AN INVESTIGATION OF SELECTED CYBERINFRASTRUCTURE AND INTEROPERABILITY ELEMENTS: DATA SHARING AND REUSE IN THE SCIENCES

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

ABSTRACT
This paper outlines the preliminary proposal plans for dissertation research to examine cyberinfrastructure and interoperability elements that facilitate or interfere with data sharing and reuse. The paper includes a brief review of the literature, a description of previous studies, a preliminary plan for dissertation research, and expected contributions to Information and Library Science. The proposed work will be conducted in the DataONE community. The DataONE provides a rich environment for studying cyberinfrastructure. Findings of the proposed work will contribute to DataONE's goals to develop and sustain an infrastructure that provides valuable support to scientists. The research proposed will also provide important feedback to other organizations creating systems for data sharing and reuse, and will provide the community an understanding of the factors that need to be addressed when creating infrastructure for scientific data sharing and reuse.

Categories and Subject Descriptors
H.2.8 [Database Applications] Scientific databases. H.3 Information Storage and Retrieval. J.2 [Computer Applications] Physical Sciences and Engineering – *aerospace, archeology, astronomy, chemistry, earth and atmospheric sciences, electronics, engineering, mathematics and statistics, physics*. J.2 [Computer Applications] Life and Medical Sciences – Biology and genetics.

General Terms
Documentation, Design, Reliability, Human Factors, Standardization

Keywords
Interoperability Elements, Data Sharing and Reuse, DataONE,

1. INTRODUCTION
This paper outlines dissertation proposal plans*[1] for examining a selected set of cyberinfrastructure and interoperability elements that facilitate or interfere with data sharing and reuse. The specific environment of interest is the US NFS-DataNets, and the test environment is DataONE.

Data sharing and reuse in the sciences has been a topic of growing attention over the last several years. This attention stems from the many changes that are occurring within scientific practices driven by the data deluge [2, 14], fourth paradigm data-intensive science [13], and changes in journal and grant agency policies [23, 25]. The sharing of data provides the ability to extract additional value from data, avoid reproducing research, enables researchers to ask new questions of existing data, and advance the state of science in general [5, 6]. These potential opportunities of data sharing and have placed pressure on the scientific community and funding agencies to provide infrastructure solutions for the changes in scientific practice.

A series of initiatives are developing long-term sustainable data infrastructures, interoperable data preservation and access, and cyberinfrastructure capabilities [24] to support how data within the sciences is changing. In 2007, the U.S. National Science Foundation announced a request for proposals for "Sustainable Digital Data Preservation and Access Network Partners" (DataNet) to address the challenges in science by "creating a set of exemplar national and global data research infrastructure organization that provide unique opportunities to communities of research to advance science and/or engineering research and learning" [24]. Specifically, these DataNet programs were created to:

- Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline.

[1] This planned proposal was written at the time of the iPRES-2013/DC-2013 Doctoral Consortium (September 2013). The research design has been refined and revised as of August 2014, but this paper remains as a record of plans at the time it was presented. Additionally, this paper outlines long-term goals and explores the benefits of multiple methods many of which are included in the revised dissertation proposal.

- 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.
- Serve as component elements of an interoperable data preservation and access network.

The Data Observation Network for Earth (DataONE), one of the initial NSF DataNets, 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" [10]. The DataONE provides this framework through cyberinfrastructure and an education and outreach program. Scientists participating in DataONE are able to deposit, search, and reuse data available through the various DataONE tools. The majority of DataONE research activities have focused on development, and it is timely to evaluate this progress. Specifically, there is a need to assess how various infrastructure elements may facilitate or interfere with data sharing and reuse.

The proposed research recognizes this need, and will investigate how selected infrastructure elements may impact data sharing and reuse. DataONE, including member nodes, will serve as the test environment. The research will study four specific infrastructure elements: 1.) provenance, 2.) data types, 3.) metadata, and 4.) tools (notebooks, software, workflow tools associated with data). A mixed method approach will use a content analysis of ONEMercury metadata (the DataONE web based search interface), a transaction log analysis from the ONEMercury, a quasi-experiment, a survey, and semi-structured interviews. The data gathered will enable the examination of these four infrastructure elements in more detail. The overriding goal is to determine how the selected infrastructure elements may facilitate or interfere with data sharing and reuse within the DataONE, and other cyberinfrastructure initiatives with similar goals.

This paper begins with a literature review of data sharing and reuse and of research conducted regarding DataONE. The third section outlines previous studies conducted by the researcher that has lead to this proposal. The fourth section outlines the planned proposal including research questions and research design. The fifth section outlines the research methods that will be employed during the dissertation. The fifth section outlines the expected contributions. Lastly, conclusions and implications for will be discussed.

2. LITERATURE REVIEW

This literature review addresses data sharing and reuse in the sciences, as well as provides a discussion of literature specific to the DataONE project. This review will begin with a discussion of the research conducted in relation to data sharing and reuse and the second portion of this review will provide an overview of research related to the DataONE project.

2.1 Data Sharing and Reuse

There is extensive literature that investigates data sharing and reuse and this literature review addresses the major themes found in the literature and provides a summary of the current knowledge of data sharing and reuse.

There have been multiple studies describing why data sharing and reuse is valuable to the scientific community. These studies have indicated that data sharing and reuse will advance scientific development through factors such as avoiding duplication of work, allowing new questions to be asked of existing data, and encouraging diversity in analysis [6, 15].

Along with making data available 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 management plans and some publishers have threatened to not publish articles if scientists do not make their data publically available [4, 17, 32].

Scientific reputation is another factor that influences scientists to share their data. Scientists unwilling to share their data could seem possibly fraudulent; therefore data sharing is considered in part of seeming reputable by fellow scientists [3, 8, 34].

Many researchers have attempted to analyze 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. Data ownership, previous assistance from coworkers, journal policies, and grant agency requirements were some of the motivations for data sharing [4, 6, 9, 35]. Disincentives for data sharing include financial, lack of time and funding, lack of organizational support, lack of documentation and complexity of metadata standards, as well as the difficulty to anticipate intended users [3, 36].

Researchers have also investigated data sharing through examination of journal data sharing 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 journal policies have indicated a refusal to publish without evidence of data deposition [7, 17, 18]. Studies have indicated that while no journal has complete compliance, much research data is deposited along with the article [26, 27]. 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 [29, 30].

The studies above demonstrate that most scientists do believe that data sharing and reuse is important to drive science forward and these studies also describe the general motivations and inhibitors for data sharing and reuse. 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 natural sciences would provide a different perspective, as scientific practices differ within these areas of study.

2.2 DataONE

DataONE is a four-year initiative that began in 2009, and has been extended to 2014. 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 DataONE in regards to it being a transdisciplinary organization [1]. Studies have also been conducted to demonstrate 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 [31]. There has also been research to investigate the infrastructure of DataONE, particular to describe how this infrastructure allows for the integration of large-scale computational runs with DataONE data, metadata, and workflow tools [11]. Lastly, 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 [32, 33]. 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 what infrastructural elements facilitate or interfere with data sharing and reuse. Also, this study looked specifically at scientists making their data available for sharing, but did not address scientists reusing data.

The studies specific to DataONE have begun preliminary research related to the DataONE.. In order to ensure that these services are being used to their fullest potential a study of cyberinfrastructure, interoperability, technical and system elements in regards to data sharing and reuse needs to be conducted. The proposed research will examine cyberinfrastructure and interoperability elements that facilitate or interfere with data sharing and reuse within the DataONE user community.

3. PREVIOUS STUDIES, RESULTS TO DATE, AND RELATIONSHIP TO GOAL

A series of earlier studies led by the researcher have informed the design of the proposed research. The below section highlights three previous studies conducted by the researcher on data sharing and reuse within the natural sciences and are relevant to framing the proposed dissertation project.

3.1 Data Sharing in the Scientific Community: A Bibliometric Study

This study analyzed a small dataset of 11,603 bibliographic records to explore data deposition in the biological sciences. This study used publically available data to examine when researchers deposited their data alongside their research articles [28].

The research questions were:

- Are researchers more likely to deposit their data in recent years than in year's prior?
- Does the number of authors contribute to researchers depositing their data along with their articles?
- Does the number of previous publications contribute to researchers depositing their data along with their articles?
- Does world region determine if data is deposited along with their data?

The dependent variable was if the article had an associated dataset deposited and the independent variables included year, number of previous publications, world region, number of authors, and sex. The researcher conducted descriptive statistics and elaboration models to analyze the data.

The findings indicated that within this dataset, data was not often deposited along with research articles and number of previous publications did not contribute to data deposition. The higher number of authors, the more likely data would be deposited and women were more likely to deposit data alongside research articles than men. Additionally, Mexico, Central and South American authors had a slight increase in data deposition than authors from other world regions.

Since this dataset was small and there was a significant amount of missing data, further investigation would need to be conducted to verify if this analysis indicated trends in the deposition of data alongside research articles. This study was conducted as an exploratory study to investigate some of the factors associated with data deposition [20].

3.2 Understanding User Motivations in Earth Science Data Reuse: Accessing Opinions on Skills, Access, and Trust

This study investigated how members of the Earth Science Information Partnership (ESIP) [12] reuse data. The study investigated skills required to find data, methods and barriers to accessing data, and how users evaluated the trustworthiness of data.

The research questions were:

- What skills do ESIP members find valuable when searching for data?
- Where did they learn these skills?
- What steps do ESIP members take to discover data?
- What do they do when they cannot get access to data they need or encounter barriers during their searches?
- How do they determine quality and trustworthiness of a source providing data?

An online survey was distributed to the ESIP community. The results indicated that 82% of the participants had reused data in the past 5 years and the majority of participant's stated they reuse data "all the time" in their work. However, 97% of the respondents also indicated that they encountered barriers or other difficulties when located data for reuse.

Scientists stated that knowledge of their field, resources available, and experience with similar data types were the most useful skills in regards to finding data for reuse. Scientists also indicated that skills they learned on the job were the most valuable skills for reusing data.

In regards to barriers, scientists stated difficulties finding acceptable datasets were in relation to poor metadata and cost. Participants stated that when they encounter a barrier they most often contact the person or organization responsible for the dataset in order to try to solve their access issues.

Lastly, scientists discussed which characteristics they felt were most important to determine confidence and trust regarding data for reuse. Scientists indicated that the quality of the metadata was the most important factor in regards to confidence and trust for data reuse.

The findings of this study have only been preliminary analyzed for presentation and the researcher continues to analyze this data as well as recruit for a larger sample of respondents [22].

3.3 Data At Risk Initiative: Examining and Facilitating the Scientific Process in Relation to Endangered Data

This study examined the scientific process in relation to endangered scientific data, data reuse and sharing. Endangered data is significant to the research process since deterioration, format obsolescence, and insufficient metadata for discovery are problems that lead to loss of scientific data. These data can be vital for long-term trend analysis.

The research questions were:

- What perceptions do scientists have on the topic of data at risk?
- What perceptions do scientists have of data reuse and sharing?

The researcher conducted four one-hour focus groups and a demographic survey with 14 scientists.

The results indicated that unavailability, lack of context, accessibility issues, and potential endangerment were all key concerns to scientists in regards to endangered data. The scientists stated that although they were able to gain access to data, without the context of metadata or proper metadata that the data itself was unable to be reused.

Scientists also stated incentives and disincentives to sharing and reusing data. The disincentives they states included: scooping/competition, the inability to share outside of their research group, equipment and technical issues, as well as metadata issues. In regards to incentives, the scientists suggested that collaboration, additional publication, and moving science forward were all reasons to share and reuse each other's data. Scientists echoed similar ideas expressed in the literature review, which indicated that many scientists believe that data sharing and reuse bring important opportunities to the scientific process to push scientific research forward and they believed that this should be the norm in science.

Preliminary results of this study were presented at the 2012 CODATA conference in Taipei, Taiwan and full results of this study were published in Data Science Journal [19, 21].

The studies described in this section, as well as the literature review demonstrate the researcher's knowledge of the domain, as well as the researcher's ability to conduct research projects using a variety of methods. These studies demonstrate the need for further investigation how interoperability elements facilitate and/or inhibit data sharing and reuse particularly within the DataONE community and have provided a framework of how to conduct dissertation level research on this topic.

4. PLANNED PROPOSAL[2]

This section describes the proposed research to investigate a selected set of cyberinfrastructure and interoperability elements that facilitate or interfere with data sharing and reuse in the DataONE user community. The following sub-sections will address the research questions, research methods, sample population, recruitment strategies, and plans for data analysis.

4.1 Research Questions

The proposed research will be guided by the following questions developed to investigate data sharing and reuse within the current infrastructure of the DataONE. These questions address two specific aspects of data sharing, making the data available for access and accessing the data to reuse. The preliminary research questions are:

- Within the DataONE environment, what infrastructure and interoperability elements facilitate or inhibit data sharing and reuse?
- Within the DataONE – ONEMercury, which results are deemed relevant for reuse?
    - What properties of these data and metadata facilitate or inhibit reuse?
- Within the DataONE community, how do these cyberinfrastructure and interoperability elements facilitate and/or inhibit data sharing and reuse?

The research design described below has been developed address and answer the above research questions.

4.2 Planned Research Methods

The proposed research will use a mixed methods approach that includes: (1) a profiling data assessment, (2) transaction log analysis, (3) quasi-experimental think-aloud, (4) online survey, and (5) semi-structured interviews.

4.2.1 Profiling Data Assessment

A profiling data assessment will be used to analyze the types of data being deposited into the DataONE infrastructure. This profiling data assessment will be conducted through an analysis of metadata records extracted from DataONE's – ONEMercury. A random representative sample of metadata records will be extracted from ONEMercury. Through a content analysis of these metadata records the researcher will determine which agencies are contributing to DataONE, which disciplines are depositing data, what data they are contributing, what types of metadata they are providing along with the data, and auxiliary information being deposited along with the data. This examination will provide details such as data standards, metadata standards, as well as discipline information regarding the data that is most likely to be deposited and accessible through the DataONE.

[2] As stated above, this planned proposal was written at the time of the iPRES-2013/DC-2013 Doctoral Consortium (September 2013). The research design has been refined and revised as of August 2014, but this paper remains as a record of plans at the time it was presented.

4.2.2 Transaction Log Analysis

Transaction logs will provide evidence to investigate the research questions. Transaction logs will be collected from ONEMercury. This will enable the researcher to assess the types of queries that are being conducted and the types of data being accessed through the ONEMercury. This will enable the researcher to assess how users search and download data. Lastly, this will enable the researcher to examine what data is being searched and downloaded and where the ONEMercury infrastructure facilitates and/or inhibits data sharing and reuse.

4.2.3 Quasi-Experimental Study

A quasi-experimental study will be conducted. The profiling data assessment and the transaction log analysis will assist in the creation of the quasi-experimental design by providing queries and records that can be emulated during the quasi-experimental study. The researcher will control the queries and the results in order to examine what elements facilitate or inhibit scientist's ability to reuse the resulting data from the query. Participants will complete a participant profile, which will provide background information. Participants will take part in the quasi-experiment by performing searches on ONEMercury. They will be asked to think aloud and note which elements facilitate or inhibit their ability to reuse the data. Participants will also complete a post-task survey so that the researcher can gather additional information regarding which elements facilitated or inhibited their ability to reuse the resulting data.

4.2.4 Online Survey

An online survey will be used to ask scientists questions about how selected cyberinfrastructure and interoperability elements facilitate or inhibit data sharing and reuse. The researcher will create an online survey based on the background literature and the results from the quasi-experiment. The researcher will make it clear in the recruitment message that all scientists whether or not they share and reuse data through the DataONE framework are encouraged to complete the survey, this way the researcher can gain data from scientists who are not actively using the framework to determine what is deterring them from doing so. Through addressing the process of accessing and depositing data, the researcher will be able to address more intricate details of what interoperability elements facilitate or inhibit the process of sharing and reuse.

4.2.5 Semi-structured Interviews

Lastly, the researcher will conduct semi-structured interviews with core members of Member Nodes to gain an understanding of the cyberinfrastructure and interoperability elements that facilitate or inhibit data sharing and reuse. This will gather more nuanced information regarding data sharing and reuse. The interview questions will be composed based on the results of the quasi-experiment and the survey to gain more in-depth knowledge of elements that are facilitating or inhibiting sharing and reuse in the sciences.

4.3 Planned Sample Population

The planned sample population for this study will be the DataONE user community. In order to gain a full understanding of the elements that facilitate or inhibit data sharing and reuse within the DataONE infrastructure, the researcher plans to keep the recruitment limited to this community. As described above, the researcher will ensure that scientists who both use the framework for data sharing and reuse and those who do not use the framework will be encouraged to participate in the survey and follow up interview, to gain an understanding of all members of the community.

4.4 Planned Recruitment Strategies

Scientists will be recruited through email listserves for the DataONE community, as well as through the DataONE website. A solicitation for completion of the online survey will be sent to the listserv twice in order to gain as many participants as possible. Another solicitation for the survey will be placed on the DataONE website for users to complete. The quasi-experiment may a take place at a DataONE users group meeting, all-hands meeting, or on-site at Member Node sites. At the end of the survey, participants will be asked to provide an email address if they are willing to be part in a Skype or telephone interview. After the survey is completed, the researcher will contact those who were willing to be part of the interview via email to set up an interview time and means of contact.

4.5 Planned Data Analysis

The transaction logs and content analysis of the metadata records will be analyzed through classification and coding. The quasi-experiment will be analyzed through statistical methods and content analysis. The online survey will be analyzed through descriptive statistics, as well as other statistical methods. Lastly, the interviews will be analyzed through inductive qualitative content analysis.

The goal upon completion of the data analysis would be to provide a framework of interoperability elements that either facilitate or interfere with data sharing and reuse that can be used by the scientific community as a whole.

5. EXPECTED CONTRIBUTIONS AND APPLICATION IMPLICATIONS

As described in much literature there are many reasons why scientists and the scientific community need the ability to share and reuse data. For example, in Lord and McDonald's 2003 [15] report on e-Science, the authors state that data sharing can extract additional value and avoid duplication of existing work. The OEDC's (2002) report indicated that sharing data reinforces scientific inquiry, encourages diversity in analysis, and promotes new ways to test hypothesis or methods of analyzing data [15]. Moreover, data sharing and reuse is important to the scientific community because it (1) makes results of publicly funded data available, (2) enables others to ask new questions, (3) advances the state of science, and (4) aids in reproducing research [5].

As described in the literature review, while there has been investigation to address this topic, much of this research has been in the biological sciences. The sample population was chosen specifically due to the lack of research of data sharing and reuse in the earth and ecological sciences. Also, much of the literature looks at motivations for deposition of data, and does not specifically address how interoperability elements inhibit or facilitate data sharing and reuse for accessing data. Lastly, there has not been investigation of the DataONE user community itself particularly with the current infrastructure to investigate the motivations and inhibitors to data sharing and reuse within the current user community and infrastructure.

Furthermore, although this research is specific to the DataONE user community it can provide a framework of best practices to encourage data sharing and reuse.

6. CONLUSION

This paper provides the overview for a planned dissertation proposal to examine a selected set of cyberinfrastructure and interoperability elements that facilitate or interfere with data sharing and reuse in the DataONE user community. As described in the literature review there is extensive literature that discusses data sharing and reuse, however, very little work that addresses technical elements and more specifically the DataONE user community. The planned dissertation research will use a mixed method, approach by examining transaction logs and metadata records within the DataONE framework, conducting an quasi-experiment, through asking scientists about their practices directly via an online survey, and through intensive interviews to gain access to the more nuanced information how interoperability elements facilitate or inhibit data sharing and reuse.

7. ACKNOWLEDGMENTS

I would like to acknowledge the support of the SILS Carnegie Fund, the UNC Center for Global Initiatives, CODATA, ESIP, and DataONE. Special acknowledgement to Dr. Jane Greenberg.

8. REFERENCES

[1] 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.

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

[3] Birnholtz, J. P., & Bietz, M. J. (2003). Data at work. *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work - GROUP '03* (p. 339). New York, New York, USA: ACM Press.

[4] 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.

[5] Borgman, C. (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/.

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

[7] Brown, C. (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.

[8] Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science Technology And Human Values*, *13*(1/2), 45–52.

[9] 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. Retrieved from http://isr.journal.informs.org/content/5/4/400.short.

[10] DataONE. (n.d.). DataONE - Data Observation Network for Earth. Retrieved May 3, 2012, from http://www.dataone.org/.

[11] Dexter, N. C., Cobb, J. W., Vieglais, D., Jones, M. B., & Lowe, M. (2011). DataONE member node pilot integration with TeraGrid? *Proceedings of the TeraGrid 2011 Conference Extreme Digital Discovery TG11*. .

[12] Federation of Earth Science Information Partners. (2013). ESIP-The Federation of Earth Science Information Partners. Retrieved from http://www.esipfed.org/.

[13] Hey, T., Tansley, S., & Tolle, K. M. (2009). *The Fourth Paradigm: Data-Intensive scientific discovery*. Redmond, Washington: Microsoft Research.

[14] 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

[15] 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*. The JISC Committee for the Support of Research (JCSR).

[16] Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. *Proc 3th UK e-Science All Hands Meeting*.

[17] McCain, K. W. (1995). Mandating sharing: Journal policies in the natural sciences. *Science Communication*, *16*(4), 403–431.

[18] 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.

[19] 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.

[20] Murillo, A. P. (2012). Data sharing in the scientific community: A preliminary examination. *Archival Education and Research Institute (AERI)*. UCLA.

[21] Murillo, A. P., Carver, N., Greenberg, J., Robertson, D. W., & Thompson, C. (2012). Data-At-Risk: Scientists' perceptions of endangered data and data reuse. *CODATA*. Taipei, Taiwan.

[22] Ramdeen, S. & Murillo, A. P. (2013). Understanding user motivations regarding earth science data reuse: Accessing opinions on skills, access, and trust. *Archival Education and Research Institute (AERI)*. University of Texas at Austin.

[23] National Institutes of Health. (2007). NIH Data Sharing Policy. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/

[24] 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.

[25] National Science Foundation. (2010, November 10). Dissemination and Sharing of Research Results. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[26] 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. doi:10.1371/journal.pbio.0040228

[27] 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.

[28] Piwowar, H. (2011). Data from: Who shares? Who doesn't? Factors associated with openly archiving raw research data. *Dryad Digital Repository*. doi:10.5061/dryad.mf1sd.

[29] Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS ONE*, *6*(7), e18657. doi:10.1371/journal.pone.0018657

[30] Piwowar, H. A., & Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. *Journal of Informetrics*, *4*(2), 148–156. doi:10.1016/j.joi.2009.11.010

[31] Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21311007

[32] Sayogo, D. S., & Pardo, T. A. (2011). Exploring the determinants of publication of scientific data in open data initiative. *Proceedings of the 5th International Conference on Theory and Practice of Electronic Governance* (pp. 97–106). ACM.

[33] 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.

[34] Sonnenwald, D. H. (2007). Scientific collaboration. (B. Cronin, Ed.) *Annual Review of Information Science and Technology*, *41*(1), 643–681.

[35] 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.

[36] Zimmerman, A. S. (2003). Data sharing and secondary use of scientific data: Experiences of ecologist. (*dissertation*)