Examining Data Sharing and Data Reuse in the DataONE Environment

Angela P. Murillo
University of North Carolina at Chapel Hill
School of Information and Library Sciences
216 Lenoir Drive • CB #3360
100 Manning Hall
Chapel Hill, NC 27599-3360
amurillo@email.unc.edu

ABSTRACT
The Data Observation Network for Earth (DataONE), a U.S. NSF DataNet Partner, seeks to provide cyberinfrastructure for “open, persistent, robust, and secure access to…earth science observational data”. Scientists participating in DataONE are able to deposit, search, and reuse data available through various DataONE tools. The research presented in this poster-paper reports on two studies examining data sharing and reuse in the DataONE environment. The two studies include 1) a profiling data assessment that examines the data and metadata being deposited into the DataONE system for data sharing, and 2) a pilot think-aloud study that examines what factors influence decisions regarding data reuse. From the profiling data assessment, preliminary results indicate that data being deposited into the DataONE for sharing have three specific types of metadata available including a) dataset, b) access, and c) additional metadata. Results also indicated that there is variation regarding the robustness and completeness of information. Additionally, through the think-aloud study results indicated that particular aspects the metadata information was useful for decision-making regarding reuse of data for scientists, while other metadata aspects were described as not useful. The results section provide specific details of these findings and demonstrate how these two studies examine both data sharing and reuse within the DataONE environment.

Keywords
Data sharing and reuse, scientific data, infrastructure, DataONE, NFS – DataNets

INTRODUCTION
This poster-paper presents preliminary analysis of a mixed-method study to gain an understanding of data sharing and reuse within a new cyberinfrastructure program, the DataONE.

Data sharing and reuse in the sciences has been a topic of growing attention in recent years. This attention stems from changes occurring in scientific practices driven by the data deluge (Bell, Hey, & Szalay, 2009; Hey & Trefethen, 2003) and fourth paradigm data-intensive science (Hey, Tansley, & Tolle, 2009); and changes in journal and grant agency policies (National Institutes of Health, 2007; National Science Foundation, 2010). The sharing of data provides the ability to extract additional value from existing data, avoid reproducing research, ask new questions of existing data, and advance the state of science in general (Borgman, 2012; Lord & Macdonald, 2003). These potential opportunities of data sharing and have placed pressure on the scientific community and funding agencies to provide infrastructure solutions for the changes occurring in scientific practice.

To address the above, in 2007 the U.S. National Science Foundation announced a request for proposals for Sustainable Digital Data Preservation and Access Network Partners (DataNet). The DataNet Partners were created to develop long-term sustainable data infrastructures, interoperable data preservation and access, and cyberinfrastructure capabilities (National Science Foundation, 2006). The Data Observation Network for Earth (DataONE), one of the initial DataNet Partners, provides cyberinfrastructure for “open, persistent, robust, and secure access to well-described and easily discovered earth science observational data” (DataONE, 2013). Scientists participating in DataONE are able to deposit, search, and reuse data available through the various DataONE tools.

The majority of studies specific to DataONE have addressed: the organization and the infrastructure, specific tools that the DataONE has created, and the DataONE community. Additionally, the majority of data sharing
literature have addressed: general reasons why scientists should share data, journal and grant policies that influence data sharing, behavioral aspects that influence data sharing, and have been conducted in the biological and biomedical sciences.

As the majority of the DataONE activities have focused on development, it is timely to evaluate the cyberinfrastructure progress. Furthermore, as the DataONE focuses on earth science data, this provides an environment for studying data sharing and reuse within the earth sciences. This research recognizes specific gaps in the literature: (1) the need to evaluate the DataONE cyberinfrastructure progress, (2) the need for studying how this infrastructure impacts data sharing, and (3) the need for studying sharing in the earth sciences. Additionally, considering the amount of time, energy, effort, and financial support the community has invested in DataONE, it is incredibly important to evaluate its usefulness in regards to data sharing and reuse.

This research will investigate data sharing and reuse within the DataONE cyberinfrastructure.

BACKGROUND LITERATURE

Data Sharing and Reuse
The below addresses provides a summary of the extensive research that has been conducted regarding data sharing and reuse.

Multiple studies have indicated that data sharing and reuse is will advance scientific development through factors such as avoiding duplication of work, allowing new questions to be asked to existing data, and encouraging diversity in analysis (Borgman, 2012; Lord & Macdonald, 2003).

Along with data sharing for the greater good of science, funding agencies and journals have put pressure on scientists to make their data available. Funding agencies are now requiring data sharing plans and publishers have threatened to not publish articles if scientists do not make their data publically available (Blumenthal et al., 2006; McCain, 1995; Sayogo & Pardo, 2013). Scientific reputation is another factor that influences scientists willingness to share data. Scientists unwilling to share their data could seem possibly fraudulent; therefore data sharing is considered a part of seeming reputable by fellow scientists (Birnholtz & Bietz, 2003; Ceci, 1988).

Many researchers have analyzed data sharing practices through multiple methods including direct report from scientists, journal policy studies, and bibliometric studies. Several direct report studies have investigated scientists’ attitude toward data sharing to gain an understanding of motivations. These studies suggest that scientists want to have data sharing as a norm in science (Borgman, 2012; Ceci, 1988; Lord & Macdonald, 2003). Data ownership, previous assistance from coworkers, journal policies, and grant agency requirements were some of the motivations for data sharing (Blumenthal et al., 2006; Constant, Kiesler, & Sproull, 1994; Tenopir et al., 2011). Reasons for not sharing data include financial concerns, lack of time, lack of organizational support, lack of documentation and complexity of metadata standards, as well as the difficulty to anticipate intended users (Birnholtz & Bietz, 2003; Tenopir et al., 2011; Zimmerman, 2003).

Researchers have also examined journal policies and data deposition. Since the early 1980s many journals have added policy statements to motivate scientists to share data, recently these policies have become stricter. Some publishers have indicated they will refuse to publish without evidence of data deposition (Brown, 2003; McCain, 1995, 2000). Additionally, studies have indicated that while no journal has complete compliance, much research data is deposited along with the article (Noor, Zimmerman, & Teeter, 2006; Ochsner, Steffen, Stoeckert, & McKenna, 2008). Lastly, several studies have investigated specific factors associated with data deposition. These studies have shown that author experience and publications associated with high-impact factor journals were more likely to have associated data deposited alongside the authors’ journal articles (Piwowar & Chapman, 2010; Piwowar, 2011).

The studies above demonstrate that scientists do believe that data sharing and reuse is important to drive science forward. As described, the motivations for data sharing include value to the scientific community, pressures from granting agencies and journal policies, and scientific reputation. Inhibitors include financial concerns, lack of time, lack of organizational support, lack of reward and inability to anticipate the intended user. However, the current research does not describe the motivations and inhibitors beyond general terms and therefore needs to be further examined to understand the intricate details of data sharing and reuse. Furthermore, the current research does not address cyberinfrastructure, interoperability, technical or system aspects but focuses mainly on human behavior. Moreover, much of the current research has been conducted in the biological sciences; therefore studies within the earth sciences would provide a different perspective, as scientific practices differ within these areas of study (Murillo, 2014).

DataONE Literature
DataONE is a four-year initiative that began in 2009, and has been extended to 2014, and has been further extended until Phase Two of the initiative. With the emphasis on infrastructure design and the new status of the DataONE, there has been limited time for assessment or investigation specific to the project itself, and few studies conducted have investigated data sharing and reuse within the DataONE community. The literature reviewed below provides an overview of the research that has been conducted in relation to DataONE.

Studies have been conducted to explore the collaborative relationship between information professionals and scientists and how this relates to the DataONE as a transdisciplinary organization (Allard & Allard, 2009).
Reichman, Jones, & Schildhauer (2011) demonstrated that DataONE provides access to well-curated, federated data repositories that can lead to improvements in sharing and reuse of data; however this study also indicated that the most effective means for data sharing would be to alter the reward system. Research has also investigated specific technical aspects of the DataONE, for example integration of large-scale computational runs with DataONE data, metadata, and workflow tools (Dexter, Cobb, Vieglais, Jones, & Lowe, 2011). Lastly, and perhaps most closely related to this research; there has been research conducted to address critical challenges facing researchers involved in data sharing and how these challenges influence researchers to share their data openly (Sayogo & Pardo, 2011). The data collected in this study was from 2009-2010, therefore a new study of the current DataONE users is particularly important to inform the DataONE. Furthermore, this study investigated scientists making their data available for sharing through deposition, but did not address scientists reusing data (Murillo, 2014).

These studies have begun preliminary research related to the DataONE. In order to ensure that the DataONE is being used to its fullest potential, additional studies regarding data sharing and reuse within the DataONE needs to be conducted (Murillo, 2014). This research provides the preliminary examination of data sharing and reuse within the DataONE environment, and will serve as preliminary analysis for future study that will examine factors that facilitate or interfere with data sharing and reuse within the DataONE user community.

**RESEARCH QUESTIONS AND RESEARCH METHODS**

The research questions for this study are:

Research Questions 1:

- What data is being deposited into the DataONE?
- What information is being provided regarding that data?

Research Questions 2:

- What data do scientists need when they search the DataONE?
- What data are deemed reusable and what information do scientists need about this data to deem it reusable?

The study uses a mixed-method approach: 1) a profiling data assessment and 2) a pilot think-aloud. Each method used in this research addresses specific aspects of the research questions, which are described below in more detail.

First, a profiling data assessment is being conducted through a quantitative content analysis of data deposited into the DataONE. This explores research question 1 including: a) what types of data are being deposited, b) what agencies are depositing data, 3) what disciplines are depositing data, and 4) what types of metadata are being provided. This provides an understanding of data sharing within the DataONE by analyzing the metadata that is being made available.

Secondly, a pilot think-aloud study explores: a) what types of data scientists need, b) what types of data scientists deem reusable, and c) what information do scientists need about that data to determine if it is reusable. This provides an understanding of data reuse within the DataONE by analyzing what is needed for scientists to deem data reusable.

Preliminary data gathering and analysis has been conducted and these results are described below. This data analysis will continue throughout the fall. It is expected to have more substantial results prior to the annual conference to present at the poster session.

**PRELIMINARY RESULTS**

**Profiling Data Assessments**

The profiling data assessment is being conducted through an analysis of metadata records extracted from DataONE. A random sample of metadata records have been extracted from the DataONE. This corpus of metadata records includes 650 XML records, which represents the complete corpus of 105,121 metadata records that were available from the DataONE at the time of extraction. To date, two representative samples of 650 records have been extracted; one is serving as a training and teaching set for content analysis and coding. This analysis provides an understanding of the data that is being shared into the DataONE environment, as well as the metadata that is being provided alongside the data. Preliminary analysis has begun using the training set on 20 of the XML records. This analysis will continue throughout the fall.

Of the 20 XML records analyzed all 20 records contained metadata regarding the dataset deposited into the DataONE. Additionally, 8 records contained metadata regarding access and 12 contained additional metadata. Only 5 of the XML records contained all three aspects (dataset, access, and additional metadata).

The additional metadata field contained clarifications regarding access; units, unit lists and definitions of units; and related datasets. The majority of this additional metadata was specifically related to units and unit lists. The access metadata provided information regarding access rights, although only 8 of the 20 records contained this information, of these 8 all data was public and several of these records contained guidance for citing the data.

The metadata for the datasets varied in regards to robustness and completeness. For example, while many records included a pubDate and language, not all records included this information. Furthermore, there were multiple formats for these metadata fields (i.e. “en” versus “English” for language). Additionally, 18 of the 20 records contained full abstracts, which scientists described as important to determining if a dataset was reusable during the pilot think-
aloud study discussed in the next section. Furthermore, some entries included information such as methods information, intellectual rights, and coverage while others did not.

Data analysis continues on these records and will be included in the poster to address the research questions described above. Additionally, in the pilot think-aloud study described below participants were asked how this information affected their ability to reuse the data available to them through the DataONE system.

Pilot Think-Aloud Study
A pilot think-aloud study was conducted during the November 2014, DataONE All-Hands Meeting in Albuquerque, NM and was approved by the UNC Institutional Review Board IRB number - 13-3499.

During the All-Hands meeting, I asked scientists to search for data in the DataONE system while I observed them. As they were searching the system, I asked them to think-aloud regarding the decisions they were making if the data was reusable or not reusable. Additionally, I conducted semi-structured interviews where participants were asked their subject expertise, if they had previously searched for data within the DataONE, and how often they searched. Participants were asked to describe their typical searches, what kind of data they were looking for, and what information they needed to determine if the data was relevant and reusable.

Participants indicated that they used the DataONE system fairly often, perhaps one or twice a month. They would search using general key terms, but then would narrow the results using the facets available through the system. Additionally, some participants indicated that they were also responsible for depositing data into the DataONE system, and therefore searched the system to make sure their data was being represented correctly.

Results indicated that scientists made decisions based on the metadata snippet they received from the system. These decisions were made through information including: if they had previous knowledge of the PI or author of the data, the actual data type, and the additional information provided regarding the data. In some instances, the topic or subject matter of the data was of interest to the participant, but they were unfamiliar with the data type and therefore the data was not reusable for their purposes. More often, the robustness of the metadata and the provenance information was key for the scientists choosing if this data was reusable. Scientists stated that information such as coverage and research methods helped them determine if the data was suitable for reuse.

Additionally, participants indicated that the key terms were really only helpful during the search and that they would prefer to see all of the PIs and not just the first author. Furthermore, they discussed how it would be helpful if all of the metadata snippets had the same stylesheet, as discussed above, not all of the records have the same information.

Data analysis continues on this data and will be included in the poster to address the research questions described above. This pre-pilot has served as a preliminary study for the creation of a second study, which will include a quasi-experimental aspect to control for specific variables.

SIGNIFICANCE OF WORK
This work has great significance for our field for many reasons. The need to create systems for scientific data sharing and reuse is ongoing and will continue due to changes in science. Therefore we need to understand if these systems are enabling sharing and reuse or not, and why. This work contributes a method and an approach for analyzing data sharing and reuse. It takes advantage of the opportunity to research on a rich environment supporting both the sharing side (making data available) and reuse side (reusing available data). Furthermore, this research provides insight into data sharing and reuse in the earth science community; to date data sharing research has been primarily focused in the biomedical, biological, and health sciences area.

ACKNOWLEDGMENTS
The author would like to thank Dr. Jane Greenberg for her guidance and support. The author would also like to thank the DataONE for their support.

REFERENCES


for Information Science and Technology, 54(10), 926–938.


The columns on the last page should be of approximately equal length.