observation Research in Data Sharing and Reuse

The Birnholtz and Beitz (2003) study observed three scientific disciplines to investigate data sharing. The three disciplines studied were: earthquake engineering, HIV/AIDS research, and space physics. Researchers observed scientists' day-to-day work activities and experiments, as well as conducted interviews. The authors focused on how data were used and how this influenced data sharing. Furthermore, the authors analyzed the data from the perspective of communities of practice, to address how data are implicated in the formation and maintenance of communities of practice. This was helpful in understanding the differences between experimentalists and theoretical modelers, as well as their different approaches to data sharing. Additionally, they analyzed how the data enabled inbound trajectories. This research used both observations and aspects of communities of practice to investigate data sharing and provided implications for design of CSCW systems for supporting data sharing (Birnholtz & Bietz, 2003).

Experimental Design Research in Data Sharing and Reuse

Constant, Kiesler, and Sproull (1994) conducted three experiments on attitudes about sharing technical work and expertise within organizations. The researchers employed exchange and expressive theories of information sharing in their experiments. In experiment 1, the researchers investigated information sharing as an exchange and used interdependence theory to derive predictions. The researchers employed a between-subject experimental design where participants were given a vignette stating that a coworker had refused to help fix a computer bug and is now asking for a copy of a computer program. The results indicated that information ownership were correlated information sharing and organizational ownership were moderately correlated with information sharing. Additionally work experience was positively correlated with information sharing. Experiment 2 explored attitudes about sharing intangible information and experiment 3 directly compared attitudes about sharing a computer program and sharing computer experience. The findings indicated that the correlation between organizational ownership and sharing are significant in computer program sharing (Constant et al., 1994).

Synthesis of Results

Synthesis of Theoretical Frameworks

Table 2 provides a summary of the theoretical framework, as well as the strengths and weaknesses for its application in the context of DataONE and data sharing and reuse.

Table 2

Synthesis of Theoretical Frameworks

Theoretical Framework	**Summary Of Topic(s) Theoretical Framework Explored**	**Strengths Of Theoretical Framework**	**Weaknesses And/Or Areas Of Opportunity Of Theory In Given Context**
Transdisciplinary Organization (TO)	Researchers are from different disciplines and they work jointly to create a shared conceptual framework	• Organizational understanding • Transdisciplinary Index determines if organization is a TO	• Organizational perspective • Does not investigate scientists perspective
Collaborative Data Sharing Network (CDSN)	Characteristics: • Collaboration of heterogeneous, autonomous, geographically dispersed, and interorganizational social actors • Members share common and compatible goals • Information may flow one-way or bi-directional; • Collaboration is mediated within a	• Provided understanding of collaboration • Provided understanding of data within collaboration • Discussed infrastructure	• Organizational perspective • Does not investigate scientists perspective

	trusted network • Collaboration is supported with an interoperable infrastructure		
Uncertainty Framework	Explored the DataONE for technological, organizational, scientific, and institutional uncertainty	• Determined any factors that could affect long-term stability in the organization	• Organizational perspective • Does address scientists perspective, however not regarding data sharing and reuse
Communities of Practice (COP) (Wenger, 1998)	• COP based on social theory of learning • Components include: Meaning, Practice, Community, and Identity	• Useful to investigate how behavior is learned in scientific settings or the scientific community as a whole	• Has not been employed to consider why individual scientists are motivated to share data and reuse • Focused too much on learning and community
Exchange and Expressive Theories (Chadwich-Jones, 1976; Meeker, 1971)	Explored topics such as: • Reciprocation • Redistribution • Obligations • Interdependencies • Contingencies	• Described the psychological and social psychological aspects of information sharing	• Has not been employed in many studies to examine information sharing of scientists • *Could serve as a possible theoretical foundation for data sharing and reuse*

Synthesis of Research Methods

Table 3 provides a summary of the research methods, as well as the strengths and weaknesses for its application in the context of DataONE and data sharing and reuse.

Table 3

Synthesis of Research Methods

Research Method	Summary Of Topics Method Explored	Strengths Of Method	Weaknesses And/Or Areas Of Opportunity Of Method In Given Context
Literature Review	These studies provided a summary of aspects of data sharing including: • Data sharing agreements • Materials sharing and reputation	• Provided an overview of important topics related to data	• Does not provide empirical research on these topics

	• Organizations that promote data sharing • Arguments for why data sharing is important	sharing and reuse	
Survey Research (Hank, Jordan, & Wildemuth, 2009)	These studies explored factors that influenced data sharing including: • Beliefs and opinions • Professional responsibility • Field, federal support, types of institution • Data withholding • Lack of time, funding, support for data management • Satisfaction with current practice	• Gathered information about beliefs, opinions, attributes, behaviors, attitudes, perceptions, and other psychological constructs	• These studies were conducted in many contexts, however, not specifically related to earth science observational data • Would be useful to get a broad indication and understanding of DataONE user's beliefs, perceptions, and practices of data sharing
Interviews (Luo & Wildemuth, 2009; Zhang & Wildemuth, 2009)	These studies explored secondary use of data amongst ecologists including: • Experiences of those who use shared data • How they locate data? • Characteristics of data received • Information about the data they receive and/or depend on to use the data? • Assess the quality of the data they receive? • Challenges do secondary data users face?	• Gathered information about people's experiences and inner perceptions, attitudes, and feelings of reality. • Provided an in-depth understanding of a phenomenon	• Very specific to one user community, in this case ecologists • *Would be useful to include in current study to gain an in-depth understanding of data sharing and reuse*
Citation and Bibliometric analysis	These studies specifically explored: • Journal Policies • Grant Policies • Data deposition • Discipline • Length of Career • Data Types • Institution Types • Journal Impact Factor	• Quantified citation and bibliographic information	• Studied mainly data deposition, whether the data are actually deposited along side the research article, but does not include whether the deposited data are reused or reusable
Observations (Wildemuth, 2009)	These studies explored: • Day to day activities in three scientific settings • Explored from perspective of communities of practice (COP) • Focused on how datasets used in these scientific practices to understand how data gets shared	• Gather accurate information about events • Gather more precise data about the timing, duration, and frequency of behaviors • *Could be useful to understanding when scientists share and reuse data*	• Many information behaviors are intermittent, the context may be difficult to observe, obtrusiveness of being watched

Experimental Design	This study explored, information sharing, information ownership, and organizational ownership. Used a 2x2 factorial design: Between Factors: - Perspective (information processor vs. information seeker) - Information type (computer program vs. computer expertise) Repeated Measures: - Attitudes (organizational ownership and attitudes about sharing)	- Experimental control and randomization allows researcher to rule out all the possible observed effects except for the effect of interest - Discerns the causes of the phenomenon under study	- Has not been used frequently to explore data sharing and reuse. - *Would be very useful to understanding the factors that effect data sharing and reuse*

Conclusion

The purpose of this chapter was to assist in making appropriate choices in the research design of this dissertation. While many of these methods are shown to be very useful in the understanding of data sharing and reuse and DataONE, a mixed method approach may be the most beneficial course of action to understand data sharing and reuse within the context of DataONE. The following chapter elaborates further by exploring the thematic and methodological gaps in current research, and the rationale for this dissertation study.

CHAPTER IV: RESEARCH PREPARATIONS AND RATIONALE FOR STUDY

Research preparations were conducted through a threefold approach (a) a literature assessment and gap analysis, (b) a pilot profiling data assessment study, and (c) a pilot think-aloud study.

Literature Assessment and Gap Analysis

The literature reviewed in Chapters 2 and 3 demonstrated a research gap that has informed and shaped this dissertation research. As described, while there has been extensive research on the topic, there are both (a) a topic/thematic research gap and (b) a methodological research gap in current literature regarding data sharing and reuse. Figure 10 provides a summary of this gap analysis.

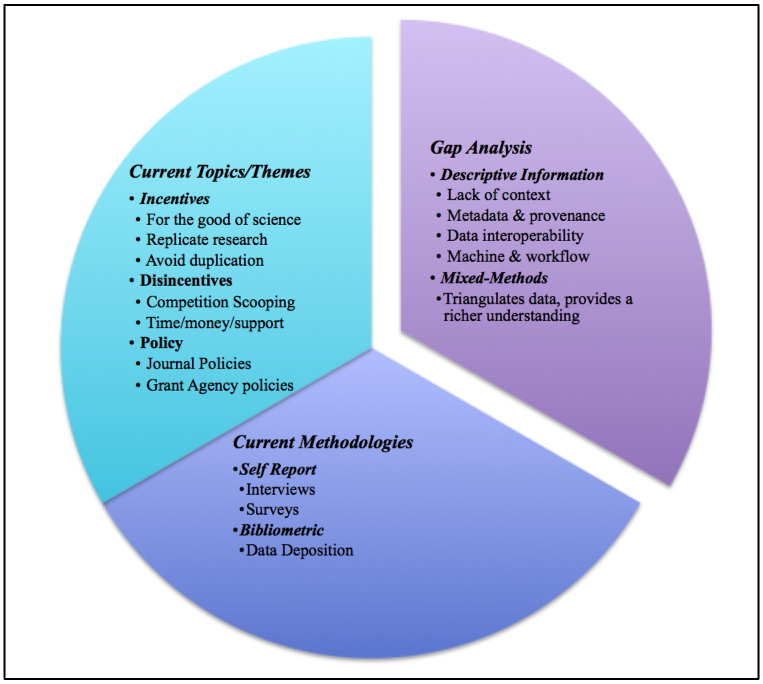


Figure 10. Thematic and Methodological Gap, and Gap Analysis

Thematic Gap

Researchers have a very good understanding of incentives for data sharing (for the good of science, avoiding duplication, etc.) and disincentives (competition, time, money, support, etc.). Researchers are also aware that data deposition policies from journal and grant agencies have placed pressure on scientists to deposit data. Additionally, there are factors such as new scientific data management tools and changes in the scientific process driven by the data deluge and fourth paradigm.

Many factors have not been addressed in previous research regarding data sharing and reuse such as types of descriptive information that facilitate or inhibit data sharing and reuse. These include: a lack of context (Baru, 2007; Zimmerman, 2003), lack of metadata (Anderson, 2004; Edwards et al., 2011), inadequate provenance information (Cheah & Plale, 2012), absence of or poor data interoperability (Borgman, 2000b; Ma & Fox, 2013), workflow inconsistencies (Ailamaki et al., 2010), and a conglomeration of these factors (Murillo, 2013; Murillo & Ramdeen, 2013). The dissertation specifically addresses this thematic gap in the current research.

Methodological Gap

Central to this dissertation is the methodological gap that is found in literature review. Research methods previously applied in this topic area include: literature reviews (Anderson, 2004; Baru, 2007; Bell et al., 2009; Borgman, 2010, 2012; Cohen, 1995; Sieber, 1988), surveys (Blumenthal et al., 2006, 1997; Ceci, 1988; Lord & Macdonald, 2003; Sayogo & Pardo, 2013, 2011a, 2011b; Tenopir et al., 2011), interviews (Zimmerman, 2003, 2008), data deposition and bibliometric research (Brown, 2003; McCain, 1995, 2000; Noor et al., 2006; Piwowar, 2011; Piwowar & Chapman, 2010), observational studies (Birnholtz & Bietz, 2003), and experimental research (Constant et al., 1994). As shown, the majority of this research has been through self-report providing opinions of sharing, or investigations of policies and data deposition. Few studies have examined types of descriptive information and how new cyberinfrastructure such as DataONE facilitates data sharing and reuse.

Pilot Studies

In addition to the literature assessment, two pilot studies were conducted including a profiling data assessment and a think-aloud study. These pilot studies were informed by the literature assessment and assisted in the development of the final research design for this dissertation.

Pilot Data Profiling Assessment

A pilot data profiling assessment was conducted through a content analysis of a random sample of 650 metadata records extracted from DataONE ONEMercury. This analysis provided an understanding of data that are being deposited in DataONE.

Table 4 provides a summary of the results of this pilot study. Forty-five XML records were analyzed from the random sample. These records indicated that there were three top-level metadata elements common in most of the records. These included dataset metadata, access metadata, and additional metadata. All of the records contained dataset metadata, while ½ contain additional metadata, and 1/3 contained access metadata. Additionally, many of these records contained metadata recording other items such as research methods, permissions, and attributes. The "top-level metadata" is specific to EML metadata, as well as the categories of dataset, access, and additional metadata.

While the robustness of these records varied greatly, the pilot study provided an opportunity to begin a preliminary codebook as shown in Appendix 2. Additionally, this analysis provided an opportunity to understand which variables to consider qualitative and qualitative variables. Furthermore, this pilot study provided information to enable the creation of a robustness scale that was used in the final codebook as shown in Appendix 3. Preliminary results of this pilot study were presented at the 2014 ASIS&T Annual Conference in Seattle (Murillo, 2014).

Table 4

Pilots Study: Data Profiling Assessment

Methods	Analysis of 45 XML Records from a random sample
Findings	• The records contained three top-level metadata elements which included: o Dataset Metadata (100%) o Access Metadata (33%) o Additional Metadata (51%) • Only 25% contained all three top-level elements • Dataset o Research methods (60%) o Metadata standard (Majority EML) o Provenance information (22%) • Access o Permissions o Intellectual property rights • Additional Metadata o Attribute and Unit list definitions o Metadata Provider/Creator o Contact Information o Keywords/Keyword Thesauri • The majority contained full abstracts • Robustness varied
Purpose	• Assists in creation of preliminary codebook located in Appendix 2.

Pilot Think-Aloud Study

A pilot think-aloud study was conducted at the DataONE All Hands meeting in November 2014. This study was conducted to gain an understanding of data reuse within

DataONE and what information users needed to determine data reusability. This study provided guidance for the research design of this dissertation.

Table 5 provides a summary of this pilot study. Six users searched the DataONE ONEMercury system for data to reuse. Participants were asked to think-aloud and describe how they determined data reusability. From this study, it was determined that decisions were based on the metadata record/snippet provided by DataONE. Some of the important factors included knowledge of the PI, robustness of metadata, data type (once downloaded), and information regarding research methods.

This pilot study reinforced that a think-aloud provided a useful way to understand how users make decisions regarding reuse. Additionally, the think-aloud provided a more accurate and natural way to examine data reusability than interviews or surveys. The study also reinforced that some metadata elements were more important to users than others, and additionally reinforced the idea that robustness of metadata played a role in how scientists choose data to reuse.

Table 5

Pilot Think-Aloud Study

Methods	• Asked participants to search DataONE system for data to reuse. • Asked participants to describe how they deemed data reusable or not. • There were a total of 6 participants.
Findings	• Scientists made decisions based on the metadata snippet they received • Important factors: o Knowledge of PI o Data Type o Robustness of metadata in general o Research methods information o Key metadata elements such as provenance information, unit and attribute lists, and full

	abstract
Purpose	• To gain an understanding of what factors influenced reuse

Rationale for Study

These research preparations provided the knowledge and experience needed to determine the final research design for this dissertation study and provided a rationale for this dissertation study. The literature assessment made the research gap apparent and the pilot studies confirmed this research gap while assisting in the development of the methods. The below summarizes this rationale.

Although many studies have led to a more thorough understanding of data sharing and reuse, there are multiple factors that have not been captured by this previous research and there is a research gap in the literature. For example, researchers have an understanding of incentives and disincentives regarding data sharing and reuse, however types of descriptive information need to be further investigated. Furthermore, while researchers have an understanding of DataONE infrastructure, they do not have a thorough understanding of how this infrastructure supports data sharing and reuse. Additionally, many of the methods to investigate these topics have been conducted through self report, literature reviews, and observations, however they have not used mixed methods approaches to test how descriptive information influences data sharing and reuse.

This mixed methods approach was selected for this dissertation because it addresses this gap in research both thematically by addressing data sharing, data reuse, and descriptive information, as well as methodologically. This approach triangulates data from several sources and provides both a quantitative and qualitative approach which will lead to a richer understanding of the research problem, as well as minimize biases through balancing data from several sources (Kennedy, 2009). The use of mixed methods is incredibly useful to assist in understanding a problem. Neuendorf suggests that "quantitative and qualitative research may be viewed as different ways of examining the same problem…this triangulation of methods strengthens the researcher's claim" (Neuendorf, 2002, p. 15). Additionally, "most scholars agree that the 'best' approach is one of triangulation...with a variety of methods – experiments, surveys, and other more qualitative methods…the various methods' strengths and weaknesses tend to balance out" (Neuendorf, 2002, p. 49). This dissertation uses a quantitative and qualitative content analysis and a quasi-experiment think-aloud study to provide this triangulation in the hope to balance out the strengths and weaknesses of each method.

CHAPTER V: RESEARCH QUESTIONS AND TERMINOLOGY

This chapter presents both the overarching and the targeted research questions, describes and operationalizes key concepts in order to ensure clarity, and provides a discussion of the assumptions of this dissertation research.

Research Questions

The overarching research question posited in this dissertation is: *How does descriptive information influence scientists' ability to share and reuse data within DataONE?* DataONE served as the test environment for this dissertation. Targeted research questions include:

RQ1: What types of descriptive information are being made discoverable through DataONE?

- How robust is the descriptive information made available regarding that data?
- How is information being provided about the data, such as information regarding metadata standards, provenance information, research methods, instrumentation?
- How is the provision of this descriptive information impacting the data-sharing infrastructure?

RQ2: What types of descriptive information could inhibit or facilitate data reuse?

- How is information about the data such as information regarding metadata standards, provenance information, research methods, and instrumentation, influencing scientists' ability to determine if that data is reusable?
- How does this information assist scientists in their ability to reuse this data?

Context and Terminology Operationalization

In order to ensure clarity this section provides a brief summary of the terms and concepts that are relevant to this dissertation. Terminology that is discussed includes: infrastructure, cyberinfrastructure, interoperability, and data sharing and reuse. Terminology specific to DataONE includes: DataONE Member Nodes, DataONE ONEMercury, EML Metadata Elements, and the robustness scale. This section does not cover all aspects of scientific data management or scientific data sharing; there are numerous online resources available including the Digital Curation Centre (Digital Curation Centre, 2004), DataONE Educational Modules (DataONE, 2013g), and International Council for Science: Committee on Data for Science and Technology (ICS CODATA, 2015).

Table 6 provides short definitions of terminology discussed in further detail in this chapter and are important for understand the context of this dissertation.

Table 6

Brief Definitions

Term	**Definition**
Infrastructure & Cyberinfrastructure	The technology, systems, practices, and people needed to communicate, access, and share information.
Interoperability	The compatibility of systems to exchange information without manipulation.
Data Sharing and Reuse	Data sharing is making data available for reuse. Data reuse is repurposing data.
DataONE Member Node	Established organizations that expose their metadata to DataONE Coordinating Nodes for replication, indexing, and search.
DataONE ONE Mercury	DataONE online search interface, open to the general public to search for data.
Metadata Element	An individual category or field that holds an individual piece of description of an information package (Taylor, Miller, & Taylor, 2006, p. 533).
“Top-Level” Metadata	Term used to describe the top-level of metadata reported in the Ecological Metadata Schema. Important to this dissertation since so much of the metadata within DataONE uses EML as

	their metadata schema.
Immutable data	Data that are no longer changeable.
Active data	Data that are changeable or can be changed.
Data grid	A data grid provides middleware services and applications to pull dispersed data together. A good example of a data grid is the DataNet Federation Consortium (DFC). Data grid is defined as: a set of structured services that provides multiple services like the ability to access, alter and transfer very large amounts of geographically separated data, especially for research and collaboration purposes. Data from different regions are pulled from administrative domains which filter data for security purposes, and present it to the user upon request by means of a middleware application (Technopedia, 2016).
Centralized Union Catalog	A union catalog represents the holdings of more than one institution or collection (Taylor, 2004, p. 35).
Robustness	Term used to describe amount of metadata provided in each element. The robustness scale is: 0 – no information 1 – adequate information 2 – comprehensive information.

<u>Infrastructure and Cyberinfrastructure</u>

At a very high level an infrastructure can be defined as “a framework in which communications networks support higher-level services for human communication and access to information” (Borgman, 2000a, p. 26). From this perspective, one could view the entire DataONE organization as an infrastructure, as well as the DataNets as a whole. Borgman suggested that the integration, interaction, and interdependence of information-related tasks and activities have led us to think in terms of an information infrastructure. Borgman added that the term infrastructure is being used to describe “a set of technologies … principles … communications networks … aggregation of people, technology, and content” (Borgman, 2000a, p. 18). Star and Ruhleder provided examples of infrastructure as railroads, telephones, energy, and banking, and stated that "with the rise of decentralized technologies used across wide geographically distance, both the need for common standards and the need for situated, tailorable, and flexible technologies grow stronger" (Susan Leigh Star & Ruhleder, 1996, p. 122).

Atkins (2003) provided a historical discussion of infrastructure and described how the term infrastructure has been used since the 1920s to refer to the roads, power grids, telephone lines, etc., that are required for an industrial economy to function. Atkins (2003) described that the newer term cyberinfrastructure refers to infrastructure based on distributed computing, information, and communications technology. As stated by Atkins, "if infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (Atkins et al., 2003, p. 5). Additionally, the authors described the base technologies needed for cyberinfrastructure including: computing, storage, communication technologies, software programs, data, social practices, and personnel are part of cyberinfrastructure (Atkins et al., 2003).

Liz Lyon described significant gaps in the current infrastructures for scientific data by stating that "data in diverse formats are at the heart of the research process but there are significant gaps in infrastructure to effectively share, manage, curate, preserve and potentially reuse the rapidly growing volumes of data" (Lyon, 2012, p. 127). The author described how an infrastructure encompasses the hardware and software components for data integration, manipulation, recombination and storage, but also requires a human infrastructure (Lyon, 2012).

There various types of cyberinfrastructures that supports data sharing and reuse. Centralized union catalogs are one type. In DataONE, Member Nodes expose their metadata and the Coordinating Nodes expose that metadata and make it available for users to access. Data

Grids are another option in that they provide middleware services and applications to pull dispersed data together. In the case of the DataNet Consortium Federation, this creates a national data infrastructure to enable collaborative research on shared data collections through managing the collection lifecycle.

For the purpose of this dissertation two specific parts of DataONE infrastructure were investigated; the data that was deposited through DataONE Member Nodes and the reusability of this data through DataONE ONEMercury.

Interoperability

Interoperability is another important term that should be described for clarity. Taylor defined interoperability as "the compatibility of two or more systems to exchange information and data and can use the exchanged information and data without any special manipulation" (Taylor, 2004, p. 369). Hodge described interoperability as "a condition under which dissimilar entities, systems, or… standards, can be interfaced" (Hodge, 2008, p. 26). Lanz et al. suggested that data interoperability allows users of information system to share information, and cited the International Organization for Standardization definition as "the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units", (Lanz, Brandli, & Baltensweiler, 2007, p. 99).

Lanz et al. (2007) described requirements for interoperability including comprehensive metadata, distributed data storage, and scalability (Lanz et al., 2007, pp. 103–107). There are factors that influence interoperability among standards including: metadata profiles, metadata registries, shared vocabularies, and crosswalks to assist with the interoperability of standards (Hodge, 2008). Bargmeyer and Gillman suggested that metadata standards and metadata registries promote interoperability between organizations, systems, and people (Bargmeyer & Gillman, 2014).

Data Sharing and Reuse

As described in the previous section, this dissertation investigates both the data sharing (making data available) and the data reuse (reusing available data) sides of data sharing and reuse. Figure 11 describes the data sharing and reuse cycle.

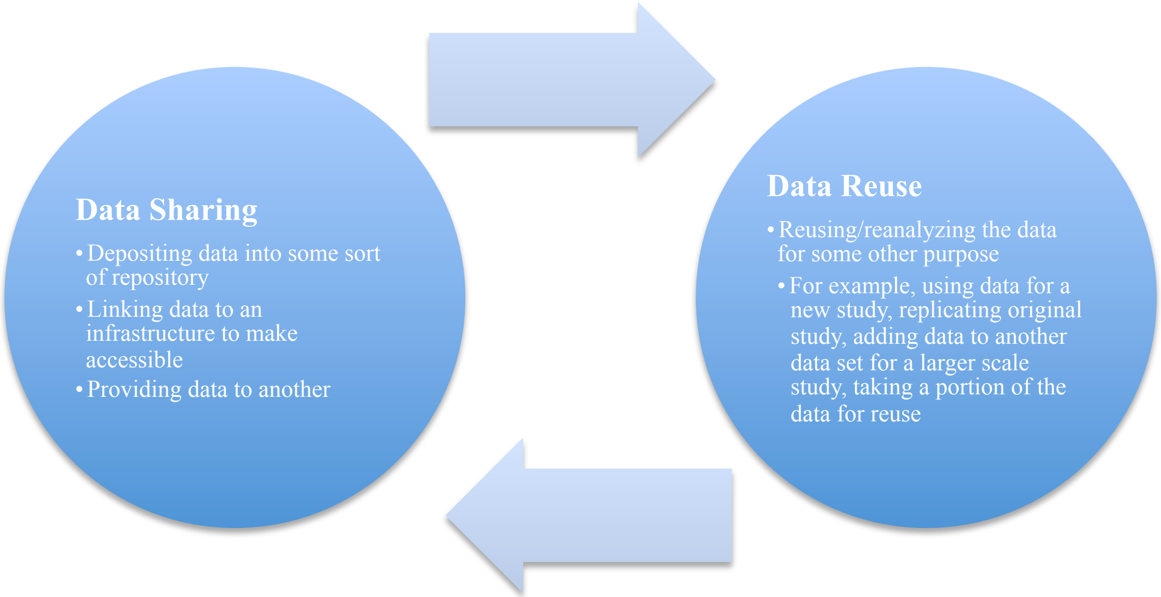

Data Sharing
• Depositing data into some sort of repository
• Linking data to an infrastructure to make accessible
• Providing data to another
Data Reuse
• Reusing/reanalyzing the data for some other purpose
• For example, using data for a new study, replicating original study, adding data to another data set for a larger scale study, taking a portion of the data for reuse

Figure 11. Data Sharing and Data Reuse Cycle

This study examines both aspects within the same environment, DataONE. For the purpose of this dissertation the phrase "made discoverable" is used to describe data that are being shared through DataONE Member Nodes. This terminology is used due to the fact that in some cases it is not the data that are shared, but a copy of the metadata is shared that describes the data. Additionally, for this dissertation, the focus is on published data. While there are scientific networks that share mutable and intermediate data products such as data grids, this is outside of the scope of this dissertation.

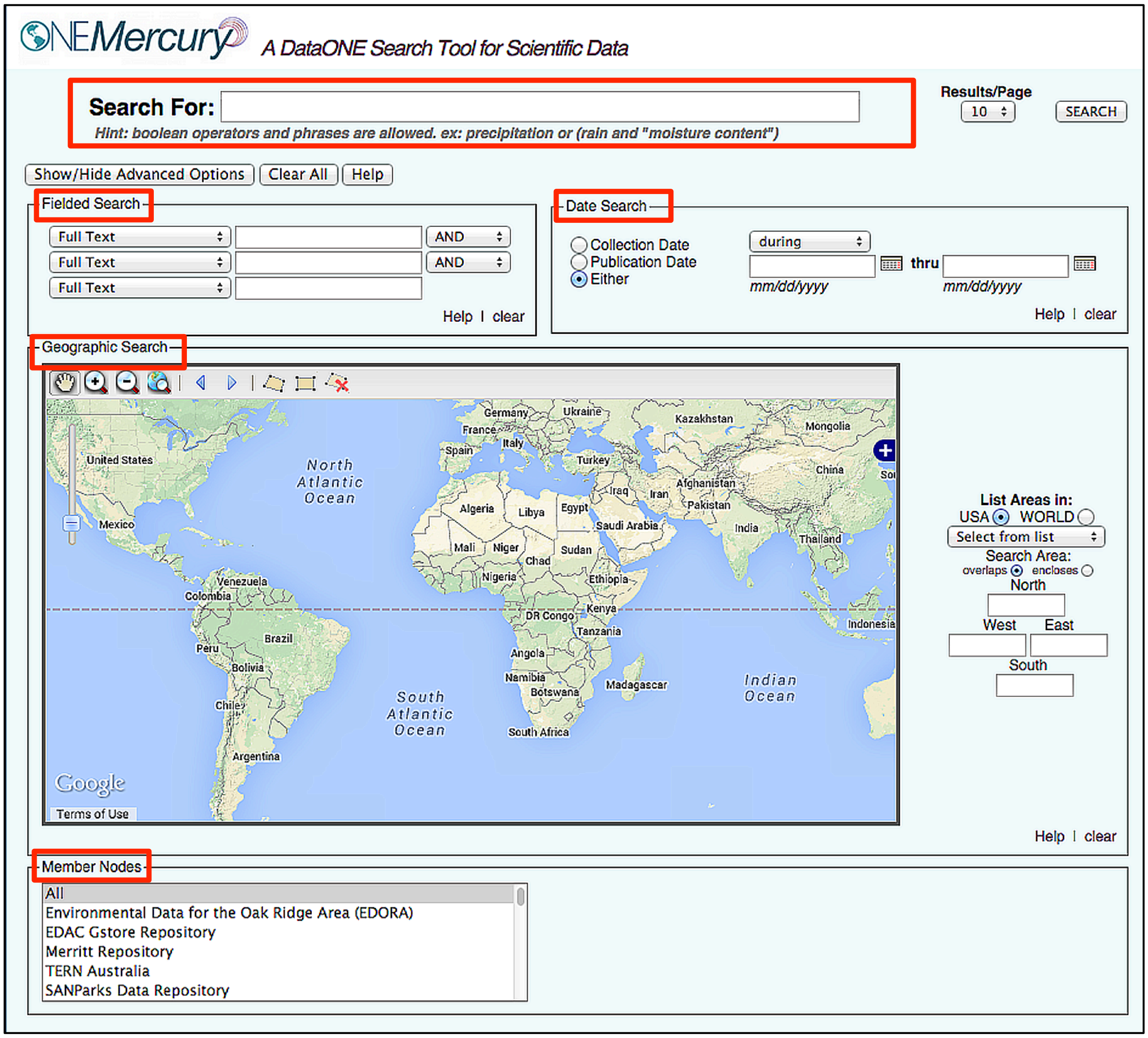
DataONE Member Nodes

As described in Chapter Two, DataONE Member Nodes exposes their data and metadata through a common set of interfaces and services, the Member Node service interface (DataONE, 2013c). Metadata and data get deposited into one of the current Member Nodes through that Member Node's submission process. Once the data are published through that Member Node, it is available in DataONE and should work with the various items in the Investigator Toolkit, including the DataONE ONEMercury described in the next section (DataONE, 2013b).

For the purpose of this dissertation the Member Nodes are considered those nodes that were in existence at the point of data collection which occurred fall through spring 2014-2015. The researcher is aware that there are new Member Nodes that have been added since data collection. See Appendices 1 and 4 for a list of the past and current Member Nodes.

DataONE ONEMercury

DataONE ONE Mercury is the DataONE’s web-based search interface that allows users to find data within DataONE. DataONE ONEMercury (https://cn.dataone.org/onemercury/) has a simple full-text search interface that allows users to search by date, geographic, content type, and Member Nodes. Figure 12 shows the initial search page and the various facets of the search page highlighted with red boxes.



Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and "moisture content")

○ Collection Date
○ Publication Date
● Either

USA ● WORLD ○

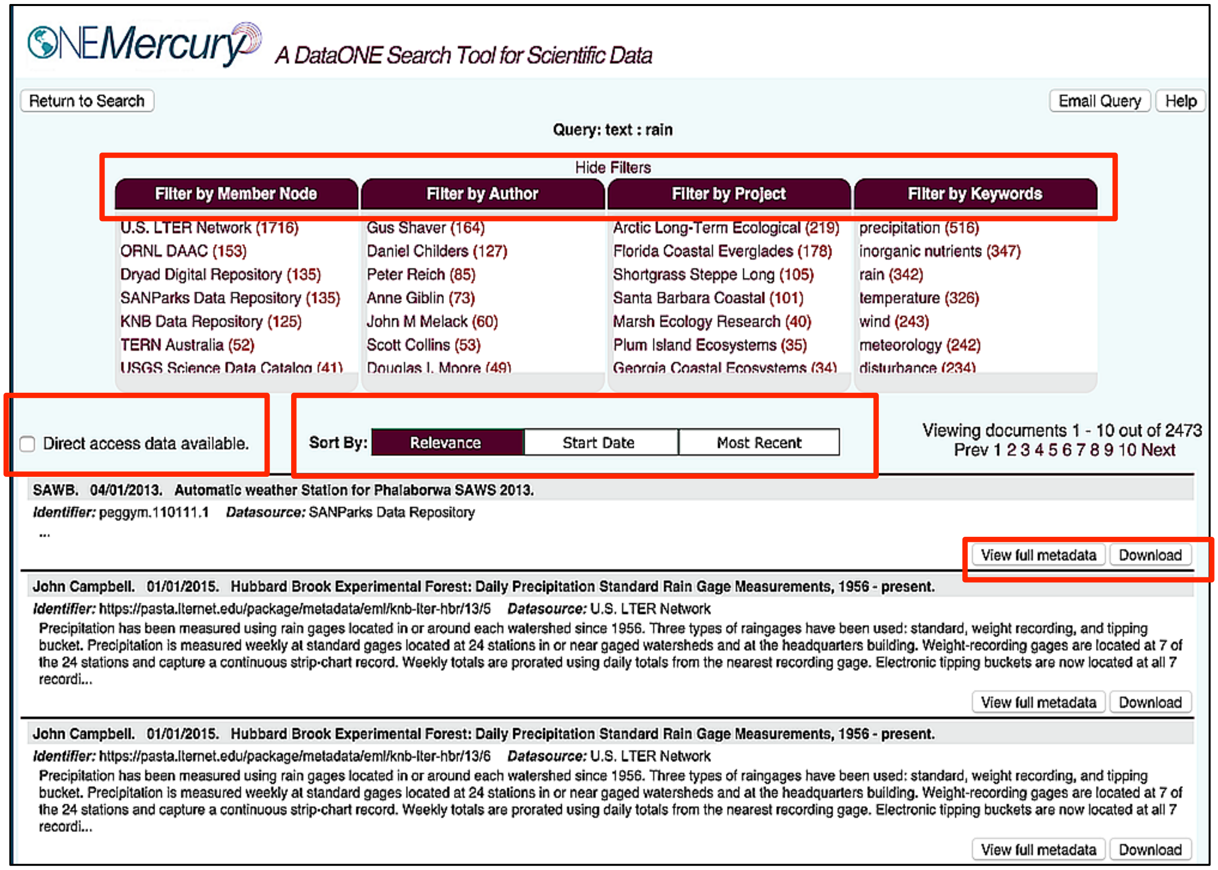
Select from list

overlaps ● encloses ○

All
Environmental Data for the Oak Ridge Area (EDORA)
EDAC Gstore Repository
Merritt Repository
TERN Australia
SANParks Data Repository

Figure 12. The DataONE ONEMercury, Search Interface

Figure 13 shows the results page of DataONE ONEMercury from an example search of "rain." As is shown in the figure below, users can filter results by (a) member node, (b) author, (c) projects, and (d) keywords. The search can be sorted by (a) relevance, (b) start data, and (c) most recent. These results can be further distinguished by results that have "direct access data available". Red boxes in Figure 13 highlight important facets of the interface that allow users to filter through their search.



Query: text : rain

Filter by Member Node	Filter by Author	Filter by Project	Filter by Keywords
U.S. LTER Network (1716)	Gus Shaver (164)	Arctic Long-Term Ecological (219)	precipitation (516)
ORNL DAAC (153)	Daniel Childers (127)	Florida Coastal Everglades (178)	inorganic nutrients (347)
Dryad Digital Repository (135)	Peter Reich (85)	Shortgrass Steppe Long (105)	rain (342)
SANParks Data Repository (135)	Anne Giblin (73)	Santa Barbara Coastal (101)	temperature (326)
KNB Data Repository (125)	John M Melack (60)	Marsh Ecology Research (40)	wind (243)
TERN Australia (52)	Scott Collins (53)	Plum Island Ecosystems (35)	meteorology (242)
USGS Science Data Catalog (41)	Douglas L. Moore (49)	Georgia Coastal Ecosystems (34)	disturbance (234)

☐ Direct access data available.

Sort By: Relevance | Start Date | Most Recent

Viewing documents 1 - 10 out of 2473
Prev 1 2 3 4 5 6 7 8 9 10 Next

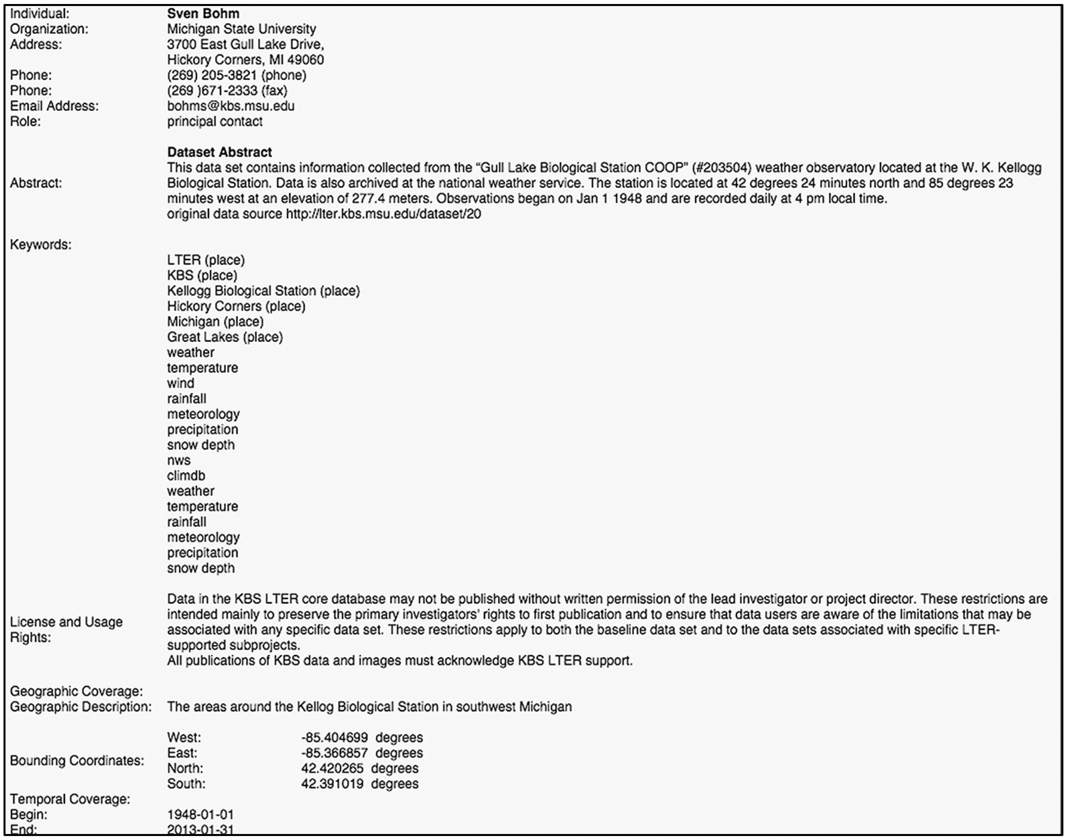
SAWB. 04/01/2013. Automatic weather Station for Phalaborwa SAWS 2013.
Identifier: peggym.110111.1 ***Datasource:*** SANParks Data Repository
...

John Campbell. 01/01/2015. Hubbard Brook Experimental Forest: Daily Precipitation Standard Rain Gage Measurements, 1956 - present.
Identifier: https://pasta.lternet.edu/package/metadata/eml/knb-lter-hbr/13/5 ***Datasource:*** U.S. LTER Network
Precipitation has been measured using rain gages located in or around each watershed since 1956. Three types of raingages have been used: standard, weight recording, and tipping bucket. Precipitation is measured weekly at standard gages located at 24 stations in or near gaged watersheds and at the headquarters building. Weight-recording gages are located at 7 of the 24 stations and capture a continuous strip-chart record. Weekly totals are prorated using daily totals from the nearest recording gage. Electronic tipping buckets are now located at all 7 recordi...

John Campbell. 01/01/2015. Hubbard Brook Experimental Forest: Daily Precipitation Standard Rain Gage Measurements, 1956 - present.
Identifier: https://pasta.lternet.edu/package/metadata/eml/knb-lter-hbr/13/6 ***Datasource:*** U.S. LTER Network
Precipitation has been measured using rain gages located in or around each watershed since 1956. Three types of raingages have been used: standard, weight recording, and tipping bucket. Precipitation is measured weekly at standard gages located at 24 stations in or near gaged watersheds and at the headquarters building. Weight-recording gages are located at 7 of the 24 stations and capture a continuous strip-chart record. Weekly totals are prorated using daily totals from the nearest recording gage. Electronic tipping buckets are now located at all 7 recordi...

Figure 13. The DataONE ONE Mercury Search Results

When users click on “View Full Metadata” they are provided with a record of the data. Figure 14 provides a brief example of a partial metadata record that users would see when clicking on the “view full metadata” button. From here users can determine if they want to download the data for use.



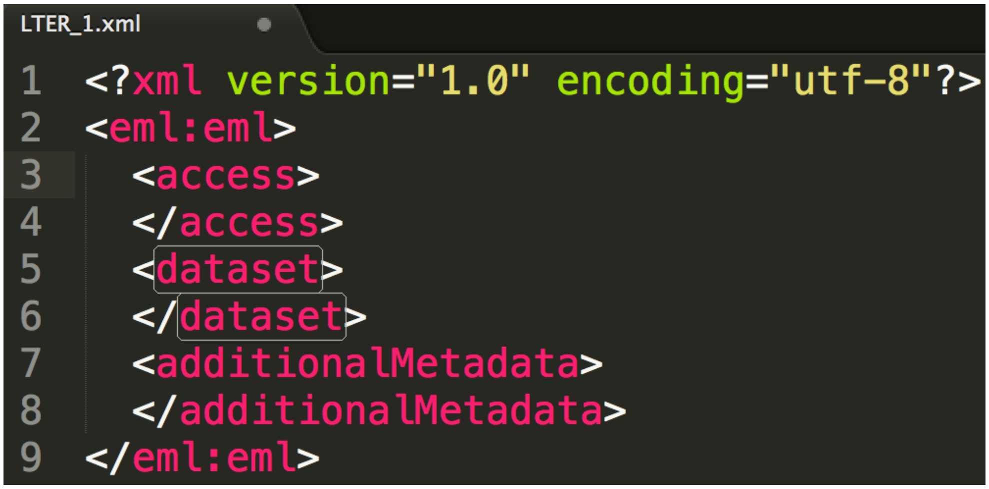
Individual:	**Sven Bohm**
Organization:	Michigan State University
Address:	3700 East Gull Lake Drive, Hickory Corners, MI 49060
Phone:	(269) 205-3821 (phone)
Phone:	(269)671-2333 (fax)
Email Address:	bohms@kbs.msu.edu
Role:	principal contact
	Dataset Abstract
Abstract:	This data set contains information collected from the “Gull Lake Biological Station COOP” (#203504) weather observatory located at the W. K. Kellogg Biological Station. Data is also archived at the national weather service. The station is located at 42 degrees 24 minutes north and 85 degrees 23 minutes west at an elevation of 277.4 meters. Observations began on Jan 1 1948 and are recorded daily at 4 pm local time. original data source http://lter.kbs.msu.edu/dataset/20
Keywords:	LTER (place) KBS (place) Kellogg Biological Station (place) Hickory Corners (place) Michigan (place) Great Lakes (place) weather temperature wind rainfall meteorology precipitation snow depth nws climdb weather temperature rainfall meteorology precipitation snow depth
License and Usage Rights:	Data in the KBS LTER core database may not be published without written permission of the lead investigator or project director. These restrictions are intended mainly to preserve the primary investigators' rights to first publication and to ensure that data users are aware of the limitations that may be associated with any specific data set. These restrictions apply to both the baseline data set and to the data sets associated with specific LTER-supported subprojects. All publications of KBS data and images must acknowledge KBS LTER support.
Geographic Coverage:	
Geographic Description:	The areas around the Kellog Biological Station in southwest Michigan
Bounding Coordinates:	West: -85.404699 degrees East: -85.366857 degrees North: 42.420265 degrees South: 42.391019 degrees
Temporal Coverage:	
Begin:	1948-01-01
End:	2013-01-31

Figure 14. Metadata Record

EML Specific Metadata Elements

During the creation of the codebook (See Appendix 2 and 3) for the data profiling assessment a selected set of metadata elements became apparent and emerged from the data. This was based on the metadata that was made discoverable through DataONE ONEMercury and took the form of the metadata standards that were used when the record was deposited into DataONE ONEMercury.

Given that the majority of the data followed the EML standard, the codebook was influenced regarding what metadata was available and established this idea of “top-level metadata” in that EML’s hierarchy included: (a) dataset metadata, (b) access metadata, and (c) additional metadata. Figure 15 below shows the example of the LTER-1 record, which follows the EML format. The top-level elements include dataset, access, and additional metadata.



```xml
<?xml version="1.0" encoding="utf-8"?>
<eml:eml>
  <access>
  </access>
  <dataset>
  </dataset>
  <additionalMetadata>
  </additionalMetadata>
</eml:eml>
```

Figure 15. EML Top-Level Metadata Structure

Secondary elements fall underneath these three top-level elements. For example, under "dataset metadata" there were the elements of "title, creator, abstract, and methods"; under "access metadata" there was "permissions"; and under "additional metadata" there was "unit list, attribute list". This significantly influenced the creation of the codebook, as it followed the EML structure described on the Knowledge Network for Biocomplexity (KNB) website (Knowledge Network for Biocomplexity, 2015).

Robustness Scale

For this dissertation, a robustness scale was created to measure the amount of information provided by each metadata element. From the Oxford English Dictionary robustness means: "strong and hardy; strongly and solidly built, sturdy; healthy" and "powerful, full" (Oxford English Dictionary, 2016c). This term has been used by others to describe the fullness and hardiness of metadata (Reichman et al., 2011).

For this dissertation a robustness scale was created: 0 – no information, 1 – adequate information, and 2 – comprehensive information. As described in the Oxford English Dictionary, adequate means "Satisfactory, but worthy of no stronger praise or recommendation; barely reaching an acceptable standard; just good enough" (Oxford English Dictionary, 2016a).

Lastly, the term comprehensive means: “having the attribute of comprising or including much; of large content or scope” (Oxford English Dictionary, 2016b). These definitions guided the researcher during the data profiling assessment to measure the robustness of each variable.

Assumptions

The research questions and the mixed methods research design provides a framework for studying how descriptive information facilitates or inhibits data sharing and reuse. Prior to this research there are several assumptions that need to be addressed. These assumptions have been informed both the literature review and by my previous research and experience with the user community of DataONE and the earth sciences in general. These assumptions are:

- Scientists, at least in some sectors, want to share and reuse each other’s data.
- New infrastructure such as DataONE need to be to examine how they inhibit and facilitate data sharing and reuse.

These assumptions are based on the literature, which suggests that scientists believe that data sharing and reuse benefits science. Scientists are required to publish data through journal and grant policies, and therefore know that it is in their best interest to try to adhere to these requirements. The literature review demonstrated that we have a good understanding of the policy and behavioral motivations for sharing, and reuse (i.e. journal policies, fear of scooping, important for science). However, in assessing the literature, it became apparent that we do not have a good understanding of whether these systems support sharing and reuse since many of them, including DataONE are quite new and are in need of testing.

CHAPTER VI: RESEARCH METHODS AND PROCEDURES

This chapter provides the selected research methods, sampling approaches, procedures, recruitment, and data analysis activities for each method used in this dissertation study. The methods include:

- A profiling data assessment conducted as a content analysis, and
- A quasi-experiment think-aloud study.

Each of the research methods was selected to examine the targeted research questions described in the previous chapter. Table 7 indicates which method addresses which targeted research question.

Table 7

Research Questions and Research Methods

Research Question	**Research Method**
Research Question 1 *1. What types of descriptive information are being made discoverable through DataONE?* *a. How robust is the descriptive information made available regarding that data?* *b. How is information being provided about the data, such as information regarding metadata standards, provenance information, research methods, instrumentation?* *c. How is the provision of this descriptive information impacting the data-sharing infrastructure?*	Data Profiling Assessment (Quantitative and Qualitative Content Analysis)
Research Question 2 *2. What types of descriptive information could inhibit or facilitate data reuse?* *a. How is information about the data such as information regarding metadata standards, provenance information, research methods, and instrumentation, influencing scientists' ability to determine if that data is reusable?*	Quasi-Experiment Think-Aloud Study

b. How does this information assist scientists in their ability to reuse this data?	

Data Profiling Assessment (Content Analysis)

The profiling data assessment analyzed the types of data and metadata being made discoverable through DataONE. This assessment was conducted through a quantitative and qualitative content analysis of metadata records extracted from DataONE's – ONEMercury; the online search interface of DataONE (https://cn.dataone.org/onemercury/).

The metadata records extracted from DataONE provided a source to address research question #1. These extracted records provided information regarding what data standards, metadata standards, robustness of the descriptive information, and additional information were provided from the data available through DataONE. Through this examination the researcher was able to discern what information was made available to scientists as they search for data, and determine which factors were included or excluded such as structured metadata, provenance information, and research methods information. The data profiling assessment provided the ability to analyze data that are currently being shared in a live and active environment. This provided the opportunity to understand the best types of examples to use during the second part of this dissertation study, the quasi-experiment think-aloud study.

Content Analysis Context

In order to address research question #1, a content analysis was conducted. The details of this method were developed from Neuendorf's "The Content Analysis Guidebook". The content analysis included both quantitative and qualitative components and was exploratory and emergent. This tactic provided the best possible and most logical way to answer and address the research question and gain an understanding of the data and metadata being made available through the DataONE. There are many uses for content analysis including description, hypothesis testing, and facilitating inference (Neuendorf, 2002), and in the case of this dissertation research, a descriptive content analysis.

Variable identification and codebook creation were a vital part of this dissertation and took an ample amount of rigorous examination through the pilot study and preliminary analysis before the final codebook was created. This followed Neuendorf's description in that "exploratory work can and should be done before a final coding scheme is 'set in stone' and the entire process may be viewed as a 'combination of induction and deduction'" (Neuendorf, 2002, pp. 11–12). In order to ensure the accurate representation of the data made available through DataONE, this codebook was created through a combination of induction and deduction analysis.

This process followed the workflow described in Neuendorf (2002) in order to ensure the appropriate scientific rigor. These steps included: (a) rationale: a literature review was conducted in order to define the research questions; (b) conceptualization: variables were conceptualized through preliminary analysis; (c) operationalization: variables were operationalized and measures were defined; (d) human coding: a coding scheme was created, as well as a codebook and coding forms; (e) sampling; (f) training and pilot reliability; (g) coding; (h) final reliability; and (i) tabulation and reporting (Neuendorf, 2002, pp. 50–51).

This dissertation considered techniques for determining variables during the variable and codebook creation process. These recommended techniques included: (a) universal variables, (b) using theory and past research for variable creations, (c) a grounded or 'emergent' process of variable identification, and (d) attempting to find medium-specific critical variables (Neuendorf, 2002, pp. 101–107). For this dissertation, some of the variables are considered universal and medium-specific; for example, a metadata standard could be seen as a medium specific and universal. Additionally, past research was used to determine the variable collection, in that a pilot study was conducted. Additionally, an emergent process was used to identify variables with preliminary analysis of DataONE records.

This dissertation also considered the importance of variable types, which is important to determine and identify during content analysis in order to ensure accurate measurement for each variable. As the codebook was created and variables were conceptualized and operationalized, variables emerged that were countable, and a robustness scale was created. The codebook includes the variable name, definitions, and measurement, as seen in Appendix 3.

Regarding sampling, in content analysis there is an attempt to measure all variables as they naturally occur, which is typically a random sample. However in the case of this dissertation a stratified sample was used. A random sample was used for the pilot study and preliminary analysis, as well as the creation of the preliminary codebook. After careful consideration of what was learned from this pilot and preliminary analysis, a stratified sample was chosen as the sampling frame for the final sample for this dissertation. A stratified sample is "segmented according to categories on some variable(s) of prime interest to the researcher, this segmentation or stratification ensures appropriate representation for the various groupings" (Neuendorf, 2002, pp. 85–88). In this case 10 units from each Member Node were selected.

Sampling for Profiling Data Assessment

Training Sample (Random Sample)

During the spring 2014, a random representative sample of 650 xml records was extracted. This sample represented ONEMercury's corpus at the time of extraction, which had a population of 105,121 metadata records. This initial sample provided the training and teaching set for developing the codebook, as well as conducting the pilot study described in Chapter 4.

Approximately 7% of the training set was fully analyzed to develop the preliminary codebook that is located in Appendix 2. Additionally, a larger portion of the first random sample was analyzed from a cursory perspective. Through observations made from this initial analysis, and discussions with other scholars, a modification to the sample methods was made for the final dissertation sample. Table 8 shows the timeline of the samples created and subsequent analysis for each sample.

Table 8

Timeline for Sampling and Data Analysis of Data Profiling Assessment

Task	**Spring 2014**	**Summer 2014**	**Fall 2014**	**Winter 2014-15**	**Spring 2015**
Extraction of First Random Sample	🟥				
Pilot Study and preliminary creation of codebook	🟫	🟫			
Continue analysis and finalizing of codebook		🟩	🟩		
Final Stratified Sample			🟦		
Final Analysis				🟦	🟦

Modification to Sampling Method - Stratified Sample

While the first random sample was being analyzed and the creation of the original codebook was being developed, it was decided to change the original sampling method to a stratified sample. This was due to a problem with the original sampling method, which did not accurately account for specific Member Nodes having significantly more records available than others. The random sample did not accurately represent the entire set of data being made discoverable through DataONE. As is seen in Table 9, certain Member Nodes, such as Dryad, PISCO, and LTER, have a much larger representation than others. Without using a stratified sample, these Member Nodes were overly represented and smaller Member Nodes were underrepresented.

Final Stratified Sample for Dissertation Study

In order to obtain the most up-to-date stratified sample for the data profiling analysis, the researcher downloaded 10 records from each of the 19 Member Nodes listed below. Table 9 shows the 19 Member Nodes and the amount of metadata records available from each during the time of extraction.

Table 9

Member Nodes and Available Metadata Records

Member Node	**Metadata Records (183,702 as of 10/21/15)**
1. CLO eBird	2 (1)
2. ESA Data Registry	157
3. Dryad Digital Repository	25,838
4. Earth Data Analysis Center (EDAC)	357
5. Europe Long-Term Ecosystem Research Network (LTER Europe)	167
6. Gulf of Alaska Data Portal	481
7. Knowledge Network for Biocomplexity	5,625
8. LTER Network Member Node	45,489
9. Merritt Repository	31,590
10. Montana IoE Data Repository	*73 (none available)*
11. ONEShare Repository	127
12. ORNL DAAC	1,226
13. PISCO MN	68,101
14. SANParks Data Repository	*1,638 (none available)*
15. SEAD Virtual Archive	12
16. Taiwan Forestry Research Institute	2,383
17. USA National Phenology Network	14 (6)
18. USGS Core Sciences Clearinghouse	250
19. University of Kansas - Biodiversity Institute	172

To obtain the sample, the researcher searched DataONE ONEMercury by Member Node. From the results list the researcher filtered for the "most recent" records available. If it was obvious that there had been a bulk upload by the same person or entity, the researcher selected one from that bulk upload then moved to the next 9 records in order to ensure that there was a diverse representation from each member node. For Member Nodes that did not have 10 records available, the researcher chose all that were available. Additionally, the researcher emailed a copy of the records for backup of the original data. Furthermore, the researcher downloaded the XML file, and any additional files associated with the record for analysis.

The final sample was downloaded during October and November 2014. Due to server errors, records were not gathered from two Member Nodes (Montana IoE Data Repository and SANParks Data Repository), as indicated in Table 9. Furthermore, there were only two records available for CLO-eBird and only six records available from USA National Phenology Network. When choosing records from the Taiwan Forestry Research Institute only records that were majority English language were chosen. However, for the qualitative analysis on occasion Google Translate was used when needed to translate textual based data that was written in Chinese in order to contribute to the qualitative analysis.

The preliminary codebook (Appendix 2) was created during the pilot study and preliminary analysis. The final codebook (Appendix 3) was created during the final analysis conducted for this dissertation study.

Data Analysis for Profiling Data Assessment

A quantitative and qualitative content analysis analyzed the final sample of the DataONE records. Since the goal of research question #1 was to gain an understanding of what types of descriptive information are being made discoverable through the DataONE. This assessment examined what information was made available and the robustness of this information.

For the pilot study, the researcher analyzed 45 records from the first random sample that was extracted in the Spring 2014. This preliminary analysis was conducted in two distinct phases. During phase one, the researcher began noting potential variables of interest, examining what information was made available in the records, and noting potential hierarchies within these variables. From this first analysis, the researcher noticed a hierarchy in the records in the form of top-level metadata (dataset metadata, access metadata, and additional metadata).

This preliminary analysis continued with the completion of 45 records. The researcher used both deduction and induction methods when analyzing the records. Additionally, the researcher began to consider the robustness of the variables found in the data. For example, in regard to the variable "data citation", some records contained a brief statement such as "please cite this data" while others provided very specific details including preferred citation text. The researcher began to realize that there needed to be a way to identify this difference in robustness and decided to include a robustness scale, as can be seen in Appendix 2. Originally this scale included 4 categories: 0 – no information, 1 – very little information, 2 – some information, and 4 – a vast amount of information. It became clear that these were too hard to distinguish from each other, and the researcher decided to change it to only three categories prior to the final analysis.

Final analysis was conducted with the qualitative and quantitative content analysis of a total of 157 records extracted from DataONE ONEMercury. These 157 records were extracted using a stratified sample method. Upon extraction, the researcher first examined the first 20 records to continue the creation of the final codebook. Additionally, the research began creating definitions for the variables in an iterative inductive process.

The researcher worked through each record in detail, making decisions on the robustness of each variable, supplying yes or no information if the variable was dichotomous, or supplying the textual information for textually based variable. The final codebook, definitions, and measurements are located in Appendix 3, as well as Table 10 for ease of reading.

Table 10

Final Codebook, Definitions, and Measurements

Name	**Definition**
File Size	The file size of the corresponding XML file for the record. *(KB)*
Data Citation	Information regarding how to cite the data such as a suggested citation, DOI, etc. *(RS)*
Dataset Metadata	Top-level category regarding the dataset. *(RS)*
Access Metadata	Top-level category regarding access. *(RS)*
Additional Metadata	Top-level category regarding any additional metadata outside of information regarding access and/or the dataset. *(RS)*
Metadata Standard	The metadata standard used for the record. *(Full text)*
Additional Metadata Standard Information	Additional information regarding the metadata standard that was used for the record. *(RS)*
Provenance Information	Information regarding changes made to the record; includes change history and maintenance. *(RS)*
Instrumentation Information	Information regarding the instruments that were used to collect the data. *(RS)*
Research Methods Information	Information regarding the methods used to collect the data. (RS)
Associated Party	If there is another party involved with the data outside of the creator or metadata provider. *(0 = no, 1 = yes)*
Creator and Metadata Provider Information	The creator or the organization that provided the metadata for the record. *(Full text)*
Contact Information	Contact information for the creator or metadata provider. *(0 = no, 1 = yes)*
Publication Date	The date the data was published. *(Full text/numeric/date)*
Abstract	The brief summary typically of the project, which gathered the associated data. *(Full text & RS)*
Keywords	The keywords listed for the data. *(Full text)*
Additional Access Information	Additional information regarding access and use rights and restrictions. (Full text & RS)
Temporal Coverage	The date range covered by the data. *(Full text/numeric/date)*

Taxonomic Information	Information regarding the taxonomical coverage. *(RS)*
Publisher Information	Information regarding the publisher. *(RS)*
Data Type	The type of data available. *(Full text)*
Keyword Thesauri	If there is a thesaurus or thesauri associated with the keywords. *(Full text)*
Attribute List	List of variables in the data. *(RS)*
Unit List	Unit for each attribute or variable. *(RS)*
Geographic Information	Information regarding geographic location, bounding coordinate, etc. *(RS)*
Funding Source	The grant or the agency that funded the project. *(RS)*
Methods (part of abstract or own section)	Description of where the methods were located within the metadata snippet. *(0 = part of abstract, 1 = own section)*
Instrumentation (part of abstract or own section)	Description of where the instrumentation information was located within the record metadata snippet. *(0 = part of abstract, 1 = own section)*
Intellectual rights (multiple steps for use or none)	Description of where the instrumentation information was located within the record metadata snippet. *(0 = no steps, 1 = multiple steps)*
Data Availability	If the data is readily available. *(0 = no link to data/contact provider, 1= direct link to data, 2 = data directly available)*
Measurements	***RS*** = Robustness Scale – 0 = no information; 1 = adequate information; 2 = comprehensive information ***Full text***: Written text

Additionally, the researcher worked with one additional coder to complete the final analysis. The additional coder assisted with the creation of the codebook, and served the purpose of coder validation. The preliminary intercoder reliability was approximately 0.82 Krippendorff's alpha and the final intercoder reliability was approximately 0.91 Krippendorff's alpha.

Quantitative Analysis

Specific variables that were analyzed quantitatively were:

- Data citation,
- Dataset metadata,
- Access metadata,
- Additional metadata,
- Additional metadata standard information,

- Provenance information,
- Instrumentation information,
- Research methods information,
- Associated party,
- Contact information,
- Abstract,
- Taxonomic information,
- Publisher information,
- Attribute list,
- Unit list,
- Funding source,
- Instrumentation (part of abstract or own section),
- Instrumentation (part of abstract or own section),
- Intellectual property rights (multiple steps for use), and
- Data availability.

For the quantitative variables, there were two measurement types, meaning either the variables were dichotomous (present or not present) or the robustness scale was used to measure the robustness of their presence. Robustness evaluation was assessed in several steps. First, each record/variable pair was assigned robustness for the entire corpus. Then the researcher compared the robustness group for each variable for the entire corpus in order to ensure that there was consistency with the robustness being assigned. Additionally, the researcher made some decisions based on word count; for example, with the variable "abstract" it was decided that abstracts with a 0 to 10 word count were considered to have "no information" with a word count of 11-100 these were considered to have "adequate information"; and lastly with a word count of 101 onward these were considered to have comprehensive information. Using the word count assisted to make more precise decisions regarding robustness. Additionally, through this analysis it became apparent that some of the variables did have a qualitative component that needed further qualitative analysis.

Qualitative Analysis

Any variable that was considered qualitative in nature was analyzed using inductive qualitative analysis of content. The variables were textual based variables that included paragraph, sentence or word-level descriptions of variable for that record. For example, the variable additional access information typically contained several sentences of instruction regarding access and use rights for the data associated with the particular records. The variables that had a qualitative aspect were: (a) metadata standard, (b) creator and metadata provider information, (c) keywords, (d) additional access information, (e) data type, and (f) keyword thesauri. In some cases, counts were appropriate such as with metadata standard, keywords, and keyword thesauri. In these cases data cleaning needed to occur in order to get an accurate count, particularly with misspellings, abbreviations, or other non-conforming way to describe these variables. After cleanup, these variables were imported into NVivo to measure counts so that frequencies could be determined.

Additionally, two variables in particular required inductive qualitative analysis: (a) additional access information and (b) abstract. The data from these variables were first placed in their own excel worksheet and imported into NVivo 10. They were analyzed individually through inductive qualitative content analysis. NVivo is a commonly used system for inductive qualitative content analysis as it assists with the identification of themes.

Quasi-Experiment Think-Aloud Study

A quasi-experiment think-aloud study was conducted to test whether specific descriptive information facilitates or inhibits data reuse. The portion of the study asked participants to search DataONE ONEMercury (quasi-experiment think-aloud interface)[6] using a predetermined

[6] The researcher created a mirrored version of DataONE ONEMercury and created a quasi-experiment think-aloud interface. The results were placed in the correct order for the quasi-experiment to test the questions described

query and results set designed specifically to study how metadata robustness, abstract information, research methods information, and attribute/unit list information influenced the participants ability to determine if the data were reusable or not. Additionally, the participants were asked to talk aloud regarding how the information presented to them assisted in their ability to determine if the data were reusable. Lastly, they were asked to take a brief post-result survey, rank-order survey, and a post-task survey. The procedures will be described in 6.2.4.

The query and results were developed through the profiling data assessment and the researcher created a version of the system to conduct the study. Ultimately, it was determined instead to use a static version of the results in a Qualtrics survey in order to keep the participant from having to move between systems, and so participants could move smoothly between the think-aloud results and surveys. This decision was made based on witnessing some confusion and frustration for the participants during piloting and losing flow by having the participants move between systems.

Quasi-Experiment Overview

The quasi-experiment design for this dissertation resembles a counterbalanced quasi-experiment, which is typically used to test different search interfaces. However in this case, the purpose was to test what information is needed to determine data reusability.

The post-result usefulness survey allowed the researcher to determine the usefulness of the result. The rank-order survey assisted in determining which result was most useful after participants saw all four results. The post-search survey assisted with providing additional information regarding what influences a scientist's ability to determine if data are reusable. The think-aloud, allowed the participant to specifically point out any additional factors influencing

below. After several tests, it was determined to move static versions of the results in a Qualtrics survey in order to keep participants in one system rather than moving between two systems.

the participants' ability to determine if the data are reusable that were not included in the manipulated results.

The query "soil moisture content" was used and four results were created. This query and results were used because it was broad enough to be applicable to many sciences and the results were robust enough to be able to develop several versions of results for the quasi-experiment. Additionally, feedback I received from scientists during the pilot study verified my above rationale regarding the query and results.

The results and query were manipulated to test the following questions:

Q1: Does metadata robustness facilitate data reuse?

Q2: Does abstract availability facilitate data reuse?

Q3: Do research methods information facilitate data reuse?

Q4: Does attribute/unit information facilitate data reuse?

The researcher chose to design the study similarly to that of a counterbalanced quasi-experiment design. In a counterbalanced design, the comparison of interest is within each subject's performance in the multiple treatment conditions and therefore multiple treatments or interventions are applied to each subject (Hank & Wildemuth, 2009). In this case, the quasi-experiment think-aloud interface was created to determine if metadata robustness, abstract availability, research methods information, and unit/attribute information assisted participants in determining that the data are reusable. Table 11 provides the counter-balanced design.

Table 11

Counter-Balanced Design

Participant 1:	X1	O	X2	O	X3	O	X4	O
Participant 2:	X2	O	X4	O	X1	O	X3	O
Participant 3:	X3	O	X1	O	X4	O	X2	O
Participant 4:	X4	O	X3	O	X2	O	X1	O

For this study:

- X1 refers to Result 1, which contained robust metadata that included an abstract, a research methods section, and a unit/attribute list.
- X2 refers to Result 2, which contained basic metadata robustness, an abstract, and a research methods section.
- X3 refers to Result 3, which contained basic metadata robustness and an abstract.
- X4 refers to Result 4, which contained basic metadata robustness and a research methods section.

Think-Aloud Method Overview

In addition to the quasi-experiment, participants thought aloud as they were reviewing the search results. The purpose of this was to gain an understanding of how participants made decisions regarding the reusability of the data presented. Think-aloud protocols are useful in providing an understanding of the cognitive processes and knowledge acquisition involved in decision making, as well as alerting the researcher to any problems participants were experiencing during task performance (Oh & Wildemuth, 2009; Someren, Barnard, & Sandberg, 1994, p. i-2). In the case of this study, the researcher was interested in understanding the decision making regarding data reuse, as well as what information in the record assisted in determining the reusability of data. The think-loud provided participants the opportunity to discuss any additional factors that were influencing their ability to reuse the data presented that were not the focus of the researcher. The observation guide and interview guide are in Appendix 6.

The think-aloud protocol was chosen to answer research questions #2 regarding what descriptive information inhibited or facilitated data reuse. To understand this decision making on a deep level it was important to consider what constitutes decision-making and problem solving. Someren (1994) described how most problem solving involves a combination of two types of reasoning: constructing solutions and constructing justification for these solutions and described how think-aloud protocols allow the researcher to understand directly how these solutions and justifications are being constructed. In the case of reusing data, scientists are constructing solutions by thinking about if the available data fits their needs, and are constructing justification for these solutions by deciding which information about the data helps them decide or "justify" if the data fit their needs (Someren et al., 1994).

The think-aloud method can be used to acquire knowledge of a system (Someren et al., 1994). As described by Someren (1994) knowledge-based systems have the following characteristics: (a) expert level performance, performance is comparable to the level reached by humans who are specialized in a task, (b) they have a narrow task domain, and (c) the system can explain or justify its outcomes. DataONE ONEMercury contains two of the three of these characteristics; it has expert-level performance and a narrow task domain. Someren provided the example of "a knowledge engineer wanting to understand how a person carries out a task to build a computer system to do the same" (Someren et al., 1994, p. 14). This can be seen as somewhat analogous to research question #2, which seeks to understand how a user decides if the data within DataONE ONEMercury are reusable and to help provide feedback to DataONE systems and other similar systems.

Additionally, the think-aloud method was specifically chosen because of the strength of the method itself. All research methods have pros and cons in regards to cognitive disturbances, memory errors, and interpretation. For the think-aloud protocol, there are no memory errors, little to no interpretation errors, and while there may be some cognitive disturbances, they are minimal (Someren et al., 1994, pp. 24–25). Lastly, it was concluded through the pilot testing and research preparation described in Chapter 4 that the combination of the quasi-experiment and think-aloud was the best way to address research question #2. This combination allows for both control in the query and presentation of the search results, while providing participants the opportunity to describe their decision making process regarding data reusability.

Quasi-Experiment Think-Aloud Study Set-Up and Procedures.

Quasi-Experiment Setup

It was decided through the discussion with the committee to conduct this is a naturalistic setting and not a laboratory setting in order to be able to recruit at scientific conferences and meetings and have an open recruitment to scientists outside of academia. Therefore, the researcher created a quasi-experiment think-aloud interface that participants could use on the researcher's laptop so that the researcher could meet participants wherever it was convenient for them.

Pilot Testing

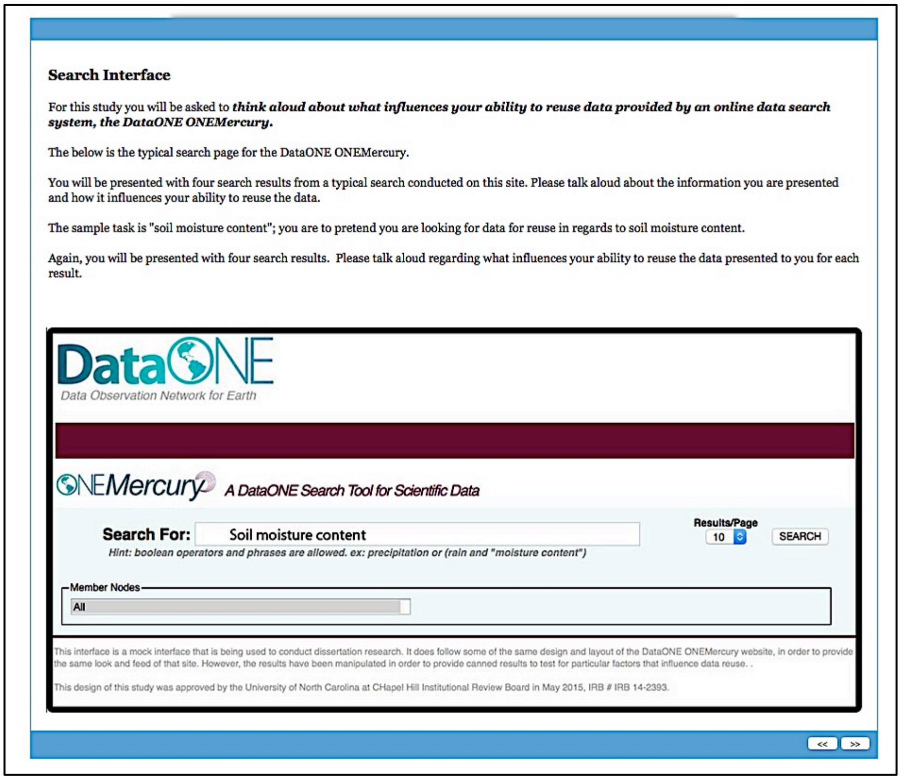
The researcher conducted five pilot tests in order to develop the quasi-experiment think-aloud interface and test the search tasks. Through the feedback from this testing, the researcher decided to move the entire interface into the Qualtrics survey system to assist with ease of participation. Additionally, through feedback from the pilot study participants, the researcher finalized the search task and sample search results for the dissertation study.

Task & Results Design

A realistic research task was used to create. The sample task was to search for "soil moisture content". This search was chosen based on the broad amount of potential results that it would provide, as well as that this search would be of interest to many scientists in various disciplines. Additionally, one of the search results was quite robust which provided the opportunity to create sample results based on the manipulation of results needed for the quasi-experiment think-aloud.

Quasi-Experiment Think-Aloud Interface with Survey in Qualtrics

Through feedback from the pilot test and the data profiling assessment the Quasi-Experiment Think-Aloud Interface was designed. As described earlier, the quasi-experiment think-aloud interface was placed within the UNC Qualtrics system, as the pilot testing indicated it was easier for participants to have all items in one system. Participants were first asked to read the consent form located in Appendix 7. Secondly, participants were provided the typical DataONE ONEMercury search interface as shown in Figure 16.



Search Interface

For this study you will be asked to ***think aloud about what influences your ability to reuse data provided by an online data search system, the DataONE ONEMercury.***

The below is the typical search page for the DataONE ONEMercury.

You will be presented with four search results from a typical search conducted on this site. Please talk aloud about the information you are presented and how it influences your ability to reuse the data.

The sample task is "soil moisture content"; you are to pretend you are looking for data for reuse in regards to soil moisture content.

Again, you will be presented with four search results. Please talk aloud regarding what influences your ability to reuse the data presented to you for each result.

DataONE
Data Observation Network for Earth

ONEMercury A DataONE Search Tool for Scientific Data

Search For: Soil moisture content

Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and "moisture content")

Results/Page: 10

Member Nodes: All

This interface is a mock interface that is being used to conduct dissertation research. It does follow some of the same design and layout of the DataONE ONEMercury website, in order to provide the same look and feed of that site. However, the results have been manipulated in order to provide canned results to test for particular factors that influence data reuse. .

This design of this study was approved by the University of North Carolina at CHapel Hill Institutional Review Board in May 2015, IRB # IRB 14-2393.

Figure 16. DataONE ONEMercury Sample Search Interface

Participants were able to practice the idea of “thinking aloud” through this example and were given more information on what they would be asked during the rest of the study. They were provided this page as both a way to understand the purpose of the study and provide context, as well as to get used to the “think aloud” process. By presenting the search interface first it seemed a more natural way to show participants the search results.

Next, participants were presented four search results based on the quasi-experiment setup. The full results are located in Appendix 11. Participants were presented the results using a counterbalanced design (Table 11).

Procedures and Data Gathered for Quasi-Experiment Think-Aloud Study

There were multiple points of data gathering for the quasi-experiment think-aloud portion of this dissertation study. Figure 17 provides a visualization of the procedures and points of data gathered. As shown below:

1. First participants were provided an overview of the study,
2. Then they read and agreed to the consent form,
3. Then the examined each result and “thought aloud”,
4. After each result they participated in the Post-Result Usefulness Survey (Appendix 9),
5. After the last result they partook in the Rank-Order Survey (Appendix 9),
6. Lastly, they answered the Post-Search Survey, which contained general questions, open-ended questions, a data reuse factors survey, and a demographic survey (Appendix 10).

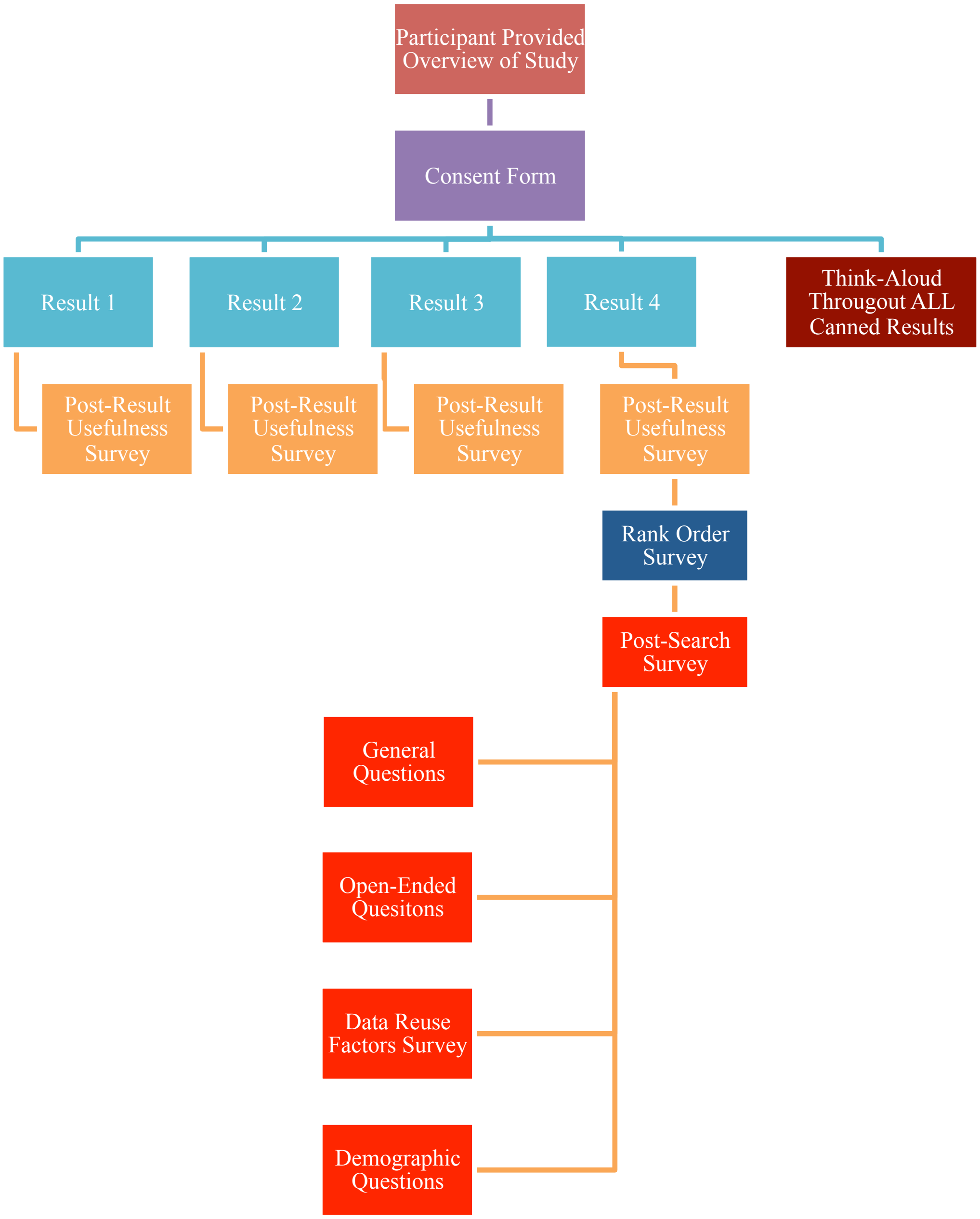

Participant Provided Overview of Study
Consent Form
Result 1
Result 2
Result 3
Result 4
Think-Aloud Througout ALL Canned Results
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Rank Order Survey
Post-Search Survey
General Questions
Open-Ended Quesitons
Data Reuse Factors Survey
Demographic Questions

Figure 17. Quasi-Experiment Think-Aloud Procedures/Data Collection

Recruitment for Quasi-Experiment Think-Aloud Study

Participants were recruited through sending emails to listservs of UNC and NC State scientific departments. Departments that were specifically targeted were geological sciences, environmental science, ecology and biology. Additionally, participants were recruited at the 2015 Annual Geological Society of America in Baltimore, Maryland. Furthermore, the CODATA listserv was used to recruit, as well as word of mouth.

Participants were recruited continuously throughout the summer 2015, fall 2015, and spring 2016. While participant recruitment was somewhat slow due to the specificity of the participants needed, the researcher was able to find 16 scientists that were interested in participating in the study. The researcher paid participants $20 in cash or an Amazon gift card for their time.

Data Analysis for Quasi-Experiment Think-Aloud Study

Basic descriptive statistics were conducted on the usefulness and rank order rankings, post-task survey, and demographic survey. Qualitative content analysis was conducted on the notes taken during the session, as well as the open ended questions that were answered during the post-task survey. The results of the data profiling assessment and quasi-experiment think-aloud were analyzed in conjunction with each other.

CHAPTER VII: RESULTS

Results Part One: Data Profiling Assessment

A sample of 157 DataONE records was analyzed for the data profiling assessment consisting of a quantitative and qualitative content analysis. Ten records were taken from each of the 19 DataONE Member Nodes in order to create the stratified sample. However, as described in Chapter 6 not every Member Node contained 10 usable records, therefore the sample of 157 were analyzed

The metadata of each record was typically chunked into three specific top-level categories. These top-level[7] categories included metadata in relation to:

- Dataset
- Access
- Additional metadata

This is due to the EML hierarchical structure as discussed in section 5.2.6. The dataset metadata has the greatest amount of lower-level categories compared to the access and additional metadata categories. Some of these lower-level categories include creator, metadata provider, publication date, abstract, maintenance, instrumentation information, and taxonomic coverage. Access metadata include permissions and other related information. The additional metadata include unit and attribute lists. The lower-level categories are described in further detail below in the detailed results.

[7] As many of the records followed the EML (Ecological Metadata Language) standard, these three categories in particular are a reflection of that standard, as can be seen in this example (https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html#N1003E)

Top-Level Metadata Results

Nearly all of the records sampled contain metadata in relation to the dataset, in fact only 1 record did not contain this information. Three-quarters of the records contain access metadata and 52% of the records contained additional metadata. These percentages, as well as the total counts are shown in Figure 18.

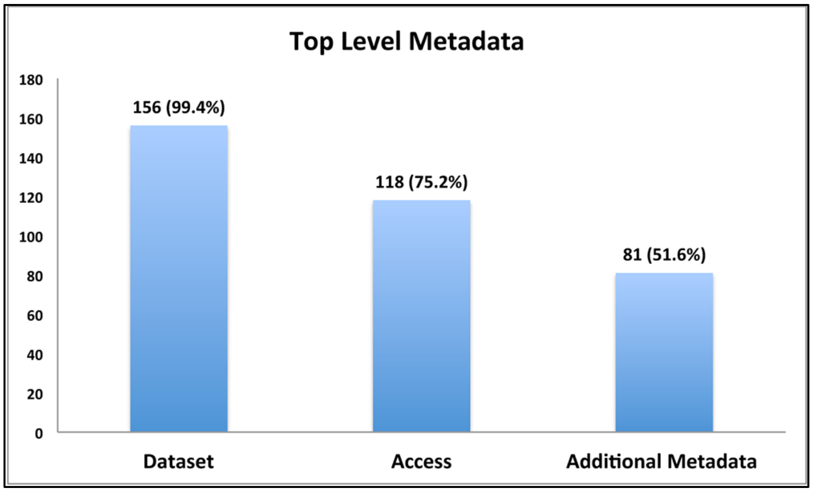


Figure 18. Percent and Count of Top-Level Metadata in Records

Robustness of Top-Level Metadata

The robustness of the dataset metadata is shown in Figure 19. Dataset metadata is the most robust of the top-level categories as it contains the most comprehensive records, 51% contains comprehensive information, and 48% contains adequate information.

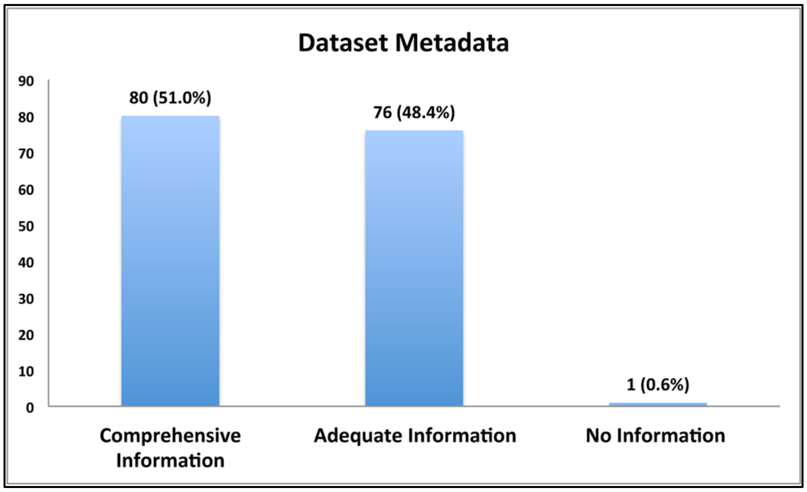


Figure 19. Percent and Count of Robustness of Dataset Metadata in Records

In regards to the top-level category of access metadata, only 4.5% contain comprehensive information, while 24.8% contain no information. From Figure 20, we can see the majority of the records (70.7%) did have adequate information.

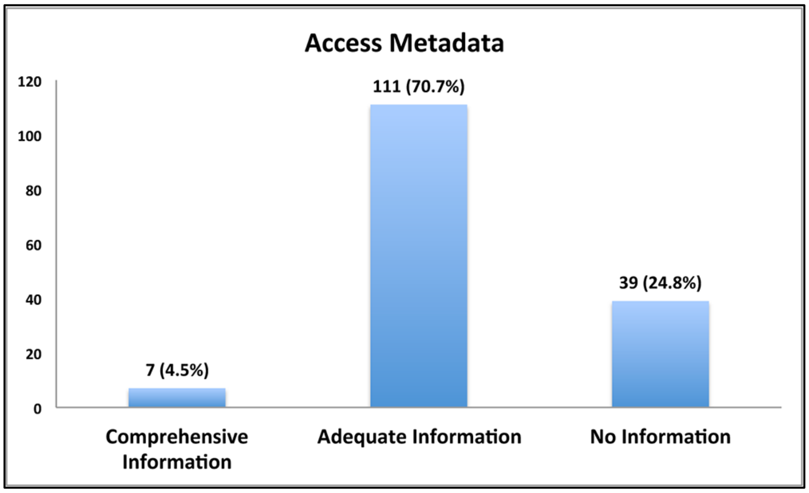


Figure 20. Percent and Count of Robustness of Access Metadata in Records

Lastly, very few of the records contain descriptive information in the additional metadata category. As shown in Figure 21, only 6.4% of the records contain comprehensive information. Additionally, 48.4% contain no additional metadata information.

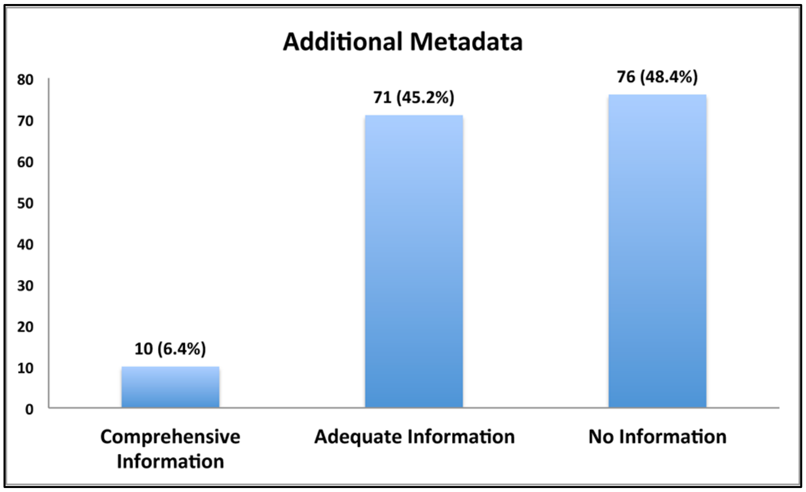


Figure 21. Percent and Count of Robustness of Additional Metadata in Records

Lower-Level/Detailed Results

To gain further insight regarding what descriptive information is being shared about the data, the records were analyzed for more detailed/granular information. This section describes the results from this analysis. The records were analyzed through iterative inductive and deductive quantitative and qualitative content analysis and the codebook was created. The findings below describe the remainder of the variables found in this process. Appendix 3 provides a list and definition of all pieces of descriptive information found in the data and analyzed, and the units of measurement.

File Size

The sample includes a range in file sizes for the source XML files from 4 KB to 2100 KBs. Several records also have multiple source files and in one case, the source did not download properly due to a corrupted file. The distribution of the files is shown in Figure 22.

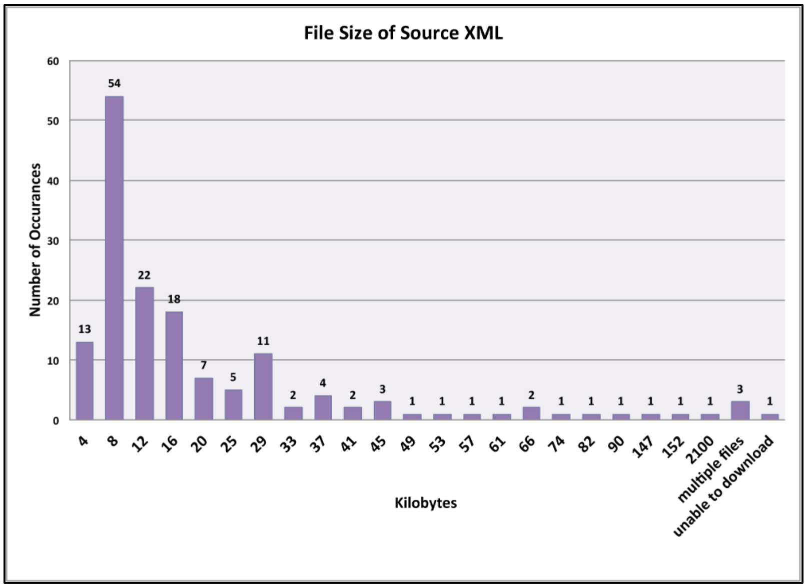


Figure 22. File Size of XML File in Records

Data Citation

All of the records contain some indication of how to cite the associated datasets through a data citation field in the metadata record. As shown in Figure 23, 65% of the records contain adequate information and 35% contain comprehensive information regarding how to cite the data associated with the record. Comprehensive records contain information consisting of fully formatted data citation, which included DOIs, as well as information on how to cite the data. Adequate records (65%), contain some of the important information on how to cite the data, but were not as precise as the comprehensive records.

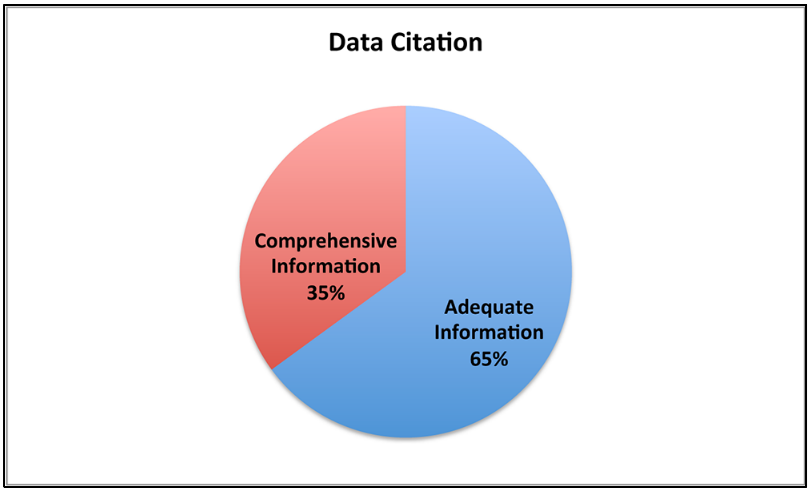


Figure 23. Robustness Percentage of Data Citation Information in Records

Metadata Standard and Additional Metadata Standard Information

Only three metadata standards are used in the records, including Dublin Core, EML, and FGDC. Of the three standards, EML is used most often at 62%. FGDC is used for 32% of the records, and 6% is used Dublin Core.

For over half of the sample, however, it is not always clear through the online metadata record alone which metadata standard was being used. In some cases, the metadata standard could only be located by looking at the related XML files if the metadata standard was not apparent through the online records. Figure 24 shows records that use FGDC as their standard always indicate this in their online record. Figure 25 provides an example of how FGDC indicates the standard in the online record.

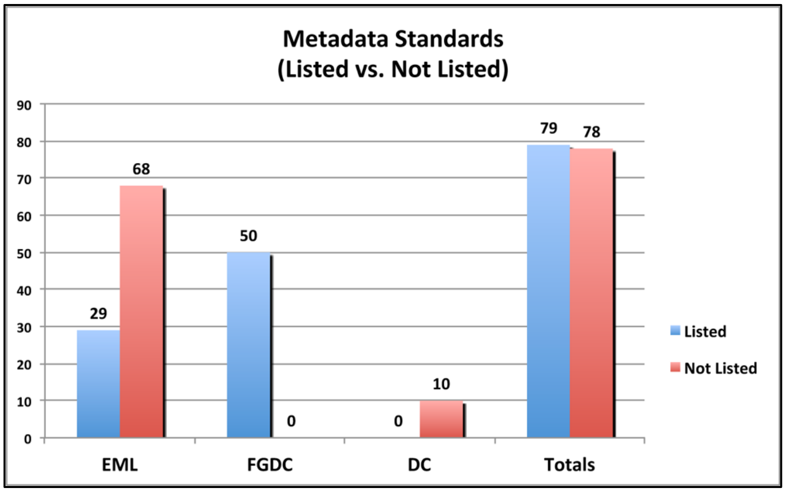
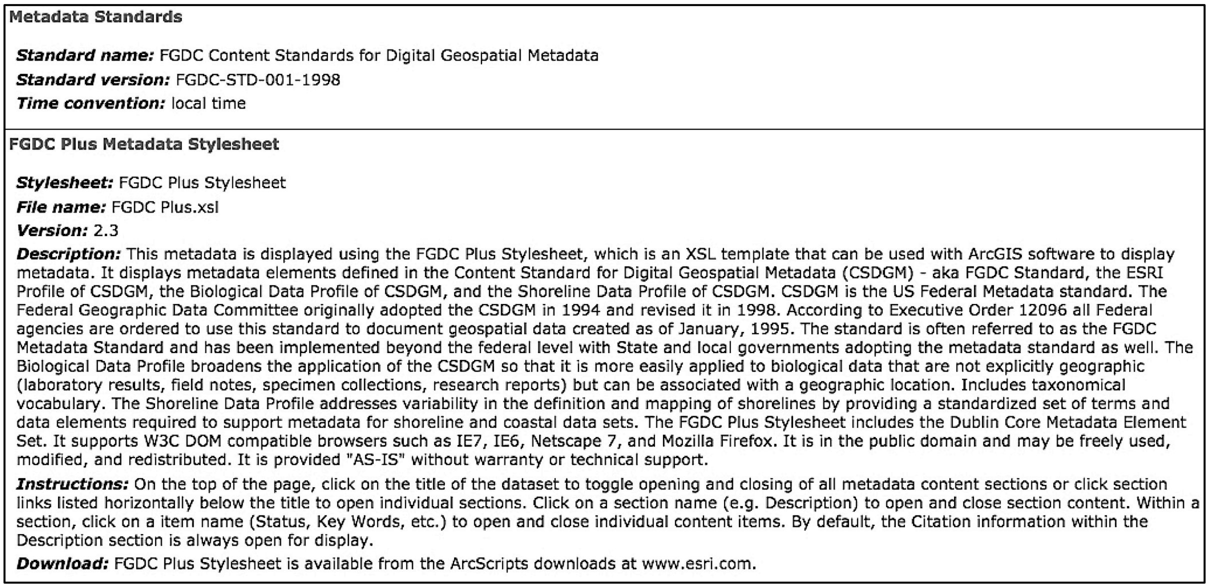


Figure 24. Count of Metadata Standards in Records (Blue indicates listed, Red indicates not listed)



Metadata Standards

Standard name: FGDC Content Standards for Digital Geospatial Metadata
Standard version: FGDC-STD-001-1998
Time convention: local time

FGDC Plus Metadata Stylesheet

Stylesheet: FGDC Plus Stylesheet
File name: FGDC Plus.xsl
Version: 2.3
Description: This metadata is displayed using the FGDC Plus Stylesheet, which is an XSL template that can be used with ArcGIS software to display metadata. It displays metadata elements defined in the Content Standard for Digital Geospatial Metadata (CSDGM) - aka FGDC Standard, the ESRI Profile of CSDGM, the Biological Data Profile of CSDGM, and the Shoreline Data Profile of CSDGM. CSDGM is the US Federal Metadata standard. The Federal Geographic Data Committee originally adopted the CSDGM in 1994 and revised it in 1998. According to Executive Order 12096 all Federal agencies are ordered to use this standard to document geospatial data created as of January, 1995. The standard is often referred to as the FGDC Metadata Standard and has been implemented beyond the federal level with State and local governments adopting the metadata standard as well. The Biological Data Profile broadens the application of the CSDGM so that it is more easily applied to biological data that are not explicitly geographic (laboratory results, field notes, specimen collections, research reports) but can be associated with a geographic location. Includes taxonomical vocabulary. The Shoreline Data Profile addresses variability in the definition and mapping of shorelines by providing a standardized set of terms and data elements required to support metadata for shoreline and coastal data sets. The FGDC Plus Stylesheet includes the Dublin Core Metadata Element Set. It supports W3C DOM compatible browsers such as IE7, IE6, Netscape 7, and Mozilla Firefox. It is in the public domain and may be freely used, modified, and redistributed. It is provided "AS-IS" without warranty or technical support.
Instructions: On the top of the page, click on the title of the dataset to toggle opening and closing of all metadata content sections or click section links listed horizontally below the title to open individual sections. Click on a section name (e.g. Description) to open and close section content. Within a section, click on a item name (Status, Key Words, etc.) to open and close individual content items. By default, the Citation information within the Description section is always open for display.
Download: FGDC Plus Stylesheet is available from the ArcScripts downloads at www.esri.com.

Figure 25. FGDC Example of "Additional Metadata Information"

As shown in Figure 24, none of the Dublin Core records indicate the namespace online, while some (29 of 97) of the EML records indicate they were using EML as their standard. Overall, 50.3% of the records list their metadata standard, while 49.7% do not have their standard listed.

In addition to the standard identified, the majority of the records that use FGDC contain additional metadata information regarding the standard itself. As can be seen in Figure 25, all

FGDC records contain information regarding the standard version, stylesheet, stylesheet version, stylesheet description, and download location for the stylesheet.

Research Methods, Provenance, and Instrumentation Information

The records were analyzed for the robustness of provenance, research methods, and instrumentation information each of the records provided. Figures 26, 27, and 28 show the results of this analysis.

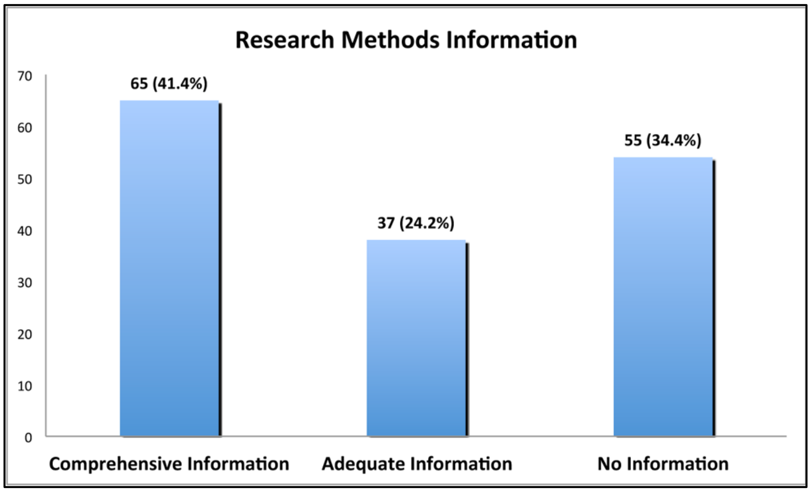


Figure 26. Robustness of Research Methods Information in Records

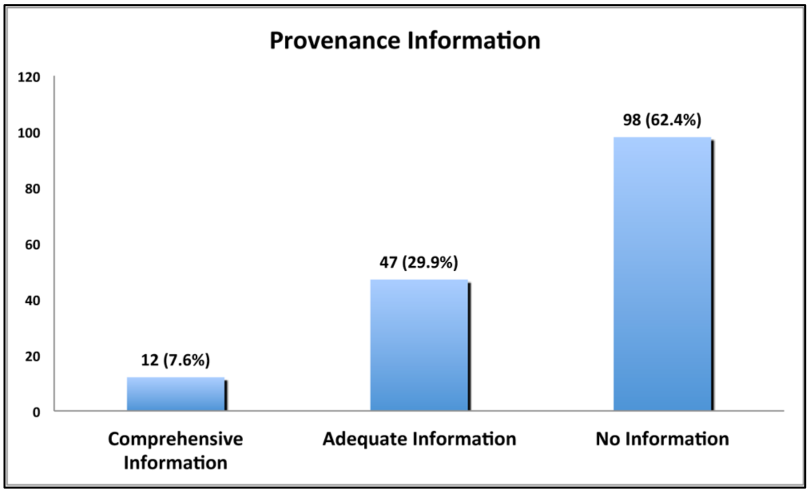


Figure 27. Robustness of Provenance Information in Records

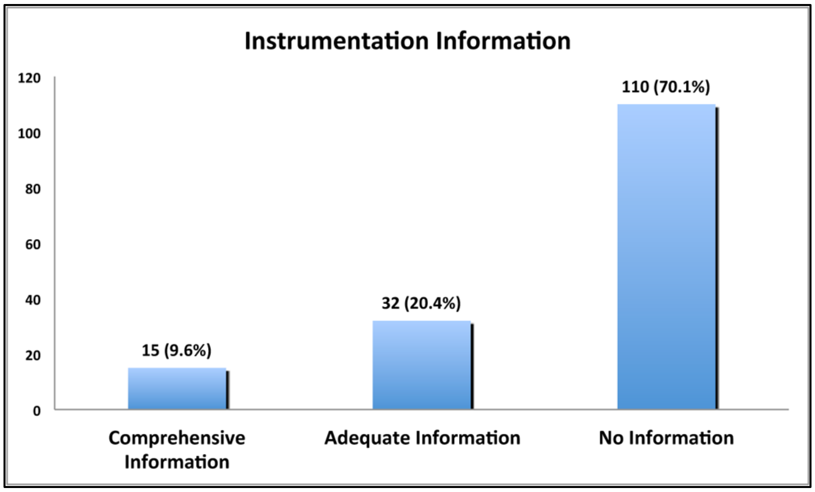


Figure 28. Robustness of Instrumentation Information in Records

Figure 26 shows that more than half of the records contain comprehensive or adequate information regarding research methods (65.6%). Provenance information is significantly lower in that only 37.2% of the records contain comprehensive or adequate information (Figure 27). Instrument information has the lowest amount of information with only 30% containing comprehensive or adequate information (Figure 28).

Creator and Metadata Provider; Contact Information; and Associated Party

The majority of the records (146 of 157, 93%) contain contact information. Contact information provides a means to contact the individual or association responsible for the data. Contact information typically includes a physical address, email address, or telephone number. Table 12 provides the organizational name for the top creators and metadata providers of the records.

Table 12

Creator or Metadata Provider

Creator or Metadata Provider	Counts
Long Term Ecological Research Network (LTER)	18
U.S. Geological Survey (USGS)	18

Lifemapper	10
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)	10
University of Kansas	10
Earth Data Analysis Center	10
Oak Ridge National Laboratory (ORNL)	10
Ecological Society of America	10
Information Center for the Environment	9
California Department of Public Health	9
Taiwan Forestry Research Institute	8

In addition to the publisher name, metadata provider, and creator, many of the records also contain associated parties. Associated parties are organizations or other scientists who work with the project, 57% of the records contain associated parties.

Publication Information (Date and Publisher Information)

The records include publication date and publisher name information. The publication date for the sample ranged from 1999 to 2014, and 106 of 157 records (67%) contain a publication date. Additionally, 31 records (20%) contain the publisher name, while 80% do not. Publisher name is a distinct and different field than metadata provider or creator, in that this information included the person or agency that made the data available. Publication date is provided for the records and is shown in Figure 29. There is a discernable increase in data published in recent years. However, 51 (32%) of the 157 records do not contain this information.

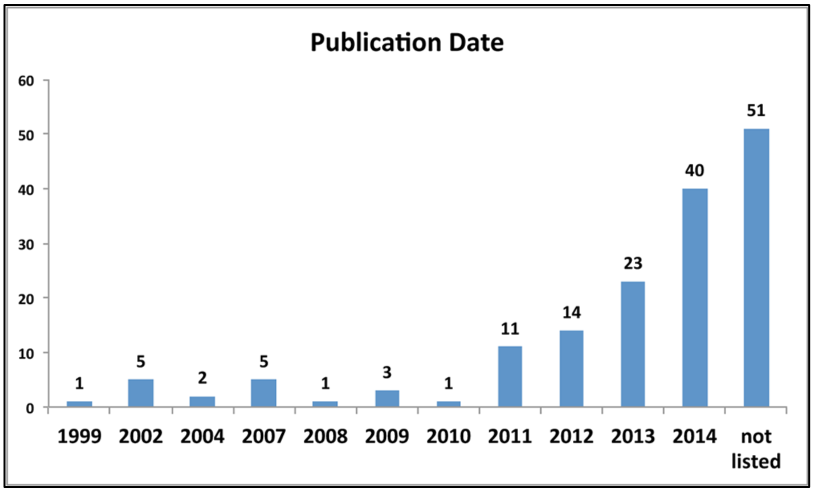


Figure 29. Publication Date Reported for Records

Keywords and Keyword Thesauri

All but three of the 157 records contain keywords (98.09%.). Keywords are drawn from a single thesaurus, multiple thesauri, a free text approach, or combining these various options. A frequency analysis was conducted using a word frequency software called "Word Counter" to determine which keywords were most often used in the records. Table 13 shows the results of this word frequency count and provides the most frequent keywords included in the records.

Table 13

Most Frequent Keywords

Keyword	Count
Field Investigation	38
Earth Science	35
California	24
Temperature	19
Analysis	16
Oceans	14
USA	12
Biosphere	11
Health	11
Land surface	11
Alaska	11

Approximately half (48.4%) of the records identify one or more thesauri for their keywords. Table 14 shows the frequency of thesauri used more than one time. In viewing this data, it is important to keep in mind that some records include controlled keywords from more than one thesaurus. For this reason, it should be noted that Table 14 does not necessarily show an accurate count of each keyword thesauri. For example, there are multiple LTER vocabulary types including LTER core research and LTER v1. For the purpose of this analysis I grouped these into one category, Long Term Ecological Research (LTER).

Table 14

Keyword Thesaurus

Keyword Thesaurus	**Occurrences**
Global Change Master Directory (GCMD)	30
ISO 19115 Topic Categories	27
EnvEurope Thesaurus	10
Parameter_Sensor_Source, Parameter, Source, Sensor, Place Keywords	9
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)	9
Spatial Reference System Identifiers	8
Gulf Watch Alaska Thesaurus	8
Long Term Ecological Research (LTER)	8
Exxon Valdez Oil Spill Trustee Council (EVOSTC)	7
Integrated Ocean Observing System (IOOS) Vocabulary	5
Experimental Program to Stimulate Competitive Research (EPSCoR) Project	2
Geographic Names Information System (GNIS)	2

Temporal Coverage

Most of the records contain information regarding temporal coverage. This information is different than publication date, in that it provides the temporal coverage of the study. For example the ESA_1 record has a temporal coverage as “Begin: 1997-04-28, End: 2007-07-01” indicating that the study ran from 1997 to 2007. Figure 30 shows the robustness of temporal coverage in the records and indicates that 85% of the records contain temporal information.

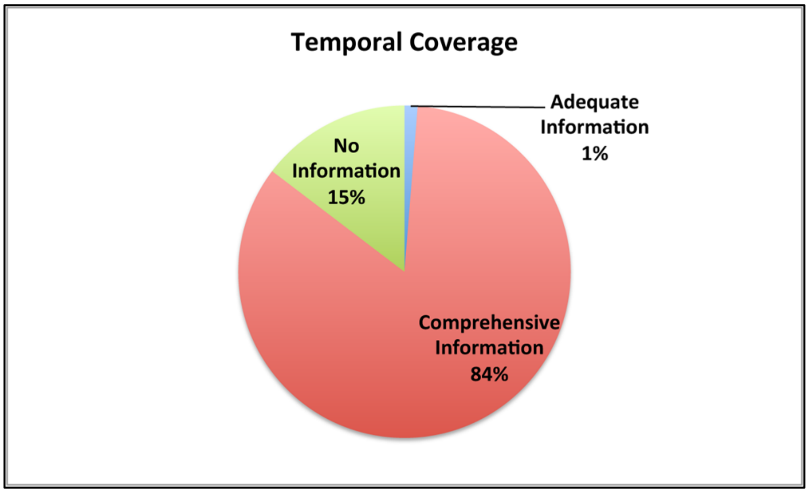


Figure 30. Robustness of Temporal Coverage in Records

Taxonomic Coverage

Few of the records contain taxonomic information regarding the dataset; only 39 (25%) of the records contained comprehensive or adequate information, while 118 (75%) contain no information. An example of a record that contains taxonomic information is the eBird_1 record, which contains the "The Clements Checklist of Birds, Sixth Edition" and includes the genus, species, and common name for citizen observers following the eBird/Clements Checklist. Figure 31 shows the percentage of records containing taxonomic information and as shown only 25% contain any information.

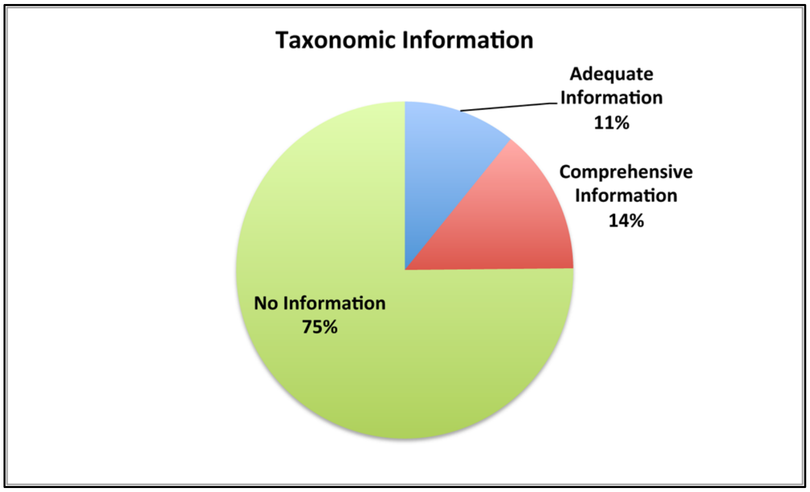
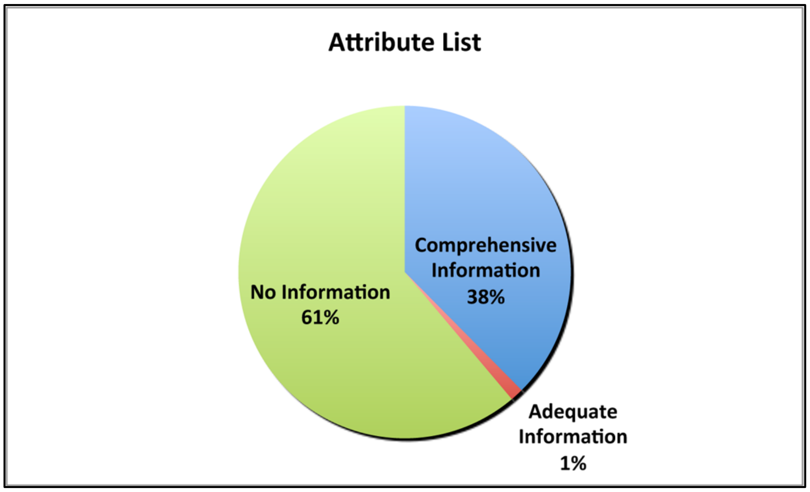


Figure 31. Robustness of Taxonomic Information in Records

Attribute and Unit Lists

Attribute lists are lists of the variables available in the data and unit lists refer to the unit or measurement of each attribute (variable). Very few records contain these lists. The robustness of the attribute and unit lists is shown in Figure 32. A little over 30% of the records contain comprehensive or adequate information for attribute and unit lists.



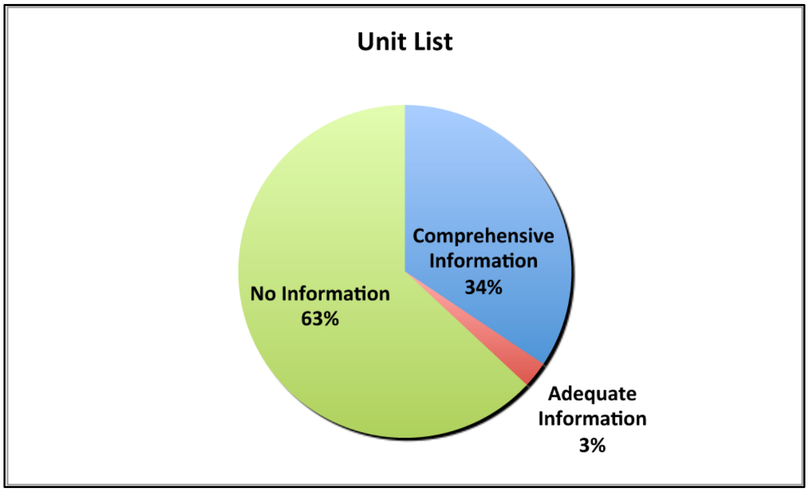


Figure 32. Robustness of Attribute and Unit Lists in Records

Geographic Information

Many of the records contained geographic information. This came in the form of either named locations, as well as bounding coordinates including longitude and latitude. Figure 33 shows the robustness of the geographic information on the records. As shown only 11% of the records did not contain geographic information, and nearly one-third of the records contained comprehensive information.

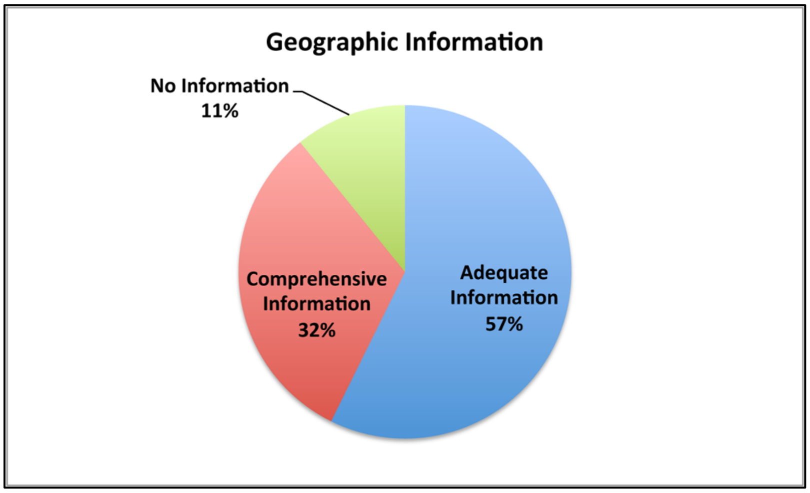


Figure 33. Robustness of Geographic Information in Records

Funding Source

Very few of the records contain funding source information; in fact 83% of the records contain no information (See Figure 34). While some funding information could be extrapolated from the abstract (see section 7.1.3), only 26 (17%) of the records contain a clearly defined "funding source" variable.

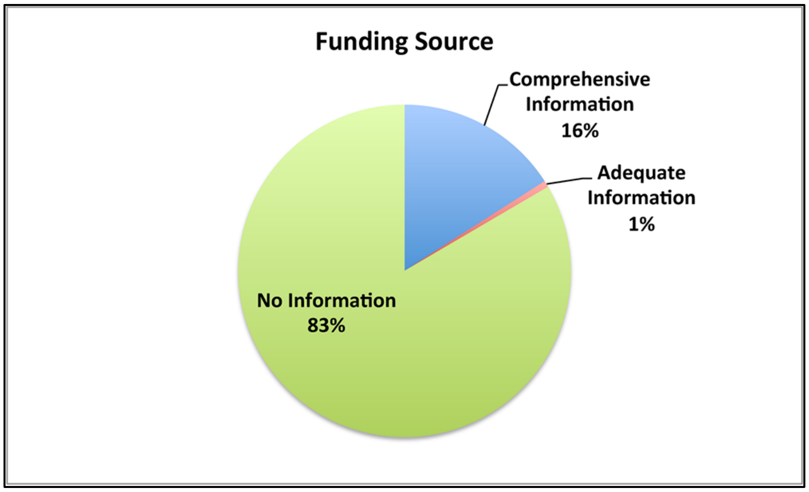


Figure 34. Robustness of Funding Source in Records

Data Types

The majority of the records (57%) did indicate the data type. These include archive files, csv/excel, text files, and image files. Figure 35 shows the data type distribution for records that contain data type information.

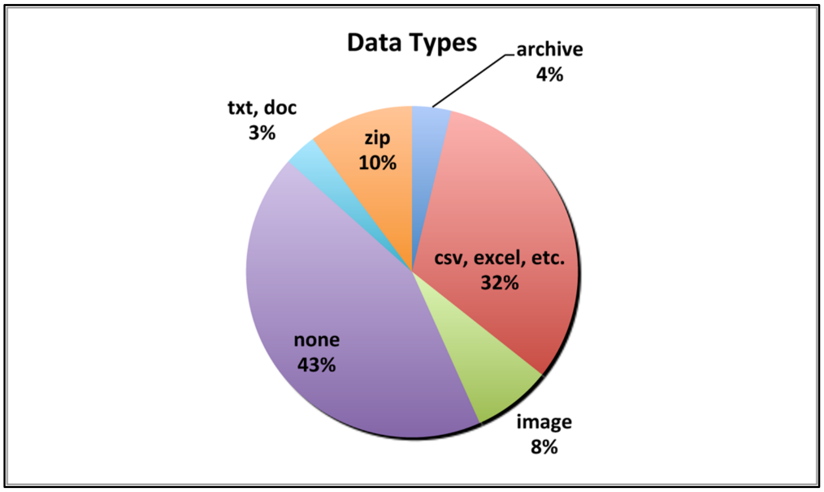


Figure 35. Data Type Distribution

Mixed-Methods Results of Data Profiling Assessment

There were several variables that were analyzed both quantitatively and qualitatively. These text-based variables were additionally analyzed first quantitatively, then through inductive qualitative content analysis. These variables include: Abstract and Additional Access Information.

Abstract

Close to 87% of the records contain an abstract. As can be seen from Figure 36, 64% of the records contain comprehensive abstracts. The abstracts include information such as how the data was gathered, the project associated with the data, and how the data was analyzed. Additionally, 22% of the records contain adequate information.

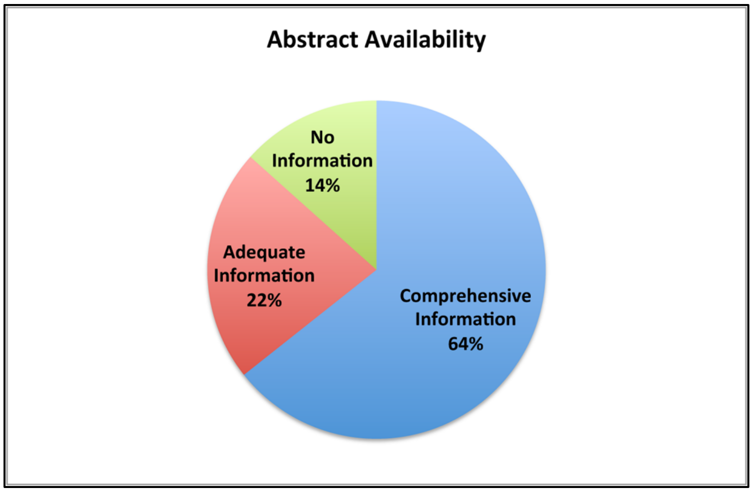


Figure 36. Abstract Availability

The abstracts were further analyzed through inductive qualitative analysis. These were imported into NVivo to examine themes within the abstracts. Figure 37 shows not only the themes, but also the distribution of the themes. The majority of the abstracts contain at least a high-level description of the project that created the data. Many of the abstracts contain data collection, temporal information, geographic information, and research methods information.

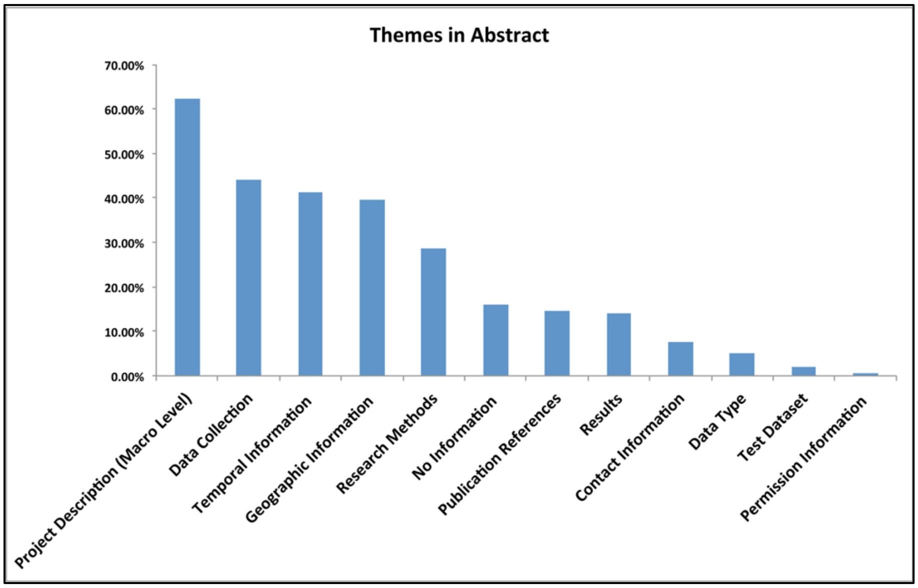


Figure 37. Themes from Abstract

Additional Access Information

Figure 38 shows the robustness of the additional access information. As shown, 33% of the records contain comprehensive information and 52.9% contained adequate information. Only 14% of the records did not contain any additional access information. The majority of the records contain textual content and provides the opportunity to qualitatively analyze this data.

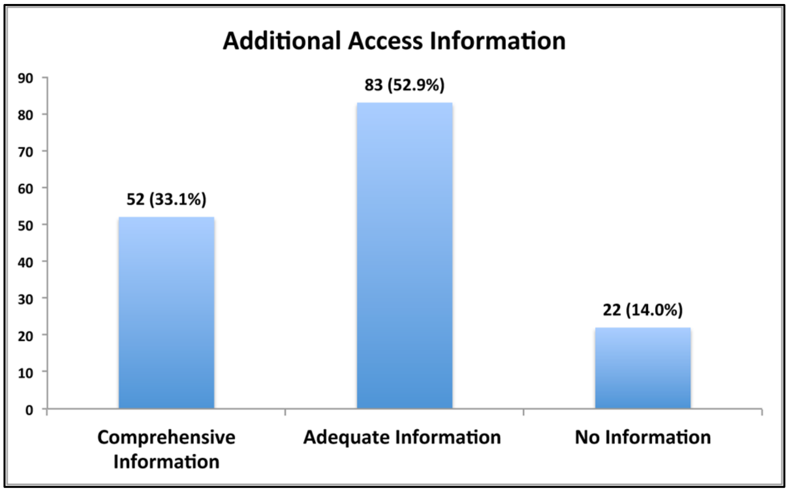


Figure 38. Robustness of Additional Access Information in Records

Many of the records include information regarding access and permissions of the data. These were recorded in the codebook as “additional access information”. For many of these records full paragraph descriptions of access and use information are provided. Figure 39 shows the themes that are in the data. High-level themes include: (a) use, (b) access-availability, (c) creative commons, (d) cite data, and (e) themes that were unsorted into higher-levels.

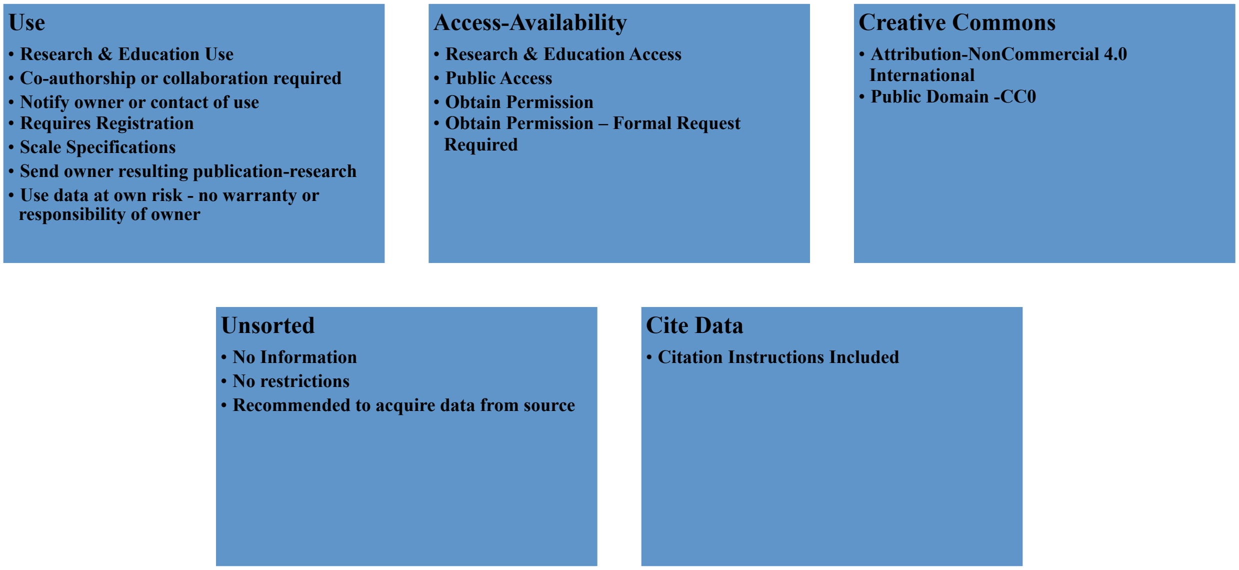

Use
• Research & Education Use
• Co-authorship or collaboration required
• Notify owner or contact of use
• Requires Registration
• Scale Specifications
• Send owner resulting publication-research
• Use data at own risk - no warranty or responsibility of owner
Access-Availability
• Research & Education Access
• Public Access
• Obtain Permission
• Obtain Permission – Formal Request Required
Creative Commons
• Attribution-NonCommercial 4.0 International
• Public Domain -CC0
Unsorted
• No Information
• No restrictions
• Recommended to acquire data from source
Cite Data
• Citation Instructions Included

Figure 39. Themes Additional Access Information

Additional Mixed-Methods Analysis

Through qualitative and quantitative analysis of the abstract and additional access information, four additional codes became part of the codebook. These are:

1. Is research methods information part of the abstract or do they have their own section?
2. Is instrumentation information part of the abstract or does it have its own section?
3. Is access (intellectual rights) a multi-step process or readily accessible?
4. Is the data available for download directly, is there a link to the data, or is no link to the data and the provider needs to be contacted?

Research Methods Information Location

Research methods information is in various locations of the record. These include: own section (52%), a reference (7%), abstract (6%), abstract and own section (1%), or no research methods information existed (34%), as shown in Figure 40 below.

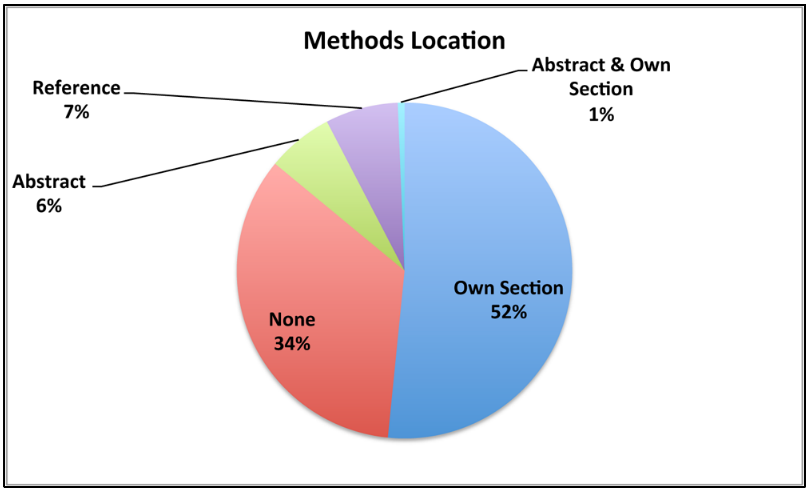


Figure 40. Research Methods Location

Instrument Information Location

Along with research methods, instrument information is found in various locations including: the abstract, methods, own section, or a combination of these options as shown in Figure 41.

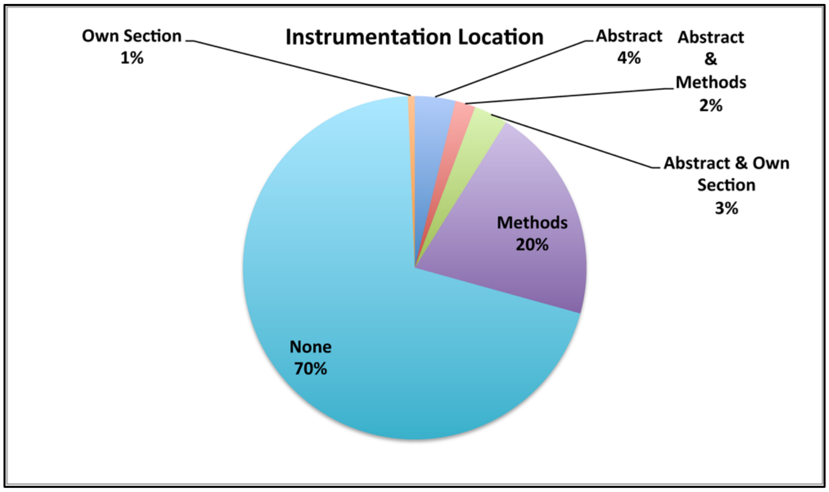


Figure 41. Instrument Location

Access/Intellectual Rights Information Steps

Additionally, the access information has either a single or multi-step process for users to access the data. As shown in Figure 42, approximately two-thirds of the records did not have any information regarding access steps, however one-third did have a single or multi-step process for users.

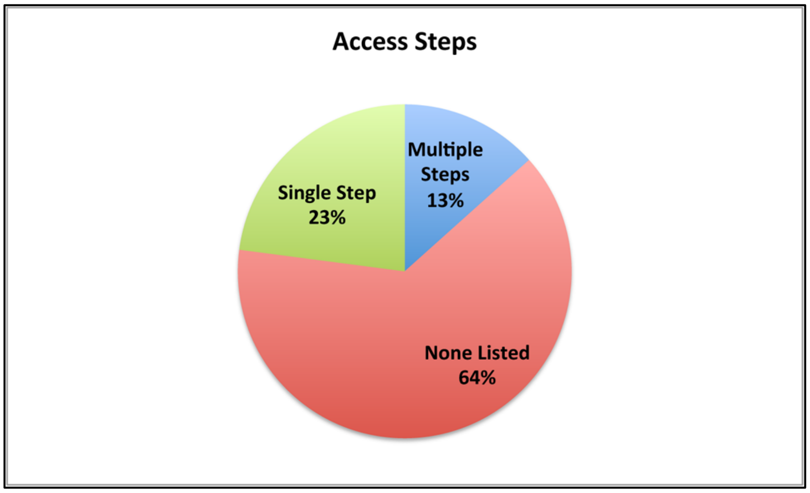


Figure 42. Access Steps

Data Availability

The data was not always made directly available from the metadata snippet. In some cases users would have to follow a link to the data, in other cases the data could be downloaded directly from the record in DataONE. However, in 32% of the records there is no obvious way to obtain the data outside of contacting the owner (Figure 43).

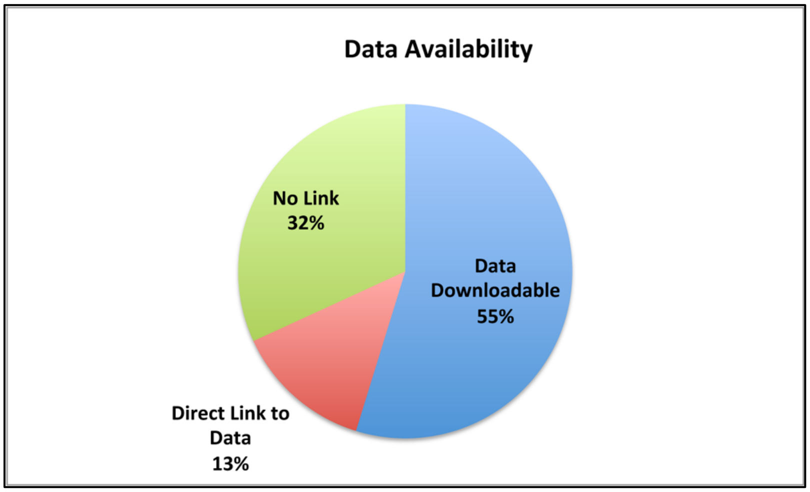


Figure 43. Data Availability

Results Part Two: Quasi-Experiment Think-Aloud

The second major data collecting effort of this dissertation was through a quasi-experiment think-aloud study. This work involved recruiting scientists to examine search results presented in a quasi-experiment and to have them "think aloud" regarding what information they needed to determine data reusability.

Sixteen participants were recruited. Participants were paid $20 for their participation either by cash or an Amazon gift card. The study was conducted either face-to-face or via screen share over Skype, and lasted approximately between 45 minutes to 1 hour, with the longest at 1 hour 30 minutes and the shorted lasting 43 minutes.

There are multiple data gathering points to this portion of the study including: a post-result usefulness survey, rank order survey, post-search survey results (general questions, open-ended questions, data reuse factors survey, and demographic survey), and the think-aloud which was gathering data throughout. The details of each of these data gathering methods can be found in Appendices 5-11. Screenshots of Results 1-4 can be found in Appendix 11.

Participants - Demographics

Of the sixteen participants 56% (9) are male and 44% (7) are female and they are all students, professors, or researchers in the earth sciences. Six participants considered their primary area of expertise as Geology, four – Ecology, two – Atmospheric Science, two – Environmental Science, 1– Physics, and 1 – Hydrology. Participant's sub-disciplines include: paleoecology, geophysics, seismology, macro-ecology, evolutionary biology, planetary geology, sedimentology, and coral reef conservation. Of the participants 62.5% are students, 31.25% are professors, and 6.25% work in a professional organization. Additionally, 31.3% have Ph.D.'s, 37.5% have master's degrees, and 31.3% have bachelor's degrees. None of the participants had used DataONE previously.

Post-Result Usefulness Survey Results

As a reminder after a participant "thought aloud" about a result and prior to seeing the next result, they were asked "On a scale of 1 to 5, with 1 being the not useful and 5 being very useful, how would you rate this results in regards to assisting you in the ability to reuse the data." The results are presented in Table 15. Result #1 is the most useful with a mean usefulness score of 3.56 and Result #3 is the least useful with a mean of 2.25. Result #1 contains the data table, Result #2 does not contain a data table but contains the abstract and methods section, Result #3 contains an abstract, but no methods section, and Result #4 contains a methods section but no abstract. There is a preference to having more information over less information, and there is a preference to having the methods section over the abstract section.

Table 15

Usefulness of Each Result – Means (SD)

Result #1	Result #2	Result #3	Result #4
3.56 (.81)	3.31 (.79)	2.25 (1.00)	2.31 (.79)

1 – Not Useful ---- 5 – Very useful

Rank Order Survey Results

Participants were asked to rank the results in order of preference from most useful to least useful in regards to data reuse. Figure 44 shows the results of these rankings. Result #1 ranks the highest and Result #3 ranks the lowest. Additionally, Result #2 ranks the second highest, while Result #4 ranks the third highest.

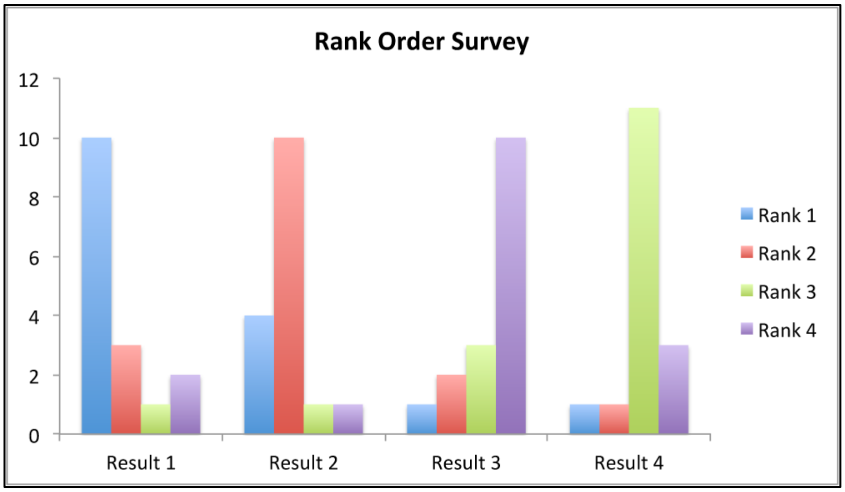


Figure 44. Rank Order of Results

Post-Search Survey (Reuse Factors)

Additionally, participants took a short survey (see Appendix 10) to determine which information they need determine data reusability. As shown, the participants were provided options for: (a) metadata standard, (b) provenance information, (c) permission and intellectual property, (d) instrumentation, (e) research methods and were asked to rank these from 1 to 7 with (1-Not at all Important, 2-Very Unimportant, 3-Somewhat Unimportant, 4-Neither Important nor Unimportant, 5-Somewhat Important, 6-Very Important, and 7-Extremely Important). Table 16 shows the mean and standard deviation from this survey.

Table 16

Post-Search Survey (Reuse Factors)

Factor	**Metadata Standard**	**Provenance Information**	**Intellectual Property Information**	**Instrument Information**	**Research Methods Information**	**Other**
Mean (SD)	4.94 (1.53)	5.25 (1.18)	4.75 (1.29)	5.88 (1.5)	6.13 (1.45)	6.6 (0.52)

Under the "other" option participants suggested: (a) attribute table, (b) attribute/units, (c) geographic location, (d) time frame of study, (e) field collection, (f) requires information regarding research methods, (g) variable/attribute table, (h) short description, (i) metadata structure, and (j) experimental setup.

Post-Search Survey (Open-Ended Questions Results)

Participants also had the opportunity to answer several open-ended questions regarding data reuse. These questions were:

1. When looking at the search results above, what information did you need to determine if the data is relevant?
2. In regard to DataONE, what information inhibits your ability to reuse data?
3. In regard to DataONE, what information facilitates your ability to reuse data?
4. When thinking about DataONE, what information did you need that the system did not provide?

These questions were asked to participants after seeing all results. For question #1, fourteen participants answered the question, for question #2 and #3 thirteen participants answered, and for question #4 ten participants answered the question.

Results Question #1: When looking at the search results above, what information did you need to determine if the data is relevant?

Ten participants indicated that the methods and the attribute table were the information needed to determine if the data was relevant. For example, participant 1 stated, “The description box and the methods data was important in determining whether the data was relevant to my needs”; by description box they were referring to the attribute table. Six participants suggested that the data description was important for determining relevance. In fact, participant 12 suggested it was the most important with

> The short summary of the dataset present in the final search result (Result #1) was perhaps the most pertinent information needed, but was unfortunately buried at the

bottom of the result. The short summary provided the contents of the dataset, and a quick look of whether or not it would be applicable.

Other items that were important for determining relevance include: abstract (3 participants), temporal information (3 participants), and provenance (2 participants). Participant #5 stated they needed to know the "who, what, when, where, and how of the data" in order to determine if the data was relevant for reuse for them. Participants also made suggestions for information that was not currently in the results example including: field collection, uncertainty information, experimental setup, instruments and calibration, and data analysis techniques.

Results Question #2: In regard to DataONE, what information inhibits your ability to reuse data?

Four participants indicated that not knowing enough about the data format inhibited their ability to reuse the data. One participant stated, "The data may be in a format that is difficult to extract. Not knowing the format of the data may lead to the user to not want to use the data." Additionally, two participants suggested that there was too much information. However one of the participants suggested that the problem was the organization of the information with "Organization of information, for example, the dataset description was vital to understanding the dataset, however, was near the bottom. Also, I would prefer more of a snapshot of the data rather than the long list."

Other factors that participants suggested inhibited their ability to reuse the data include: unknown data quality and no links to secondary publications. Two participants stated that there were no factors that inhibited their ability to reuse the data and three participants left the box blank or put dashes in the box implying they did not consider there were any factors that inhibited their ability to reuse the data.

Results Question #3: In regard to DataONE, what information facilitates your ability to reuse data?

Several participants indicated the layout of the page facilitated their ability to reuse the data. They stated that they “liked that all of the information was on one page” and the “easy to follow layout” of the page. Additionally, five participants stated that the attribute and unit list table facilitated their ability to reuse the data; four stated the methods section facilitated their ability to reuse the data, and three stated that the data description/summary facilitated their ability to reuse the data. Participants also stated their appreciation for the licensing information, the abstract, the geographic information (particularly the coordinates), and the instrument information. One participant stated they appreciated that DataONE provided the “ability to quickly see most important aspects of study instead of having to read an entire article.”

Only two notes by participants suggested areas of improvement, which include having a more distinct download location for the data and a clearer licensing summary. It was suggested that these both should be simple such as, “download data here” and “data can be used” buttons, respectively.

Results Question #4: When thinking about DataONE, what information did you need that the system did not provide?

Lastly, in regards to information that DataONE did not provide that they would have wanted to have, four participants stated they would like more information about the actual data itself, a snapshot or description of the data. One participant suggested that they “would like to get a preview of the raw data. The dataset may contain other information that is not displayed that can be useful for my study.” Another participant suggested more information about the actual data was so important with:

> knowing the format and size of a dataset could be critical. For instance, if I needed image files in .PNG format, it would save time if I knew that a dataset were .JPG only. Similarly, knowing the size of the dataset could be critical if I don't want to download 10 TBs worth of seed storage temperature data. (P12)

Two participants stated that along with the bounding coordinates they would like a map. Other suggestions include: more information on the history of the data manipulation and provenance, sample size, publications, storage, uncertainty information, and naming conventions. One participant also suggested it would be helpful to be able to control the information that was provided through the use of drop down menus. Additionally, six participants either did not answer this question or drew a dash, suggesting they were pleased with the information provided by DataONE.

Think-Aloud Results

While participants were thinking aloud, the researcher took notes regarding their thoughts of the results presented and what descriptive information was useful for determining data reusability.

Think-Aloud Results: Result #1

The majority of the participants stated that they appreciated the data description, attribute table, and research methods information and stated that these items were all important for them to determine data reusability. As shown in the usefulness and rank order survey most participants agreed that Result #1 was the most useful. In general most participants preferred having too much information to not enough information. One participant stated they preferred this because it provided the “who, what, when, where, and how” of the data,” and that was the information they needed to determine if the data was reusable.

On two occasions participants stated that there was too much information, however, this was not the opinion of the majority. Some suggestions include providing the type of data (e.g. experiment, field, sensor), adding a drop down menu so user can determine what they want to look at. Other suggestions include moving the data description to the top of the page to make it more prominent. Nearly all participants saw the data description, attribute table, and research methods as the most vital pieces of information for their determination of reusability.

Think-Aloud Results: Result #2

The majority of participants indicated that they appreciated having the methods and abstract (particularly the methods), which were useful for determining if they were able to reuse the data. However, seven participants stated that they wanted more information with a description of the data or a snippet of the data. Even those participants who had not seen the data description and attribute table stated they "really want a short description of the actual dataset" (P4). Without the data description and attribute table, participants stated that they did appreciate the conciseness of this result. They stated that the abstract, and methods were all very useful. Other items that the helpful were the bounding coordinates, contact information, and keywords.

Think-Aloud Results: Result #3

Most participants found Result #3 hard to work with. Participant #2 stated "not enough information" and participant 5 stated "mostly secondary information". In general participants stated the abstract was too much text to parse through and it made it difficult to determine if the data was reusable or not. They stated they really did not prefer the "wall of text" organization. Additionally, they stated that a description of the data and an abstract was not the same thing and wanted to know if the abstract was a paper abstract or a data abstract. These results were similar to the results from the usefulness and rank order surveys. For participants, this was the least useful of all of the results.

Think-Aloud Results: Result #4

For result #4 the majority of participants agreed that the most helpful item was the research methods. Participants who had already seen Result #3 suggested that they preferred having the methods to having the abstract indicating that having the methods was more important. One of the reasons why participants preferred the methods to the abstract is that it was easier for them to parse the information from the method. They were able to find out information such as data collection and experimental set-up if the methods were available, however this information would not always be available in an abstract. Those who had already seen Result #1 did state that they still preferred having the attribute table and data descriptions, but found the methods valuable. These results were similar to the results in the usefulness and rank order survey, participants ranked this #3 overall.

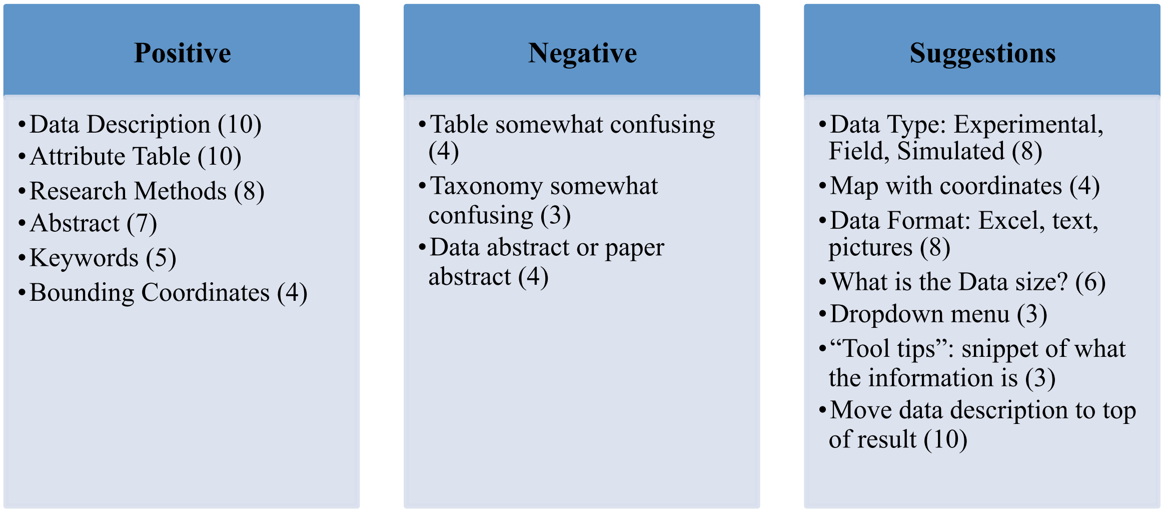
Additional Think-Aloud Results

Participants also provided a number of important observations regarding the results. In general they considered the data description, attribute table, and research methods the most useful information. They also considered the abstract useful, but found the research methods to be more useful. Regarding secondary information they really appreciated the keywords and bounding coordinates. Participants suggested that the bounding coordinates have a map and that the keywords be linked within the system so that they can pull all the data from the same keyword.

While participants agreed that the data description and attribute table were the most useful elements, there were importance items missing from these. For example, participants wanted to know the data type, meaning: field, experimental, simulated, or sensor. Participants also wanted to know the data format: text, excel, pictures, and the data size. As described by participant 11, “I don't want to download it and have my computer practically die just from opening the data.” Furthermore, participants suggested that they really wanted a snippet of the data and variables.

Many participants suggested that they usually look for articles first prior to looking for a dataset. Participants discussed how it would be helpful to have all associated publications linked and listed and additionally any other associated datasets or other studies. Participants also stated that while having the abstract was useful, that it really depended on the quality of the abstract.

Participants also provided technical suggestions. Several participants recommended having a side-table, a dropdown menu, or some way for users to select what information was provided to them. Participants suggested having some sort of “tip tools” or help so that when they hovered over an item they would receive a definition as to what that element was. The participants wanted it to be more obvious where to download the data and what was actually being downloaded when they clicked the download button, and a snippet of the dataset. Figure 45 provides an overview of these results.



Positive

- Data Description (10)
- Attribute Table (10)
- Research Methods (8)
- Abstract (7)
- Keywords (5)
- Bounding Coordinates (4)

Negative

- Table somewhat confusing (4)
- Taxonomy somewhat confusing (3)
- Data abstract or paper abstract (4)

Suggestions

- Data Type: Experimental, Field, Simulated (8)
- Map with coordinates (4)
- Data Format: Excel, text, pictures (8)
- What is the Data size? (6)
- Dropdown menu (3)
- “Tool tips”: snippet of what the information is (3)
- Move data description to top of result (10)

Figure 45. Syntheses of Think-Aloud Results

CHAPTER VIII. OBSERVATIONS AND DISCUSSION

This chapter discusses the results from this dissertation study in relation to the research questions, literature review, and gap analysis. Throughout this chapter there may be repetition of results, contextual secondary analysis, and observations to provide context for the discussion. Additionally, this chapter provides recommendations for DataONE and similar organizations to assist in data sharing and reuse.

Discussion Part One: Data Profiling Assessment

Research Question #1

To review, research question 1 asks, "What types of descriptive information are being made discoverable through DataONE?" While there are variations within the data packages that are being made discoverable through DataONE ONEMercury, there are also patterns and consistencies in the records.

High-Level Metadata

The majority of the records contain high-level metadata regarding the dataset (99.4%) and access information (75.2%). Of the dataset metadata 51% is comprehensive, however only 4.5% of the access metadata is comprehensive indicating the majority of metadata providers did try to have robust metadata at the least at a high-level regarding the dataset. This is important because previous literature has indicated a need for adequate metadata in scientific data (Baru, 2007; Edwards et al., 2011), and this result demonstrates that the metadata providers did try to provide adequate metadata at least at a high-level.

Half of the records contain additional metadata (51.6%). This “additional metadata” category exists likely because these items did not meet any other categories of metadata and contains information including: related datasets, unit lists, attribute tables, and occasionally additional information regarding access. However, this information is not very robust; only 45.2% of the records contain adequate information and 48.4% of the records contain no information. It is not surprising that many records did not contain related datasets since many datasets are from standalone studies. However, it was unfortunate to see that so many records did not contain attribute and unit lists because these are important to scientist.

The records contain a fair amount of nuanced metadata. Appendix 2 contains the information made discoverable in the records. This information points to consistencies within the three-metadata schemas that were used to describe the records (EML, FGDC, and Dublin Core), and is indicative of directions or instructions provided by the Member Nodes. Some of available items included: data citation information, provenance information, research methods, creator and metadata provider, and keywords. In total there are 30 unique pieces of descriptive information found in records.

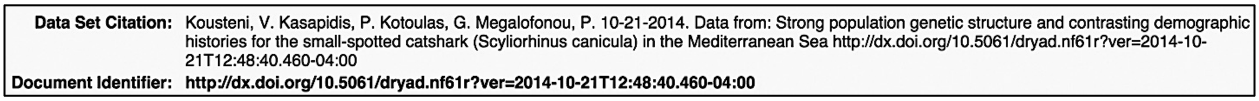
High robustness in high-level metadata and the amount of variables provided by the records indicate that Member Nodes aim to share both a wide breadth and depth of information throughout the records within DataONE.

Data Citation and DOI

Relationships in data citation information, DOI, and requirements for reuse found in the additional access information indicates the importance of this information being available for certain Member Nodes, scientific communities, and metadata standards.

All of the records contain some indication as to how to cite the associated dataset. This demonstrates that data providers believed it important to provide users a way to cite the data. However, there are variations in the presentation pointing toward constraints of the metadata standard, guidance from the Member Node, and disciplinary norms. As described in the literature review, disciplines and organizations have gone to great lengths to develop standards for their scientific disciplines. For example, Edwards and colleagues (2011) discussed how the LTER sites adopted the EML standard in order to assist with consistency throughout the multiple LTER sites. Jones and colleagues (2001) examined Metcat, a framework built for ecologists and biologists to assist with metadata creation and editing. And lastly, Michener and colleagues (1997) discussed the creation of a metadata standard for nongeospatial ecological data. The rigorous use and creation of metadata standards for the ecological and biological communities is seen throughout the results of the data profiling assessment.

All Dryad's records provide a very precise data citation. Figure 46 is an example from the Dryad Member Node that the majority of the records emulate.



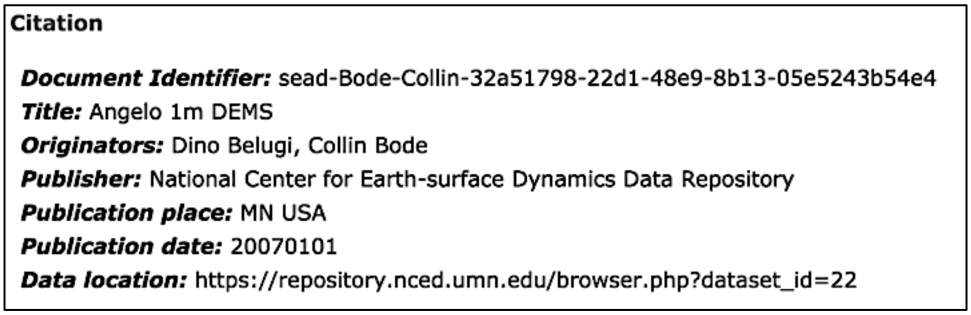
Data Set Citation:	Kousteni, V. Kasapidis, P. Kotoulas, G. Megalofonou, P. 10-21-2014. Data from: Strong population genetic structure and contrasting demographic histories for the small-spotted catshark (Scyliorhinus canicula) in the Mediterranean Sea http://dx.doi.org/10.5061/dryad.nf61r?ver=2014-10-21T12:48:40.460-04:00
Document Identifier:	**http://dx.doi.org/10.5061/dryad.nf61r?ver=2014-10-21T12:48:40.460-04:00**

Figure 46. Dryad Dataset Citation Suggestion and DIO

Dryad data always includes data DOIs. This information is indicative of data citation being valued strongly by the Dryad creators. Dryad has very defined format instructions that are integrated into their scholarly communication workflow and are communicated on the website and may be enforced through Dryad data curators. Dryad's examples on their website are a key factor contributing to the clearness of the data citation information (Dryad, 2016). These instructions would make it very easy for anyone using Dryad data to reuse and these strong

Dryad recommendations and procedures likely contribute to the very well-structured data citation found in the records (Dryad, 2016).

Other Member Nodes provide very detailed data citation information, however, not as prescriptive as Dryad. All of the FGDC records (EDAC, Merritt, ORLN, SEAD, and USGS Member Nodes) also provide very explicit data citation information. Although the citation information is very specific, it is not formatted as formally as the Dryad data citation information. Figure 47 shows an example from the SEAD Member Node and all Member Nodes that use the FGDC metadata format have similar data citation information and structure.



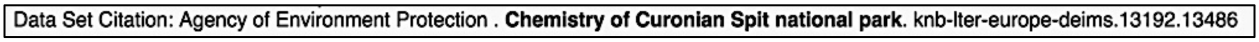
Citation

Document Identifier: sead-Bode-Collin-32a51798-22d1-48e9-8b13-05e5243b54e4
Title: Angelo 1m DEMS
Originators: Dino Belugi, Collin Bode
Publisher: National Center for Earth-surface Dynamics Data Repository
Publication place: MN USA
Publication date: 20070101
Data location: https://repository.nced.umn.edu/browser.php?dataset_id=22

Figure 47. SEAD Example for FGDC

As shown in Figure 47, the SEAD citation contains much of the same information as the Dryad citation, such as title, author, and date. However, the Dryad citation uses the traditional citation format that most scientists are accustomed to. It is implied that the FGDC metadata format highly values data citation information since all of the FGDC records the information shown in Figure 47. However, it would be the responsibility of the user to format appropriately.

The remainder of the records contain some data citation information, however are not as robust as Dryad or FGDC. Figure 48 shows a less robust record from the LTER member node. While this record provides the agency and title, as well as an identifier, it does not have a date or DOI. Additionally, it is not apparent what the identifier is referring to. A data user would have an incomplete data citation or have to piece together information from the record to create a more complete data citation which would be a time consuming processes. As seen in the literature review, time constraints are a key concern to scientists regarding data sharing and reuse (Sayogo & Pardo, 2011b; Tenopir et al., 2011).



Data Set Citation: Agency of Environment Protection . **Chemistry of Curonian Spit national park.** knb-lter-europe-deims.13192.13486

Figure 48. Less Robust Data Citation Information

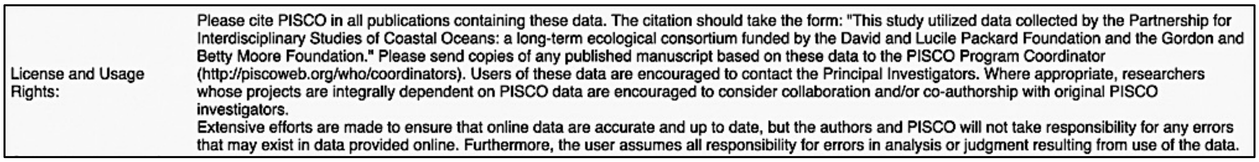
Each of the examples shows the variation within the records in regards to data citation information and DOI availability. As shown, the Dryad Member Node provides the most straightforward approach to providing this information for users of the data. Additionally, the Member Nodes using the FGDC format provide a comprehensive amount of information regarding data citation, as well as the DOI, however their lack of use of traditional citation style makes it more difficult for users to simply copy and paste the data citation, which ultimately burdens the user to have to format the citation. Lastly, many of the records that use the EML standard only contained adequate robustness indicating that these Member Nodes and perhaps community have placed less emphasis on data citation and DOI information

Through looking at these examples, it is clear that the Dryad Member Node provides the best choice for data citation and DOI information and should be considered a best practice for what information should be available, how to format this information, and how to ensure that data sharers provide data citation information. Additionally, having data citation instructions on the Member Node website for data sharers to review prior to submitting their data should be considered a best practice and the specific instructions provided by the Dryad Member Node should be an example of this best practice.

Instructions for Reuse

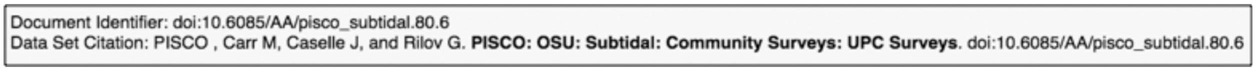
Some Member Nodes provide specific user instructions for data reuse. These are located in several places in the record and are not mutually exclusive in regards to data citation and DOI instructions.

All PISCO records contain detailed access and use requirements as shown in Figure 49 and include: an in-text citation, the exact language to be used, and where copies should be sent of any subsequent published materials. Additionally, PISCO records typically provide a well-formatted dataset citation and a DOI to be cited in the reference list as seen in Figure 50.



License and Usage Rights:	Please cite PISCO in all publications containing these data. The citation should take the form: "This study utilized data collected by the Partnership for Interdisciplinary Studies of Coastal Oceans: a long-term ecological consortium funded by the David and Lucile Packard Foundation and the Gordon and Betty Moore Foundation." Please send copies of any published manuscript based on these data to the PISCO Program Coordinator (http://piscoweb.org/who/coordinators). Users of these data are encouraged to contact the Principal Investigators. Where appropriate, researchers whose projects are integrally dependent on PISCO data are encouraged to consider collaboration and/or co-authorship with original PISCO investigators. Extensive efforts are made to ensure that online data are accurate and up to date, but the authors and PISCO will not take responsibility for any errors that may exist in data provided online. Furthermore, the user assumes all responsibility for errors in analysis or judgment resulting from use of the data.

Figure 49. PISCO instructions for citation format



Document Identifier: doi:10.6085/AA/pisco_subtidal.80.6
Data Set Citation: PISCO , Carr M, Caselle J, and Rilov G. **PISCO: OSU: Subtidal: Community Surveys: UPC Surveys**. doi:10.6085/AA/pisco_subtidal.80.6

Figure 50. PISCO Dataset Citation and DOI

Not only does PISCO provide a well-formatted data citation and DOI, the Member Node keeps track of who is using their data making it much easier to track their data. The major disadvantage to this tactic is that it relies on users noticing information from two different locations in the record. An advantage is that data owners are able to track how their data is being reused. This is not only useful information for the user of the data, but also is helpful to ensure proper citation of the data, and the possibility of tracking the use of the data through DOIs (Piwowar & Vision, 2013)

Other Member Nodes provide data citation instructions in either the additional access information or the abstract. The LTER Member Node states,

> The Principal Investigator of the dataset be sent a copy of the report or manuscript prior to submission and be adequately cited in any resultant publications. A copy of any resultant publications should be sent to the McMurdo data manager and principal investigator,

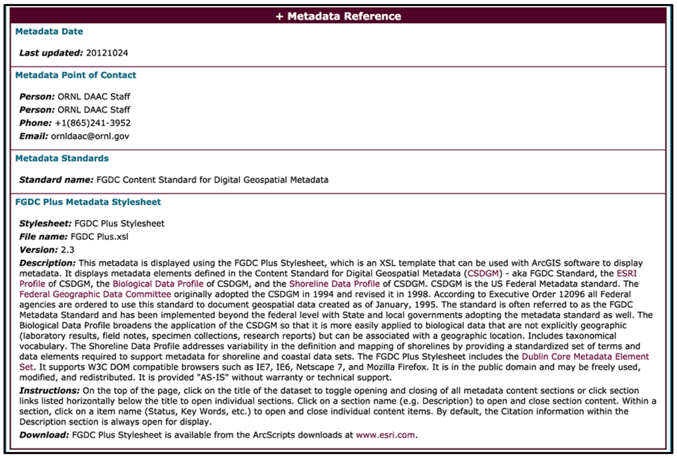
for all of their records. It is clear that the data owners would like to keep track of what publications resulted from their data. This shows use of their datasets and keeps track of the type of research occurring with their data. This question of why data owners ask for copies of published works has not been answered in current literature. However, it is known that scientists want attribution and acknowledgement for their work (Acord & Harley, 2013; Piwowar & Vision, 2013; Sayogo & Pardo, 2013; Tenopir et al., 2011). Data citations could contribute to the tenure process. Additionally, there are other non-tenure related advantages in knowing what research is being produced from your data, including potential collaborations and ensuring the data is not used in error.

The various locations for data citation, DOIs, and additional information for use could easily confuse users. This information was found in the "dataset citation", "abstract", and "additional access information" fields and are generally not located in the same area of the record. This may be part of the structure of EML, in that these fields are generally not located near each other. However, having these items located in separate fields puts an unnecessary burden on the user. It is recommended to have this information located either near each other or include a cross-referenced so that users have all of the information they need to both cite the data and follow any additional requirements.

Metadata Standards

The majority of the records use EML for their metadata standard (62%). The rest of the records use FGDC (32%) and Dublin Core (6%). Having only three metadata standards provides some consistency and is useful because users would become accustomed to the standard. However, while it is always apparent when the records use FGDC, it is not obvious when EML and Dublin Core are used.

There are some advantages to FGDC in that it provides users the ability to choose information through dropdown menus. Additionally, FGDC records indicate that they followed the FGDC standard as shown in Figure 51. This indicates that the FGDC standard believes it important for users to have information regarding the standard, stylesheet, and version. No other standard provided this information.



+ Metadata Reference

Metadata Date

Last updated: 20121024

Metadata Point of Contact

Person: ORNL DAAC Staff
Person: ORNL DAAC Staff
Phone: +1(865)241-3952
Email: ornldaac@ornl.gov

Metadata Standards

Standard name: FGDC Content Standard for Digital Geospatial Metadata

FGDC Plus Metadata Stylesheet

Stylesheet: FGDC Plus Stylesheet
File name: FGDC Plus.xsl
Version: 2.3

Description: This metadata is displayed using the FGDC Plus Stylesheet, which is an XSL template that can be used with ArcGIS software to display metadata. It displays metadata elements defined in the Content Standard for Digital Geospatial Metadata (CSDGM) - aka FGDC Standard, the ESRI Profile of CSDGM, the Biological Data Profile of CSDGM, and the Shoreline Data Profile of CSDGM. CSDGM is the US Federal Metadata standard. The Federal Geographic Data Committee originally adopted the CSDGM in 1994 and revised it in 1998. According to Executive Order 12096 all Federal agencies are ordered to use this standard to document geospatial data created as of January, 1995. The standard is often referred to as the FGDC Metadata Standard and has been implemented beyond the federal level with State and local governments adopting the metadata standard as well. The Biological Data Profile broadens the application of the CSDGM so that it is more easily applied to biological data that are not explicitly geographic (laboratory results, field notes, specimen collections, research reports) but can be associated with a geographic location. Includes taxonomical vocabulary. The Shoreline Data Profile addresses variability in the definition and mapping of shorelines by providing a standardized set of terms and data elements required to support metadata for shoreline and coastal data sets. The FGDC Plus Stylesheet includes the Dublin Core Metadata Element Set. It supports W3C DOM compatible browsers such as IE7, IE6, Netscape 7, and Mozilla Firefox. It is in the public domain and may be freely used, modified, and redistributed. It is provided "AS-IS" without warranty or technical support.

Instructions: On the top of the page, click on the title of the dataset to toggle opening and closing of all metadata content sections or click section links listed horizontally below the title to open individual sections. Click on a section name (e.g. Description) to open and close section content. Within a section, click on a item name (Status, Key Words, etc.) to open and close individual content items. By default, the Citation information within the Description section is always open for display.

Download: FGDC Plus Stylesheet is available from the ArcScripts downloads at www.esri.com.

Figure 51. FGDC Standard Information Example

DataONE provides support for EML through the Morpho Metadata Editor in the Investigator Toolkit. Morpho allows users to enter metadata information, create local catalogs, and edit metadata. Additionally, much of the data made accessible through DataONE is environmental data; therefore having most records use EML as a standard was not surprising given that the user community is accustomed to EML (“Morpho” DataONE, 2016). Between the user community and DataONE actively supporting the Morpho Metadata Editor it is likely that many data providers would use EML for their records.

Lastly, only 10 records used Dublin Core for their records, these came from the Dryad Member Node. There is an advantage to only having three metadata standards utilized for the records, as well as having the metadata standards being as well-established as these in that many users would either have some background or they would be able to build some working knowledge of the standard as their looked at records due to the consistence.

Creator/Contact Information

Most of the records contain creators/contact information. However, there are several locations for this information in the records and which could be confusing to users.

The creator/contact metadata fields include: Associated Party, Creator, Metadata Provider, and Contact Information. These are not uniform throughout the records. 98.1% contain creator information and 66.2% contain metadata provider. Additionally, many records contain associated party (57%). It was hard to determine the precise differences for these fields. Lastly, the majority (93%) contain contact information. This is usually in the form of physical addresses or email addresses. It is very convenient that the majority of records provide a way to contact the provider. The inconsistency as to how these fields are titled is likely due to variations in metadata standard, as well as that creator, provider, and contact can differ. However, the amount of information indicates the many potential points of contact for users.

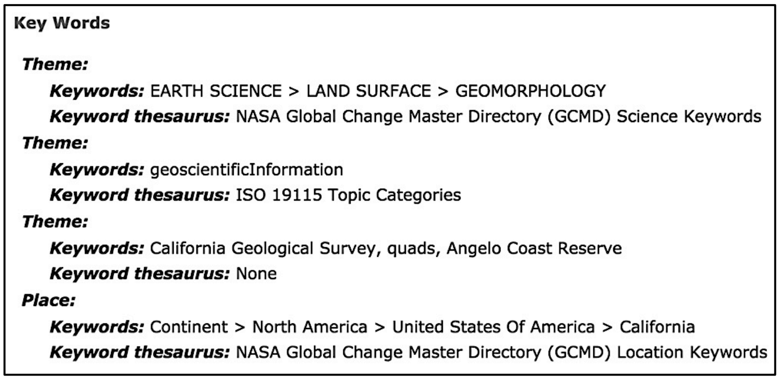
Publication date and publisher information are provided in some of the records. While the majority of the records contain a publication date (67.52%), only 20% of the records contain a publisher. However, this is not surprising, because publisher does likely not exist for all records. Additionally, the funding source is provided in less than 17% of the records.

Additionally, the funding source was either provided in its own field or was described in the license and usage rights.

The above indicates how the majority of the Member Nodes did provide a way for users to contact data providers, which is incredibly useful to users, especially if they have any questions regarding the data. The main recommendation is for streamlining and consistency among the records.

Keywords & Keyword Thesauri

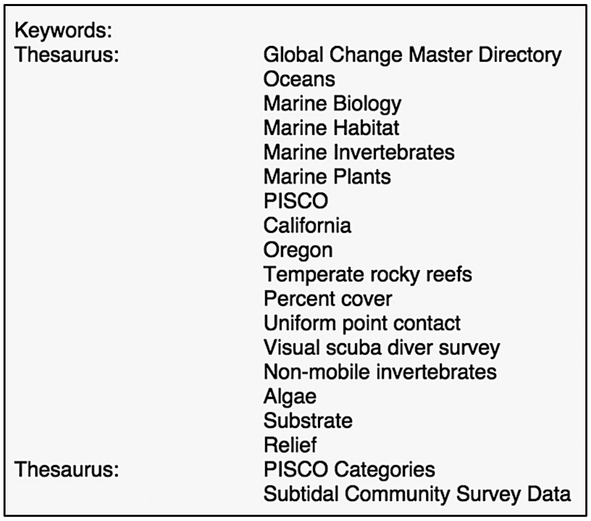
Keywords and keyword thesauri are provided for most records; only three records did not contain keywords. This is not surprising given that scientists are accustomed to providing keywords for their publications. Figure 52 and 53 provide two examples of how keywords and thesauri are presented in the records. These are organized indicate which keywords is associated with which thesaurus. This keyword association is clear in the FGDC example (Figure 52). For example the keyword "geoscientificInformation" is associated with the keyword thesaurus "ISO 19115 Topic Categories". Additionally, Figure 52 shows when there is no thesaurus associated with a keyword in the example "California Geological Survey". In Figure 53 the keyword "Sub tidal Community Survey Data" is associated with the thesaurus "PISCO Categories" and the rest are associated with the Global Change Master Directory.



Key Words

Theme:
Keywords: EARTH SCIENCE > LAND SURFACE > GEOMORPHOLOGY
Keyword thesaurus: NASA Global Change Master Directory (GCMD) Science Keywords
Theme:
Keywords: geoscientificInformation
Keyword thesaurus: ISO 19115 Topic Categories
Theme:
Keywords: California Geological Survey, quads, Angelo Coast Reserve
Keyword thesaurus: None
Place:
Keywords: Continent > North America > United States Of America > California
Keyword thesaurus: NASA Global Change Master Directory (GCMD) Location Keywords

Figure 52. Keywords and Keyword Thesauri (FGDC)



Keywords:	
Thesaurus:	Global Change Master Directory
	Oceans
	Marine Biology
	Marine Habitat
	Marine Invertebrates
	Marine Plants
	PISCO
	California
	Oregon
	Temperate rocky reefs
	Percent cover
	Uniform point contact
	Visual scuba diver survey
	Non-mobile invertebrates
	Algae
	Substrate
	Relief
Thesaurus:	PISCO Categories
	Subtidal Community Survey Data

Figure 53. Keywords and Keyword Thesauri (EML)

Disciplines that have very well established keyword thesauri and keywords are more likely to have established controlled vocabularies. Therefore, it is not surprising to see how well structured the keywords and keyword thesauri are in the record using the NASA Global Change Master Directory Keywords, given that these are very well established keywords and associated thesauri (NASA, 2016a, 2016b).

Since many scientists are accustomed to these keywords and other controlled vocabulary, it is recommended to continue having high usage in the records. These provide scientists with a simple and quick way to understand the dataset and would likely make it easy for scientists to determine relevance. It is recommended that DataONE continue to encourage Member Nodes to include keywords and keyword thesauri.

Abstract Information

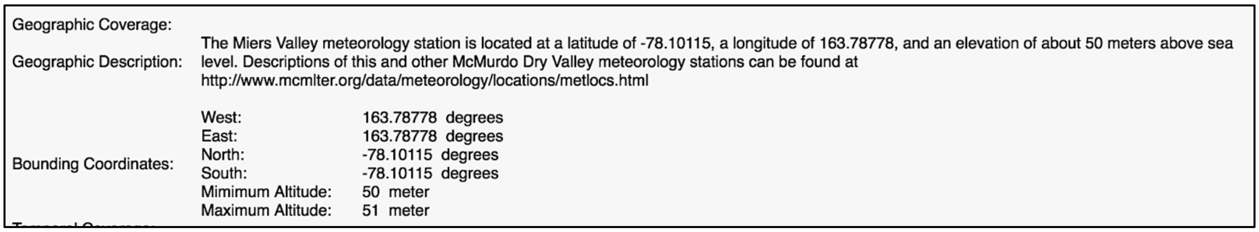
Most of the records (86%) contain abstracts providing valuable information regarding the data. These abstracts contain project description and data collection information, which is not surprising given that abstracts typically provide this information. Temporal, geographic, and research methods information are also found in the abstracts. This could be seen as a: who (project description – macro level), what (data collection), when (temporal), where (geographic), and how (research methods), technique of providing information through an abstract, which is also quite common for publication abstracts. Results are also provided, which again is common in publication abstracts.

There are several themes that are unusual for abstracts including: (a) publication references, (b) contact information, (c) data type, and (d) permissions information. Considering these are data abstracts including data type is useful for users. However, publications references, permissions information, and contact information are odd to be included in a data abstract since these are not common in publication abstracts.

Overall, the records show a comprehensive breadth and depth of information located in the data abstracts and it is recommended to continue this breath and depth so that data users have this information available to them in a format that they are accustomed to. Additionally, it is recommended to ensure that all data abstracts include: who (project description – macro level), what (data collection), when (temporal), where (geographic), and how (research methods), as well as the data type since this is the information that would be most useful to the users.

Geographic & Temporal Information

The majority of the records contain geographic information outside of the abstract. Only 11% did not contain this information indicating that geographic information is something that most data providers consider as primary for users. Comprehensive records contain a description of the location and the longitudinal/latitudinal coordinates (see Figure 54).

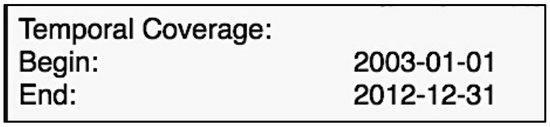


Geographic Coverage:		
Geographic Description:	The Miers Valley meteorology station is located at a latitude of -78.10115, a longitude of 163.78778, and an elevation of about 50 meters above sea level. Descriptions of this and other McMurdo Dry Valley meteorology stations can be found at http://www.mcmlter.org/data/meteorology/locations/metlocs.html	
Bounding Coordinates:	West:	163.78778 degrees
	East:	163.78778 degrees
	North:	-78.10115 degrees
	South:	-78.10115 degrees
	Mimimum Altitude:	50 meter
	Maximum Altitude:	51 meter

Figure 54. Geographic Information (Comprehensive Robustness)

It is not surprising that the majority did have geographic information because any field study will have a geographic location.

The majority of the records (85%) contain temporal information and provide the time span of the research (Figure 55). Temporal coverage is not unusual because most scientific research projects have a begin and end date, or if the data is being gathered automatically by a sensor or instrument these will usually have a time stamp.



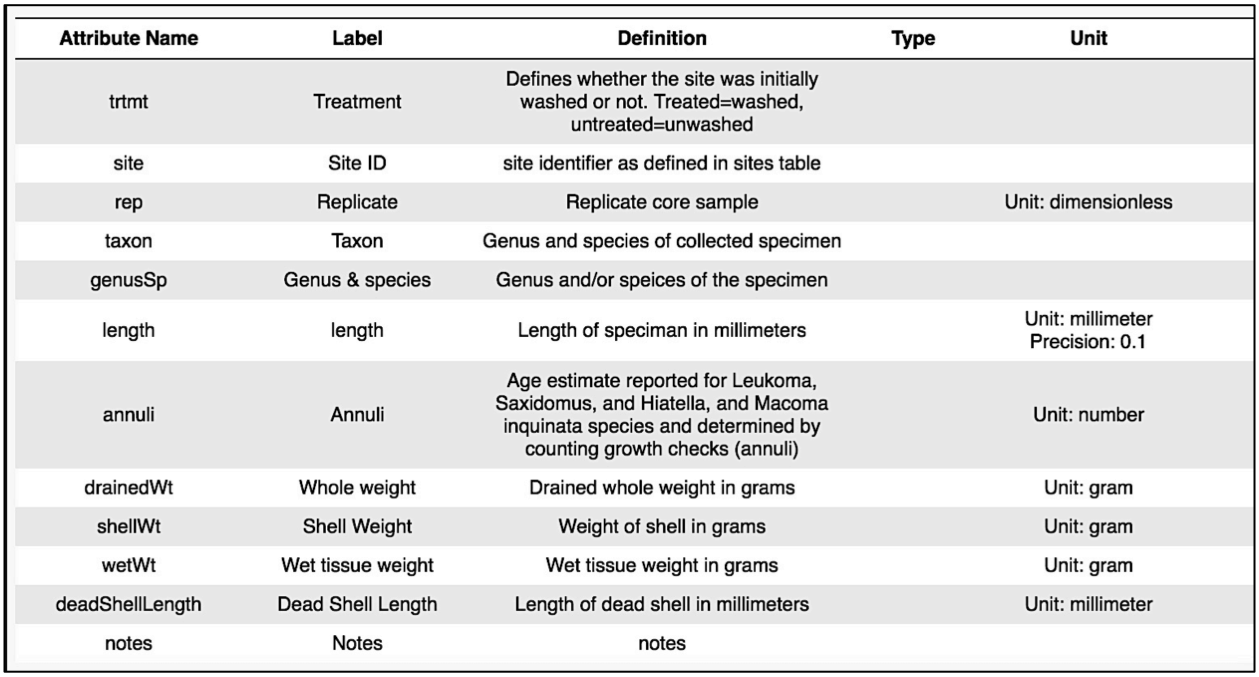
Temporal Coverage:
Begin: 2003-01-01
End: 2012-12-31

Figure 55. Temporal Coverage Example

Both geographic and temporal information are not seen as particularly unique as most studies and data collection events have a geographic location and time duration. It is recommended that DataONE Member Nodes continue to include this information in their records. Additionally, it is recommended that the Geographic Description be included only if needed. In the case of Figure 54, some of this is repeated information, therefore it is recommended to remove any duplicate information and only include new information in the Geographic Description in order to provide more streamlined information and less repetition.

Attribute and Unit Lists

Perhaps the most important piece of information in the records are the attribute and unit lists. Typically if a record contained an attribute list, it also contained the units for those attributes, as shown in Figure 56. Attributes often require a unit to measure them, therefore having the attribute and unit list together is a practical way to provide this information. In some cases there are no units along with an attribute, particularly if these are text-based attributes. For example, the “trtmt” and “taxon” attributes do not have units associated.



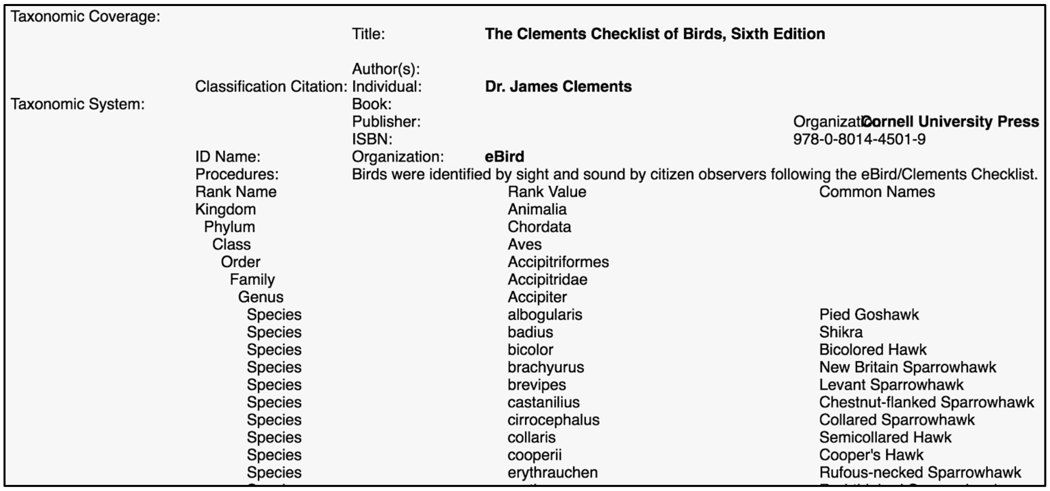
Attribute Name	Label	Definition	Type	Unit
trtmt	Treatment	Defines whether the site was initially washed or not. Treated=washed, untreated=unwashed		
site	Site ID	site identifier as defined in sites table		
rep	Replicate	Replicate core sample		Unit: dimensionless
taxon	Taxon	Genus and species of collected specimen		
genusSp	Genus & species	Genus and/or speices of the specimen		
length	length	Length of speciman in millimeters		Unit: millimeter Precision: 0.1
annuli	Annuli	Age estimate reported for Leukoma, Saxidomus, and Hiatella, and Macoma inquinata species and determined by counting growth checks (annuli)		Unit: number
drainedWt	Whole weight	Drained whole weight in grams		Unit: gram
shellWt	Shell Weight	Weight of shell in grams		Unit: gram
wetWt	Wet tissue weight	Wet tissue weight in grams		Unit: gram
deadShellLength	Dead Shell Length	Length of dead shell in millimeters		Unit: millimeter
notes	Notes	notes		

Figure 56. Attribute and Unit List (Comprehensive)

This is vital information for scientists to understand what the actual data represents. Unfortunately, only 1/3 of the records contain attribute and unit lists. It is strongly recommended that DataONE and other similar organization reinforce the importance of including this type of information for data shared within these systems. Additionally, it is strongly recommended that Member Nodes provide a statement for data depositors to include attribute and unit lists or analogous information. As seen in the quasi-experiment think-aloud results, this information is of vital importance for determining data reusability for scientists.

Taxonomic Information

Only 25% of the records contain taxonomic information including: Kingdom, Phylum, Class, Order, Family, Genus, and Species, Common Names, and sometimes reference information (Figure 57).



Taxonomic Coverage:				
		Title:	**The Clements Checklist of Birds, Sixth Edition**	
		Author(s):		
	Classification Citation:	Individual:	**Dr. James Clements**	
Taxonomic System:		Book:		
		Publisher:		Organizat**Cornell University Press**
		ISBN:		978-0-8014-4501-9
	ID Name:	Organization:	**eBird**	
	Procedures:	Birds were identified by sight and sound by citizen observers following the eBird/Clements Checklist.		
	Rank Name		Rank Value	Common Names
	Kingdom		Animalia	
	Phylum		Chordata	
	Class		Aves	
	Order		Accipitriformes	
	Family		Accipitridae	
	Genus		Accipiter	
	Species		albogularis	Pied Goshawk
	Species		badius	Shikra
	Species		bicolor	Bicolored Hawk
	Species		brachyurus	New Britain Sparrowhawk
	Species		brevipes	Levant Sparrowhawk
	Species		castanilius	Chestnut-flanked Sparrowhawk
	Species		cirrocephalus	Collared Sparrowhawk
	Species		collaris	Semicollared Hawk
	Species		cooperii	Cooper's Hawk
	Species		erythrauchen	Rufous-necked Sparrowhawk

Figure 57. Taxonomic Information (Comprehensive)

This information is relevant to only certain studies; therefore having such low rates of reporting in the records is not surprising. It is recommended that DataONE continue to encourage data providers to include this information when appropriate.

<u>Research Question #1a</u>

Research question 1a asks, "How robust is the descriptive information made available regarding that data?" In order to address this question the results were grouped into five categories shown in the tables below. Table 17 addresses the robustness of the top-level metadata, Table 18 the robustness of the data collection information, Table 19 the robustness of the creator and contact information, Table 20 the robustness of the information regarding the physical features of the data, and Table 21 the robustness of other descriptive information regarding the data.

Robustness Averages for Top-Level Metadata

Table 17 provides the average robustness for top-level metadata. As shown the robustness levels of the top-level metadata is fairly average with 20.6% of the records containing comprehensive information, 54.8% of records containing adequate information, and 24.6% of the records containing no information.

Table 17

Robustness Averages for Top-Level Metadata

Robustness	**Dataset**	**Access**	**Additional**	**Average**
Comprehensive Information	51.0%	4.5%	6.4%	**20.6%**
Adequate Information	48.4%	70.7%	45.2%	**54.8%**
No Information	0.6%	24.8%	48.4%	**24.6%**

The robustness of the dataset metadata indicates that the Member Nodes and data providers consider this information important to share with data users. This is not surprising considering that this category of information supplied data specific information. Additionally, it is refreshing to see that only 24.8% of the records did not contain access metadata. Considering that data users need to know how to access the data and if there are any barriers to use this information is vital. Lastly, it is quite unfortunately to see that so many records (48.4%) contained no information regarding Additional Information, considering that the attribute table is key to users ability to determine data reusability. It is recommended that DataONE focus on ensuring that the additional metadata is included in the records.

Robustness Averages for Data Collection Metadata

The robustness of research methods, provenance, and instrument information is much less robust than the top-level metadata. Only 19.5% of the records contain comprehensive information and 24.8% of the records contain adequate information, while 55.6% contain no information regarding research methods, provenance, or instrument information of the data (see Table 18).

Table 18

Robustness Averages of Research Methods, Provenance, and Instrument Information (Data Collection)

Robustness	**Research Methods**	**Provenance Information**	**Instrument Information**	**Averages**
Comprehensive Information	41.4%	7.6%	9.6%	**19.5%**
Adequate Information	24.2%	29.9%	20.4%	**24.8%**
No Information	34.4%	62.4%	70.1%	**55.6%**

This indicates that data providers are much less likely to provide the lower-level/detailed metadata about the data shared. Additionally, research methods contain the highest robustness, which is important because as indicated in the quasi-experiment think-aloud, scientists found research methods as important in determining data reusability. It is unfortunate to see provenance information and instrument information having such a low robustness levels, given that scientists do prefer to have this information to determine data reusability.

It is recommended that DataONE encourage Member Nodes to include more adequate information for the provenance and instrument information, and strive to have more comprehensive information for research methods.

Robustness Averages of Creator and Contact Information

Creator and contact information is provided through multiple metadata fields including: (a) creator, (b) metadata provider, (c) associated party, (d) contact information, (e) publisher information, and (f) funding source, as shown in Table 19. More than half (58.6%) of the records have creator or contact information.

Table 19

Robustness of Creator and Contact Information

Robustness	Creator	Metadata Provider	Associated Party	Contact Information	Publisher Information	Funding Source	Averages
Available	98.1%	66.2%	57.3%	93%	20%	17%	**58.6%**
Not Available	1.9%	33.8%	42.7%	7%	80%	83%	**41.4%**

The amount of different information sources could confuse users and there is no clarification in the records as to how these categories differed. For example, there is no way to understand the difference between the creator and metadata provider. While the associated party could be implied to be some outside party involved with the record, it is still not clear. Several participants from the quasi-experiment think-aloud indicated that having definitions could be useful in determining what the categories mean. It is recommended that these categories be streamlined.

Robustness Averages of Data Related Information

The data related information has the greatest robustness, for both physical features and descriptive features of the data. The physical features of the data, shown in Table 20 are well described with 40.4% containing comprehensive metadata and 14.6% containing adequate metadata. The description of the data has the greatest percentage of comprehensive (63%) and adequate metadata (18%), with only 19% not containing any information (see Table 21).

Table 20

Robustness of Data Related Information (Physical Features of Data)

Robustness	Geographic Information	Temporal Information	Taxonomic Information	Attribute List	Unit List	Averages
Comprehensive Information	32%	84%	14%	38%	34%	**40.4%**
Adequate Information	57%	1%	11%	1%	3%	**14.6%**
No Information	11%	15%	75%	61%	63%	**45.0%**

Table 21

Robustness of Data Related Information (Description of Data)

Robustness	Data Citation	Publication Date	Abstract	Keywords	Keyword Thesauri	Additional Access Information	Averages
Comprehensive Information	65%	68%	64%	98%	48%	33%	**63%**
Adequate Information	35%	0%	22%	0%	0%	53%	**18%**
No Information	0%	32%	14%	2%	52%	14%	**19%**

Tables 20 and 21 show where Member Nodes are providing robust information such as keywords, temporal information, and publication date. However, this also indicates where Member Nodes need to focus on including information such as taxonomic information (where applicable) and attribute and unit lists. As discussed earlier, it is recommended to include taxonomic information where applicable, however it is strongly recommended to include attribute and unit lists to all data records given that these provide very valuable information to data users.

Research Question #1b

Research question 1b asks, "How is information being provided about the data, such as information regarding metadata standards, provenance information, research methods, instrumentation?"

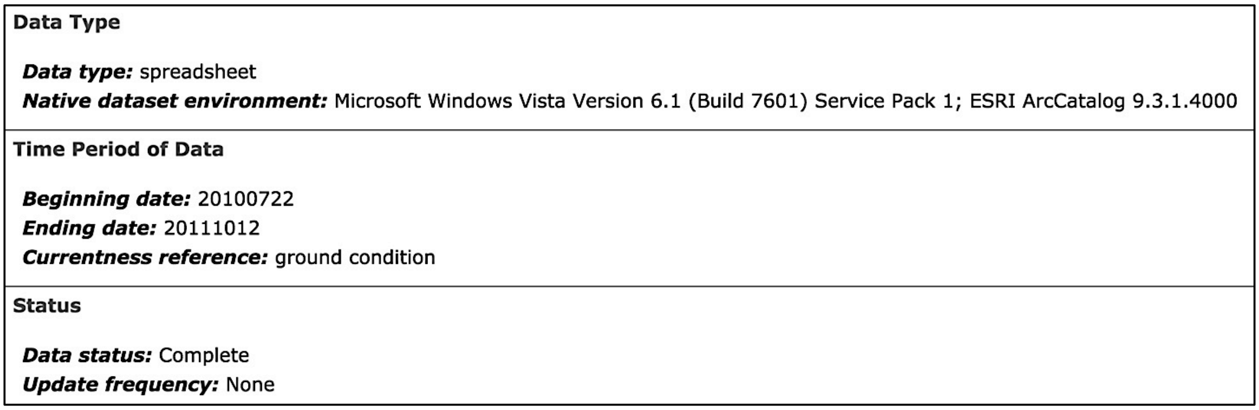
Metadata Standard

Information regarding metadata standard is not clear, even though half (50.3%) of the records indicate which metadata standard the record used. The reference to the metadata standard is in relation to the record and data. This causes confusion on what standard is used for the data. This confusion was a topic of discussion during the quasi-experiment think-aloud portion of the study. Furthermore, the results indicate that participants do not care which standard is used. It is recommended that the metadata standard of the data be provided and/or a snapshot or codebook.

Provenance

A little over one-third of the records contain provenance information (37.6%). However the robustness of this information is not very comprehensive or easily understood in many cases. As shown in the results only 7.6% of the records have comprehensive provenance information, while 29.9% have adequate information, and 62.4% have no information. In the records, the provenance information is labeled as “change history”, “maintenance”, or “status” to prompt users to know this was provenance information. None of the records examined actually used the term “provenance.”

FGDC records use a combination of fields to indicate provenance information, data type and status. As shown in Figure 58, users have a variety of information regarding provenance including native dataset environment, beginning and ending time period of the data, as well as a status update and if there are any upcoming updates to the data.



Data Type

Data type: spreadsheet
Native dataset environment: Microsoft Windows Vista Version 6.1 (Build 7601) Service Pack 1; ESRI ArcCatalog 9.3.1.4000

Time Period of Data

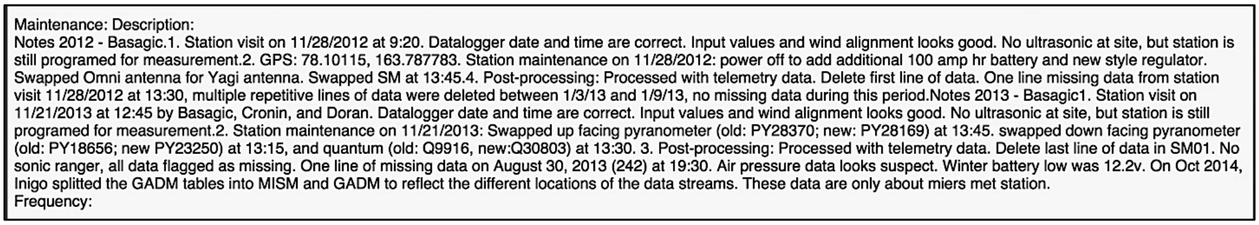
Beginning date: 20100722
Ending date: 20111012
Currentness reference: ground condition

Status

Data status: Complete
Update frequency: None

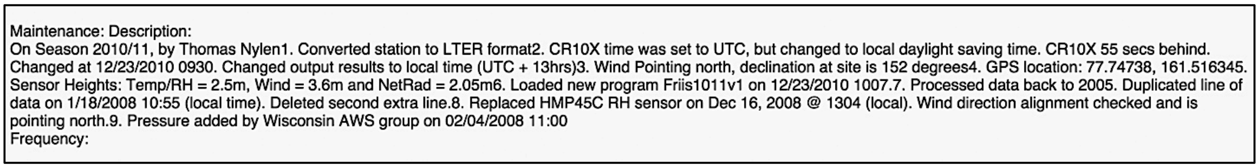
Figure 58. FGDC Provenance Example

EML records use the terms “maintenance” or “change history” to refer to provenance information for the data. In some cases, as shown in Figure 59 and 60, information is provided regarding the equipment used, how this affected the data, and any manipulation of the data.



Maintenance: Description:
Notes 2012 - Basagic.1. Station visit on 11/28/2012 at 9:20. Datalogger date and time are correct. Input values and wind alignment looks good. No ultrasonic at site, but station is still programed for measurement.2. GPS: 78.10115, 163.787783. Station maintenance on 11/28/2012: power off to add additional 100 amp hr battery and new style regulator. Swapped Omni antenna for Yagi antenna. Swapped SM at 13:45.4. Post-processing: Processed with telemetry data. Delete first line of data. One line missing data from station visit 11/28/2012 at 13:30, multiple repetitive lines of data were deleted between 1/3/13 and 1/9/13, no missing data during this period.Notes 2013 - Basagic1. Station visit on 11/21/2013 at 12:45 by Basagic, Cronin, and Doran. Datalogger date and time are correct. Input values and wind alignment looks good. No ultrasonic at site, but station is still programed for measurement.2. Station maintenance on 11/21/2013: Swapped up facing pyranometer (old: PY28370; new: PY28169) at 13:45. swapped down facing pyranometer (old: PY18656; new PY23250) at 13:15, and quantum (old: Q9916, new:Q30803) at 13:30. 3. Post-processing: Processed with telemetry data. Delete last line of data in SM01. No sonic ranger, all data flagged as missing. One line of missing data on August 30, 2013 (242) at 19:30. Air pressure data looks suspect. Winter battery low was 12.2v. On Oct 2014, Inigo splitted the GADM tables into MISM and GADM to reflect the different locations of the data streams. These data are only about miers met station.
Frequency:

Figure 59. EML Provenance Example 1



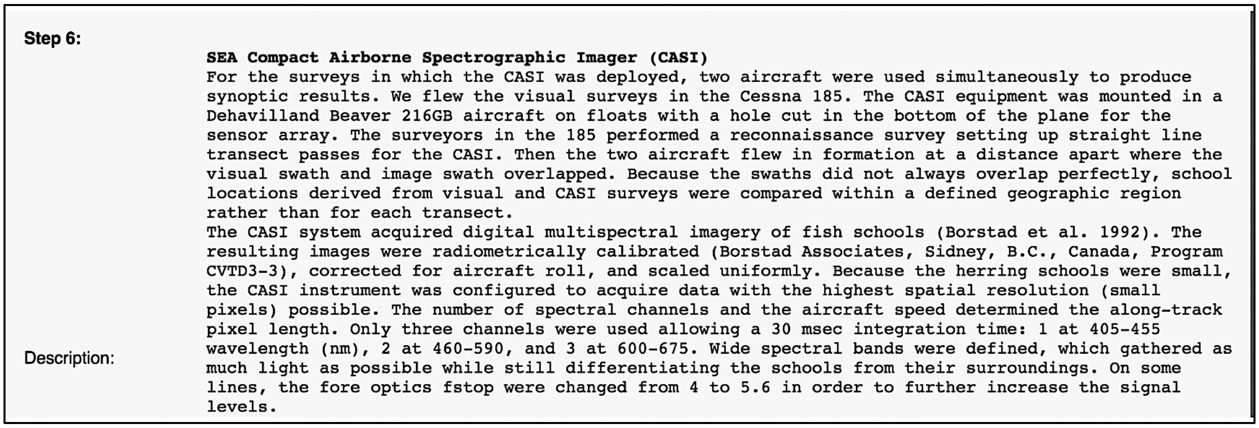
Maintenance: Description:
On Season 2010/11, by Thomas Nylen1. Converted station to LTER format2. CR10X time was set to UTC, but changed to local daylight saving time. CR10X 55 secs behind. Changed at 12/23/2010 0930. Changed output results to local time (UTC + 13hrs)3. Wind Pointing north, declination at site is 152 degrees4. GPS location: 77.74738, 161.516345. Sensor Heights: Temp/RH = 2.5m, Wind = 3.6m and NetRad = 2.05m6. Loaded new program Friis1011v1 on 12/23/2010 1007.7. Processed data back to 2005. Duplicated line of data on 1/18/2008 10:55 (local time). Deleted second extra line.8. Replaced HMP45C RH sensor on Dec 16, 2008 @ 1304 (local). Wind direction alignment checked and is pointing north.9. Pressure added by Wisconsin AWS group on 02/04/2008 11:00
Frequency:

Figure 60. EML Provenance Example 2

From the above examples, several items become clear. Provenance information is not labeled as “provenance”; it is labeled “change history”, “maintenance”, or “status”, which could cause confusion for the users. Using the term provenance would be helpful so that users understood what this information was referring to. Additionally, there is some advantage to how the EML records provide the maintenance description, because these are quite detailed and provide contextual information to understand the data. It is recommended to include this type of contextual information whenever possible for provenance information and well as use the term provenance.

Instrument Information

Most records did not contain instrument information (70.1%). Only 9.6% of the records contain comprehensive information and 20.4% contain adequate information.



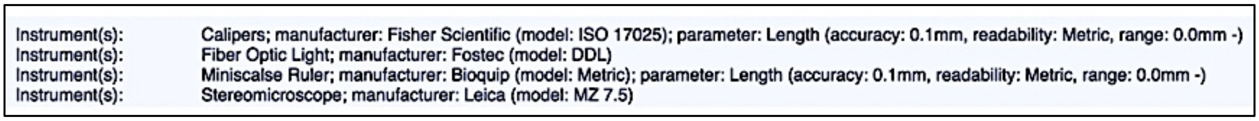
Step 6:

SEA Compact Airborne Spectrographic Imager (CASI)
For the surveys in which the CASI was deployed, two aircraft were used simultaneously to produce synoptic results. We flew the visual surveys in the Cessna 185. The CASI equipment was mounted in a Dehavilland Beaver 216GB aircraft on floats with a hole cut in the bottom of the plane for the sensor array. The surveyors in the 185 performed a reconnaissance survey setting up straight line transect passes for the CASI. Then the two aircraft flew in formation at a distance apart where the visual swath and image swath overlapped. Because the swaths did not always overlap perfectly, school locations derived from visual and CASI surveys were compared within a defined geographic region rather than for each transect.
The CASI system acquired digital multispectral imagery of fish schools (Borstad et al. 1992). The resulting images were radiometrically calibrated (Borstad Associates, Sidney, B.C., Canada, Program CVTD3-3), corrected for aircraft roll, and scaled uniformly. Because the herring schools were small, the CASI instrument was configured to acquire data with the highest spatial resolution (small pixels) possible. The number of spectral channels and the aircraft speed determined the along-track pixel length. Only three channels were used allowing a 30 msec integration time: 1 at 405-455 wavelength (nm), 2 at 460-590, and 3 at 600-675. Wide spectral bands were defined, which gathered as much light as possible while still differentiating the schools from their surroundings. On some lines, the fore optics fstop were changed from 4 to 5.6 in order to further increase the signal levels.

Description:

Figure 61. Instrument Information (Comprehensive)

As shown in Figure 61, the instrument information is incorporated into the research methods section. Additionally, in Figure 61 important contextual information to understanding how the instrument is used is provided. Outside of the research methods section, instrument information is found in its own section as shown in Figure 62. This made it difficult to know precisely where to look for the instrumentation information.



Instrument(s): Calipers; manufacturer: Fisher Scientific (model: ISO 17025); parameter: Length (accuracy: 0.1mm, readability: Metric, range: 0.0mm -)
Instrument(s): Fiber Optic Light; manufacturer: Fostec (model: DDL)
Instrument(s): Miniscalse Ruler; manufacturer: Bioquip (model: Metric); parameter: Length (accuracy: 0.1mm, readability: Metric, range: 0.0mm -)
Instrument(s): Stereomicroscope; manufacturer: Leica (model: MZ 7.5)

Figure 62. Instrumentation Information (Own Section)

The total amount of instrumentation information available is 30%; with 20% in the methods section, 4% in the abstract section, 3% both in the abstract and own section, 2% both in the abstract and methods, and 1% in its own section. This could easily cause confusion for users because they would have to read large sections of text and multiple places of the record. However, as shown in Figure 61, providing instrument information in the context of the method step is useful. It is recommended to choose one location for instrument information and to provide contextual information to assist in understanding how the instruments were used.

Research Methods Information

Research methods are found in various locations including: own section (52%), abstract (6%), reference (7%), and abstract & own section (1%). This could easily cause confusion to the users as to where the research methods information is located. For those records that provided a reference, the user would have to access this reference outside of DataONE, which leads to an additional step.

It is recommended that research methods have their own section so that the methods steps can be provided in a methodical step-by-step way. Having the research methods included with the abstract dilutes the understandability of the research methods. Additionally having the methods in more than one location or in a reference causes confusion and potentially time-wasted for the user to determine how to locate the reference or parse through an abstract to understand the methods. It is recommended that the research information and placement is consistent and streamlined across all records.

Research Question #1c

Research question 1c asks, "How is the provision of this descriptive information impacting the data-sharing infrastructure?" It is apparent through the analysis that the data-sharing infrastructure of DataONE impacted the information that is made available and vice versus.

The EML and FGDC metadata structures impact the information that is shared in DataONE. For example, the FGDC structure incorporates information regarding the FGDC standard and the EML standard incorporates the three "top-level" metadata elements (dataset, access, and additional metadata). Since the majority of the data shared in DataONE is ecological data, it makes sense that the majority of the records use EML, just as it makes sense for any geospatial data to use FGDC (FGDC, 2016; Knowledge Network for Biocomplexity, 2015).

The metadata standard is not the same as the actual data and the record did not provide standard information for the data. Therefore, the data itself may or may not have followed any actual metadata structure. Not all data is available for download. Only 55% of the records have downloadable data, while 13% link to the data, and 32% do not include the data either through the record or through a link. Therefore, a user will not know which metadata standard is used for the data from the record. It is recommended to: (a) include the metadata standard for the data and (b) include either the data itself with the record or a link to the data.

The Member Node influences the information being shared. For example, all of the Dryad records contain a data citation and DOI. When scientists shared data into the Dryad repository, Dryad provides a DOI for the data and data citation instructions are provided. Additionally, the PISCO Member Node has specific instructions for usage including that copies of the manuscripts sent to the data managers. It is recommended that all Member Nodes provide DOIs. Additionally, it is recommended that if there are instructions for use to include these near the data citation information on the record.

The way the Member Nodes are structured also affect the data that is provided. Table 12 shows the organizational name for the top creators and metadata providers of the records. The top five contributors include: LTER, USGS, Lifemapper, PISCO, and the University of Kansas. This count could be seen as inaccurate because the Lifemapper project is part of the University of Kansas Biodiversity Institute. Additionally, considering that the LTER has two Member Nodes represented in the sample (LTER Network and LTER Europe) it is very unsurprising that the LTER has some of the highest Creator and Metadata Provider counts.

Another unsurprising result is the publication date. As shown in Figure 29, the publications date range is from 1999 to 2014, with the majority of the data published in the later years. This is not surprising considering that it has become more commonplace for researchers to deposit their data alongside their research (Piwowar & Chapman, 2010). Additionally, it is not surprising that the first set of deposited data happened in 1999, this is consistent with the literature (McCain, 2000).

Another unsurprising result are the Keyword Thesauri results which included many of the typical thesauri that are used to describe earth science data (see Table 14). The NASA Global Change Master Directory contains one of the most used sets of Earth Science Keywords and are used by prominent scientific organizations including: ESIP, USGS, RPI, and ORNL (NASA Global Change Master Directory, 2016b, 2016a). The ISO 19115 Topic Categories is also a very well known and commonly used standard (International Organization for Standardization, 2016). Lastly, the EnvEurope Thesaurus combined existing data with new data generated throughout multiple LTER sites in Europe. During the project this controlled vocabulary was built which provides terms for categorizing LTER information (EnvEurope, 2014a, 2014b).

While there are over 1,500 keywords in the records the results were not noteworthy. The keyword used most often is "field investigation" which indicates that many of the investigations happened in the field. The second most frequent keyword is "earth science" which was not surprising considering that the majority of Member Nodes contain earth science data. California, Alaska, and USA are all in the top results, again not surprising considering how many of the Member Nodes are US-centric. Other keywords provide some information regarding scientific topics such as ocean, biosphere, land surface, and health. However, some are rather unhelpful, such as "analysis" and "temperature". In this way, many of the keywords do not provide enough context to be useful. However, when analyzing the list more carefully, there are many keywords that do provide sufficient description and can be useful to a scientist. For example, “forest dynamics monitoring” and “prey consumption rate” are both very specific and have potentially useful information for scientists using the system. These examples provide enough context for the researcher to understand what the data is about. Additionally, perhaps it would have been a better idea to analyze the keywords within the context of each other rather than just a quantitative approach to looking at the keywords individually. However, this would be an additional study.

Summary of Recommendations from the Data Profiling Assessment

Data citation instructions and DOIs should be provided for all data. Location of information for data citations and DOIs should be kept near each other in the record, rather than spread across the record. Metadata standards of the data should be clearly labeled and should be included for the data itself not just for the record. Contact and creator information should be streamlined instead of having so many different fields to indicate contact/creator. Abstracts should continue to be as thorough as seen in DataONE records and should include: who (project description-macro level information), what (data collection), when (temporal), where (geographic), and how (research methods), as well as data type. Attribute and unit lists should be provided for each record. Provenance, instrument, and research methods information should be streamlined so that these are not in multiple locations. Additionally, research methods information should be in its own section and precise so that users can understand how the data was collected and processed. Lastly, there should be some listing of definition for each item so that the user understands precisely what each field in the record means.

Discussion Part Two: Quasi-Experiment Think-Aloud

Research Question #2

No participant had reused data through DataONE, however most had heard of DataONE. Most participants described their data reuse experiences as first looking at literature and then acquiring the data through the data owner. This tactic is consist with the literature (Zimmerman, 2003, 2008). Some participants described using data libraries including NOAA and USGS. As discussed in the literature review, there are many data repositories where scientists can acquire data. For example Marcial and Hemminger (2010) described 100 online scientific data repositories. Although no participants had used DataONE in the past, they were able to provide valuable information as to how they have reused data and how they would reuse data within DataONE.

To gather more information, the researcher asked participants to describe their past data reuse experiences. The majority of participants stated that they had reused data in the past and their experiences varied in how easy it was for them to obtain the data, as well as the quality of the data they received, and what information they needed to reuse the data. This is consistent with the literature which discussed the need for quality control and complete metadata (Baru, 2007), improved access and discovery (Beran et al., 2010), and lack of time and support to share date (Tenopir et al., 2011).

Participants who used government agency website such as USGS and NOAA described their straightforward and uncomplicated experiences reusing data. They stated that they trusted the quality and understood the format of the data they received. The literature supports these findings since many government agencies have a long history of providing well-establish and well-maintained systems for scientific data (Marcial & Hemminger, 2010; United States of

America, 2016). Additionally literature supports this observation in that certain disciplines (biomedical, chemistry, astronomy) have been quite prolific about creating systems and data structures for ease of use and access (Bussard, 1990; Kuznetsov et al., 1990; Lide, 1981).

Participants suggested that these websites did not always have the data they needed and therefore asked researchers directly for data. Participants who gathered data directly from researchers described a difficult and time-consuming process. Participants would ask for data, wait for a response, once data was acquired ask for assistance to understand the data, sometimes give up on a source, or decided that the data was not suitable. Access and use issues were described such as unresponsive data owners, poor quality, poor or no metadata, and poor formatting. This is consistent with the literature which discussed: the need for proper metadata/documentation (Edwards et al., 2011), poor data management or the complexity of data management (Agarwal et al., 2010), lack of time and effort for scientists to share data (Tenopir et al., 2011), lack of incentive (Reichman et al., 2011; Sayogo & Pardo, 2013).

Additionally, participants described the need to understand how the data were created in order to duplicate a study and how important it was for them to have a specific, precisely, and clear knowledge of the research methods. This research methods discussion has not been included in data sharing and reuse literature to my knowledge, but has touched upon in the scientific data reproducibility literature and focused on workflows (Lifschitz et al., 2011). The majority of participants stated that although DataONE was not perfect it did provide an easy way for scientists to find reusable data.

Research question 2 asks, "What types of descriptive information could inhibit or facilitate data reuse?" Participants described four ways in which the descriptive information facilitated their data reuse. Participants discussed: (a) the descriptive information provided, (b) the amount of this information, (c) the layout, and (d) information needs and suggestions to improve data reusability.

Participants provided input as to which pieces of descriptive information were particularly helpful in their ability to reuse data. Participants suggested that although the abstract was too much information, these were close enough to a publication abstracts that they were easy to understand. Participants had similar comments for the keywords and suggested these provided a breadth and depth of information to understand the topic range of the data. Keywords are very common for scientists to use and understand particularly if these are from thesauri or controlled vocabulary from specific scientific domains. On a similar note, participants suggested that the taxonomic information was also what they were accustomed to and found it helpful to understanding the data. However, several participants stated they felt the species information was hidden. Additionally, participants stated the importance of the bounding coordinates and many of them were very pleased to see this included in the results. These pieces of descriptive information (abstract, keywords, taxonomic information, and bounding coordinates) were all particularly useful. Additionally, these have not been discussed in previous data reuse literature and provide a better understanding as to what information facilitates data reuse.

Additionally, the majority of participants stated that the research methods information, attribute table, and data description were the most critical pieces of information to facilitate data reuse. These are described in Section 8.2.2 and will only be mentioned here. Participants stated that these pieces of information provided them with a basic understanding of the data (data description), as well as the rigor that was used in creating the data (attribute table and research methods), which were vital in their understanding data creation, data quality, and overall data reusability.

Secondly, participants discussed the amount of information that was provided and how this assisted or did not assist in their reuse of the data within DataONE. In regards to Result #3, which had the least amount of information, one participant stated, “This is what I’m used to, which is not nearly enough information…to determine reusability”. Participants in general felt that Results #3 & #4 provided less information than they were comfortable with and stated that it was be hard for them to know if they could reuse the data just based on this information. However, they also stated that this was what they were used to and had grown accustomed to when looking for data. In fact, participant #2 stated how nice it was to have “more than just a title”, which is what he was accustomed to when searching for data. Another factor was amount of certain types of information. For example, most participants did not find the abstract very helpful, as described by one participant “It’s very hard to parse these wall of text abstracts”. Participants described the “wall of text” as a common problem not just for DataONE, but also for data and research abstracts in general. Additionally, participants did not feel they needed all the variations in the attribute table, and that just having a snapshot of the data would have been enough information. Lastly, participants were generally pleased with the amount of methods information, as well as the presentation of this methods information, which included the step-by-step methodology. The results provide new information regarding data reuse and is not discussed or described in the literature and gives greater context regarding the information needed for scientists determine data reusability.

Participants discussed the importance of the layout for facilitating or inhibiting data reuse. They very much appreciated that the information was all on one page and did not mind that they had to scroll so much with Result #1. In fact nearly all participants stated that they would rather have all of the information on one page than have to go hunt for the information. Additionally, participants really appreciated the ease of access through having the download buttons on the top and bottom of the page and preferred a direct download. Lastly, they did have problems with placement of information. They preferred to have the data description at the top of the page, as well as the methods information, abstract, and attribute table. The other information they felt was more secondary to understanding the data.

Participants stated they were fine with scrolling through a lot of information. The sentiment does go against some of the literature that suggested that users do not want to scroll through a lot of information (Nielsen Norman Group, 2016). However some literature suggests that search stopping behavior is dependent on task (Browne, Pitts, & Wetherbe, 2007). The results indicate that scientists would prefer to scroll through all of the information rather than not have the information, supporting Browne's et al., suggestion that this behavior is task dependent. Many participants suggested they would rather have too much information than not enough information to determine data reusability. One stated that they would rather know if the data was reusable by reading through the entire result then to make a request for the data only to find out it was not usable.

Lastly, participants had many suggestions to facilitate data reuse. Several participants suggested the idea of having "tool tips", meaning some way of knowing what the descriptive information is referring to. This could be incorporated as a mouse over with a question mark so that users may reference if they are unsure about a piece of information. Additionally, many participants suggested incorporating drop down menus so that participants could decide which information to focus on. This is actually already incorporated into the FGDC results in DataONE. Participant also stated that there was information that they needed that DataONE did not provide. This information included: data format and data type. Furthermore, participants wanted to know what type of study the data was from, meaning was it a field study, experimental, sensor, or simulated. As described in the literature there are many different type of data and many different ways to gather scientific data which vary from quite simple to quite complex (Ailamaki et al., 2010; Anderson, 2004).

Research Question #2a

Research question 2a asks, "How is information about the data such as information regarding metadata standards, provenance information, research methods, and instrumentation, influencing scientists' ability to determine if that data is reusable?"

Three pieces of descriptive information are imperative for scientists to determine data reusability. These are (a) the data description, (b) the attribute table, and (c) the research methods, as shown in the results from the post-result usefulness survey, the rank order survey, the post-search survey, and the think-aloud discussion.

The data description was not originally noted in the data profiling assessment because it was seen as part of the attribute table. However, participants viewed these as two different items. Figure 63 provides an example an attribute table. The table name and description are provided above the attribute list. All participants found this to be one of the most useful items in the result to help them determine data reusability. Participants suggested that these data descriptions, or something similar, should be moved higher on the results page. They suggested that this piece of information be the first they receive. Participants made it clear that they wanted a short, succinct, description of the data, rather than a long abstract.

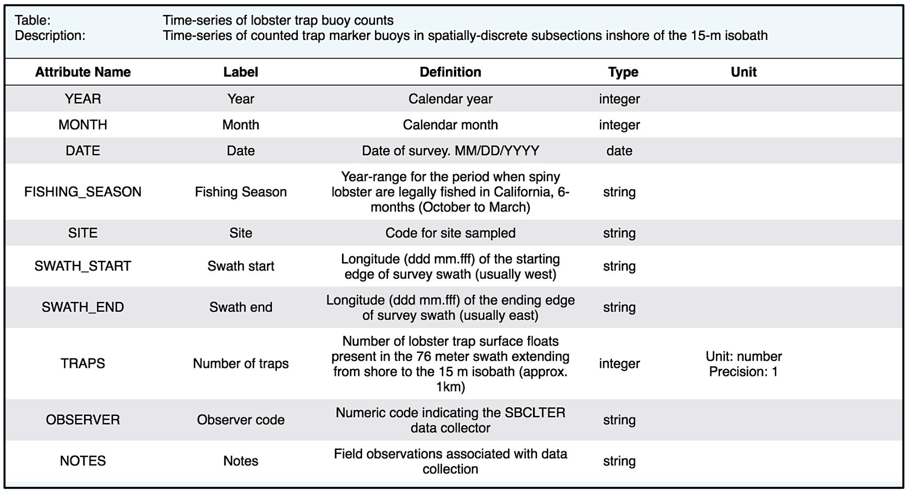


Table: Time-series of lobster trap buoy counts
Description: Time-series of counted trap marker buoys in spatially-discrete subsections inshore of the 15-m isobath

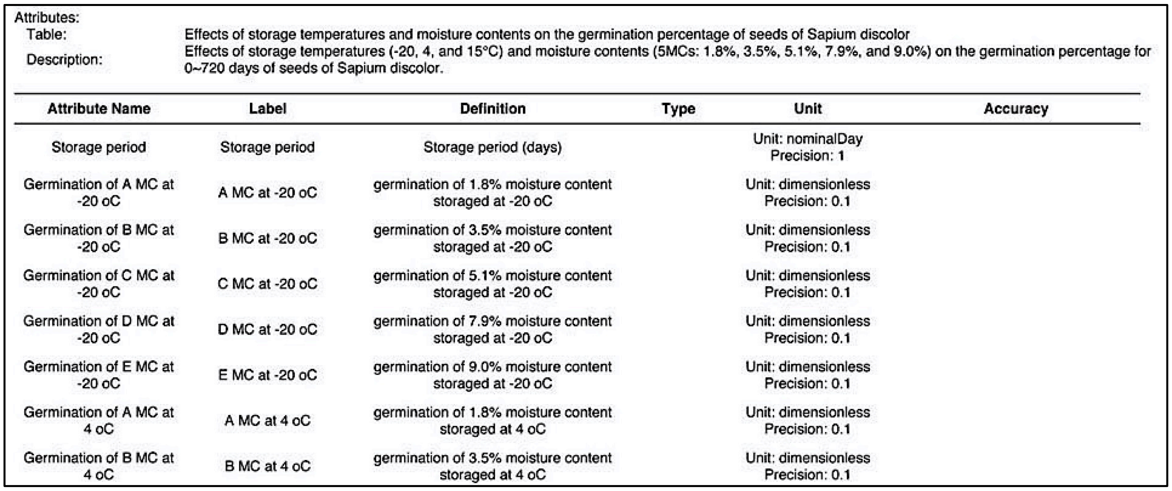
Attribute Name	Label	Definition	Type	Unit
YEAR	Year	Calendar year	integer	
MONTH	Month	Calendar month	integer	
DATE	Date	Date of survey. MM/DD/YYYY	date	
FISHING_SEASON	Fishing Season	Year-range for the period when spiny lobster are legally fished in California, 6-months (October to March)	string	
SITE	Site	Code for site sampled	string	
SWATH_START	Swath start	Longitude (ddd mm.fff) of the starting edge of survey swath (usually west)	string	
SWATH_END	Swath end	Longitude (ddd mm.fff) of the ending edge of survey swath (usually east)	string	
TRAPS	Number of traps	Number of lobster trap surface floats present in the 76 meter swath extending from shore to the 15 m isobath (approx. 1km)	integer	Unit: number Precision: 1
OBSERVER	Observer code	Numeric code indicating the SBCLTER data collector	string	
NOTES	Notes	Field observations associated with data collection	string	

Figure 63. Attribute Table with Data Description

These attribute tables are typically a codebook of the variables in the data and included: attribute name, label, definitions, type, and unit, as well as other items depending on the record. The usefulness of the attribute table are consistent with the literature in that scientists have stated they need good metadata, good record keeping, and codebooks in order to reuse data (Baru, 2007; Edwards et al., 2011). Participants described how they really appreciated seeing the attribute name, label, definition, and unit. However in Result #1, as shown in Figure 64, type and accuracy are not available. Participants stated they preferred to have all of the information.

The Post-Search Survey, particularly the open-ended questions and the Data Reuse Factors Survey reinforce these findings. As shown in Table 16, Post-Search Survey (Reuse Factors), the other category ranked the highest with a mean of 6.6. The attribute table was suggested several times as a factor that assists data reuse. Additionally, when asked what facilitated their data reuse in the open-ended questions, many scientists stated the attribute table is a facilitator.

Some participants stated that although they prefer having more information than not enough information Result #1 took this too far. Participants did not need all of the variations of the variables to understand what type of data they would be receiving (see Figure 64. Participants suggested that a few of the variables is enough information and that including a note to state that there are more variations would be enough information.

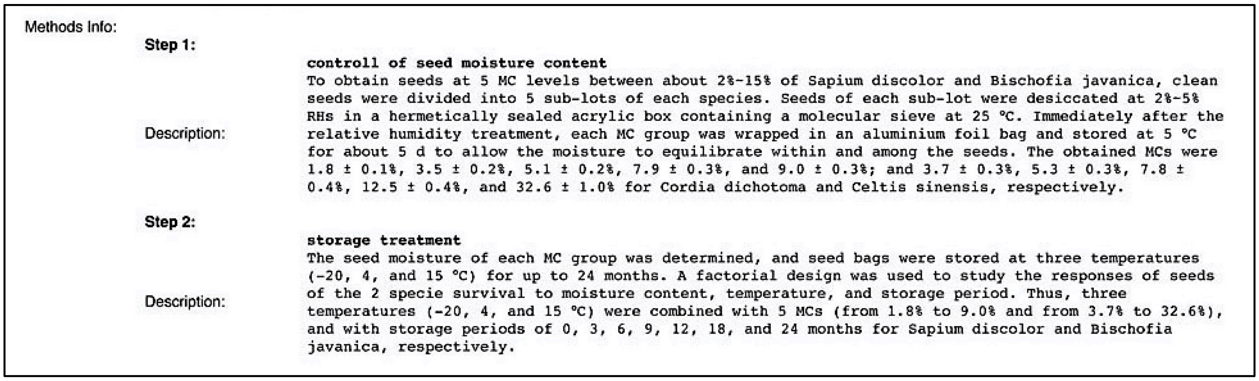


Attributes:
Table: Effects of storage temperatures and moisture contents on the germination percentage of seeds of Sapium discolor
Description: Effects of storage temperatures (-20, 4, and 15°C) and moisture contents (5MCs: 1.8%, 3.5%, 5.1%, 7.9%, and 9.0%) on the germination percentage for 0~720 days of seeds of Sapium discolor.

Attribute Name	Label	Definition	Type	Unit	Accuracy
Storage period	Storage period	Storage period (days)		Unit: nominalDay Precision: 1	
Germination of A MC at -20 oC	A MC at -20 oC	germination of 1.8% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at -20 oC	B MC at -20 oC	germination of 3.5% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at -20 oC	C MC at -20 oC	germination of 5.1% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at -20 oC	D MC at -20 oC	germination of 7.9% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at -20 oC	E MC at -20 oC	germination of 9.0% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of A MC at 4 oC	A MC at 4 oC	germination of 1.8% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at 4 oC	B MC at 4 oC	germination of 3.5% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	

Figure 64. Attribute Table from Result #1

Participants suggested that research methods information facilitates data reusability. The reasons are because they need to understand how the data was gathered, manipulated, and/or analyzed in order to understand if it is relevant to their needs. Research methods information assisted participants in understanding the data more systematically and participants particularly appreciated how the methods were provided in a step-by-step account of the data process. As shown in Figure 65, participants noted the manipulation of the data, how the data were gathered, stored, processed, were crucial for them to understand the data. From here participants could determine more easily if the data were reusable for their purposes.



Methods Info:

Step 1:

Description:

controll of seed moisture content

To obtain seeds at 5 MC levels between about 2%-15% of Sapium discolor and Bischofia javanica, clean seeds were divided into 5 sub-lots of each species. Seeds of each sub-lot were desiccated at 2%-5% RHs in a hermetically sealed acrylic box containing a molecular sieve at 25 °C. Immediately after the relative humidity treatment, each MC group was wrapped in an aluminium foil bag and stored at 5 °C for about 5 d to allow the moisture to equilibrate within and among the seeds. The obtained MCs were 1.8 ± 0.1%, 3.5 ± 0.2%, 5.1 ± 0.2%, 7.9 ± 0.3%, and 9.0 ± 0.3%; and 3.7 ± 0.3%, 5.3 ± 0.3%, 7.8 ± 0.4%, 12.5 ± 0.4%, and 32.6 ± 1.0% for Cordia dichotoma and Celtis sinensis, respectively.

Step 2:

Description:

storage treatment

The seed moisture of each MC group was determined, and seed bags were stored at three temperatures (-20, 4, and 15 °C) for up to 24 months. A factorial design was used to study the responses of seeds of the 2 specie survival to moisture content, temperature, and storage period. Thus, three temperatures (-20, 4, and 15 °C) were combined with 5 MCs (from 1.8% to 9.0% and from 3.7% to 32.6%), and with storage periods of 0, 3, 6, 9, 12, 18, and 24 months for Sapium discolor and Bischofia javanica, respectively.

Figure 65. Method Step Result #2

These findings were reinforced by the post-search survey results (see Table16), 6.13 is the mean for research methods from this survey. Additionally, participants wrote in research methods and experimental set-up under "other" which has a 6.6 mean result. Furthermore, the importance of the research methods is reinforced in the open-ended questions and was often included as a factor that assisted in the facilitation of determining data reusability.

Participants also considered provenance and instrument information important for data reusability, as these have a 5.25 and 5.88 mean, respectively. Additionally, during the quasi-experiment think-aloud several participants discussed provenance and instrument information as important. Participant 14 stated that they needed to know "all arguments of how names evolved, provenance information of names, change history of dataset provenance, as well as who changed the data." Only 37.2% of the records contain provenance information and only 30% contain instrument information. Figure 59 shows an example of provenance information and many of the details participants want in the provenance information, such as who changed the data and logic for changes are located in the example. The information in Figure 59 contains the post processing information, as well as when data lines were deleted, logic/reasons for the deletion/update, date, and by whom. Additionally, these descriptions provide instrument information. We learned from the post-search survey results shown in Table 16, as well as the quasi-experiment think-aloud results, participants do want instrument, calibration, and setting information.

Lastly, as shown in Table 16, metadata standard is not particularly influential in regards to users' ability to reuse data (4.94 mean). During the quasi-experiment think-aloud most participants stated that they wanted at least a good codebook to understand the data, some participants also indicated that the use of a community standard was important. For example, participant 11 stated that in general scientists "want good metadata and to follow a community standard, or at the least a good codebook." Additionally, participant 12 discussed the importance of structure and standards when it came to metadata and described efforts from the Nuclear Data Physics Centre of India to promote high quality metadata through workshops where professional metadata compilers are trained for the EXFOR library/database (Otuka et al., 2014). These examples indicate that although the metadata standard did not rank particularly high in the survey as important, that depending on the scientific community these are important. Scientific communities have come together to help establish standards. The EML community and the FGDC community are quite well represented in DataONE records, and it is clear from these standards are well established in the community. Given that DataONE caters to earth and environmental science data, seeing a large representation from these two standards is not surprising. Additionally, much of this information is new information and the current literature does not provide this level of detail as to what is important for scientists to determine reusability.

Research Question #2b

Research question 2b asks, "How does this information assist scientists in their ability to reuse this data?"

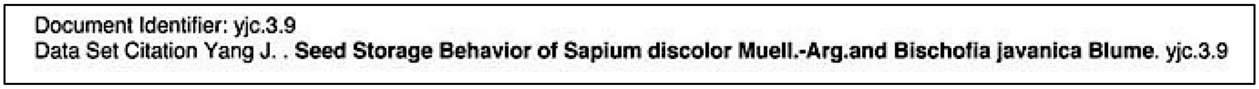
Overall, the participants found the more information provided the more likely they would be able to reuse the data. It was clear that participants preferred Result #1 from the results of the usefulness surveys, the rank order survey, the think-aloud notes, and the post-search survey and open-ended questions, which contained the most robust metadata. There were pieces of information that participants found particularly useful in Result #1 to facilitate data reusability. As described previously, these included the data description, the attribute table, and the research methods information. Participants felt that having this brief description of the data, seeing the variables, units, and definitions from the attribute table, as well as seeing how the data was collected and processed from the methods steps provided them a thorough understanding of the data. Although Result #1 was quite long, only two participants suggested it was overwhelming.

The abstract, keywords, bounding coordinates, licensing information, contact information were also considered useful. All participants stated that although some of this information was somewhat secondary, it was still helpful in their ability to determine if the data was reusable and again, preferred having the information available to them. One participant suggested that they appreciated the "ability to quickly see most important aspects of study before reading the entire article."

Additionally, participants stated that they generally liked the result format. They stated that the information was easy to follow, was what they expected (and more), and appreciated it all being on one page. One participant emphasized he was used to seeing a much less robust result, such as Result #4. He stated that this was what was typical when it came to searching for data. Therefore, he really appreciated seeing Result #1 because it provided him so much more information. Another participant stated it was preferable to know from the one page result if the data was useful or not. She stated she would rather look through a long and thorough result, than "having to search for the data, or wait for the data to be sent, to then find out it was not useful."

An item to note here is that participants very much preferred Result #1 to all of the results. Unfortunately from the data profiling assessment, we learn that only 38% of the results contain a comprehensive attribute table, and only 41% of the records contained comprehensive research methods. Furthermore, the data description was only available with those records that also contained an attribute table, so approximately only 38% of the records contained the data description.

While participants spent much time discussing what facilitated their ability to reuse the data, they also did discuss factors that inhibited their ability to reuse the data. There are only two cases where participant stated the information was not useful. The first case is in regard to the data identifier, as shown in Figure 66. The participants indicated they did not know what this document identifier was referring to and wondered if it was a DataONE identifier. Additionally, participants wanted this stated clearly in the search result and to know if it was only for internal use or had any purpose for users outside of DataONE.



Document Identifier: yjc.3.9
Data Set Citation Yang J. . **Seed Storage Behavior of Sapium discolor Muell.-Arg.and Bischofia javanica Blume**. yjc.3.9

Figure 66. Document Identifier

A second item that confused participants was the download buttons available on the top and bottom of the search result page. The download button was shown at the top and bottom of each result page. Participants indicated that they did not know what they were downloading. They stated that they were unsure if it was the actual data they would download, the metadata record, or perhaps both.

Participants needed to know both what kind of data they were be downloading, as well as the size and format of the data. Participant 1 when asked what inhibited them from reusing the data stated, “The data may be in a format that is difficult to extract. Not knowing the format of the data may lead to the user to not want to use the data”. Additionally, participants were concerned that they did not know what type of data they were downloading. As described by Participant 12,

> if I needed image files in .PNG format, it would save time if I knew that a data-set were .JPG only. Similarly, knowing the size of the data set could be critical (if I don't want to download 10 TBs worth of seed storage temperature data).

Participant 11 described another consideration of not knowing the data type by stating “I don't want to download it and have my computer practically die just from opening the data”. This participant discussed how important it was to know the data size and data type to ensure that their equipment was able to handle the data. Lastly, one participant stated, “Am I about to download a million gigabytes of photos OR a simple csv file?”

Additionally, participants stated their need to know what type of study it was, they want to know the collection method (Field, Experiment, Observational, Sensor, Simulated/Computer Generated, Hybrid) and would have preferred to see this immediately and notes precisely. Participant 3 stated, “I would like to know if it was a controlled experimental, field, or lab work” Additionally, participant 4 suggested that the type of study be included as a keyword. From Table 13, the keyword “Field Investigation” is the most frequent keyword in the records. A secondary analysis could be conducted on the keywords to see if any more of these categories could be ascertained based on more context. This complexity in data types is noted in the literature previously and therefore it was not surprising to see how scientists would want this information (Ailamaki et al., 2010; Anderson, 2004; Borgman et al., 2007; Lide, 1981). However, it was surprising to see how unclear this information was in the results.

Summary of Recommendations from the Quasi-Experiment Think-Aloud

Scientists want a data description, a short succinct statement that describes the data. Additionally, scientists want an attribute and unit list for the data. It is recommended that DataONE and other organizations consider Figure 63 as a best practice. Additionally, scientists want research methods to be clear, precise, and thorough and presented in a step-by-step format so that they are able to understand fully how the data was created and manipulated. It is recommended that DataONE and other such organizations use Figure 65 as an exemplar for how to format research methods information. Additionally, it is recommended that research methods be kept in there own section of the record, rather than incorporated into other areas of the record. These three elements: data description, attribute and unit list, and research methods should be considered the highest priority for DataONE and other similar organizations.

Scientists stated that they do want to have instrument and provenance information available. Unfortunately, as shown in the results and discussion these were not made widely available in the records. It is recommended that this information be made available. It is also recommended that the provenance and instrument information be blended together so that there is context to understand how and why any data changes occurred, such as shown in Figure 59. Additionally, it is recommended that the labeling of provenance information streamlined to be termed "provenance" so that users know where to find this information. Additionally, it is recommended to continue including data citation and DIO information, keywords & keyword thesauri, and geographic/temporal information.

Data type, format, and size should be made clear in the records so that scientists can understand what they are download prior to downloading. Additionally, the document identifier should either be removed or explained since it was unclear what this was referring to.

In general participants stated they needed to know the "who, what, when, where, and how of the data", and it is recommended to use consider all of these when sharing data. Baru (2007) provides a similar analysis of what metadata should be provided to properly understand geoscience data with "it should described not only the "what" (descriptive metadata) but also the "how" (lineage, provenance) and "why" (contextual information)" (Baru, 2007, p. 115). This suggestion from the participants takes Baru's idea and expands upon it. By using the participants suggestion, organizations such as DataONE could provide: who (data creator), what (data type), when (when collected), where (where collected), and how (research methods). The "why" of contextual information should also be included in the records. Lastly, it is recommended that the records be more streamlined so that they all have a similar format and have items located together rather than dispersed.

CHAPTER IX. CONTRIBUTIONS AND CONCLUSIONS

This dissertation contributes to the current body of research by expanding on the work of previous research in multiple areas of study. Additionally, this dissertation provides a methodological contribution not only to mixed methods studies, but also to content analysis studies. Furthermore, it provides both a practical and a theoretical contribution to multiple areas of study. And lastly, it provides specific feedback to DataONE to help improve their current infrastructure, the broader community of organizations working on similar systems, as well as search systems in general. Additionally, this chapter will examine limitations of this study and discuss considerations for future research.

Contributions

Contribution - Expansion of Previous Research

As shown in the literature review, this research is connected to multiple areas of study including: (a) research related to the DataNets and DataONE; (b) research related to data sharing and reuse in the sciences; (c) data management in the sciences; and (d) scientific infrastructure and interoperability factors. This dissertation adds to this literature by providing new information and research to the current body of research.

In regards to research related to the DataNets and DataONE, this research provides feedback on the data made accessible by DataONE; the data that is being shared within DataONE, and additionally provides direct feedback to DataONE ONEMercury as to what scientists need to determine data reusability. Additionally this feedback was provided by earth and environmental scientists, which are the type of scientists that DataONE supports.

For data sharing, this research describes what data and information about that data is being shared into DataONE, one of the most important and well-established environments for sharing and reusing earth and environmental science data. This research provides a snapshot of the data being made available and provides not only the high-level details of this data, but also the more intricate details including the robustness of descriptive information, as well as a detailed analysis of the 30 unique pieces of descriptive information that were found in the records.

Additionally, this research provides details regarding what scientists need to reuse shared data and provides the opportunity to make recommendations as to what descriptive information should be included with shared data. This expands the current literature by answering and addressing the granular and day-to-day needs of data sharing and reuse.

Contribution - Methodological

This research contributes a method and an approach for investigating how descriptive information impacts data sharing and reuse. There are many studies on the topic of data sharing and reuse. These studies use methodologies including interviews, surveys, observations, and bibliometric studies. While these studies have provided a vast amount of understanding of the topic, further research still needed to be conducted. Specifically, research that includes other methods outside of self-report could be particularly beneficial for the understanding of what factors influence data sharing and reuse. This dissertation provides such a study.

This research, by employing a mixed methods approach, triangulates data from various sources and provides a richer and deeper understanding of data sharing and reuse. Furthermore, the data used for this dissertation includes instance of data sharing and potential reuse. Through the profiling data assessment, actual instances of data that are shared through DataONE are analyzed. Through the quasi-experiment think-aloud, data that is potential reused are analyzed. These are methods that have not been utilized in the current research in this area of study (to this researchers knowledge). Additionally, many of the past studies have only looked at one aspect of data sharing and reuse, meaning they have focused only on the data sharing side or only on the data reuse side.

The profiling data assessment utilized the content analysis method in a unique way. The content analysis not only analyzing the sending and receiving of the messages, but also analyzes the information about the data (messages). This content analysis considered both quantitative and qualitative variables, as well as variables that were only able to be determined in the context of each other, as shown in section 7.1.4 by investigating location of research methods, location of instrument information, access and intellectual rights information, and data availability. To address these items, a mixture of qualitative and quantitative of content analysis methods needed to be used.

Contribution - Practical and Theoretical

This research takes advantage of the opportunity to examine and explore a rich environment, DataONE, which supports both the sharing side (making data available) and the reuse side (reusing available data). This provides insight for the factors that influence data sharing and reuse for the earth and environmental science community, funders, the research data management community, scientific data repositories, developers of repositories and other tools, and developers of general search systems.

This dissertation provides insights on data sharing and reuse in the earth and environmental science community. While there have been many studies conducted in the biomedical and health sciences areas, the earth sciences are somewhat understudied in regards to data sharing and reuse. From the results, specific findings may be more pertinent to the earth and environmental scientific community than to other communities. For example, the participants were all very concerned about having knowledge of the data collection method (i.e. field, experimental, simulated) and the data type and format. This finding may be more relevant to earth and environmental scientists based on the highly interdisciplinary nature of these sciences and may be indicative of the heterogeneity of data types found in the earth and environmental sciences.

This dissertation contributes to the scientific community specifically with regard to data lifecycles and data management plans. This dissertation provides knowledge as to what information about the data is being shared and what descriptive information users need to reuse this data. The findings and recommendations could be used for researchers to understand the data lifecycle in more detail. By examining shared data and what users need to reuse data, researchers could determine if current data lifecycle models need to be revised. Additionally, this research provides knowledge for creators of data management plans. Funders such as the NSF and the NIH could examine if their data management policies align with user needs. Additionally, the data management plan templates could be revised based on what is learned with regards to what scientists need to determine reusability. These types of revisions to data management plan templates will help assist data managers in guiding data depositors and providing the types of descriptive information and metadata most relevant to scientists. Additionally this feedback helps developers of data management plan tools, such as the DMPTool.

This dissertation assists organizations such as RDA, EarthCube, and ESIP. These organizations have focused much effort to understand the data lifecycle, the needs of scientist, and creation of tools to assist in data management. Additionally, these organizations develop tools for scientists and provide outreach and education. These findings from this dissertation could assist these organizations develop tools more in-line with scientists needs. Additionally, these organizations could create educational and outreach modules that align with the needs of scientists based on the recommendations from this dissertation. Lastly, organization such as ESIP has research and development efforts. The findings from this dissertation provide guidance for these R&D efforts.

The findings and recommendations could assist scientific data repositories outside of the natural sciences. Social science data repositories such as the ICPSR and the Odum Institute could use the recommendations to ensure that social scientists have the descriptive information they need to reuse data. Findings such as wanting to have a data description and the "who, what, when, where, and how" of the data are cross-disciplinary and can be implemented in all data repositories. These pieces of descriptive information that were found most important to the participants could become required for data sharers and could be included as required in data management plan templates. Additionally, there are data sharing organizations in the sciences such as CIESEN and NEON that could benefit from the recommendations of this dissertation by ensuring they are providing the descriptive information that scientists find more pertinent for data reusability in their current data repositories, as well as use these recommendations as best practices for anyone sharing data with these data repositories.

Furthermore, general search systems could consider the recommendations from this dissertation. These recommendations are important to all people searching the open web. Searchers want succinct descriptions of the website they are about to open. Searchers also want the major categories of information who is the author, what does the website provide them, when was the website created, where is the information from, and how was the information created and how can the user trust the quality. Searchers also want to be able to choose the information they see and therefore the suggestion of having drop-down menus is relevant for all search systems. Lastly, searchers sometimes need help understanding what they are reading, and the suggestion of having a help or hover for any information that could be confusing is useful to all websites.

Table 22 provides a summary of the previous research and the contributions of this dissertation research.

Table 22

Summary of Previous Research and New Contributions from Dissertation

Previous Sharing/Reuse Research	**New Contributions from Dissertation**
Policy-based studies: • Addressed journal and funding agency studies, data management policies, and data deposition policies.	Shared Data: • Determines what data and information is being shared. • Analyzes a live environment • Focuses on detailed information such as research methods, provenance information, data citation & DOI, attribute and unit lists. • Addresses infrastructure impact on shared data. • Addresses descriptive information impact on sharing and reusing data.
Motivators: • Explored scientific norms, scientific reputation, data value, and duplication. Inhibitors: • Described time and effort factors, work experience, ownership, financial concerns, and colleague helpfulness.	Reused Data: • Focuses on user information needs. • Focuses on detailed information provided to user and users response to this information. • Determines what users needed and wanted to determine reusability.
Answered and addressed broad questions about data sharing and reuse.	***Answers and addresses granular and day-to-day practices of data sharing and reuse.*** ***Provides recommendation and implications for broader audience, including:*** • *Data sharing organizations (e.g. ICSPR, CIESIN, NEON)*

	• *Research data management & scientific community (e.g. RDA, EarthCube, ESIP)* • *Data management plan creators and Funding agencies (e.g. DMP templates and creators of resources such as DMPTool)*

Limitations

This study does have several limitations including: 1) focusing only on the earth sciences and earth scientists, and DataONE, 2) small participant size for quasi-experiment think-aloud, 3) hypothetical searches and results rather than queries provided by participants.

Focusing on the earth science and DataONE was a consideration when designing this study. It was decided to focus mainly on one scientific subject area and that the earth sciences provided a particularly interesting area of study for several reason. DataONE provides an effective environment of study given that it provides both the data being shared into the environment, as well as the ability for scientists to reuse that data. Additionally, the earth sciences provide a unique perspective because, as shown in the literature review because the earth sciences are somewhat understudied. As seen in Figure 5, only a small amount of literature focused predominately on the earth sciences. Additionally, the earth sciences are highly interdisciplinary sciences, and therefore provided a unique perspective for study.

A second major limitation to this study is a small sample size for the quasi-experiment think-aloud. While multiple recruitment efforts were made in the summer 2015, fall 2015, and spring 2016 only 16 participants responded to the many calls for participation through the UNC listserv, the NCSU listserv, face-to-face recruitment at earth science conferences including AGU and GSA, and targeted email recruitment efforts. While more participants could have potentially provided more information, there was a fairly clear consensus from the sixteen participants. It became apparent that the participants were in clear agreement as to what inhibited and facilitated their ability to reuse the data, which pieces of descriptive information were particularly useful, and ultimately what determined data reusability.

The last limitation of the study was that I provided the query and search results for the participants of the quasi-experiment think-aloud. This decision was made when designing the study. I considered having participants determine their own search terms, but it became apparent through the pilot studies that this would lead to too many uncontrolled search results which would have created too much variation in the robustness of the results and would not provide the researcher the ability to test which descriptive information was most pertinent.

Future Research and Research Agenda

While conducting the quasi-experiment think-aloud portion of this study, one particular study came to mind. The idea is to have participants draw out their own ideas of how they would like the record organized and what descriptive information they prefer to have and where. While open-ended questions did ask what information inhibited and facilitated their ability to reuse, as well as what information they wanted was not provide, there was a lot of conversation in the think-aloud about placement of information. Providing participants the ability to draw out their placement preference would have provide further information, as well as potentially provided information that was not originally listed in the search result that the participant may have thought of while drawing their "preferred result".

Another possibility would be to talk with data providers to examine how they make decisions regarding what descriptive information to make available and what data to make available. This would provide information regarding why data providers made the information available that they made. An examination of the tools that data providers use to input data such as Dash, the DMPTool, and Morpho could be cross-compared to scientists needs. These tools could be examined to determine how well they correspond with actual needs of both data providers and scientists. Lastly, these tools could be examined for biases, meaning how does the nature of these tools affect what information is provided and how does this correspond with what the researchers actually need to determine reusability. Expanding on the above, an examination of the data sharing and reuse lifecycles could be insightful. Through keeping in mind the various tools and funder requirements, an examination could compare the actual practices of data providers and scientists in relation data sharing and reuse lifecycles, current tools, and funder requirements. How these tools and funder requirements match up with actual needs could be useful in assisting tool creation.

Additionally, a similar mixed-method study could be used to examine other well-established data sharing and reuse organizations to see if there are similar results to this study. A comparison study would be useful to see if these results can be duplicated in a similar environment and to provide generalizable results and recommendations for these types of organizations. There are many other similar environments such as the National Snow & Ice Data Center, SEDAC, NASA Earth Data Search, the GEONGRID, and NEON and conducting a duplicate study with a similar environment could assist in understanding if the recommendations from this dissertation hold true in other environments.

Lastly, stories of actual sharing and actual reuse events could be examined. This could be conducted either an exploratory qualitative study or a critical incident study to have participants talk about their data sharing and reuse experiences. Or having scientists walk through their data reuse experience and talk through step-by-step how they acquired and reused the data, what descriptive information they needed, any inhibitors that prevented them from using the data and any facilitators that ultimately assisted them with reusing the data. Rather than examining hypothetical data sharing and reuse events, these studies would focus on examining true reuse events.

Final Conclusions

This dissertation examined data sharing and reuse through: 1) conducting a quantitative and qualitative content analysis of data made accessible through DataONE and 2) conducting a quasi-experiment think-aloud study with results from DataONE ONEMercury. Table 23 provides an overview of the research questions, methods, and sample from each part of this dissertation

Table 23

Summary of Study

Part 1: Data Profiling Assessment	**Part 2: Quasi-Experiment Think-Aloud**	**Part 3: Comparison of results**
What types of descriptive information are being made discoverable through DataONE? • How robust is the descriptive information made available regarding that data? • How is information being provided about the data, such as information regarding metadata standards, provenance information, research methods, instrumentation? • How is the provision of this descriptive information impacting the data-sharing infrastructure?	What types of descriptive information could inhibit or facilitate data reuse? • How is information about the data such as information regarding metadata standards, provenance information, research methods, and instrumentation, influencing scientists' ability to determine if that data is reusable? • How does this information assist scientists in their ability to reuse this data?	Exploration of how both part 1 and part 2 can address research questions, as well as how the results from both parts inform the study.
Method: Quantitative and qualitative content analysis	Method: Quasi-Experiment Think-Aloud	Data: Comparison of results from first two data collections
Sample: 157 DataONE Records were analyzed	Sample: 16 participants	Sample: All results from

While there were recommendations made throughout Chapter 8, the most vital recommendations are reiterated here. It is recommended that the metadata fields in DataONE be streamlined so that the field names are consistent throughout the records. For example, there were many ways that users could find information in regards to creator/contact. It would be beneficial to remove the duplication of this information and provide more succinct information regarding creator and contact so that users are not looking for information in multiple places.

It is recommended that all records contain data citation information and DOI information in the format that Dryad supplied with a traditionally formatted citation and DOI. Additionally, it is recommended that any use requirements be placed in the record near the data citation information so that users can find this all in one location.

It is recommended that data descriptions, attribute, and unit lists are made a priority and that DataONE recommend that all Member Nodes include this information in their records. The data description should be located near the top of the record. Additionally, it is recommended that attribute and unit lists provide a streamlined summary of the attributes, as shown in Figure 63. Additionally, it is recommended that data size, type, and format be incorporated more clearly into the records so that users are aware of what they are downloading. Furthermore, it is recommended that records indicate how the data was collected (field study, experimental study, simulated study) so that users understand what type of data they are looking at. Lastly, it is recommended that research methods be kept in their own section and are thoroughly and precisely written out with methods steps as shown in Figure 65.

In conclusion, this research contributed a better understanding of data sharing and reuse. The findings have practical implications in that they provide a vast amount of recommendations to assist in making data sharing and reuse environments more useful to users. Additionally, these findings demonstrate the type of information that is being shared, as well as demonstrate the type of information needed for scientists to determine reusability. This ultimately addresses the changes that need to be made to data sharing and reuse infrastructures to make them the most valuable to scientists.

APPENDIX 1: ALL MEMBER NODES AS OF 7/2015

Member Nodes not included in dissertation highlighted in grey.

Member Node	**Metadata Records (241,138 as of 7/31/2015)**
CLO eBird	2
Dryad	40,003
Earth Data Analysis Center (EDAC)	357
Environmental Data for the Oak Ridge Area	28
ESA Data Registry	157
Europe Long-Term Ecosystem Research Network (LTER Europe)	167
Global Lake Ecological Observatory Network (GLEON)	20
Gulf of Alaska Data Portal	487
International Arctic Research Center (IARC) Data Archive	352
Knowledge Network for Biocomplexity	5,786
LTER Network Member Node	84,581
Merritt Repository	31,604
Minnesota Population Center (MPC)	258
Montana IoE Data Repository	73
NM EPSCoR	7
ONEShare Repository	127
ORNL DAAC	1,237
PISCO MN	68,588
Regional and Global Biogeochemical Dynamics Data (RGD)	273
SANParks Data Repository	1,640
SEAD Virtual Archive	13
Taiwan Forestry Research Institute	2,513
Terrestrial Ecosystem Research Network	2,382
University of Kansas-Biodiversity Institute	172
USA National Phenology Network	14
USGS Core Sciences Clearinghouse	250

APPENDIX 2: PRELIMINARY CODEBOOK

Robustness Scale	**0 – No Information**	**1 – Adequate Information**	**2 – Comprehensive Information**
Top-Level Elements			
Dataset			
Access			
Additional Metadata			
Additional Attributes			
Metadata Standard			
Provenance Information			
Instrumentation Information			
Research Methods Information			
Associated Party			
Creator			
Metadata Provider			
Additional Access Information			
Data Type			

APPENDIX 3: FINAL CODEBOOK AND DEFINITIONS

Name	**Definition**
File Size	The file size of the corresponding XML file for the record. *(KB)*
Data Citation	Information regarding how to cite the data such as a suggested citation, DOI, etc. *(RS)*
Dataset Metadata	Top-level category regarding the dataset. *(RS)*
Access Metadata	Top-level category regarding access. *(RS)*
Additional Metadata	Top-level category regarding any additional metadata outside of information regarding access and/or the dataset. *(RS)*
Metadata Standard	The metadata standard used for the record. *(Full text)*
Additional Metadata Standard Information	Additional information regarding the metadata standard that was used for the record. *(RS)*
Provenance Information	Information regarding changes made to the record; includes change history and maintenance. *(RS)*
Instrumentation Information	Information regarding the instruments that were used to collect the data. *(RS)*
Research Methods Information	Information regarding the methods used to collect the data. (RS)
Associated Party	If there is another party involved with the data outside of the creator or metadata provider. *(0 = no, 1 = yes)*
Creator and Metadata Provider Information	The creator or the organization that provided the metadata for the record. *(Full text)*
Contact Information	Contact information for the creator or metadata provider. *(0 = no, 1 = yes)*
Publication Date	The date the data was published. *(Full text/numeric/date)*
Abstract	The brief summary typically of the project, which gathered the associated data. *(Full text & RS)*
Keywords	The keywords listed for the data. *(Full text)*
Additional Access Information	Additional information regarding access and use rights and restrictions. (Full text & RS)
Temporal Coverage	The date range covered by the data. *(Full text/numeric/date)*
Taxonomic Information	Information regarding the taxonomical coverage. *(RS)*
Publisher Information	Information regarding the publisher. *(RS)*
Data Type	The type of data available. *(Full text)*
Keyword Thesauri	If there is a thesaurus or thesauri associated with the keywords. *(Full text)*
Attribute List	List of variables in the data. *(RS)*
Unit List	Unit for each attribute or variable. *(RS)*
Geographic Information	Information regarding geographic location, bounding coordinates, etc. *(RS)*
Funding Source	The grant or the agency that funded the project. *(RS)*
Methods (part of abstract	Description of where the methods were located within the

or own section)	metadata snippet. *(0 = part of abstract, 1 = own section)*
Instrumentation (part of abstract or own section)	Description of where the instrumentation information was located within the record metadata snippet. *(0 = part of abstract, 1 = own section)*
Intellectual rights (multiple steps for use or none)	Description of where the instrumentation information was located within the record metadata snippet. *(0 = no steps, 1 = multiple steps)*
Data Availability	If the data is readily available. *(0 = no link to data/contact provider, 1= direct link to data, 2 = data directly available)*
Measurements	***RS*** = Robustness Scale – 0 = no information; 1 = adequate information; 2 = comprehensive information ***Full text***: Written text

APPENDIX 4: MEMBER NODES AT THE TIME OF DATA COLLECTION

Member Node	Metadata Records (183,702 as of 10/21/14)
CLO eBird	2
ESA Data Registry	157
Dryad Digital Repository	25,838
Earth Data Analysis Center (EDAC)	357
Europe Long-Term Ecosystem Research Network (LTER Europe)	167
Gulf of Alaska Data Portal	481
Knowledge Network for Biocomplexity	5,625
LTER Network Member Node	45,489
Merritt Repository	31,590
Montana IoE Data Repository	73
ONEShare Repository	127
ORNL DAAC	1,226
PISCO MN	68,101
SANParks Data Repository	1,638
SEAD Virtual Archive	12
Taiwan Forestry Research Institute	2,383
USA National Phenology Network	14
USGS Core Sciences Clearinghouse	250
University of Kansas - Biodiversity Institute	172

APPENDIX 5: THINK-ALOUD EMAIL RECRUITMENT

SUBJECT: [Call for Participation] Factors Influencing Data Reuse within DataONE

I am a doctoral candidate at the University of North Carolina at Chapel Hill leading a research investigation of factors influencing data reuse within DataONE. DataONE provides users access to earth science observational data.

In order to be eligible, you will need to be a faculty, research scientists, or doctoral student in the sciences.

You will be asked to search the system and think aloud about the results you receive in regards to the reusability of the data. You will also be asked to fill out a survey about data reuse. You will be compensated with a $20 visa gift card for your time, which will take approximately 30 - 60 minutes.

Please contact me via email at amurillo@email.unc.edu if you are interested in participating.

Sincerely,
Angela Murillo
Doctoral Candidate
School of Information and Library Science,
University of North Carolina at Chapel Hill
Email: amurillo@email.unc.edu

APPENDIX 6: QUASI-EXPERIMENT THINK-ALOUD OBSERVATION GUIDE

Researcher:

Thank you for agreeing to meeting with me. What I would like you to do is to search DataONE ONEMercury system and talk aloud about what you are looking for and the why you are making the choices you are making. At certain times, I may stop and ask you questions regarding your search.

Here is the system URL (researcher will provide URL for interface created for study)

- While the participant is searching, the researcher will take notes of the search terms that they used and the facets that they used.
- The researcher will also take notes about what the participant is saying regarding which choices they've made during their search.
- The researcher will make notes regarding which results they felt were most relevant and why those results were most relevant.
- The researcher will ask the participant to discuss why certain results were more relevant than others and they them specifically about what information the participant needed in order to determine if the result was relevant or not.

APPENDIX 7: THINK-ALOUD CONSENT FORM

Consent to Participate in a Research Study

Title of Study: Data Sharing and Reuse in the Sciences: An Investigation of Infrastructure Factors
Principal Investigator: Angela P. Murillo, amurillo@email.unc.edu, (919) 962-8366
Faculty Adviser: Jane Greenberg, jg3243@drexel.edu, (215) 895-2490

What are some general things you should know about research studies?
You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, without penalty. Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study.

What is the purpose of this study?
The purpose of this research study is to gain a better understanding of the factors that influence data reuse.

How many people will take part in this study?
If you decide to be in this study, you will be one of approximately 40 people in this research study.

Are there any reasons you should not be in this study?
You should not be in this study if you are under 18 years of age or are not a scientist.

What will happen if you take part in the study?
Your part in this study will last approximately 30 - 60 minutes. During this study, you will conduct searches on an online search interface and think aloud regarding what influences your ability to reuse the data provided by the search interface. Additionally, you will answer a brief online survey at the end of your search.

What are the possible benefits from being in this study?
Research is designed to benefit society by gaining new knowledge.

What are the possible risks or discomforts involved from being in this study?
We anticipate few risks in this study. You should report any problems to the researcher.

How will your privacy be protected?
All of the data you provide will be stored anonymously. This means that there will be no way for anybody to ever link your data or the results of the study to your identity.

What if you want to stop before your part in the study is complete?

You can withdraw from this study at any time, without penalty and skip any question for any reason. The investigators also have the right to stop your participation if you have an unexpected reaction, have failed to follow instructions, etc.

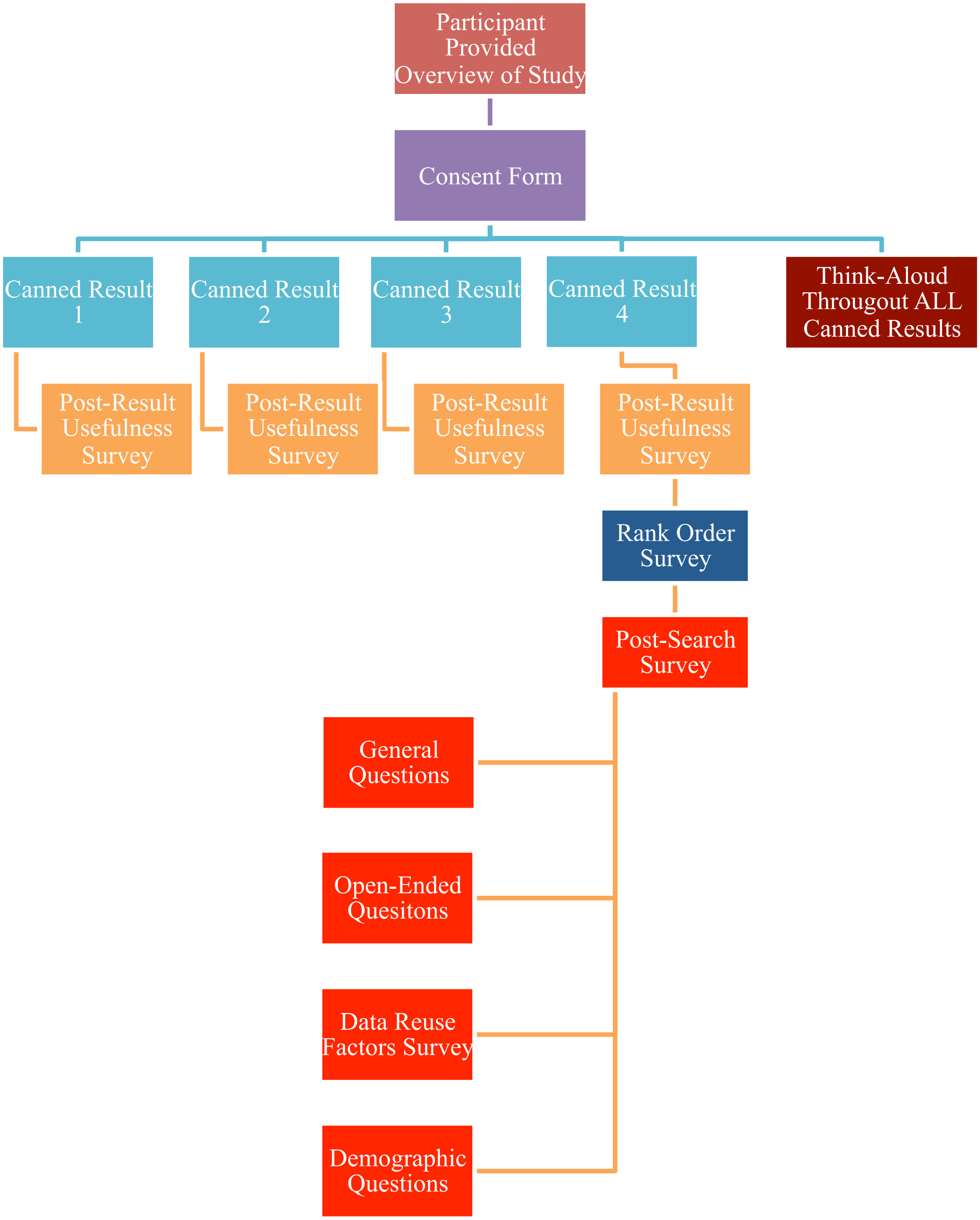
Will you receive anything for being in this study? Will it cost anything?
You will receive a $20 gift card for participating in this study. There are no costs associated with being in the study.

What if you have questions about this study?
You have the right to ask, and have answered, any questions you may have about this research. Contact the principal investigator, Angela Murillo, amurillo@email.unc.edu, with any questions, complaints, or concerns you may have.

What if you have questions about your rights as a research participant?
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns, or if you would like to obtain information or offer input, please contact the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

❍ I consent
❍ I do not consent

APPENDIX 8: QUASI-EXPERIMENT THINK-ALOUD PROCEDURES/DATA COLLECTION


Participant Provided Overview of Study
Consent Form
Canned Result 1
Canned Result 2
Canned Result 3
Canned Result 4
Think-Aloud Througout ALL Canned Results
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Post-Result Usefulness Survey
Rank Order Survey
Post-Search Survey
General Questions
Open-Ended Quesitons
Data Reuse Factors Survey
Demographic Questions

APPENDIX 9: POST RESULT USEFULNESS SURVEY & RANK ORDER SURVEY

Post Result Usefulness Survey

On a scale of 1 to 5, with 1 being the not useful and 5 being very useful, how would you rate this results in regards to assisting you in the ability to reuse the data?

Asked after the examination of each result by participant.

Post-Search Rank Order Survey

Please rank all results in the order of most useful to least useful in regards to assisting you in the ability to reuse the data?

Asked after participant examined all results prior to the post-search survey.

APPENDIX 10: POST-SEARCH SURVEY

General Questions

What is your subject expertise in?

Had your ever searched DataONE ONEMercury system?

❍ Yes
❍ No
❍ Unsure

How often have you searched?

❍ Never
❍ Rarely
❍ Sometimes
❍ Quite Often
❍ Very Often

Open Ended Questions

When looking at the search results about, what information did you need to determine if the data is relevant?

In regard to DataONE, what information inhibits your ability to reuse data?

In regard to DataONE, what information facilitates your ability to reuse data?

When thinking about DataONE, what information did you want that the system did not provide?

Data Reuse Factors Survey

(IN GENERAL) When looking at search results for data, what information do you need in order to determine if the data is relevant?

Not at all Important (1)
Very Unimportant (2)
Somewhat Unimportant (3)
Neither Important nor Unimportant (4)
Somewhat Important (5)
Very Important (6)
Extremely Important (7)

1. The data follows a specific metadata standard.
2. The data contains metadata regarding provenance information.
3. The data contains metadata regarding permission and intellectual property rights.
4. The data contains metadata regarding instrumentation.
5. The data contains metadata regarding instrumentation.
6. The data contains metadata regarding the research methods used to collect the data.
7. Other: Please specify

Demographic Questions

Sex
❍ Female
❍ Male
❍ Prefer not to answer
❍ I identify as/In another way (please specify if you wish): ____________________

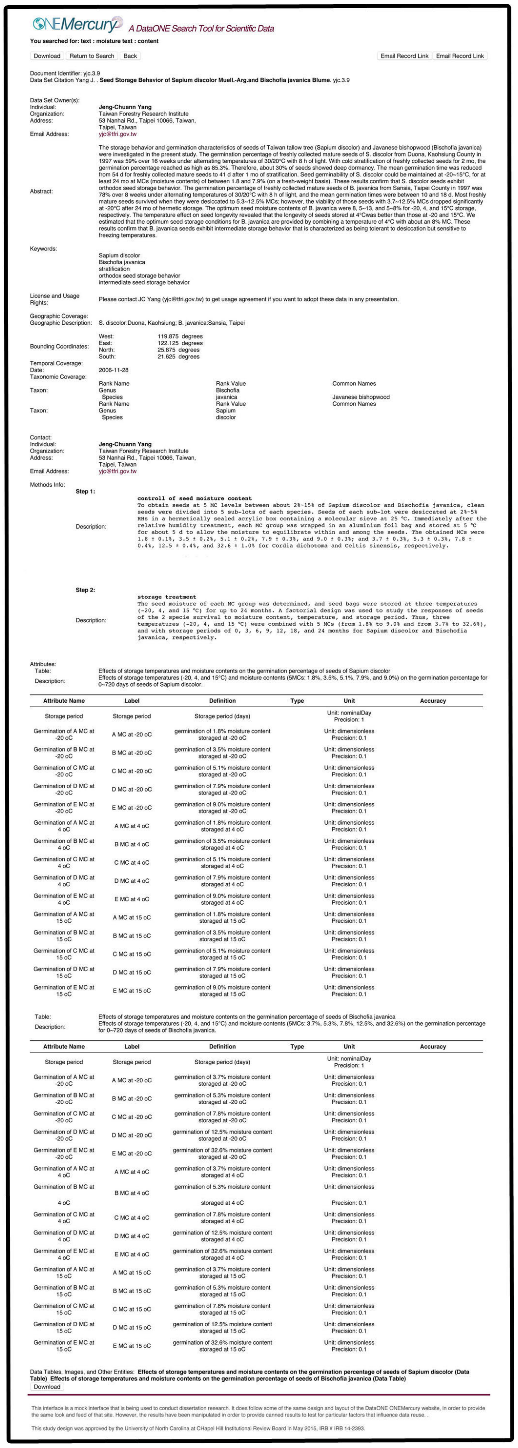
Years of professional experience
❍ 0 - 5 years
❍ 6 - 10 years
❍ 11 - 20 years
❍20 + years

Educational background
❍ High School
❍ BA/BS
❍ MA/MS
❍ PhD
❍ Other

Q7 Area of Expertise (Please select all that apply)
❍ Ecology
❍ Geology
❍ Biology
❍ Atmospheric Science
❍ Environmental Sciences
❍ Hydrology
❍ Soil Science
❍ Chemistry
❍ Physics
❍ Computational/Computer Sciences
❍ Other: Please Specify ____________________

APPENDIX 11: QUASI-EXPERIMENT THINK-ALOUD INTERFACE SEARCH RESULTS

Result #1



You searched for: text : moisture text : content

Document Identifier: yjc.3.9
Data Set Citation Yang J. . **Seed Storage Behavior of Sapium discolor Muell.-Arg.and Bischofia javanica Blume**. yjc.3.9

Data Set Owner(s):
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Abstract: The storage behavior and germination characteristics of seeds of Taiwan tallow tree (Sapium discolor) and Javanese bishopwood (Bischofia javanica) were investigated in the present study. The germination percentage of freshly collected mature seeds of S. discolor from Duona, Kaohsiung County in 1997 was 59% over 16 weeks under alternating temperatures of 30/20°C with 8 h of light. With cold stratification of freshly collected seeds for 2 mo, the germination percentage reached as high as 85.3%. Therefore, about 30% of seeds showed deep dormancy. The mean germination time was reduced from 54 d for freshly collected mature seeds to 41 d after 1 mo of stratification. Seed germinability of S. discolor could be maintained at -20~15°C, for at least 24 mo at MCs (moisture contents) of between 1.8 and 7.9% (on a fresh-weight basis). These results confirm that S. discolor seeds exhibit orthodox seed storage behavior. The germination percentage of freshly collected mature seeds of B. javanica from Sansia, Taipei County in 1997 was 78% over 8 weeks under alternating temperatures of 30/20°C with 8 h of light, and the mean germination times were between 10 and 18 d. Most freshly mature seeds survived when they were desiccated to 5.3~12.5% MCs; however, the viability of those seeds with 3.7~12.5% MCs dropped significantly at -20°C after 24 mo of hermetic storage. The optimum seed moisture contents of B. javanica were 8, 5~13, and 5~8% for -20, 4, and 15°C storage, respectively. The temperature effect on seed longevity revealed that the longevity of seeds stored at 4°Cwas better than those at -20 and 15°C. We estimated that the optimum seed storage conditions for B. javanica are provided by combining a temperature of 4°C with about an 8% MC. These results confirm that B. javanica seeds exhibit intermediate storage behavior that is characterized as being tolerant to desiccation but sensitive to freezing temperatures.

Keywords:
Sapium discolor
Bischofia javanica
stratification
orthodox seed storage behavior
intermediate seed storage behavior

License and Usage Rights: Please contact JC Yang (yjc@tfri.gov.tw) to get usage agreement if you want to adopt these data in any presentation.

Geographic Coverage:
Geographic Description: S. discolor:Duona, Kaohsiung; B. javanica:Sansia, Taipei

Bounding Coordinates:		
	West:	119.875 degrees
	East:	122.125 degrees
	North:	25.875 degrees
	South:	21.625 degrees

Temporal Coverage:
Date: 2006-11-28
Taxonomic Coverage:

	Rank Name	Rank Value	Common Names
Taxon:	Genus	Bischofia	
	Species	javanica	Javanese bishopwood
	Rank Name	Rank Value	Common Names
Taxon:	Genus	Sapium	
	Species	discolor	

Contact:
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Methods Info:

Step 1:

Description: **controll of seed moisture content**
To obtain seeds at 5 MC levels between about 2%~15% of Sapium discolor and Bischofia javanica, clean seeds were divided into 5 sub-lots of each species. Seeds of each sub-lot were desiccated at 2%~5% RHs in a hermetically sealed acrylic box containing a molecular sieve at 25 °C. Immediately after the relative humidity treatment, each MC group was wrapped in an aluminium foil bag and stored at 5 °C for about 5 d to allow the moisture to equilibrate within and among the seeds. The obtained MCs were 1.8 ± 0.1%, 3.5 ± 0.2%, 5.1 ± 0.2%, 7.9 ± 0.3%, and 9.0 ± 0.3%; and 3.7 ± 0.3%, 5.3 ± 0.3%, 7.8 ± 0.4%, 12.5 ± 0.4%, and 32.6 ± 1.0% for Cordia dichotoma and Celtis sinensis, respectively.

Step 2:

Description: **storage treatment**
The seed moisture of each MC group was determined, and seed bags were stored at three temperatures (-20, 4, and 15 °C) for up to 24 months. A factorial design was used to study the responses of seeds of the 2 specie survival to moisture content, temperature, and storage period. Thus, three temperatures (-20, 4, and 15 °C) were combined with 5 MCs (from 1.8% to 9.0% and from 3.7% to 32.6%), and with storage periods of 0, 3, 6, 9, 12, 18, and 24 months for Sapium discolor and Bischofia javanica, respectively.

Attributes:
Table: Effects of storage temperatures and moisture contents on the germination percentage of seeds of Sapium discolor
Description: Effects of storage temperatures (-20, 4, and 15°C) and moisture contents (5MCs: 1.8%, 3.5%, 5.1%, 7.9%, and 9.0%) on the germination percentage for 0~720 days of seeds of Sapium discolor.

Attribute Name	Label	Definition	Type	Unit	Accuracy
Storage period	Storage period	Storage period (days)		Unit: nominalDay Precision: 1	
Germination of A MC at -20 oC	A MC at -20 oC	germination of 1.8% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at -20 oC	B MC at -20 oC	germination of 3.5% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at -20 oC	C MC at -20 oC	germination of 5.1% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at -20 oC	D MC at -20 oC	germination of 7.9% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at -20 oC	E MC at -20 oC	germination of 9.0% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of A MC at 4 oC	A MC at 4 oC	germination of 1.8% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at 4 oC	B MC at 4 oC	germination of 3.5% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at 4 oC	C MC at 4 oC	germination of 5.1% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at 4 oC	D MC at 4 oC	germination of 7.9% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at 4 oC	E MC at 4 oC	germination of 9.0% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of A MC at 15 oC	A MC at 15 oC	germination of 1.8% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at 15 oC	B MC at 15 oC	germination of 3.5% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at 15 oC	C MC at 15 oC	germination of 5.1% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at 15 oC	D MC at 15 oC	germination of 7.9% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at 15 oC	E MC at 15 oC	germination of 9.0% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	

Table: Effects of storage temperatures and moisture contents on the germination percentage of seeds of Bischofia javanica
Description: Effects of storage temperatures (-20, 4, and 15°C) and moisture contents (5MCs: 3.7%, 5.3%, 7.8%, 12.5%, and 32.6%) on the germination percentage for 0~720 days of seeds of Bischofia javanica.

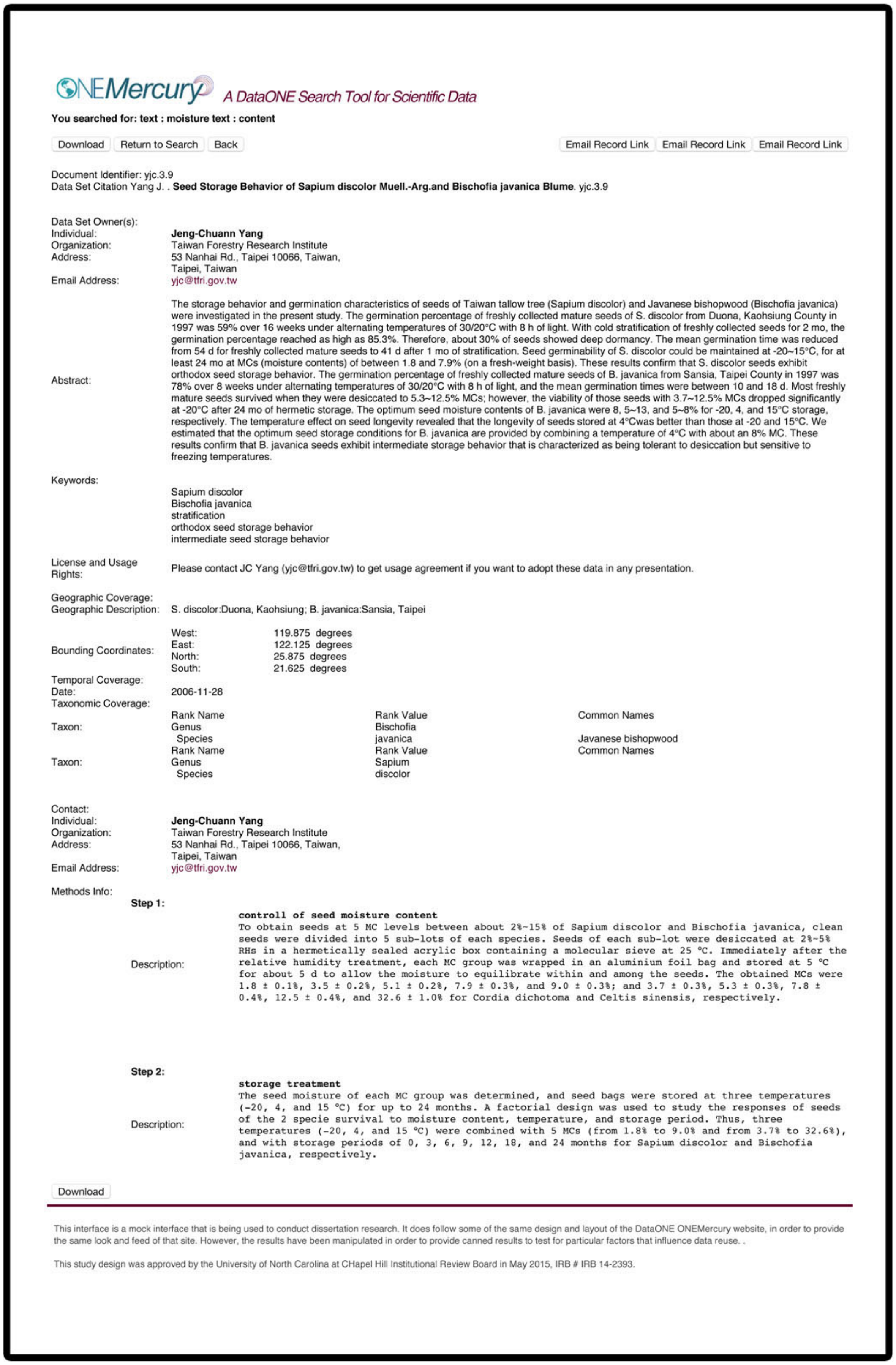
Attribute Name	Label	Definition	Type	Unit	Accuracy
Storage period	Storage period	Storage period (days)		Unit: nominalDay Precision: 1	
Germination of A MC at -20 oC	A MC at -20 oC	germination of 3.7% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at -20 oC	B MC at -20 oC	germination of 5.3% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at -20 oC	C MC at -20 oC	germination of 7.8% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at -20 oC	D MC at -20 oC	germination of 12.5% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at -20 oC	E MC at -20 oC	germination of 32.6% moisture content storaged at -20 oC		Unit: dimensionless Precision: 0.1	
Germination of A MC at 4 oC	A MC at 4 oC	germination of 3.7% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at 4 oC	B MC at 4 oC	germination of 5.3% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at 4 oC	C MC at 4 oC	germination of 7.8% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at 4 oC	D MC at 4 oC	germination of 12.5% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at 4 oC	E MC at 4 oC	germination of 32.6% moisture content storaged at 4 oC		Unit: dimensionless Precision: 0.1	
Germination of A MC at 15 oC	A MC at 15 oC	germination of 3.7% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of B MC at 15 oC	B MC at 15 oC	germination of 5.3% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of C MC at 15 oC	C MC at 15 oC	germination of 7.8% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of D MC at 15 oC	D MC at 15 oC	germination of 12.5% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	
Germination of E MC at 15 oC	E MC at 15 oC	germination of 32.6% moisture content storaged at 15 oC		Unit: dimensionless Precision: 0.1	

Data Tables, Images, and Other Entities: **Effects of storage temperatures and moisture contents on the germination percentage of seeds of Sapium discolor (Data Table) Effects of storage temperatures and moisture contents on the germination percentage of seeds of Bischofia javanica (Data Table)**

This interface is a mock interface that is being used to conduct dissertation research. It does follow some of the same design and layout of the DataONE ONEMercury website, in order to provide the same look and feed of that site. However, the results have been manipulated in order to provide canned results to test for particular factors that influence data reuse. .

This study design was approved by the University of North Carolina at CHapel Hill Institutional Review Board in May 2015, IRB # IRB 14-2393.

Result #2



A DataONE Search Tool for Scientific Data

You searched for: text : moisture text : content

Document Identifier: yjc.3.9
Data Set Citation Yang J. . **Seed Storage Behavior of Sapium discolor Muell.-Arg.and Bischofia javanica Blume**. yjc.3.9

Data Set Owner(s):
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Abstract: The storage behavior and germination characteristics of seeds of Taiwan tallow tree (Sapium discolor) and Javanese bishopwood (Bischofia javanica) were investigated in the present study. The germination percentage of freshly collected mature seeds of S. discolor from Duona, Kaohsiung County in 1997 was 59% over 16 weeks under alternating temperatures of 30/20°C with 8 h of light. With cold stratification of freshly collected seeds for 2 mo, the germination percentage reached as high as 85.3%. Therefore, about 30% of seeds showed deep dormancy. The mean germination time was reduced from 54 d for freshly collected mature seeds to 41 d after 1 mo of stratification. Seed germinability of S. discolor could be maintained at -20~15°C, for at least 24 mo at MCs (moisture contents) of between 1.8 and 7.9% (on a fresh-weight basis). These results confirm that S. discolor seeds exhibit orthodox seed storage behavior. The germination percentage of freshly collected mature seeds of B. javanica from Sansia, Taipei County in 1997 was 78% over 8 weeks under alternating temperatures of 30/20°C with 8 h of light, and the mean germination times were between 10 and 18 d. Most freshly mature seeds survived when they were desiccated to 5.3~12.5% MCs; however, the viability of those seeds with 3.7~12.5% MCs dropped significantly at -20°C after 24 mo of hermetic storage. The optimum seed moisture contents of B. javanica were 8, 5~13, and 5~8% for -20, 4, and 15°C storage, respectively. The temperature effect on seed longevity revealed that the longevity of seeds stored at 4°Cwas better than those at -20 and 15°C. We estimated that the optimum seed storage conditions for B. javanica are provided by combining a temperature of 4°C with about an 8% MC. These results confirm that B. javanica seeds exhibit intermediate storage behavior that is characterized as being tolerant to desiccation but sensitive to freezing temperatures.

Keywords:
Sapium discolor
Bischofia javanica
stratification
orthodox seed storage behavior
intermediate seed storage behavior

License and Usage Rights: Please contact JC Yang (yjc@tfri.gov.tw) to get usage agreement if you want to adopt these data in any presentation.

Geographic Coverage:
Geographic Description: S. discolor:Duona, Kaohsiung; B. javanica:Sansia, Taipei

Bounding Coordinates:

West:	119.875 degrees
East:	122.125 degrees
North:	25.875 degrees
South:	21.625 degrees

Temporal Coverage:
Date: 2006-11-28
Taxonomic Coverage:

	Rank Name	Rank Value	Common Names
Taxon:	Genus	Bischofia	
	Species	javanica	Javanese bishopwood
	Rank Name	Rank Value	Common Names
Taxon:	Genus	Sapium	
	Species	discolor	

Contact:
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Methods Info:

Step 1:

Description:

controll of seed moisture content
To obtain seeds at 5 MC levels between about 2%~15% of Sapium discolor and Bischofia javanica, clean seeds were divided into 5 sub-lots of each species. Seeds of each sub-lot were desiccated at 2%~5% RHs in a hermetically sealed acrylic box containing a molecular sieve at 25 °C. Immediately after the relative humidity treatment, each MC group was wrapped in an aluminium foil bag and stored at 5 °C for about 5 d to allow the moisture to equilibrate within and among the seeds. The obtained MCs were 1.8 ± 0.1%, 3.5 ± 0.2%, 5.1 ± 0.2%, 7.9 ± 0.3%, and 9.0 ± 0.3%; and 3.7 ± 0.3%, 5.3 ± 0.3%, 7.8 ± 0.4%, 12.5 ± 0.4%, and 32.6 ± 1.0% for Cordia dichotoma and Celtis sinensis, respectively.

Step 2:

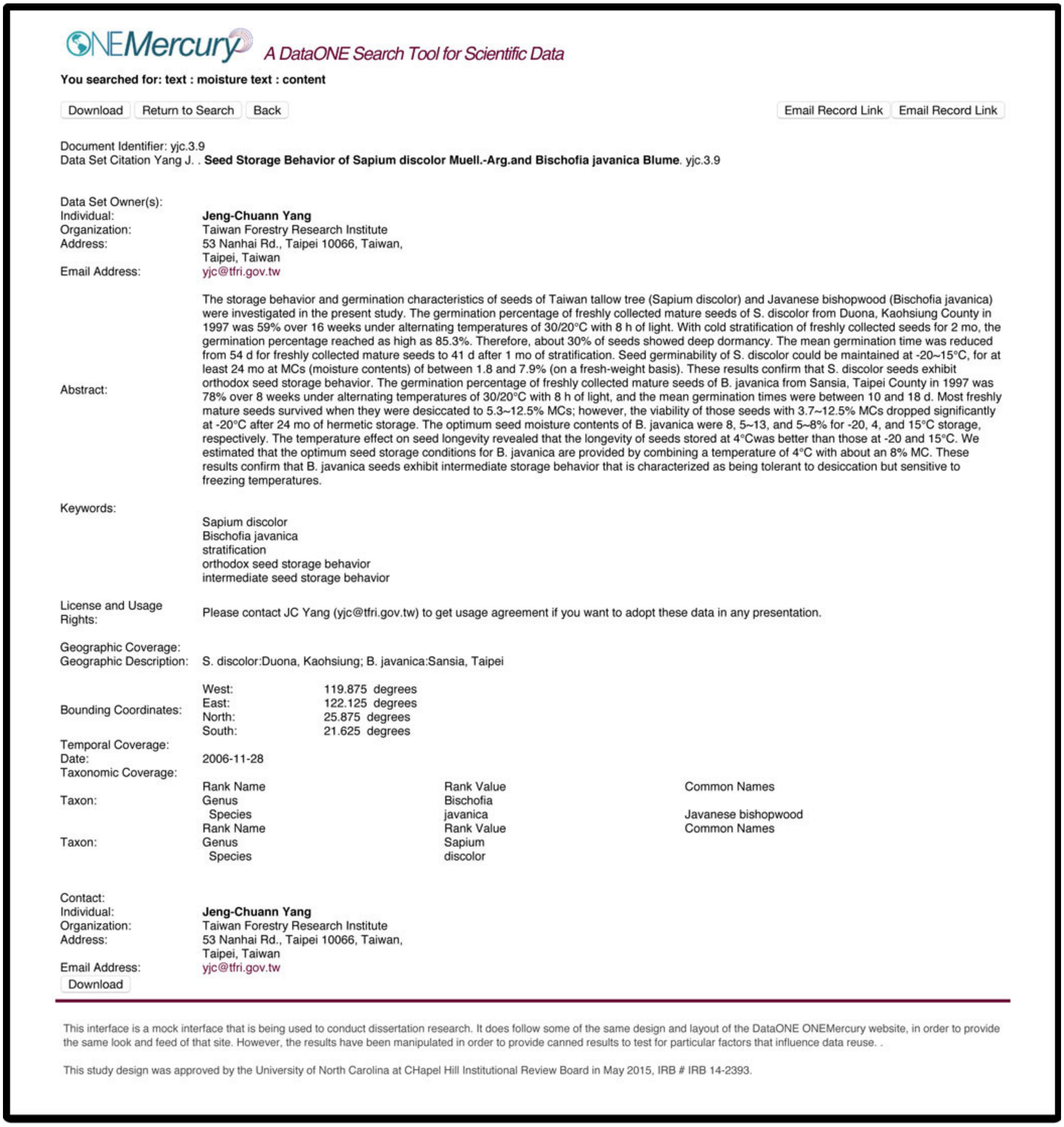
Description:

storage treatment
The seed moisture of each MC group was determined, and seed bags were stored at three temperatures (-20, 4, and 15 °C) for up to 24 months. A factorial design was used to study the responses of seeds of the 2 specie survival to moisture content, temperature, and storage period. Thus, three temperatures (-20, 4, and 15 °C) were combined with 5 MCs (from 1.8% to 9.0% and from 3.7% to 32.6%), and with storage periods of 0, 3, 6, 9, 12, 18, and 24 months for Sapium discolor and Bischofia javanica, respectively.

This interface is a mock interface that is being used to conduct dissertation research. It does follow some of the same design and layout of the DataONE ONEMercury website, in order to provide the same look and feed of that site. However, the results have been manipulated in order to provide canned results to test for particular factors that influence data reuse. .

This study design was approved by the University of North Carolina at CHapel Hill Institutional Review Board in May 2015, IRB # IRB 14-2393.

Result #3



You searched for: text : moisture text : content

Document Identifier: yjc.3.9
Data Set Citation Yang J. . **Seed Storage Behavior of Sapium discolor Muell.-Arg.and Bischofia javanica Blume**. yjc.3.9

Data Set Owner(s):
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Abstract: The storage behavior and germination characteristics of seeds of Taiwan tallow tree (Sapium discolor) and Javanese bishopwood (Bischofia javanica) were investigated in the present study. The germination percentage of freshly collected mature seeds of S. discolor from Duona, Kaohsiung County in 1997 was 59% over 16 weeks under alternating temperatures of 30/20°C with 8 h of light. With cold stratification of freshly collected seeds for 2 mo, the germination percentage reached as high as 85.3%. Therefore, about 30% of seeds showed deep dormancy. The mean germination time was reduced from 54 d for freshly collected mature seeds to 41 d after 1 mo of stratification. Seed germinability of S. discolor could be maintained at -20~15°C, for at least 24 mo at MCs (moisture contents) of between 1.8 and 7.9% (on a fresh-weight basis). These results confirm that S. discolor seeds exhibit orthodox seed storage behavior. The germination percentage of freshly collected mature seeds of B. javanica from Sansia, Taipei County in 1997 was 78% over 8 weeks under alternating temperatures of 30/20°C with 8 h of light, and the mean germination times were between 10 and 18 d. Most freshly mature seeds survived when they were desiccated to 5.3~12.5% MCs; however, the viability of those seeds with 3.7~12.5% MCs dropped significantly at -20°C after 24 mo of hermetic storage. The optimum seed moisture contents of B. javanica were 8, 5~13, and 5~8% for -20, 4, and 15°C storage, respectively. The temperature effect on seed longevity revealed that the longevity of seeds stored at 4°Cwas better than those at -20 and 15°C. We estimated that the optimum seed storage conditions for B. javanica are provided by combining a temperature of 4°C with about an 8% MC. These results confirm that B. javanica seeds exhibit intermediate storage behavior that is characterized as being tolerant to desiccation but sensitive to freezing temperatures.

Keywords:
Sapium discolor
Bischofia javanica
stratification
orthodox seed storage behavior
intermediate seed storage behavior

License and Usage Rights: Please contact JC Yang (yjc@tfri.gov.tw) to get usage agreement if you want to adopt these data in any presentation.

Geographic Coverage:
Geographic Description: S. discolor:Duona, Kaohsiung; B. javanica:Sansia, Taipei

Bounding Coordinates:	West:	119.875 degrees
	East:	122.125 degrees
	North:	25.875 degrees
	South:	21.625 degrees

Temporal Coverage:
Date: 2006-11-28
Taxonomic Coverage:

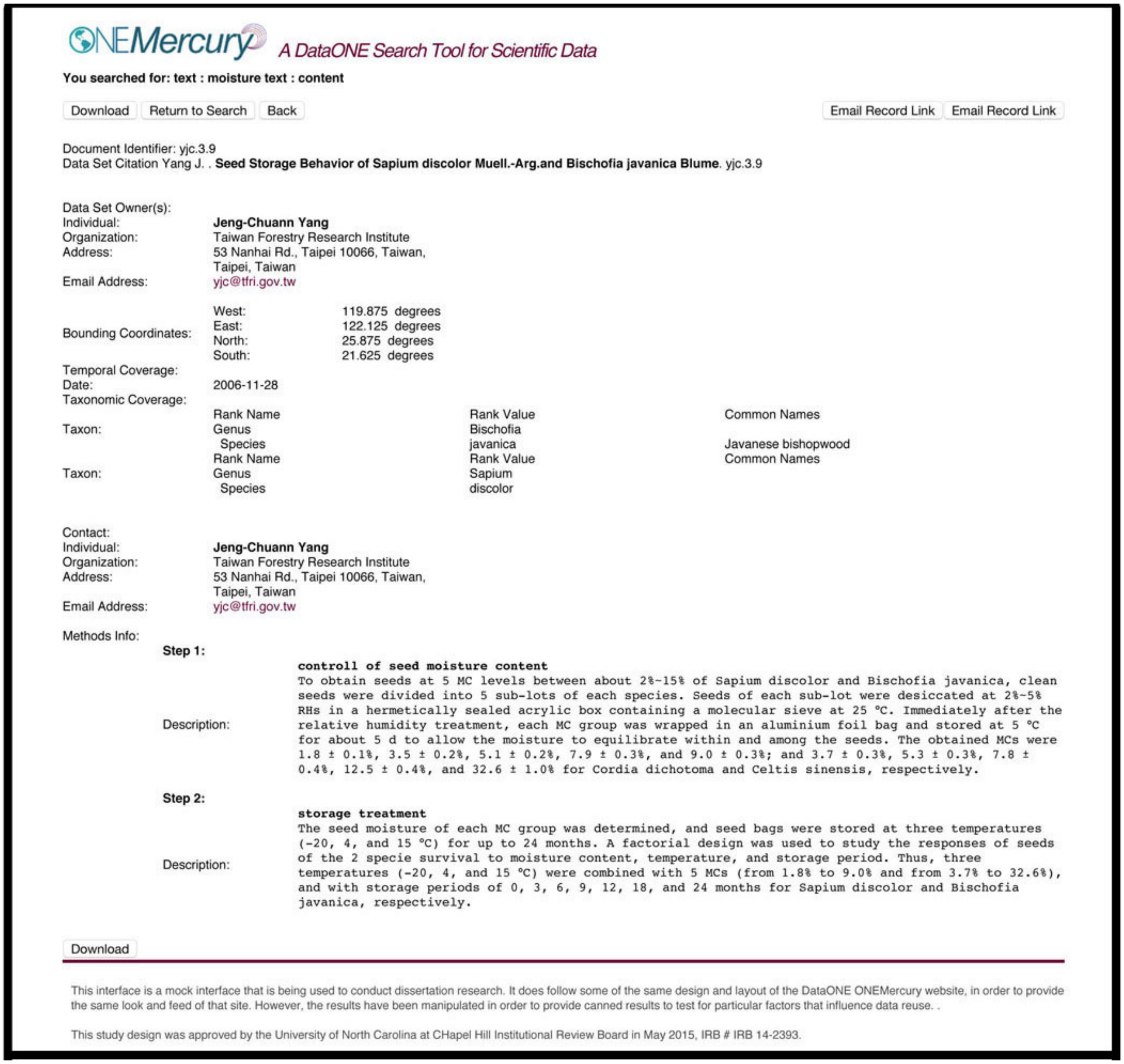
	Rank Name	Rank Value	Common Names
Taxon:	Genus	Bischofia	
	Species	javanica	Javanese bishopwood
	Rank Name	Rank Value	Common Names
Taxon:	Genus	Sapium	
	Species	discolor	

Contact:
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

This interface is a mock interface that is being used to conduct dissertation research. It does follow some of the same design and layout of the DataONE ONEMercury website, in order to provide the same look and feed of that site. However, the results have been manipulated in order to provide canned results to test for particular factors that influence data reuse. .

This study design was approved by the University of North Carolina at CHapel Hill Institutional Review Board in May 2015, IRB # IRB 14-2393.

Result #4



You searched for: text : moisture text : content

Document Identifier: yjc.3.9
Data Set Citation Yang J. . **Seed Storage Behavior of Sapium discolor Muell.-Arg.and Bischofia javanica Blume**. yjc.3.9

Data Set Owner(s):
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Bounding Coordinates:	West:	119.875 degrees
	East:	122.125 degrees
	North:	25.875 degrees
	South:	21.625 degrees

Temporal Coverage:
Date: 2006-11-28
Taxonomic Coverage:

	Rank Name	Rank Value	Common Names
Taxon:	Genus	Bischofia	
	Species	javanica	Javanese bishopwood
	Rank Name	Rank Value	Common Names
Taxon:	Genus	Sapium	
	Species	discolor	

Contact:
Individual: **Jeng-Chuann Yang**
Organization: Taiwan Forestry Research Institute
Address: 53 Nanhai Rd., Taipei 10066, Taiwan,
Taipei, Taiwan
Email Address: yjc@tfri.gov.tw

Methods Info:

Step 1:

Description:

controll of seed moisture content
To obtain seeds at 5 MC levels between about 2%-15% of Sapium discolor and Bischofia javanica, clean seeds were divided into 5 sub-lots of each species. Seeds of each sub-lot were desiccated at 2%-5% RHs in a hermetically sealed acrylic box containing a molecular sieve at 25 °C. Immediately after the relative humidity treatment, each MC group was wrapped in an aluminium foil bag and stored at 5 °C for about 5 d to allow the moisture to equilibrate within and among the seeds. The obtained MCs were 1.8 ± 0.1%, 3.5 ± 0.2%, 5.1 ± 0.2%, 7.9 ± 0.3%, and 9.0 ± 0.3%; and 3.7 ± 0.3%, 5.3 ± 0.3%, 7.8 ± 0.4%, 12.5 ± 0.4%, and 32.6 ± 1.0% for Cordia dichotoma and Celtis sinensis, respectively.

Step 2:

Description:

storage treatment
The seed moisture of each MC group was determined, and seed bags were stored at three temperatures (-20, 4, and 15 °C) for up to 24 months. A factorial design was used to study the responses of seeds of the 2 specie survival to moisture content, temperature, and storage period. Thus, three temperatures (-20, 4, and 15 °C) were combined with 5 MCs (from 1.8% to 9.0% and from 3.7% to 32.6%), and with storage periods of 0, 3, 6, 9, 12, 18, and 24 months for Sapium discolor and Bischofia javanica, respectively.

This interface is a mock interface that is being used to conduct dissertation research. It does follow some of the same design and layout of the DataONE ONEMercury website, in order to provide the same look and feed of that site. However, the results have been manipulated in order to provide canned results to test for particular factors that influence data reuse. .

This study design was approved by the University of North Carolina at CHapel Hill Institutional Review Board in May 2015, IRB # IRB 14-2393.

REFERENCES

Abramson, D., Enticott, C., & Peachey, T. (2008). Parameter estimation using scientific workflows. In *2008 IEEE Fourth International Conference on eScience* (pp. 392–393). Indianapolis, IN: IEEE. https://doi.org/10.1109/eScience.2008.97

Acord, S. K., & Harley, D. (2013). Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0. *New Media & Society*, *15*(3), 379–397. https://doi.org/10.1177/1461444812465140

Agarwal, D. A., Humphrey, M., Beekwilder, N. F., Jackson, K. R., Goode, M. M., & van Ingen, C. (2010). A data-centered collaboration portal to support global carbon-flux analysis. *Concurrency and Computation: Practice and Experience*, *22*(17), 2323–2334. https://doi.org/10.1002/cpe.1600

Ailamaki, A., Kantere, V., & Dash, D. (2010). Managing scientific data. *Communications of the ACM*, *53*(6), 68. https://doi.org/10.1145/1743546.1743568

Allard, S. (2012a). DataONE: Facilitating eScience through collaboration. *Journal of eScience Librarianship*, *1*(1), 4–17. https://doi.org/10.7191/jeslib.2012.1004

Allard, S. (2012b). National data management initiative and the U.S. Exemplar: DataONE. In N. Xiao & L. R. McEwen (Eds.), *Special Issues in Data Management* (Vol. 1110, pp. 47–67). Washington, DC: American Chemical Society.

Allard, S., & Allard, G. (2009). Transdisciplinarity and information science in earth and environmental science research. *Proceedings of the American Society for Information Science and Technology*, *46*, 1–9. https://doi.org/10.1002/meet.2009.1450460346

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*, 191–202.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Messerschmitt, D. G., Messina, P., … Wright, M. H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*.

Atkins, D. E., Hey, T., & Hedstrom, M. (2011). *Data and visualization final report* (No. Final Report) (pp. 1–40). National Science Foundation - Advisory Committee for Cyberinfrastructure. Retrieved from http://www.nsf.gov/od/oci/taskforces/

Aydinoglu, A. U. (2011). *Complex adaptive systems theory applied to virtual scientific collaborations: The case of DataONE*. University of Tennessee, Knoxville. Retrieved from http://trace.tennessee.edu/utk_graddiss/1054

Bargmeyer, B. E., & Gillman, D. W. (2014). Metadata standards and metadata registries: An overview. Retrieved from www.bls.gov/ore/pdf/st000010.pdf

Barkstrom, B. R. (2010). A mathematical framework for earth science data provenance tracing. *Earth Science Informatics*, *3*(3), 167–196. https://doi.org/10.1007/s12145-010-0057-0

Baru, C. (2007). Sharing and caring of eScience data. *International Journal on Digital Libraries*, *7*, 113–116. https://doi.org/10.1007/s00799-007-0029-2

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

Beran, B., van Ingen, C., & Fatland, D. R. (2010). SciScope: A participatory geoscientific web application. *Concurrency and Computation: Practice and Experience*, *22*(17), 2300–2312. https://doi.org/10.1002/cpe.1597

Berman, F., Fox, G. C., & Hey, A. J. G. (Eds.). (2003). *Grid computing: making the global infrastructure a reality*. New York: J. Wiley.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting and sharing in science and engineering. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 339–348). New York, New York, USA: ACM Press. https://doi.org/10.1145/958160.958215

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., & Louis, K. S. L. (1997). Withholding research results in academic life science. Evidence from a national survey of faculty. *JAMA: The Journal of the American Medical Association*, *277*(15), 1224–1228.

Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., & Hilgartner, S. (2006). Data withholding in genetics and the other life sciences: Prevalences and predictors. *Academic Medicine*, *81*(2), 137–145.

Borgman, C. L. (2000a). *From Gutenberg to the global information infrastructure: Access to information in the networked world*. Cambridge, Mass: MIT Press.

Borgman, C. L. (2000b). The Presmise and the Promise of a Global Information Infrastructure. In C. L. Borgman, *From Gutenberg to the global information infrastructure: access to information in the networked world* (pp. 1–31). Cambridge, Mass: MIT Press.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? *Fifth China – North America Library Conference 2010*, (September). Retrieved from http://works.bepress.com/borgman/238/

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

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

Brazier, P., Chebotko, A., Gates, A. Q., & Salayandia, L. (2009). GEO-SEED: A metadata repository for geosciences web service discovery. *2009 Congress on Services - I*, 356–359. https://doi.org/10.1109/SERVICES-I.2009.43

Brown, C. M. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology*, *54*(10), 926–938. https://doi.org/10.1002/asi.10289

Browne, G. L., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive Stopping Rules for Terminiating Information Search in Online Tasks. *MIS Quarterly*, *31*(1), 88–104.

Bussard, A. E. (1990). Obtaining and storing knowledge: Scientific data banks. In P. S. Glaeser (Ed.), *Scientific and Technical Data in a New Era* (pp. 6–9). New York, New York, USA: Hemisphere Publisher Corporation.

Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science, Technology, and Human Values*, *13*(No. 1/2), 45–52.

Chadwich-Jones, J. K. (1976). *Social Exchange Theory: Its structure and influence in social psychology*. (H. Tajfel, Ed.). Halifax, Canada: Academic Press.

Cheah, Y.-W., & Plale, B. (2012). Provenance analysis: Towards quality provenance. In *2012 IEEE 8th International Conference on E-Science*. Chicago: IEEE. https://doi.org/10.1109/eScience.2012.6404480

CODATA. (2002). *CODATA workshop on archiving scientific and technical (S&T) data* (pp. 1–14). Pretoria South Africa: CODATA.

Cohen, J. (1995). Share and share alike isn't always the rule in science. *Science*, *268*(5218), 1715–1718.

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, D.C: National Academic Press.

Constant, D., Kiesler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research*, *5*(4), 400–421.

Consultative Committee for Space Data Systems. (2002). *Blue Book: Reference model for an Open Archival Information System (OAIS)* (No. CCSDS 650.0-B-1) (pp. 1–148). CCSDS. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Consultative Committee for Space Data Systems. (2011). CCSDS Recommendations and Reports - Magenta Books: Recommended Practices. Retrieved July 15, 2015, from http://public.ccsds.org/publications/MagentaBooks.aspx?RootFolder=

DataNet Federation Consortium. (2014). DFC – DataNet Federation Consortium. Retrieved December 24, 2013, from http://datafed.org/

DataONE. (2013a). Benefits of becoming a Member Node | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/benefits-becoming-member-node

DataONE. (2013b). Contribute Data | DataONE. Retrieved May 31, 2015, from https://www.dataone.org/contribute-data

DataONE. (2013c). Coordinating Nodes | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/coordinating-nodes

DataONE. (2013d). DataONE - Data Observation Network for Earth. Retrieved from http://www.dataone.org/

DataONE. (2013e). DataONE Overview — v1.2.0. Retrieved February 15, 2014, from http://mule1.dataone.org/ArchitectureDocs-current/overview.html

DataONE. (2013f). DataONE Users Group | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/dataone-users-group

DataONE. (2013g). Education Modules | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/education-modules

DataONE. (2013h). Graduate Courses | DataONE [Graduate Courses]. Retrieved January 2, 2014, from http://www.dataone.org/graduate-courses

DataONE. (2013i). Internships | DataONE. Retrieved February 14, 2014, from http://www.dataone.org/internships

DataONE. (2013j). Investigator Toolkit | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/investigator-toolkit

DataONE. (2013k). Protege | DataONE. Retrieved May 2, 2014, from http://www.dataone.org/software-tools/protégé

DataONE. (2013l). Training Activities | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/training-activities

DataONE. (2013m). What is DataONE? | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/what-dataone

DataONE. (2013n). Working Groups | DataONE. Retrieved January 2, 2014, from http://www.dataone.org/working_groups

DataONE. (2016). Morpho | DataONE. Retrieved July 25, 2016, from https://www.dataone.org/software-tools/morpho

Devarakonda, R., Palanisamy, G., Green, J. M., & Wilson, B. E. (2010). Data sharing and retrieval using OAI-PMH. *Earth Science Informatics*, *4*(1), 1–5. https://doi.org/10.1007/s12145-010-0073-0

Dexter, N. C., Cobb, J. W., Vieglais, D., Jones, M. B., & Lowe, M. (2011). DataONE member node pilot integration with TeraGrid? In *Proceedings of the TeraGrid 2011 Conference Extreme Digital Discovery TG11*. https://doi.org/10.1145/2016741.2016756

Digital Curation Centre. (2004, 2016). DCC Curation Lifecycle Model. Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle-model

Dryad. (2016). Frequently Asked Questions - Dryad. Retrieved May 15, 2016, from https://datadryad.org/pages/faq

Edwards, P. N. (2011). *A Vast Machine: Computer models, climate data and the politics of global warming* (Vol. 28). MIT Press. Retrieved from http://www.amazon.com/Vast-Machine-Computer-Climate-Politics/dp/0262013924 http://doi.wiley.com/10.1111/j.1541-1338.2011.00522_3.x

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

Enriquez, V., Judson, S. W., Weber, N. M., Allard, S., Cook, R. B., Piwowar, H. A., … Wilson, B. (2010). Data citation in the wild. *Nature Precedings*, (713), 10–10. https://doi.org/10.1038/npre.2010.5452.1

EnvEurope. (2014a). EnvEurope Project — ENVeurope. Retrieved February 18, 2016, from http://www.enveurope.eu/

EnvEurope. (2014b). EnvThes - Environmental Thesaurus — ENVeurope. Retrieved February 18, 2016, from http://www.enveurope.eu/news/envthes-environmental-thesaurus

FGDC. (2016). The Federal Geographic Data Committee — Federal Geographic Data Committee. Retrieved May 18, 2016, from http://www.fgdc.gov/

Fredje, G., & Meinard, M. (1990). Biological Data Banks: A Preliminary Study. In P. S. Glaeser (Ed.), *Scientific and Technical Data in a New Era* (pp. 28–36). New York, New York, USA: Hemisphere Publisher Corporation.

Glaeser, P. S. (1990). *Scientific and technical data in a new era*. (P. S. Glaeser, Ed.). New York: Hemisphere Publisher Corporation.

Goranova, M., Shishedjiev, B., & Georgieva, J. (2011). Research on building scientific data ontology. In *4th International Conference Developments in eSystems Engineering* (pp. 541–546). IEEE. https://doi.org/10.1109/DeSE.2011.71

Gray, J. (1996). Evolution of data management. *Computer*, *29*(10), 38–46.

Greenberg, J. (2003). Metadata and the World Wide Web. *Encyclopedia of Library and Information Science*, 1876–1888.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*(5–6), 907–928.

Hank, C., Jordan, M. W., & Wildemuth, B. M. (2009). Survey research. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 256–270). Westport, Conn: Libraries Unlimited.

Hank, C., & Wildemuth, B. M. (2009). Quasi-experimental studies. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 93–104). Westport, Conn: Libraries Unlimited.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The Fourth Paradigm: Data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

Hey, T., & Trefethen, A. E. (2003). The data deluge: An e-Science perspective. In F. Berman, A. J. G. Hey, & G. C. Fox (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 809–824). Wiley and Sons. Retrieved from http://en.scientificcommons.org/2325382

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

Hodge, G. (2008). Toward interoperability: A report from the 11th open forum on metadata registries and related standards. *Bulletin of the American Society for Information Science and Technology*, *35*(1), 25–31.

Home - Fluxdata.org. (n.d.). Retrieved September 22, 2012, from http://www.fluxdata.org/default.aspx

HUBzero | DataONE. (n.d.). Retrieved May 2, 2014, from https://www.dataone.org/software-tools/hubzero

HUBzero - Platform for Scientific Collaboration. (n.d.). Retrieved September 24, 2012, from http://hubzero.org/

ICS CODATA. (2015). CODATA, The Committee on Data for Science and Technology. Retrieved February 12, 2011, from http://www.codata.org/

International Organization for Standardization. (2016). ISO 19115-1:2014 - Geographic information -- Metadata -- Part 1: Fundamentals. Retrieved February 18, 2016, from http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798

Jeffery, K. G., Asserson, A., & Houssos, N. (2013). A 3-Layer model for metadata. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2013*. Lisbon, Portugal.

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *Internet Computing*, (September-October), 59–68.

Kelper/Core. (n.d.). The Kepler Project — Kepler. Retrieved September 22, 2012, from https://kepler-project.org/

Kennedy, P. (2009, August 20). Johnny Holland. Retrieved July 23, 2014, from http://johnnyholland.org/2009/08/practical-triangulation/

King, T., Merka, J., Narock, T., Walker, R., & Bargatze, L. (2010). A registry framework and rosetta attributes for distributed science. *Earth Science Informatics*, *3*, 127–133. https://doi.org/10.1007/s12145-010-0047-2

Knowledge Network for Biocomplexity. (2015). Ecological Metadata Language. Retrieved from https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

Kuznetsov, F. A., Titov, V. A., Borisov, S. V., & Vetroprakhov, V. N. (1990). Databases for properties of electronic materials. In P. S. Glaeser (Ed.), *Scientific and Technical Data in a New Era* (pp. 73–75). New York, New York, USA: Hemisphere Publisher Corporation.

Lagoze, C., & Patzke, K. (2011). A research agenda for data curation cyberinfrastructure. In *Proceedings of the 2011 ACM/IEEE Joint Conference on Digital Libraries* (pp. 373–382). Ottawa, ON. Retrieved from http://dl.acm.org/citation.cfm?id=1998145

Lanz, A., Brandli, M., & Baltensweiler, A. (2007). A Large-scale, Long-term view on collecting and sharing landscape data. In F. Kienast, O. Wildi, & S. Ghosh (Eds.), *A changing world: challenges for landscape research* (pp. 93–111). Dordrecht, The Netherlands: Springer.

Latour, B., & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, N.J: Princeton University Press.

Lee, J. W., Zhang, J., Zimmerman, A. S., & Lucia, A. (2009). DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. *AIChE Journal*, *55*(11), 2757–2764. https://doi.org/10.1002/aic.12085

LeFurgy, B. (2012, February 21). Life cycle models for digital stewardship | The signal: Digital preservation. Retrieved from http://blogs.loc.gov/digitalpreservation/2012/02/life-cycle-models-for-digital-stewardship/

Lide, D. R. (1981). Critical data for critical needs. *Science*, *212*(4501), 1343–1349.

Lifschitz, S., Gomes, L., & Rehen, S. K. (2011). Dealing with reusability and reproducability for scientific workflows. In *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops* (pp. 625–632). Atlanta, Georgia: IEEE.

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* (pp. 1–84). Twickenham, England: The JISC Committee for the Support of Research (JCSR). Retrieved from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. In *Proc 3th UK e-Science All Hands Meeting*. https://doi.org/10.1.1.111.7425

Luo, L., & Wildemuth, B. M. (2009). Semistructured interviews. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 232–241). Westport Conn.: Libraries Unlimited.

Lyon, L. (2012). The informatics transform: Re-engineering libraries for the data decade. *International Journal of Digital Curation*, *7*(1), 126–138. https://doi.org/10.2218/ijdc.v7i1.220

Ma, X., & Fox, P. (2013). Recent progress on geologic time ontologies and considerations for future works. *Earth Science Informatics*, *6*(1), 31–46. https://doi.org/10.1007/s12145-013-0110-x

Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, *61*(10), 2029–2048. https://doi.org/10.1002/asi.21339

McCain, K. W. (1995). Mandating sharing: Journal policies in the natural sciences. *Science Communication*, *16*, 403–431. https://doi.org/10.1177/1075547095016004003

McCain, K. W. (2000). Sharing digitized research-related information on the World Wide Web. *Journal of the American Society for Information Science*, *51*(14), 1321–1327.

Meeker, B. F. (1971). Decisions and Exchange. *American Sociological Review*, *36*(3), 485. https://doi.org/10.2307/2093088

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., … Vieglais, D. A. (2011). Participatory design of dataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, *11*, 5–15. https://doi.org/10.1016/j.ecoinf.2011.08.007

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, *7*(1), 330–342.

Michener, W. K., Vieglais, D., Vision, T., Kunze, J., Cruse, P., & Janée, G. (2011). DataONE: Data Observation Network for Earth, preserving data and enabling innovation in the biological and environmental sciences. *DLib Magazine*, *17*(1/2), 1–12. https://doi.org/10.1045/january2011-michener

Murillo, A. P. (2013). Data At Risk Initiative: Examining and facilitating the scientific process in relation to endangered data. *Data Science Journal*, *12*, 207–219. https://doi.org/10.2481/dsj.12-048

Murillo, A. P. (2014). Examining Data Sharing and Data Reuse in DataONE Environment. In *ASIST 2014*. Seattle, Washington.

Murillo, A. P., Greenberg, J., Kunze, J., & Boone, J. (2012). *Components of a successful metadata registry framework*. Presented at the DataONE All Hands Meeting, University of New Mexico.

Murillo, A. P., & Ramdeen, S. (2013). Understanding user motivations regarding earth science data reuse: Accessing opinions on skills, access, and trust. In *Archival Education and Research Institute (AERI)*. University of Texas at Austin.

NASA. (2016a). GCMD Science Keywords and Associated Directory Keywords. Retrieved May 15, 2016, from http://gcmd.nasa.gov/learn/keyword_list.html

NASA. (2016b). Global Change Master Directory (GCMD) Mission. Retrieved May 15, 2016, from http://gcmd.nasa.gov/learn/mission.html

NASA Global Change Master Directory. (2016a). Global Change Master Directory (GCMD) Mission. Retrieved February 18, 2016, from http://gcmd.gsfc.nasa.gov/learn/mission.html

NASA Global Change Master Directory. (2016b). Keyword Community Page. Retrieved February 18, 2016, from http://gcmd.nasa.gov/learn/keywords.html

National Center for Biotechnology Information. (2013). GenBank Overview [National Library of Medicine]. Retrieved April 26, 2014, from http://www.ncbi.nlm.nih.gov/genbank/

National Institutes of Health. (2007). NIH Data Sharing Policy. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/

National Research Council. (1994). *Realizing the information future: The internet and beyond*. Washington DC: National Academy Press.

National Science Foundation. (2006, November 7). Sustainable Digital Data Preservation and Access Network Partners (DataNet). Retrieved February 4, 2014, from http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

National Science Foundation. (2010a). Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans. Retrieved December 13, 2010, from http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation. (2010b, November 10). Dissemination and Sharing of Research Results. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, Calif: Sage Publications.

Nielsen Norman Group. (2016). Infinite Scrolling is Not for Every Website. Retrieved September 10, 2016, from https://www.nngroup.com/articles/infinite-scrolling/

Noor, M. A. F., Zimmerman, K. J., & Teeter, K. C. (2006). Data sharing: How much doesn't get submitted to genBank? *PLoS Biology*, *4*(7), e228. https://doi.org/10.1371/journal.pbio.0040228

Ochsner, S. A., Steffen, D. L., Stoeckert, C. J., & McKenna, N. J. (2008). Much room for improvement in deposition rates of expression microarray datasets. *Nature Methods*, *5*(12), 991. https://doi.org/10.1038/nmeth1208-991

Oh, S., & Wildemuth, B. M. (2009). Think-aloud Protocols. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 178–188). Westport Conn.: Libraries Unlimited.

Otuka, N., Dupont, E., Semkova, V., Pritychenko, B., Blokhin, A. I., Aikawa, M., … Zhuang, Y. (2014). Towards a More Complete and Accurate Experimental Nuclear Reaction Data Library (EXFOR): International Collaboration Between Nuclear Reaction Data Centres (NRDC). *Nuclear Data Sheets*, *120*, 272–276. https://doi.org/10.1016/j.nds.2014.07.065

Oxford English Dictionary. (2016a). adequate, adj. : Oxford English Dictionary. Retrieved June 4, 2016, from http://www.oed.com.libproxy.lib.unc.edu/view/Entry/2299?rskey=jeIz1Q&result=1#eid

Oxford English Dictionary. (2016b). comprehensive, adj. : Oxford English Dictionary. Retrieved June 4, 2016, from http://www.oed.com.libproxy.lib.unc.edu/view/Entry/37859#eid8604868

Oxford English Dictionary. (2016c). robust, adj. and n. : Oxford English Dictionary. Retrieved June 4, 2016, from http://www.oed.com.libproxy.lib.unc.edu/view/Entry/166651?redirectedFrom=robust

Parsons, M. A. (2011). Making data useful for modelers to understand complex Earth systems. *Earth Science Informatics*, *4*(4), 197–223. https://doi.org/10.1007/s12145-011-0089-0

Patel, J., Okamoto, S., Dascalu, S. M., & Harris, Jr., F. C. (2012). Web-Enabled Toolkit for Data Interoperability Support. In *Proceedings of the 2012 ISCA International Conference on Software Engineering and Data Engineering*. Los Angeles, CA.

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*, *6*(7, e18657), 1–13. https://doi.org/10.1371/journal.pone.0018657

Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics*, *4*, 148–156. https://doi.org/10.1016/j.joi.2009.11.010

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705.

Riley, J. (2009). *Glossary of metadata standards* (pp. 1–18). Indiana University Libraries: Indiana University Libraries White Professional Development Award.

Sayogo, D. S., & Pardo, T. a. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19–S31. https://doi.org/10.1016/j.giq.2012.06.011

Sayogo, D. S., & Pardo, T. A. (2011a). Exploring the determinants of publication of scientific data in open data initiative. In *Proceedings of the 5th International Conference on Theory and Practice of Electronic Governance* (pp. 97–106). Tallinn, Estonia: ACM. https://doi.org/10.1145/2072069.2072087

Sayogo, D. S., & Pardo, T. A. (2011b). Understanding the capabilities and critical success factors in collaborative data sharing network: The case of DataONE. In *Proceedings of the 12th Annual International Digital Government Research Conference dgo 2011* (pp. 74–83). ACM. https://doi.org/10.1145/2037556.2037568

SEAD. (2013). SEAD. Retrieved December 24, 2013, from http://sead-data.net/

Selkov, E. E., Goryanin, I. I., Kaimachnikov, N. P., Shevelev, E. L., & Yunus, Y. A. (1990). Data and knowledge banks on enzymes and metabolic pathways. In P. S. Glaeser (Ed.), *Scientific and Technical Data in a New Era* (pp. 22–27). New York, New York, USA: Hemisphere Publisher Corporation.

Sieber, J. E. (1988). Data Sharing: Defining Problems and Seeking Solutions. *Law and Human Behavior*, *12*(2), 199–206.

Someren, M. W. van, Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London; San Diego: Academic Press.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, “Translations,” and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907 – 1939. *Social Studies of Science*, *19*, 387–420.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large Information spaces. *Information Systems Research*, *7*(1), 111–135.

Stodden, V., Guo, P., & Ma, Z. (2013). Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLoS ONE*, *8*(6), e67111. https://doi.org/10.1371/journal.pone.0067111

Taylor, A. G. (2004). *The organization of information* (2nd ed). Westport, Conn: Libraries Unlimited.

Taylor, A. G., Miller, D. P., & Taylor, A. G. (2006). *Introduction to cataloging and classification* (10th ed). Westport, Conn: Libraries Unlimited.

Technopedia. (2016). What is a Data Grid? - Definition from Techopedia. Retrieved June 9, 2016, from https://www.techopedia.com/definition/26335/data-grid

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6, e21101), 1–21.

The Sheridan Libraries at Johns Hopkins University. (2013). Data Conservancy. Retrieved December 24, 2013, from https://dataconservancy.org/

United States of America. (2016). Data.gov. Retrieved May 1, 2014, from http://www.data.gov/

University of California - California Digital Library. (2016). EZID: EZID Home. Retrieved September 29, 2016, from http://ezid.cdlib.org/

University of Minnesota. (2013). Terra Populus. Retrieved December 24, 2013, from http://www.terrapop.org/

University of Minnesota. (2014). Environmental Information Management Institute. Retrieved January 2, 2014, from http://library.unm.edu/services/instruction/eimi.php

VisTrails. (2014). VisTrailsWiki. Retrieved May 2, 2014, from http://www.vistrails.org/index.php/Main_Page

W3C. (2013). PROV-Overview. Retrieved May 9, 2014, from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Wenger, E. (1998). *Communities of Practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Wickett, K. M., Sacchi, S., Dubin, D., & Renear, A. H. (2013). Identifying content and levels of representation in scientific data. In *ASIST 2013*. Baltimore MD.

Wildemuth, B. M. (2009). Direct observation. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 189–198). Westport Conn.: Libraries Unlimited.

Zhang, Y., & Wildemuth, B. M. (2009). Unstructured interviews. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 222–231). Westport Conn.: Libraries Unlimited.

Zimmerman, A. S. (2003). Data sharing and secondary use of scientific data: Experiences of ecologists.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values*, *33*(5), 631–652.