Thu-Mai L. Christian. Tracking Institutional Data Assets Using OAI-PMH. A Master's Paper for the M.S. in I.S degree. April, 2012. 60 pages. Advisor: Helen Tibbo

This master's paper explores the impetus for data stewardship policies and guidelines and describes an empirical study conducted to assess the feasibility of employing OAI-PMH as a tool to enable institutions to track data stewardship activities.

Quantitative analyses of Dublin Core metadata harvested from twelve data-specific repositories were used to make conclusions about the current state of OAI-PMH implementation, to consider ways in which the unique properties of scholarly domains and their data might be reflected in metadata values, and to suggest steps repositories can take to enable the development and implementation of a federated index of distributed data records as a tool to support the sustainability of the research enterprise.

Headings:

Data Libraries

Dublin Core

Information retrieval

Data policy

Scientific Archives

Social Science Archives

TRACKING INSTITUTIONAL DATA ASSETS USING OAI-PMH

by
Thu-Mai L. Christian

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2012

Approved by

Helen Tibbo

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Table

Figure

ABBREVIATIONS

CISL RDA	Computational and Information Systems Laboratory Research Data Archive
DC	Dublin Core
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
DOI	Digital Object Identifier
HDL	Handle
IQSS DVN	Institute for Quantitative Social Science Dataverse Network
NCAR CDP	National Center for Atmospheric Research Community Data Portal
NEH	National Endowment for the Humanities
NIH	National Institutes of Health
NSF	National Science Foundation
OAI-PMH	Open Access Initiative Protocol for Metadata Harvesting
OHS	Open Harvester Systems
URL	Uniform Resource Locator

CHAPTER 1

INTRODUCTION

A great deal of attention has recently been given to research data stewardship as a result of mandates issued by major funding agencies that acknowledge both the affordances of the technology that has enabled the substantial increase in research data production as well as the complexities that come along with digital scholarship. In May 2010, the National Science Foundation (NSF) issued a press release announcing its new data management plan requirement for all grant applications. As of January 18, 2011, all researchers who submit grant applications must include in their proposals a one- to two-page description of data types, data and metadata formats, access and sharing policies, provisions for reuse, and plans for archiving and preserving data. Taking its cue from NSF, the National Endowment for the Humanities (NEH) (2011) also issued its own data management plan policy for Digital Humanities Implementation Grants. These policies were preceded by the National Institutes of Health (NIH), which issued their Final NIH Statement on Sharing Research Data in February 2003. This statement emphasized NIH's promotion of data sharing by requiring that grant proposals requesting funds totaling $500,000 or more to include a data sharing plan that details how data will be shared, or justifies why data will not be shared.

NSF, NEH, and NIH are not alone in their initiatives to hold researchers accountable for data stewardship practices. The SHERPA (Securing a Hybrid

Environment for Research Preservation and Access) Project's (2009) Juliet database contains a listing of 92 funding agencies around the world (not counting NSF and NEH) that have issued publication and/or data archiving policies. One can expect this number to increase in the coming years, which points toward the significance of data stewardship as an important component of the research enterprise that is changing in response to advances in technology.

Likewise, journal publishers have impressed upon authors the need for research results to be made accessible. Hanson, Sugden, and Alberts (2011) of the journal *Science* published an editorial that justified extensions to their data policies as a means to ensure proper citation, description, standardization, preservation, and access of data underlying published research results. The *American Naturalist* (2010) published an editorial that introduced the journal's new data archiving policy, which outlines expectations for data archiving for purposes of supporting conclusions in published articles and allowing for interpretation and reuse. This policy also applies to *Evolution*, *Journal of Evolutionary Biology*, *Molecular Ecology*, *Hereditary*, and other journals owned or sponsored by the Society of American Naturalists. For all of their journals, the Public Library of Science (PLoS) (2011) has specific policies on sharing materials, methods, and data, which are based on principles outlined in the National Academies Press (2003) report "Sharing Publication-Related Data and Materials." The report highlights the authors' responsibility to share data as part of a community standard that is essential to research in the life sciences. The American Psychological Association (APA) has implemented a practice of sending an "Open Letter to Authors for APA Journals" to the authors of its 34

journals. This letter declares APA's requirement that authors share their data upon requests for validation of published results.

Aside from funding agency and publisher mandates, the push for research data preservation and access is in response to a growing emphasis on cross-disciplinary research and public interest in research data, including from non-scientists outside of the discipline in which the data were originally collected and used (Faniel & Zimmerman, 2010; National Research Council, 2010). Supporting the viability of research beyond the active primary research phase by providing access to data for secondary uses extends the utility of data to new communities, makes novel solutions to human problems possible, and enables "system-level" science (Faniel & Zimmerman, 2011).

Dozier and Gail (2009) use the topic of water as an example of the necessary integration of disciplines to conquer problems of water supply shortages. The types of data required to adequately address this issue include, but are not limited to, satellite data to identify water sources over an extended geographic area, sensor data to determine ground moisture levels, social science data to analyze human behaviors in relation to hydrologic phenomena—all of which require the cyberinfrastructure to integrate and provide real-time access to several forms of data (Dozier & Gail, 2009). In the medical field, the combination of electronic health records, biotechnology, and scientific research outputs (i.e., journal publications) enable researchers to ascertain the complexity of specific illnesses. These resources allow scientists to investigate patterns and associations among symptoms and outcomes described in the medical records of affected patients, the known genetic factors that predispose individuals to the illness, and the epidemiological "problem space" determined by the body of literature on the specific

illness (Buchan, Winn, & Bishop, 2009). These examples show ways in which the value of data assets are further extended when made accessible beyond their initial purpose, thus providing funding agencies with greater returns on their investments (OECD, 2007; Rusbridge et al., 2005).

As a result of increasing pressures from funding agencies and publishers to manage their data in ways that enable sharing, discovery, and reuse, researchers are forced to consider the adequacy of their research data stewardship practices to achieve these mandated goals. Moreover, the institutions that host scholarly research are obliged to provide oversight of data stewardship practices to guarantee sustained funding and promotion of their research enterprise—a duty that is not easily fulfilled.

The scope of data being produced within any single institution is difficult to capture. In sites around the world, massive volumes of data in various forms are being generated in laboratories and by sensors. After they have served their primary research purpose, the data may be stored in any of a countless number of repositories in yet more locations spanning the globe. Each of these repositories operates according to the culture and norms of the scholarly community it represents. The burden, therefore, is placed on the institution to apprehend the data diversity, magnitude, locations, and research domains to be able to track and support data stewardship activities that funding agencies and publishers require.

Tracking and supporting data stewardship requires a mechanism for the search and discovery of data sets across disciplines and across geographically distributed repositories. The Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) was established to allow for such an apparatus. In 2001, Carl Lagoze and Herbert Van de

Sompel introduced the Open Archives Initiative and its development of a "low-entry and well-defined interoperability framework applicable across domains…[as an] appropriate catalyst for the federation of a broad cross-section of content providers" (p. 54). Though few articles about OAI-PMH have been published in the past few years, which might suggest that acceptance of the framework has waned, current trends in the research enterprise demand another look at OAI-PMH as a possible tool for tracking institutional data assets.

This master's paper explores the impetus for data stewardship policies and guidelines and presents the analysis of an experiment to assess the feasibility of employing OAI-PMH as a tool to enable institutions to track data stewardship activities to ensure the sustainability of their research enterprise.

CHAPTER 2

REVIEW OF THE LITERATURE

Data and the Research Enterprise

Today's research enterprise is one that is characterized by the *data deluge*. Hey and Trefethen used this oft-cited term in 2003 to describe the magnitude of data being generated in the new age of e-Science[1] for which current database capacity cannot accommodate or soon will not be able to accommodate. Pointing to the petabytes of data produced in astronomy facilities each year and the anticipation of genome sequencing to increase four-fold per year, the authors do not lack compelling examples of data deluge. The National Research Council (2009) described the technological innovations that have enabled this enormous scale of data production as profound. Digital technology has allowed researchers to process and analyze these massive quantities of data while at the same time introducing new approaches to scientific inquiry enhanced by the integration of simulation, observation, and experimentation data. Furthermore, virtual platforms have made possible interdisciplinary collaborations that connect researchers from different parts of the world. New advances in technology for data access, sharing, and reuse have transformed the way scholars conduct research—and transformed the challenges associated with these advances.

---

[1] *e-Science* is a term that was coined in 1999 by Dr. John Taylor, Director General of Research Councils in the UK Office of Science and Technology, to refer to the evolution of scientific research as a result of developments in cyberinfrastructure: "e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" (Hey & Trefethen, 2002).

Despite all of the affordances of technology, Bell, Hey, and Szalay (2009) find the data deluge to be "burdensome," particularly in regard to the proliferation of born-digital data in laboratories that engage in data-intensive research (p. 1297). The authors blame these burdens on the shortcomings of required tools, which have been slow to develop in comparison to the volume of data outputs. Lord, Macdonald, Lyon, and Giaretta (2004) state more plainly that the technology puts data at risk. Berman (2008) uses the term "fragile" to describe digital data, and as such requires infrastructure that is able to handle technical, policy, and economic issues among others (p.50). Brase (2004) notes that "project data is often poorly documented, therefore badly accessible and not maintainable over long time periods. Large amounts of data are unused as they are only known and accessible to a small group of scientists" (p. 488).

Funding agencies and publishers are attempting to overcome these challenges by issuing specific requirements for data access, preservation, and sharing. Funding agencies want to ensure that funded research provides the greatest returns on their investments. The NSF's (2011) Grant Proposal Guide requires that a data management plan document be included in the grant proposal. This document must contain descriptions of the data type along with any other types of research output (e.g., software, samples); data and metadata format and content standards; access and sharing policies with information on measures to protect privacy, confidentiality, and intellectual property rights; policies for secondary analysis; and provisions for data archiving and preservation. NSF makes clear that the data management plan is a fundamental part of the grant proposal and will be reviewed accordingly. In a press release distributed in May 2010, NSF acknowledged the growth in data-intensive research and justified the data

management plan as a way to promote collaboration. The press release included a quote from Jeanette Wing, assistant director for NSF's Computer & Information Science & Engineering directorate, underscoring the value of data: "digital data are both the products of research and the foundation for new scientific insights and discoveries that drive innovation" (NSF, 2010). Appended to the announcement of the NEH new digital Humanities Implementation Grants program was a statement on the NSF-based data management plan requirement, which NEH adopted (Bobley, 2011). Other major funding agencies that have issued guidelines and/or policy on data management and sharing include the NIH (2003), Centers for Disease Control and Prevention (2005), the Department of Defense (1998), and the National Aeronautics and Space Administration (2011).

Meanwhile, publishers want to ensure that research results published in articles can be replicated as a means to maintain the integrity of their scholarly journals. An editorial written by the editors of *Science* echoed the sentiments about data-driven science in the NSF press release (Hanson et al., 2011). The *Science* editors point out that new technologies not only have ushered in advances in research, but also have brought about challenges to making data accessible. The strengthening of their policies on data include, among other things, a requirement similar to funding agencies' data management plan. *Science* now requires that authors submit a statement describing data access and archiving as part of the author's acknowledgments. PLoS (2011) has not issued a specific data policy, but does include in their editorial and publishing policy a declaration stating that authors who do not comply with best practices for data sharing established in their respective disciplinary domain will see publication decisions affected.

These are but few examples of policies that are now applicable to researchers seeking funding or publication. Among these policies are requirements that vary widely, even within a particular research domain. Weber, Piwowar, and Vision (2010) performed an analysis of data citation and sharing policies affecting the environmental science domain. Within this field of study alone, the authors found that various stakeholders commonly issue a diverse range of policies, many of which do not provide clear guidance on how to fulfill stated requirements. Beagrie, Beagrie, and Rowlands (2009) reported similar findings from a survey conducted of 179 researchers and face-to-face interviews of 37 researchers from four UK universities. Based on researchers' responses to questions covering data storage requirements, legal requirements, and data access, the authors concluded that funding agency requirements for data preservation and availability of services to enable archiving support for policy requirements are not consistent in each discipline.

Roles and Responsibilities

Where requirements differ, so do the research approaches, disciplinary culture, community best practices for data management, and relevant laws and policy dictating data stewardship responsibilities within the disciplines that produce data and publish articles based on those data. Thus, institutions have further reason to feel obligated to provide the services, infrastructure, and policies to ensure researchers are able to navigate the constellation of rules and policies that can affect their funding.

Moreover, because the university is often named owner of research data funded by government agencies like NSF and NEH, it must ensure that researchers actually perform the tasks described in the data management plan, particularly those involving

data archiving and sharing (Culliton, 1988; Fishbein, 1991). Institutions assert ownership rights because of the responsibility it bears for oversight of research ethics. This is a "work-for-hire" model that protects the reputation of the institution in instances of scientific misconduct (Fishbein, 1991, p. 131). Thus, several universities have been explicit about their ownership rights. Duke University (2007), Johns Hopkins University (2008), Stanford University (1997), University of Kentucky (2011), University of Pittsburgh (2009) and many others have distributed institutional policy documents that put forth an assertion of institutional ownership in some version of a statement that reads, "the University owns all research data generated by research projects conducted at or under the auspices of the University" (Johns Hopkins University, 2008, p. 2). It is a tall order for the institution to protect its data assets and fulfill its owner obligations by supporting and overseeing data stewardship practices that are as heterogeneous as the data themselves.

Fortunately, many scholarly communities already have established mechanisms for archiving and sharing data. Domain-specific repositories such as PubChem for chemistry, Dryad for biological sciences, Dataverse Network for social science, and PANGAEA for geoscientific and environmental data each provide an effective infrastructure for data archiving and storage for data producers in their respective disciplines. These repositories have received data submissions consistently for years, which can be attributed to established standards in disciplines in which data formats are constant (e.g., genome sequencing, social science) (Nelson, 2009). It makes little sense for institutions to replicate these existing services.

Still, the institution must be able to track their researchers' archiving and sharing activities among these and many other distributed data repositories. In 2009, the National Research Council held a workshop to ascertain the issues and challenges of integrating scientific research data. Workshop attendees acknowledged the need for a "homogenous logical view of data that is physically distributed over heterogeneous data sources" (Zigeler & Dittrich, 2004 as cited in National Research Council, 2010, p. 2). The Council determined that this view could be seen using a data registry, which would allow users to "obtain a single view of collection all over the world" (National Research Council, 2010, p. 12). Such a view would allow institutions to track data stewardship activities for the purpose of accountability to data management plan mandates and to facilitate interdisciplinary research that funding agencies promote.

OAI-PMH as a Tool for Institutional Data Asset Tracking

One potential tool for tracking institutions' data assets is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which enables the harvest and aggregation of metadata from distributed repositories into a single searchable interface. Such a tool could act as a registry of existing datasets, which would allow institution administrators tasked with data management oversight to track archiving and sharing of the institution's data assets and to ensure compliance with funding agency and publisher mandates.

At a 2001 conference, Carl Lagoze and Herbert Van de Sompel (2001) introduced the Open Archives Initiative (OAI) and its development of a "low-entry and well-defined interoperability framework applicable across domains…[as an] appropriate catalyst for the federation of a broad cross-section of content providers" (p. 54). The impetus for

such a development, they explained, is the rapid increase of data being produced in the sciences that requires mechanisms for sharing results along with the rising use of Internet technology for delivery of those results (Lagoze & Van de Sompel, 2001). These, along with increasingly prohibitive commercial publishing costs that have been cited as cause for a crisis in the traditional model of scholarly communication have instigated a movement towards open access initiatives (Lagoze & Van de Sompel, 2001; Yiotis, 2005).

OAI-PMH provides the technical framework to support a federated repository model for which scholars are able to engage with datasets across disciplines though an interoperable network of digital repositories (Lagoze & Van de Sompel, 2001). Such an apparatus also enables the development of a service that "might use information from various repositories and process that information to link citations, create cross-repository query interfaces, or maintain current awareness services" (Lagoze & Van de Sompel, 2001, p. 55). These services that OAI-PMH creators described over a decade ago is one that institutions are now seeking in light of the new and growing demands being placed on researchers to engage in data management practices—and on the institutions who are obligated to provide the resources and infrastructure required to sustain the value of their data assets. This is also the service that the current project hopes to demonstrate is feasible by exploiting the OAI-PMH framework as Lagoze and Van de Sompel described.

Dublin Core Metadata

OAI-PMH acts as a vehicle that enables distributed repositories, or data providers, to allow web access to their metadata records. To do so, OAI-PMH makes specific

requirements for metadata implementation that is based on the goal of interoperability, which is dependent on metadata standardization (Lagoze & Van de Sompel, 2001, p. 57). Hey & Trefethen (2003) stressed the importance of the quality of metadata associated with the data to allow for the mining of metadata though search engines (p. 9). Citing the astronomy domain as an example, the authors assert that "[t]he existence of such standards for metadata will be vital for the interoperability and federation of [data] held in different formats in file systems, databases, or other archival systems" (Hey & Trefethen, 2003, p. 10).

The OAI-PMH requirement to use the Dublin Core Metadata Element Set (DCMES) fulfills the goal of interoperability. The Dublin Core Metadata Initiative (DCMI) emerged in the 1990s as a means to build core metadata vocabularies that support interoperability among content stewards. One of its principles of operation is "the discovery of resources across the boundaries of information silos on the Web and within intranets" (DCMI, n.d.) DCMI describes four "levels of interoperability" that defines the scope of metadata, or the "metadata landscape": 1) shared term definitions, 2) formal semantic operability, 3) description set syntactic interoperability, and 4) description set profile interoperability. DCMI provides the graphic below (fig. 1) to further illustrate these four levels (DCMI, n.d.).

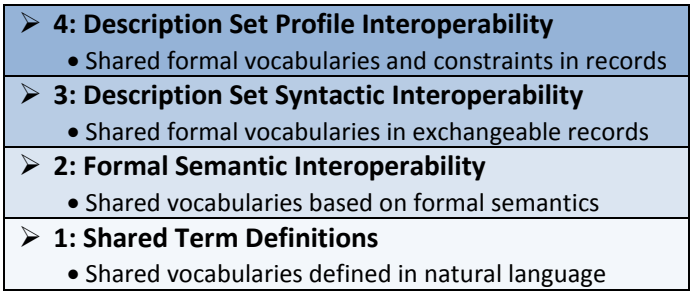

Figure 1. Four Levels of Metadata Interoperability

For the purpose of a federated index of datasets (as opposed to a federated catalog of the datasets themselves), only levels 1 and 2 are required, and are of interest in the current paper.

DCMES is comprised of 15 elements that provide generic descriptions applicable to a diversity of resource types and domains. The table below lists the elements (table 1).

contributor	publisher
coverage	relation
creator	rights
date	source
description	subject
format	title
identifier	type
language	

Table 1. Dublin Core Metadata Elements

To achieve the 1st and 2nd levels of interoperability that enable the harvesting of metadata values necessary to provide enough information to allow users to search for and locate a data set, a minimum of five elements are required (Altman & King, 2007). The first three—dc.creator, dc.date, and dc.title—correspond to document-type objects and provide sufficient information to locate the intended record. According to the DCMI standard, dc.creator is the author, dc.date indicates the date the data were generated or published, and dc.title is the descriptive name used to refer to the record. However, the perpetually changing Internet landscape makes persistent, unique identifiers a necessity for locating a digital object, particularly when the object's physical location or owner changes. Thus, the dc.identifier element has become increasingly important for the discovery of data sets (Rajasekar & Moore, 2001; Brase, 2004; Paskin, 1999; Paskin,

2005). Because the proposed registry focuses specifically on data sets, the dc.type element is necessary to distinguish among different types of repository content that may be present in any individual repository. Certainly, additional elements may further assist users in interpreting the nature of the data set. However, the degree to which a searchable registry is successful depends on the consistent and standard implementation of the five Dublin Core (DC) metadata elements identified above.

In 2001, the developers of OAI-PMH made the decision to make all DC elements optional rather than requiring a minimum set of elements (Lagoze & Van de Sompel, 2001). They argued that the purpose of unqualified DCMES as the common metadata set was for discovery, while detailed description of the resource would rely on local metadata specific to the community (Lagoze & Van de Sompel, 2001). This decision has drawn criticism from several fronts. Researchers have demonstrated that repositories' implementation of DC metadata has been inadequate and undermines any of the technical affordances of OAI-PMH (Ward, 2004; Van de Sompel, Nelson, Lagoze, & Warner, 2004; Shreeves, Kaczmarek, & Cole, 2003; Jackson, Han, Groetsch, Musafoff, & Cole, 2008).

The Illinois OAI-PMH project, a Mellon Foundation funded project to test the efficacy of OAI-PMH for federated search across cultural heritage repositories, found that the flexibility of DCMES introduced wide variability in metadata implementation, which significantly affected the ability of the system to discover resources across distributed repositories (Shreeves et al., 2003, p. 162). In 2004, Ward published results of a content analysis of DC metadata for 100 repositories harvested using the OAI-PMH protocol. The author found that repositories used only a handful of the 15 DC elements,

which prompted her call for further research into the underutilization of DC elements. Jackson et al.'s (2008) examination of metadata harvested via OAI-PMH by the Institute of Museum and Library Services (IMLS) Digital Collections and Content Project and the Committee on Institutional Cooperation Metadata Portal supports Ward's findings that use of DC is limited at best; the only elements used consistently among harvested repository metadata were the dc.title and dc.identifier elements. Moreover, the authors found that the harvested DC metadata were not generated natively in DC but rather mapped from other schemas used in the repositories' local context (Jackson et al., 2008, 2003). The result was incorrect mapping or mapping that ignored semantics values.

Yet, very little criticism specifically of the technical framework of OAI-PMH was found in the literature. Shortcomings primarily pointed to inadequacies of metadata implementation rather than to the protocol itself. Shreeves, Kaczmarek, and Cole (2003) conceded that OAI-PMH has the potential to support search and discovery; it is the implementation of the metadata that wants for further analysis. Where issues with OAI-PMH metadata harvesting have been identified, authors often suggest new approaches having to do with changes in how metadata are applied (Liu et al., 2002; Van de Sompel et al., 2004; Haslhofer & Schandl, 2008).

It should be noted that active federated repository indexes do exist that effectively employ OAI-PMH. The Research Library of the Los Alamos National Laboratory (LANL) uses OAI-PMH to store and access content (Jerez, Liu, Hochstenbach, & Van de Sompel, 2004). The Data Preservation Alliance for the Social Sciences (Data-PASS) group has successfully replicated archival content across a distributed network of repositories using OAI-PMH (Altman et al., 2009). A case study of the National Science

Digital Library's use of OAI-PMH found the OAI approach to be appropriate for the development of a large-scale digital library (Arms, Dushay, Fulker, & Lagoze, 2003). Just seven years after its release, the number of metadata records harvested using the OCLC OAIster harvester tool was almost 20 million (Beisler & Willis, 2009).

Despite the lackluster use of a minimum set of DCMI elements, the examples of successful OAI-PMH implementation in the literature warrant a consideration of OAI-PMH as a tool for supporting a federated index of data sets. Therefore, the current paper examines OAI-PMH implementation by digital repositories dedicated to storing and providing access to datasets to determine whether or not the current state of OAI-PMH in terms of tool functionality and implementation of a minimum set of Dublin Core metadata elements is able to effectively support a middleware service that aggregates discovery metadata from distributed data repositories. To do so, I conducted a research project to test the following hypotheses:

- OAI's harvesting protocol is a viable option for the obtaining the metadata necessary to populate a federated index of data sets that are stored in spatially and disciplinary disparate data repositories.
- Any constraints to the development of a registry employing OAI-PMH would present themselves in an analysis of metadata values.

To test these hypotheses, I employed OAI-PMH to harvest and analyze metadata records from data repositories serving a diversity of scientific domains. Success in harvesting the metadata was meant to demonstrate the adequacy of the OAI-PMH tool to perform such a task, thus supporting the first hypothesis. To support the second hypothesis, a content analysis revealed whether or not the harvested metadata provided sufficient information to populate a searchable index of distributed datasets. Support of both hypotheses was necessary to suggest that OAI-PMH is a viable option for tracking

data stewardship activities. The analysis was then used to make conclusions about the state of OAI-PMH implementation in data-specific repositories, to consider ways in which the unique properties of the scholarly domain and its data might be reflected in the metadata, and to suggest steps repositories can take to enable the development of a federated index of distributed data records.

CHAPTER 3

METHOD

Identification of OAI-PMH Compliant Data Repositories

Identifying OAI-PMH compliant repositories that primarily contain data sets proved more difficult than expected. No current list or registry of OAI-PMH compliant repositories exists on the World Wide Web. The Open Archives Initiative website does provide a list of registered data providers; however, dates of last validation of registration records are generally no more current than 2007 (OAI, 2011). Moreover, a repository's inclusion in the list is voluntary, and many have not taken steps to register their repository information. Therefore, the OAI list of 1,602 registered data providers is not comprehensive. The University of Illinois also hosts a web-accessible OAI-PMH Data Provider Registry (University of Illinois, 2011). With a significantly higher number of repositories listed at 2,895, the list appeared promising. However, the site indicates that validity checks for many repositories had not been completed in the last two to three years. The Open Access Directory (OAD) wiki hosted by the Graduate School of Library and Information Science at Simmons College (2011) provides links to data repositories' web sites, which are conveniently categorized by disciplinary domain. To determine whether or not each repository implemented OAI-PMH and provided a valid base URL, I had to visit each web site and browse for the required information. Another resource I consulted was the Directory of Open Access Repositories, or OpenDOAR (2010). This

service includes a search function, which allowed me to obtain a list of repositories filtered by the "datasets" content type. A review of the resulting list showed that very few of the repositories contained primarily data, and fewer had implemented OAI-PMH. Thus, the repositories selected to participate in the investigation were few despite having used several sources to identify those meeting the research project criteria (i.e., data content, OAI-PMH compliance). Still, the repositories selected for the project contain data produced in a variety of research disciplines including the social sciences, biology, environmental sciences, archeology, and geography—each representative of their respective domains (see table 2).

Repository	Domain	Host
CISL Research Data Archive (RDA) http://dss.ucar.edu/	Atmospheric and Geosciences	Computational and Information Systems Library (CISL) at the National Center for Atmospheric Research
DataCite Metadata Store (MDS) http://oai.datacite.org/	Multidisciplinary	DataCite members representing several data providers
Dryad http://datadryad.org/	Biological Sciences	National Evolutional Synthesis Center; University of North Carolina Metadata Research Center
eCrystals http://ecrystals.chem.soton.ac.uk/	Biology and Chemistry	University of Southampton
Edinburgh DataShare http://datashare.is.ed.ac.uk/	Multidisciplinary	University of Edinburgh
Institute for Quantitative Social Science (IQSS) Dataverse Network (DVN) http://dvn.iq.harvard.edu/	Social Science	Harvard University
National Center for Atmospheric Research (NCAR) Community Data Portal http://cdp.ucar.edu/	Atmospheric and Geosciences	Computational and Information Systems Library (CISL) at the National Center for Atmospheric Research
Odum Institute Dataverse Network (DVN) http://arc.irss.unc.edu/dvn/dv/odvn	Social Science	University of North Carolina
Open Context http://opencontext.org/	Archaeology	Alexandria Archive Institute
PANGAEA http://www.pangaea.de/	Environmental and Geosciences	Alfred Wegener Institute for Polar and Marine Research; Center for Marine Environmental Sciences
ShareGeo http://www.sharegeo.ac.uk/	Geography	EDINA, JISC National Data Centre
VizieR Catalogue http://vizier.u-strasbg.fr/viz-bin/VizieR	Physics and Astronomy	Centre de Données Astronomiques de Strasbourg

Table 2. List of Harvested Repositories

Installation and Configuration of the OAI-PMH Harvester

After download and attempted installation of several OAI-PMH harvesters, I selected the Public Knowledge Project's Open Harvester Systems (OHS) for its user-friendly graphical user interface (GUI), which made the need for advanced command prompt programming unnecessary. The Public Knowledge Project website also provides comprehensive user documentation to support installation, configuration, and harvesting applications, which was useful during software set-up.

Installation of the current OHS version 2.3.1 requires several software applications to be present on the system:

- PHP
- MySQL
- Apache
- Operating system supporting PHP, MySQL, and Apache

For the current project, OHS was installed in a Mac environment. To fulfill prerequisites for OHS installation and operation, I downloaded MAMP on the project machine. As its name suggests, MAMP (an acronym for Macintosh, Apache, MySQL, and PHP) allows for a single, integrated download of the applications required to run OHS. MAMP sets up a local web server environment that supports the OHS GUI and the database that stores harvest data. A script included in the installation package files created the database, installed the database tables, and provided initial data. After exploring the newly-installed tool, I was able to begin harvesting repository metadata.

The OHS harvester tool includes functionality for browsing the metadata records for each repository. Using the GUI, users are able to view and browse the records of

each harvested repository (fig. 2) and view the exposed metadata for individual records (fig. 3).

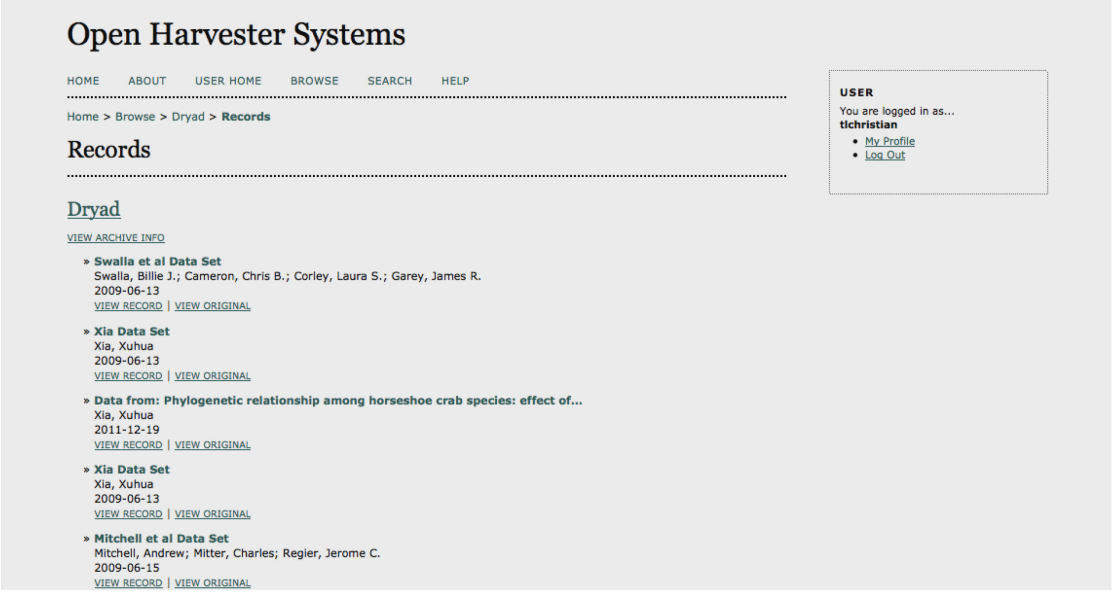

Figure 2. Open Harvester Systems GUI: Browsing repository records

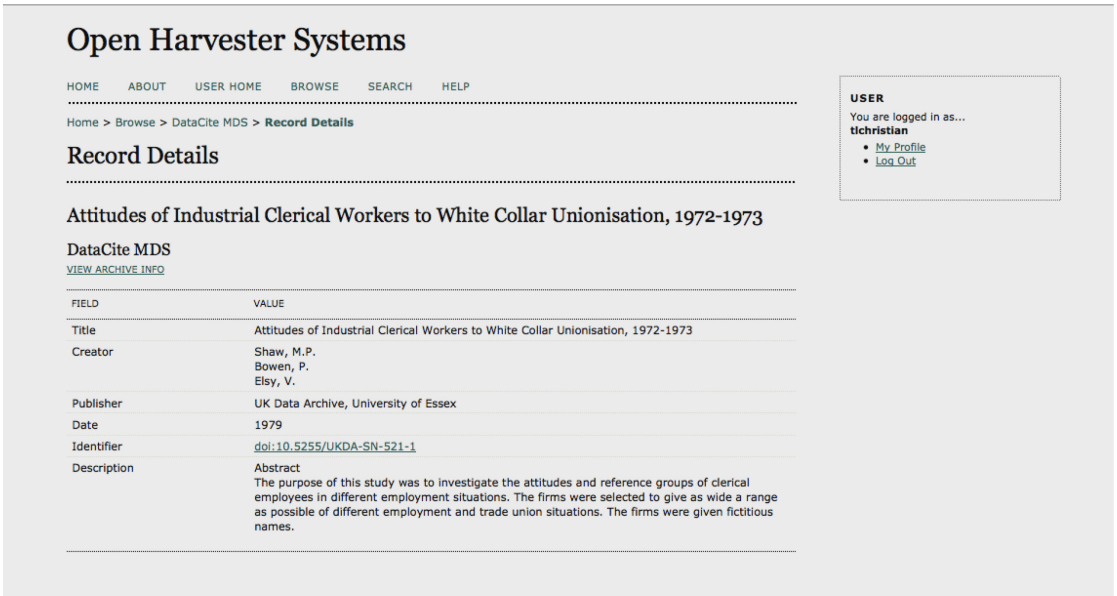

Figure 3. Open Harvester Systems GUI: Viewing a Metadata Record

To perform analyses of metadata content, the harvested metadata records were exported from the MySQL database containing the OHS records. This "data dump" was then imported into several database and spreadsheet applications to generate aggregations of metadata content to determine frequencies of metadata element usage. Aggregations provided a complete accounting of DC metadata elements used by each repository to determine whether or not inputs exist for the minimum metadata elements required for resource discovery and access. In addition, I conducted content analyses of the dc.identifier and dc.type elements by running frequencies of element field values to determine the quality of metadata inputs. Quality metadata contains the necessary information required to support interoperability that is critical to the operation of the federated index. For the dc.identifier element, quality is based on the use of widely adopted, standard persistent identifiers that link the metadata record to the actual data set. For dc.type, the use of controlled vocabularies indicates quality. Where controlled vocabularies are not used, the consistent use of meaningful terms in the dc.type element is considered of higher quality than inconsistent and/or non-descriptive terms used to describe dataset objects.

CHAPTER 4

RESULTS

Harvesting metadata from each repository using OHS was generally without incident. Harvests of larger collections took as long as one hour to complete with some failing initially. I found that successful harvest of the largest collections required a stable network using a wired connection. Any other harvest failures were resolved by editing incorrect base URLs or updating obsolete base URLs, which was expected based on Ward's (2004) experiment finding that metadata harvests for approximately 25% of repositories were unsuccessful due to various errors in base URLs.

Implementation of Minimum Dublin Core Metadata Elements

The analysis of the harvested metadata corroborated the conclusions from authors previously mentioned who have conducted similar projects. That is, implementation of unqualified Dublin Core metadata is variable and may preclude the development of a federated index of research data (Shreeves et al., 2003; Jackson et al., 2008). The table below shows the degree of implementation of the five minimum DC elements by repository (table 3).

Repository	Number of Records	Minimum Dublin Core Elements				
		dc.creator	dc.date	dc.identifier	dc.title	dc.type
IQSS DVN	**2913**	775 (26.6%)	706 (24.2%)	2913 (100.0%)	2913 (100.0%)	395 (13.6%)
CISL RDA	**613**	613 (100.0%)	70 (11.4%)	613 (100.0%)	613 (100.0%)	613 (100.0%)
PANGAEA	**16120**	16119 (100.0%)	16120 (100.0%)	16120 (100.0%)	16120 (100.0%)	16120 (100.0%)
NCAR CDP	**7904**	179 (2.3%)	7904 (100.0%)	7904 (100.0%)	7904 (100.0%)	0 (0.0%)
Dryad	**16151**	2042 (12.6%)	15989 (99.0%)	16067 (99.5%)	16136 (99.9%)	13436 (83.2%)
eCrystals	**501**	501 (100.0%)	501 (100.0%)	501 (100.0%)	501 (100.0%)	501 (100.0%)
Share Geo	**164**	47 (28.7%)	164 (100.0%)	164 (100.0%)	164 (100.0%)	161 (98.2%)
Edinburgh DataShare	**18**	4 (22.2%)	18 (100.0%)	18 (100.0%)	18 (100.0%)	18 (100.0%)
VizieR Catalogue	**461**	461 (100.0%)	460 (99.8%)	460 (99.8%)	461 (100.0%)	0 (0.0%)
DataCite	**16508**	16508 (100.0%)	16508 (100.0%)	16508 (100.0%)	16508 (100.0%)	15855 (96.0%)
Odum Institute DVN	**3254**	2974 (91.4%)	3240 (99.6%)	3254 (100.0%)	3254 (100.0%)	2841 (87.3%)
Open Context	**23**	23 (100.0%)	23 (100.0%)	23 (100.0%)	23 (100.0%)	0 (0.0%)
TOTAL	**64630**	**40246 (62.3%)**	**61703 (95.5%)**	**64545 (99.9%)**	**64615 (100.0%)**	**49940 (77.3%)**

Table 3. Repository Implementation of Minimum Dublin Core Elements

The data in the table suggests that the metadata elements required to make connections to objects in their respective repositories are not used consistently among the vast majority of the repositories studied. For each repository, the table displays the number of metadata records that include each of the five DC elements along with the percentage of element use among all repository records. For example, of the 2913 metadata records harvested from IQSS DVN, only 755, or 26.6%, included values present in the dc.creator element field. While IQSS DVN does employ all five minimum DC elements, not every record includes all five. PANGAEA and eCrystals are the only two repositories in which the five minimum DC elements are used for virtually all of their metadata records. The only metadata element that is consistently used in the metadata

records of all repositories studied is the dc.title element. More detailed results of the remaining four elements follow.

Dublin Core Metadata Content Analysis

dc.type

An aggregation of the values present in the dc.type element produced an exhaustive list of 148 unique terms used to describe the record type. Although the repositories were selected for the study based on the assumption that collections were comprised primarily of data sets, the analysis indicates that repositories also include in their collections a variety of other digital object types. Also, rather than describing the object type, some records provided the name of the file format used to render the data or aspects of the data collection method. Only 53% of the records use the DCMI (2010) Type vocabulary term "dataset" to describe the object, while others use more specific or specialized terms to describe the data type.

Because of the large number of unique terms used in the metadata records, each was placed into categories that define the manner in which the dc.type element was interpreted and used by the repository. Though most metadata inputs had straightforward correspondence to one of five different categories, some fell into two different categories. The categories are listed and defined below.

- Data Kind: Describes the data type such as *survey data*, *aggregate data*, *coded data*, and *observational data*.
- Data Collection Mode: Refers to the data collection method. Examples include *interview*, *field study*, and *survey*.
- Time Method: Refers to the time dimension associated with the collected data. *Longitudinal*, *panel*, *cross-sectional* are all terms in this category.

- File Format: Indicates the file format or software application used to render the data. Examples include *SAS*, *SPSS*, *HTML*, and *Excel*.

- "Dataset": This is the DCMI Type Vocabulary term prescribed by DCMI to refer to data. Values that use this term are counted in this category.

Some of the metadata indicated that the repository collection contains materials other than datasets. Also, some values could not be categorized into any of the five data type categories listed above. In those instances, the following categories were assigned:

- Non-Data: Terms referring to digital objects that are not traditionally considered data. These include *thesis*, *conference paper*, *book*, and *journal article*.

- Other: This category was assigned to terms that could not be interpreted or did not provide a meaningful description of the nature of the record. Some examples include *Archive*, *Digital*, and *untilArticleAppears*.

- Null: The type element field was left blank.

The table below (table 4) shows the distribution of the 148 unique dc.type values into the eight categories that reflect how the metadata element was interpreted and used. A complete listing of all dc.type values appears in Appendix 2.

Repository	Number of Unique Values	Data Kind	Data Collection Mode	Time Method	File Format	Non-Data	Other	Null	"Dataset"
IQSS DVN	**83**	101	162	84	8	0	4	2616	4
CISL RDA	**1**	0	0	0	0	0	0	0	613
PANGAEA	**1**	0	0	0	0	0	0	0	16120
NCAR CDP	**N/A**	0	0	0	0	0	0	7904	0
Dryad	**9**	63	0	0	0	544	753	2785	12006
eCrystals	**1**	0	0	0	0	0	501	0	0
Share Geo	**6**	157	0	0	0	0	4	3	0
Edinburgh DataShare	**1**	0	0	0	0	0	0	0	18
VizieR Catalogue	**N/A**	0	0	0	0	0	0	461	0
DataCite	**18**	1	0	0	0	14879	378	684	566
Odum Institute DVN	**36**	540	1	0	2647	0	0	458	12
Open Context	**N/A**	0	0	0	0	0	0	23	0
TOTAL	**148**	**1402**	**164**	**84**	**5302**	**15423**	**1640**	**15392**	**29351**

Table 4. Categories of dc.type Element Usage by Repository

IQSS DVN had the greatest number of unique values in the dc.type element field (83), followed by Odum Institute DVN (36). CISL RDA, PANGAEA, and Edinburgh DataShare consistently applied the term "dataset" for each record in their collections. Though Dryad did use the term "dataset" to refer to the object type, the number of terms assigned to the Non-Data category suggests (whether incorrectly or not) that its collection contains materials other than datasets. The same assumption may be made about DataCite, which applied the most Non-Data terms in the dc.type element field. NCAR CDP, VizieR Catalogue, and Open Context do not employ the dc.type element.

dc.identifier

An examination of the dc.identifier element shows that the majority of repositories have adopted current standards for persistent identification and access. Of

the 12 repositories studied, 7 use either the DOI or HDL standard for persistent identification. Of the remaining 5 repositories, 3 make use of some type of identifier scheme, while the other 2 rely solely on URLs for identification. The table below (table 5) shows examples of dc.identifier values used by each repository.

Repository	Examples of dc.identifier Element Value(s)
IQSS DVN	iqss//**hdl**:1902.1/00001
CISL RDA	ds010.0
PANGAEA	http://doi.pangaea.de/10.1594/PANGAEA.50003 **doi**:10.1594/PANGAEA.50003
NCAR CDP	cgd.cdas https://cdp.ucar.edu/getCatalog.do?ID=cgd.cdas
Dryad	**doi**:10.5061/dryad.100 http://hdl.handle.net/10255/dryad.100
eCrystals	http://ecrystals.chem.soton.ac.uk/1000/
ShareGeo	http://**hdl**.handle.net/10672/10 http://www.sharegeo.ac.uk/bitstream/handle/10672/10/QBSatimagerySYRIA.zip?sequence=1
Edinburgh DataShare	http://**hdl**.handle.net/10283/10 Andrew, Theo; Greig, Morag; Ashworth, Susan. (2008). IRI-Scotland senior management survey [Dataset].
VizieR Catalogue	ivo://CDS/VizieR/I/100A/w10 http://vizier.u-strasbg.fr/cgi-bin/Cat?I/100A/w10
DataCite Metadata	**doi**:10.5284/1000389
Odum Institute DVN	**hdl**:1902.29/030423
Open Context	http://opencontext.org/projects/3/Animal Bone

Table 5. Examples of dc.identifier Element Values

dc.creator

Results of the metadata harvest show that the dc.creator element was used for only 62.3% of all harvested records (see table 6). While half of the repositories studied

used the dc.creator element for all of their records, others used this element for as few as 2.3% (NCAR CDP). In this case, NCAR CDP used the dc.contributor element significantly more often (88.1%) and the dc.publisher element for the vast majority of its records (97.2%). The same tendency is seen in the Edinburgh DataShare repository, in which 22.2% of its records include values in the dc.creator element field, while 77.8% of records use the dc.contributor and dc.publisher fields. With the exception of IQSS DVN, the "data producer" is represented in at least one of the three elements that provide authorship information.

Repository	Number of Records	Dublin Core Elements		
		dc.creator	dc.contributor	dc.publisher
IQSS DVN	**2913**	775 (26.6%)	297 (10.2%)	430 (14.8%)
CISL RDA	**613**	613 (100.0%)	610 (99.5%)	613 (100.0%)
PANGAEA	**16120**	16119 (100.0%)	0 (0.0%)	16120 (100.0%)
NCAR CDP	**7904**	179 (2.3%)	6962 (88.1%)	7686 (97.2%)
Dryad	**16151**	2042 (12.6%)	9177 (56.8%)	5005 (31.0%)
eCrystals	**501**	501 (100.0%)	0 (0.0%)	501 (100.0%)
Share Geo	**164**	47 (28.7%)	164 (100.0%)	0 (0.0%)
Edinburgh DataShare	**18**	4 (22.2%)	14 (77.8%)	14 (77.8%)
VizieR Catalogue	**461**	461 (100.0%)	0 (0.0%)	460 (99.8%)
DataCite	**16508**	16508 (100.0%)	3426 (20.8%)	16508 (100.0%)
Odum Institute DVN	**3254**	2974 (91.4%)	0 (0.0%)	2978 (91.5%)
Open Context	**23**	23 (100.0%)	3 (13.0%)	0 (0.0%)
TOTALS	**64630**	**40246 (62.3%)**	**20653 (32.0%)**	**50315 (77.9%)**

Table 6. Use of dc.creator, dc.contributor, and dc.publisher Metadata Elements by Repository

dc.date

While the majority of repositories include values in the dc.date element field for virtually all of their metadata records, IQSS DVN, and CISL RDA use the dc.date element for 24.2% and 11.4% of their metadata records, respectively (see table 7).

Repository	Number of Records	Dublin Core Element
		dc.date
IQSS DVN	**2913**	706 (24.2%)
CISL RDA	**613**	70 (11.4%)
PANGAEA	**16120**	16120 (100.0%)
NCAR CDP	**7904**	7904 (100.0%)
Dryad	**16151**	15989 (99.0%)
eCrystals	**501**	501 (100.0%)
Share Geo	**164**	164 (100.0%)
Edinburgh DataShare	**18**	18 (100.0%)
VizieR Catalogue	**461**	460 (99.8%)
DataCite	**16508**	16508 (100.0%)
Odum Institute DVN	**3254**	3240 (99.6%)
Open Context	**23**	23 (100.0%)
TOTAL	**64630**	**61703 (95.5%)**

Table 7. Use of dc.date Metadata Element by Repository

CHAPTER 5

DISCUSSION & CONCLUSION

The results of the study support the hypotheses—with caveats. For the first hypothesis—*OAI's harvesting protocol is a viable option for the obtaining the metadata necessary to populate a federated index of data sets that are stored in spatially and disciplinary disparate repositories*—successful harvest of the data repositories with few obstacles demonstrated the ease at which OAI-PMH does fulfill the intended purpose of harvesting exposed metadata from distributed repositories. The key term, however, is "exposed." The limited number of known data repositories that have exposed their metadata via OAI-PMH precludes any possibility of discovering all archived data sets from a single search portal. A comprehensive federated data repository index that represents the body of knowledge available requires that data repositories prescribe to open access initiatives such as OAI-PMH.

The reason for the small number of repositories supporting OAI-PMH is unknown; however, one might surmise that participation in the open access movement (or lack thereof) is a reflection of disciplinary cultures in which data are not afforded value in the same way as journal publication and other traditional modes of scholarly communication. Also, one might ask if repositories lack incentives to make their metadata available for harvesting. Further study would be required to confirm these

assumptions and to determine the constraints that prevent full-fledged adoption of OAI-PMH.

More profound insights were derived from the analysis to support the second hypothesis—*considering the literature describing inconsistencies in the implementation of metadata among data providers, any constraints to a registry employing OAI-PMH would present themselves in an analysis of metadata values*—corroborated the previous literature on the inadequacies of DC metadata implementation to fully support the efficacy of OAI-PMH and its uses. Quoting Ward (2002), DC "is not used to the fullest extent possible." Not only was that the case here, but also the interpretation of the metadata elements varied among repositories. The criticism of OAI-PMH in its use of unqualified DC metadata is warranted, but not because OAI-PMH fails to provide the framework for harvesting metadata records. Rather, it is warranted because of the lack of consistency of the interpretation and content of metadata elements. In the ten-year period since OAI-PMH was introduced, information professionals have made great strides to promote the importance of metadata. Yet, full implementation of DCMES has not yet been (if it will ever be) embraced by data repositories.

A consideration of the individual research communities served by the repository may reveal ways in which the use of DCMES reflects the unique nature of the discipline and how they "do research." Furthermore, the emergence of e-Science has re-introduced an examination of the definition of "data," according to Cole (2008), for which unified notions across and within disciplines are elusive. He states that data "reveals itself as open to multiple and continuing interpretation through its deployment in different contexts, at different levels of abstraction, and in line with various working policies"

(Cole, 2008, p. 247). This was noted in the results of the harvesting project. The number of unique values in the dc.type element was largest among the two social science repositories, both of which employ the Data Documentation Initiative (DDI) metadata specification. DDI proclaims itself as "an effort to create an international standard for describing data from the social, behavioral, and economic sciences" (DDI, 2011). When it comes to describing data, the DDI schema attempts to describe data throughout their lifecycle from conceptualization to archiving. It defines over 341 global elements, with another 231 "complexTypes" components used to further express the nature of the global elements (DDI, 2012). The analysis of the dc.type element reveals the number of ways the element has been interpreted by curators and the even larger number of terms used to characterize the object type, which is likely a reflection of the heterogeneity of data and their manifestations.

The bottom line remains that interoperability across and within disciplines is possible only with the adoption of a single metadata standard, or one that allows for crosswalks between natively generated metadata and the established metadata standard for interoperability. It is evident from the literature and results of the study that OAI-PMH does indeed provide a viable means of federating search and discovery of distributed research datasets. The onus for making this possible, however, is on each repository to comply with OAI-PMH and to implement DC metadata based on widely-accepted standards. Until, then, the utility of OAI-PMH cannot be realized.

Appendix I:

Implementation of the 15 Dublin Core Metadata Elements by Repository

Repository	Dublin Core Elements							
	dc.contributor	dc.coverage	dc.creator	dc.date	dc.description	dc.format	dc.identifier	dc.language
IQSS DVN	297 (10.2%)	524 (18.0%)	775 (26.6%)	706 (24.2%)	2913 (100.0%)	439 (15.1%)	2913 (100.0%)	16 (0.5%)
CISL RDA	610 (99.5%)	21 (3.4%)	613 (100.0%)	70 (11.4%)	613 (100.0%)	384 (62.6%)	613 (100.0%)	613 (100.0%)
PANGAEA	0 (0.0%)	16117 (100.0%)	16119 (100.0%)	16120 (100.0%)	254 (1.6%)	16120 (100.0%)	16120 (100.0%)	16120 (100.0%)
NCAR CDP	6962 (88.1%)	7069 (89.4%)	179 (2.3%)	7904 (100.0%)	7741 (97.9%)	3963 (50.1%)	7904 (100.0%)	0 (0.0%)
Dryad	9177 (56.8%)	12210 (75.6%)	2042 (12.6%)	15989 (99.0%)	12960 (80.2%)	10784 (66.8%)	16067 (99.5%)	48 (0.3%)
eCrystals	0 (0.0%)	0 (0.0%)	501 (100.0%)	501 (100.0%)	0 (0.0%)	0 (0.0%)	501 (100.0%)	0 (0.0%)
Share Geo	164 (100.0%)	164 (100.0%)	47 (28.7%)	164 (100.0%)	164 (100.0%)	131 (79.9%)	164 (100.0%)	164 (100.0%)
Edinburgh DataShare	14 (77.8%)	13 (72.2%)	4 (22.2%)	18 (100.0%)	18 (100.0%)	8 (44.4%)	18 (100.0%)	15 (83.3%)
VizieR Catalogue	0 (0.0%)	0 (0.0%)	461 (100.0%)	460 (99.8%)	0 (0.0%)	0 (0.0%)	460 (99.8%)	460 (99.8%)
DataCite	3426 (20.8%)	27 (0.2%)	16508 (100.0%)	16508 (100.0%)	13012 (78.8%)	15737 (95.3%)	16508 (100.0%)	15858 (96.1%)
Odum Institute DVN	0 (0.0%)	2112 (64.9%)	2974 (91.4%)	3240 (99.6%)	3254 (100.0%)	747 (23.0%)	3254 (100.0%)	15 (0.5%)
Open Context	3 (13.0%)	13 (56.5%)	23 (100.0%)	23 (100.0%)	0 (0.0%)	12 (52.2%)	23 (100.0%)	12 (52.2%)
TOTALS	**20653 (32.0%)**	**38270 (59.2%)**	**40246 (62.3%)**	**61703 (95.5%)**	**40929 (63.3%)**	**48325 (74.8%)**	**64545 (99.9%)**	**33321 (51.6%)**

Repository	Dublin Core Elements						
	dc.publisher	dc.relation	dc.rights	dc.source	dc.subject	dc.title	dc.type
IQSS DVN	430 (14.8%)	479 (16.4%)	338 (11.6%)	304 (10.4%)	511 (17.5%)	2913 (100.0%)	395 (13.6%)
CISL RDA	613 (100.0%)	251 (40.9%)	68 (11.1%)	51 (8.3%)	613 (100.0%)	613 (100.0%)	613 (100.0%)
PANGAEA	16120 (100.0%)	14557 (90.3%)	15342 (95.2%)	1332 (8.3%)	16120 (100.0%)	16120 (100.0%)	16120 (100.0%)
NCAR CDP	7686 (97.2%)	7783 (98.5%)	301 (3.8%)	238 (3.0%)	44 (0.6%)	7904 (100.0%)	0 (0.0%)
Dryad	5005 (31.0%)	2737 (16.9%)	12183 (75.4%)	1627 (10.1%)	13366 (82.8%)	16136 (99.9%)	13436 (83.2%)
eCrystals	501 (100.0%)	500 (99.8%)	0 (0.0%)	1 (0.2%)	501 (100.0%)	501 (100.0%)	501 (100.0%)
Share Geo	0 (0.0%)	10 (6.1%)	164 (100.0%)	77 (47.0%)	164 (100.0%)	164 (100.0%)	161 (98.2%)
Edinburgh DataShare	14 (77.8%)	12 (66.7%)	7 (38.9%)	10 (55.6%)	18 (100.0%)	18 (100.0%)	18 (100.0%)
VizieR Catalogue	460 (99.8%)	460 (99.8%)	460 (99.8%)	26 (5.6%)	0 (0.0%)	461 (100.0%)	0 (0.0%)
DataCite	16508 (100.0%)	2741 (16.6%)	13084 (79.3%)	813 (4.9%)	14865 (90.0%)	16508 (100.0%)	15855 (96.0%)
Odum Institute DVN	2978 (91.5%)	243 (7.5%)	3223 (99.0%)	3211 (98.7%)	3228 (99.2%)	3254 (100.0%)	2841 (87.3%)
Open Context	0 (0.0%)	0 (0.0%)	12 (52.2%)	2 (8.7%)	22 (95.7%)	23 (100.0%)	0 (0.0%)
TOTALS	**50315 (77.9%)**	**29773 (46.1%)**	**45182 (69.9%)**	**7692 (11.9%)**	**49452 (76.5%)**	**64615 (100%)**	**49940 (77.3%)**

Appendix II:

List of Unique dc.type Values

dc.type Value	Number of Records
NULL	14934
Academic Test Score data	1
Administrative	1
administrative data	1
Administrative records data	23
Aerial or Satellite Imagery	26
aggregate data	33
aggregate data, and survey data; administrative records data	1
aggregate macropolitical and economic data	1
Archive	373
Article	552
Book	4
case study, survey	1
case study/oral history	11
case study/oral history, field study	1
case study/oral history, longitudinal	2
census data	3
census/enumeration	1
census/enumeration data	35
Collection of Datasets	4
commercial	1
Conference full text	607
Conference paper	962
Conference presentation	968
Conference proceedings	1
ConferencePaper	1412
Content Analysis Data	1
country-year observations	1
Cross-country data	1

Cross-section, and Cross-Section Time-Series	1
cross-sectional	2
cross-sectional, field study	1
cross-sectional, longitudinal	3
cross-sectional, longitudinal, survey data	1
CSTS, country-year	1
data dictionaries + data definition statements + data files	1
Dataset	28986
Digital	5
Digital Terrain Model	8
Documentation + Data file (Stata, ASCII Codebook)	1
Documentation + data files	1
enter data type here	1
ESRI shapefile	1
Excel	2
Experimental data	1
field experiment	1
field experiment, longitudinal	1
field study	55
field study and institutional records	1
field study, case study/oral history	1
field study, follow-up	1
field study, hereditary, institutional	1
field study, longitudinal	8
field study, longitudinal, replication	1
field study, replication, follow-up	1
Firm, project, client, and contract information	1
follow-up	5
follow-up, field study	1
formulars and derivatives	1
Geographic	1
Geographic reference	6
Geographic reference file	7
GIS vector data	105

Health history and access to health care facilities	1
Html files and I-View files	3
Image	62
institutional records	1
interview	2
interview qx, urinary samples	1
interviews	1
JournalArticle	19
KIND OF DATA HERE	1
lab data	1
laboratory experiment	2
laboratory experiment, longitudinal	2
longitudinal	18
longitudinal, case study/oral history	1
longitudinal, cross-sectional	1
longitudinal, cross-sectional, field study	1
longitudinal, field experiment	2
longitudinal, field study	18
longitudinal, field study, cross-sectional	1
longitudinal, field study, oral history	1
longitudinal, survey	6
Map	1
Metadata document	1
Micro level	1
Minnesota Multiphasic Personality Inventory (MMPI)	1
model results	1
Network Data on Political Science Journal Authors	1
none	143
NonPeerReviewed	501
Numeric	822
Numeric (Aggregate data)	1
Numeric (Aggregate)	4
Numeric (Excel File for Windows 95 Version 7.0 and Comma-delimited text file)	1

Numeric (geographic reference)	1
Numeric (Micro data)	2
numeric (micro)	2
Numeric (Microdata)	9
Numeric (microdata; unit of operation is individuals, families, and households)	1
Numeric (Microdata; units of observation are housing units and persons within housing units)	1
Numeric (SPSS and SAS portable files)	1
Numeric (Summary statistics)	171
Numeric (Summary statistics), 1 User's Manual	1
Numeric (Summary statistics, unit of observation in Files RC77-AT1 through T4 is kind-of-business. In Files RC77-A-TA5 through T8 the unit of observation is individual geographic areas)	1
Numeric (Survey and existing academic records)	1
Numeric (Survey data)	151
Numeric (survey)	1391
Numeric (Survey) data	10
Numeric (Survey) data. SPSS portable file.	51
Numeric data	14
Numeric statistics	1
Observational macro-political/economic	1
oneyear	44
Other	4
Panel and Longitudinal Sampling	1
protocol	7
questionnaire	3
questionnaire, tests	1
Rectangular	1
Report	10886
Research Data	1
Scanned map	17
Simulated data for 5000 two-dimensional policy spaces	1
Social Science Data File	12
source code	1
Stata	1

STATA and SPSS	1
Stata Dataset	1
STATA, SPSS, EXCEL	1
Summary statistics	14
Summary statistics in PDF format, with hyperlinks to Lotus and Excel worksheets.	6
Summary statistics in PDF, Microsoft Excel (.xls or .xlw), and Lotus (.wk1 or .wk4) format	1
Supplementary Collection of Datasets	194
Supplementary Dataset	139
survey	31
survey and clinical data	1
survey data	71
Survey data (summary statistics)	3
survey data, and administrative records data	1
survey, longitudinal	2
survey, replication	1
Tabular data	1
Thesis	11
time series cross sectional macroeconomic/political data	1
Time-Series Cross-Section	1
untilArticleAppears	559
Village and GP Pradhan information	1
TOTAL	**64630**

Appendix 3:

List of dc.type Values by Repository

IQSS DVN	
dc.type Value	**Number of Records**
[NULL]	2616
Academic Test Score data	1
Administrative	1
aggregate data	17
aggregate data, and survey data; administrative records data	1
aggregate macropolitical and economic data	1
case study, survey	1
case study/oral history	11
case study/oral history, field study	1
case study/oral history, longitudinal	2
census data	3
census/enumeration	1
census/enumeration data	35
commercial	1
Content Analysis Data	1
country-year observations	1
Cross-country data	1
Cross-section, and Cross-Section Time-Series	1
cross-sectional	2
cross-sectional, field study	1
cross-sectional, longitudinal	3
cross-sectional, longitudinal, survey data	1
CSTS, country-year	1
data dictionaries + data definition statements + data files	1
Documentation + Data file (Stata, ASCII Codebook)	1
Documentation + data files	1
enter data type here	1
ESRI shapefile	1

Excel	2
Experimental data	1
field experiment	1
field experiment, longitudinal	1
field study	55
field study and institutional records	1
field study, case study/oral history	1
field study, follow-up	1
field study, hereditary, institutional	1
field study, longitudinal	8
field study, longitudinal, replication	1
field study, replication, follow-up	1
Firm, project, client, and contract information	1
follow-up	5
follow-up, field study	1
formulars and derivatives	1
Health history and access to health care facilities	1
institutional records	1
interview	2
interviews	1
KIND OF DATA HERE	1
lab data	1
laboratory experiment	2
laboratory experiment, longitudinal	2
longitudinal	18
longitudinal, case study/oral history	1
longitudinal, cross-sectional	1
longitudinal, cross-sectional, field study	1
longitudinal, field experiment	2
longitudinal, field study	18
longitudinal, field study, cross-sectional	1
longitudinal, field study, oral history	1
longitudinal, survey	6
Micro level	1

Minnesota Multiphasic Personality Inventory (MMPI)	1
model results	1
Network Data on Political Science Journal Authors	1
Observational macro-political/economic	1
Panel and Longitudinal Sampling	1
questionnaire	3
questionnaire, tests	1
Research Data	1
Simulated data for 5000 two-dimensional policy spaces	1
Stata	1
STATA and SPSS	1
Stata Dataset	1
STATA, SPSS, EXCEL	1
survey	31
survey and clinical data	1
survey data	7
survey data, and administrative records data	1
survey, longitudinal	2
survey, replication	1
time series cross sectional macroeconomic/political data	1
Time-Series Cross-Section	1
Village and GP Pradhan information	1

CISL RDA	
dc.type Value	**Number of Records**
Dataset	613

PANGAEA	
dc.type Value	**Number of Records**
Dataset	16120

NCAR CDP	
dc.type Value	**Number of Records**
[NULL]	7904

DRYAD	
dc.type Value	**Number of Records**
[NULL]	2785
Article	542
Book	2
dataset	12006
Image	62
Map	1
none	143
oneyear	44
protocol	7
untilArticleAppears	559

eCrystals	
dc.type Value	**Number of Records**
NonPeerReviewed	501

ShareGeo	
dc.type Value	**Number of Records**
[NULL]	3
Aerial or Satellite Imagery	26
Digital Terrain Model	8
GIS vector data	105
Other	4
Scanned map	17
Tabular data	1
Edinburgh DataShare	
Dataset	18

VizieR Catalog	
dc.type Value	**Number of Records**
[NULL]	461

DataCite Metadata	
dc.type Value	**Number of Records**
[NULL]	684
Archive	373
Article	10
Book	2
Collection of Datasets	4
Conference full text	607
Conference paper	962
Conference presentation	968
Conference proceedings	1
ConferencePaper	1412
Dataset	229
Digital	5
JournalArticle	19
Metadata document	1
Report	10886
source code	1
Supplementary Collection of Datasets	194
Supplementary Dataset	139
Thesis	11

Odum Institute DVN	
dc.type Value	**Number of Records**
[NULL]	458
administrative data	1
Administrative records data	23
Aggregate data	16
Geographic	1
Geographic reference	6
Geographic reference file	7
Html files and I-View files	3
interview qx, urinary samples	1
Numeric	822

Numeric (Aggregate data)	1
Numeric (Aggregate)	4
Numeric (Excel File for Windows 95 Version 7.0 and Comma-delimited text file)	1
Numeric (geographic reference)	1
Numeric (Micro data)	2
numeric (micro)	2
Numeric (Microdata)	9
Numeric (microdata; unit of operation is individuals, families, and households)	1
Numeric (Microdata; units of observation are housing units and persons within housing units)	1
Numeric (SPSS and SAS portable files)	1
Numeric (Summary statistics)	171
Numeric (Summary statistics), 1 User's Manual	1
Numeric (Summary statistics, unit of observation in Files RC77-AT1 through T4 is kind-of-business. In Files RC77-A-TA5 through T8 the unit of observation is individual geographic areas)	1
Numeric (Survey and existing academic records)	1
Numeric (Survey data)	151
Numeric (Survey)	1391
Numeric (Survey) data	10
Numeric (Survey) data. SPSS portable file.	51
Numeric data	14
Numeric statistics	1
Rectangular	1
Social Science Data File	12
Summary statistics	14
Summary statistics in PDF format, with hyperlinks to Lotus and Excel worksheets.	6
Summary statistics in PDF, Microsoft Excel (.xls or .xlw), and Lotus (.wk1 or .wk4) format	1
Survey data	64
Survey data (summary statistics)	3

Open Context	
dc.type Value	**Number of Records**
[NULL]	23

REFERENCES

Altman, M., Adams, M., Crabree, J., Donakowski, D., Maynard, M. Pienta, A., & Young, C. (2009). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist, 72*, 170-184.

Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine, 13*(3/4). doi:10.1045/march2007-altman.

Arms, W., Dushay, N., Fulker, D., Lagoze, C. (2003). A case study in metadata harvesting: The NDSL. *Library Hi Tech, 21*(2), 228-237. Retrieved from http://www.emeraldinsight.com/journals.htm?articleid=861371&show=abstract

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

Beisler, A. & Willis, G. (2009). Beyond theory: Preparing Dublin Core metadata for OAI-PMH harvesting. *Journal of Library Metadata*, *9*, 65-97.

Bell, G., Hey, T. & Szalay, A. (2009). Beyond the data deluge. *Science, 323*(5919), 1297-1298.

Berman, F. (2008). Got data? A guide to data preservation in the information age. *Communications of the ACM, 51*(12), 50-56.

Bobley, B. (2011). Announcing a new grant program: Digital Humanities Implementation Grants. Retrieved from

http://www.neh.gov/ODH/ODHUpdate/tabid/108/EntryId/162/Announcing-a-New-Grant-Program-Digital-Humanities-Implementation-Grants.aspx.

Brase, J. (2004). Using digital library techniques—Registration of scientific primary data. *Lecture Notes in Computer Science, 3232*(2004), 488-494. doi: 10.1007/978-3-540-30230-8_44.

Centers for Disease Control and Prevention. (2005). Releasing and sharing data. Retrieved from http://www.cdc.gov/od/foia/policies/sharing.htm

Cole, F. (2008). Taking "data" (as a topic): The working policies of indifference, purification and differentiation. *19th Australasian Conference on Information Systems*. Christchurch, New Zealand.

Culliton, B. (1998). Authorship, data ownership examined. *Science, 242*, 658.

Data Documentation Initiative (DDI). (2011). What is DDI? Retrieved from http://www.ddialliance.org/what/

Data Documentation Initiative (DDI). (2012). DDI specification. Retrieved from http://www.ddialliance.org/Specification/

Department of Defense. (1998). *DoD Scientific and Technical Information (STI) Program (STIP)* (DoD Directive No. 3200.12). Washington, DC: U.S. Government Printing Office. Retrieved from http://www.dtic.mil/dtic/pdf/customer/STINFOdata/DoDD_320012.pdf

Dozier, J. & Gail, W. (2009). The emerging science of environmental applications. In Hey, T., Tansley, S. & Tolle, K. (Eds.), *Fourth Paradigm* (13-20). Redmond, WA: Microsoft Research.

Dublin Core Metadata Initiative (DCMI). (n.d.). Metadata basics. Retrieved from http://dublincore.org/metadata-basics/

Duke University. (2007). *Policy on research records: Sharing, retention and ownership*. Durham, NC: Duke University. Retrieved from http://www.provost.duke.edu/pdfs/fhb/FHB_App_P.pdf

Fishbein, E. (1991). Ownership of research data. *Academic Medicine*, *66*(3), 129-133.

Hanson, B., Sugden, A., & Alterts, B. (2011). Making data maximally available. *Science, 331*(6018), 649.

Haslhofer, B. & Schandl, B. (2008). The OAI2LOD server: Exposing OAI-PMH metadata as linked data. *International Workshop on Linked Data on the Web (LDOWS 2008)*. Beijing, China. Retrieved from http://www.mendeley.com/research/oai2lod-server- exposing-oaipmh-metadata-linked-data/#page-1

Hey, T. & Trefethen, A. (2002). The UK e-Science core programme and the grid. *Future Generation Computer Systems, 18*(2002), 1017-1031. Retrieved from http://eprints.soton.ac.uk/257644/1/UKeScienceCoreProg.pdf.

Hey, T. & Trefethen, A. (2003). The data deluge: An e-science perspective. In Berman, F., Fox, G., & Hey, A. (Eds.), *Grid Computing – Making the global Infrastructure a Reality* (809-824). UK: Wiley and Sons.

Hunt, J. Baldocci, D., & van Ingen, C. (2009). Redefining ecological science using data. In Hey, T., Tansley, S. & Tolle, K. (Eds.), *Fourth Paradigm* (21-26). Redmond, WA: Microsoft Research.

Jackson, A., Han, M., Groetsch, K, Mustafoff, M., & Cole, T. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata*, *8*(1), 5-21.

Jerez, H., Liu, X., Hochstenbach, P, Van de Sompel, H. (2004). The multi-faceted use of the OAI-PMH in the LANL repository. *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM.

Johns Hopkins University. (2008). Policy on access and retention of research data and materials. Retrieved from http://jhuresearch.jhu.edu/Data_Management_Policy.pdf

Lagoze, C. & Van de Sompel, H. (2001). The open archives initiative: Building a low-barrier interoperability framework. *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '01).* New York: ACM.

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M., Knudson, F., & Holtkamp, I. (2002). Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information*, *2*(4). Retrieved from http://journals.tdl.org/jodi/article/viewArticle/55/58

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. *Proceedings of the 3rd UK e-Science All Hands Meeting*, *August 31-September 3, 2004*. Nottingham, UK.

National Aeronautics and Space Administration. (2011, June 23). Data & information policy. Retrieved from http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/

National Endowment for the Humanities. (2011, June 22). Data management plans for NEH Office of Digital Humanities proposals and awards. Retrieved from

http://www.neh.gov/ODH/LinkClick.aspx?fileticket=jQ44xoe2ZjU%3D&tabid=108

National Institutes of Health (2003). *Final NIH statement on sharing research data* (NOT-OD-03-032). Bethesda, MD: NIH. Retrieved from http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html

National Research Council. (2002). *Preparing for the Revolution: Information Technology and the Future of the Research University.* Washington, DC: The National Academies Press.

National Research Council. (2003). *Sharing publication-related data and materials: Responsibilities of authorship in the life sciences.* Washington, DC: The National Academies Press.

National Research Council. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age.* Washington, DC: The National Academies Press.

National Science Foundation. (2010, May 10). Scientists Seeking NSF funding will soon be required to submit data management plans (Press release 10-077). Retrieved from http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation. (2011, January). Grant proposal guide. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

Nelson, B. (2009). Data sharing: Empty archives. *Nature*, *461*, 160-163.

Open Archives Initiative. (Accessed 2011, December 11). Registered Data Providers. Retrieved from http://www.openarchives.org/Register/BrowseSites

Organisation for Economic Co-operation and Development (OECD). (2007). OECD Principles and guidelines for access to research data from public funding. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Paskin, N. (1999). Toward unique identifiers. *Proceedings of the IEEE, 87*(7), 1208-1227.

Paskin, N. (2005). Digital object identifiers for scientific data. *Data Science Journal, 4*, 12-20.

Piwowar, H., Day, R., & Fridsma, D. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, *2*(3). Retrieved from http://dl.acm.org/citation.cfm?id=273035.273043

Public Knowledge Project. (Accessed 2011, September 2). Open Harvester Systems. Retrieved from http://pkp.sfu.ca/?q=harvester

Public Library of Science (PLoS) (2011). PLoS editorial and publishing policies. Retrieved from http://www.plosone.org/static/policies.action#sharing

Rajasekar, A. & Moore, R. (2001). Data and metadata collections for scientific applications. *Lecture Notes in Computer Science, 2110*(2001), 72-80. doi: 10.1007/3-540-48228-8_8.

Rusbridge, C., Burnhill, P. Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005). The Digital Curation Centre: A vision for digital curation. *Proceedings from Local to Global: Data Interoperability—Challenges and Technologies*. Sardinia, Italy. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

SHERPA. (Accessed October 24, 2011). Research funders' open access policies. http://www.sherpa.ac.uk/juliet/index.php.

Shreeves, S., Kaczmarek, J., & Cole, T. (2003). Harvesting cultural heritage metadata using the OAI Protocol. *Library Hi Tech*, *21*(2), 159-169.

Stanford University. (1997). Research policy handbook: Retention of and access to research data (RPH 2.10). Retrieved from http://rph.stanford.edu/2-10.html

University of Illinois. (Updated 2011, December 9).OAI-PMH Data Provider Registry. Retrieved from http://gita.grainger.uiuc.edu/registry

University of Kentucky. (2011). Data retention and ownership policy. Retrieved from http://www.rgs.uky.edu/ori/data.htm

University of Nottingham. (Updated 2010, July 11). The Directory of Open Access Repositories – *Open*DOAR. Retrieved from http://www.opendoar.org/

University of Pittsburgh. (2009, November 25). University of Pittsburgh guidelines on research data management. Retrieved from http://www.provost.pitt.edu/documents/RDM_Guidelines.pdf

Van de Sompel, H., Nelson, M., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-Lib Magazine 10*(12). Retrieved from http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html#Lagoze-etal-2002-1.

Ward, J. (2004). Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems & Services, 20*(1), 40-47.

Weber, N., Piwowar, H., & Vision, T. (2010). Evaluating data citation and sharing policies in the environmental sciences. *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*. Pittsburgh, PA.

Whitlock, M., McPeek, M, Rausher, M., Riesenberg, L, & Moore, A. (2010). *The American Naturalist, 175*(2), 145-146.

Yiotis, K. (2005). The Open Access Initiative: A new paradigm for scholarly communications. *Information Technology & Libraries*, *24*(4), 157-162.