Title
Researchers' Attitudes Towards Data Discovery: Implications for a UCLA Data Registry

Author
Mandell, Rachel Alyson

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Researchers' Attitudes Towards Data Discovery:

Implications for a UCLA Data Registry

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Library and Information Science

in Information Studies

by

Rachel Alyson Mandell

2012

ABSTRACT OF THE THESIS

A New Tool for Managing and Discovering Data:

Creating the UCLA Data Registry

by

Rachel Alyson Mandell

Master of Library and Information Science in Information Studies

University of California, Los Angeles, 2012

Professor Christine L. Borgman, Chair

Research output is becoming increasingly digital. In the sciences research output now takes the form of large and small datasets, three-dimensional images and sensor readings. In the social sciences research output includes GIS data, quantitative survey and demographic data and also qualitative ethnographic data and interview transcripts. And in the humanities, scholars now use tools such as three-dimensional maps, social networks, and text analysis that allow them to ask traditional humanist questions in completely new ways. To harness the potential of the data and digital research output being produced in all fields, information professionals and scholars need to make research data discoverable and accessible to other scholars and students. This thesis focuses on understanding the various definitions that scholars use to characterize their data

and research output, as well as the methods and tools they use and need to disseminate, manage, and make their work discoverable.

The UCLA Data Registry is a tool designed to serve the greater UCLA research community by collecting and making available surrogate records of research datasets. To figure out how to build this system in accordance with the needs of the community, a total of 20 researchers from disparate disciplines were interviewed about their data and metadata practices. The results indicate that researchers' attitudes and behaviors towards making their work discoverable depend on their concept and definition of data. Given that the UCLA Library will build the UCLA Data Registry, it is important to consider the other possible tools that researchers could use in conjunction with the registry to enhance the discoverability of their data. The Data Registry will be built utilizing a basic metadata schema rather than very specific descriptive fields. The interviews also demonstrated that the culture of publishing and venues for data dissemination are shifting away from the traditional journal article publication, especially in emerging areas such as the digital humanities. As information professionals, we must continue to develop new tools and methods for managing and maintaining access to these news types of scholarship. The UCLA Data Registry is one step towards providing the support and venues for making visible and accessible the diversity of research being conducted by UCLA researchers.

The thesis of Rachel Mandell is approved.

Jonathan Furner

Christopher Kelty

Christine L. Borgman, Committee Chair

University of California, Los Angeles

2012

Table of Contents

Acknowledgements

First and foremost, I would like to thank Christine Borgman, my advisor and mentor, who has helped me to get the most out of my graduate school experience and prepared me for my future as an information professional. I am also so thankful for the rest of my amazing thesis committee, Christopher Kelty and Jonathan Furner, you have both taught me so much about research and scholarship. Our numerous conversations throughout this process have really helped to shape my ideas and encouraged me to critically engage with my work. I also want to thank Todd Grappone for introducing me to the academic library world. I truly appreciate all of the time you have spent teaching me so many valuable skills such as wire framing, collecting requirements, and navigating the best path for a new library tool. I am so grateful to be able to carry these experiences with me as I continue my career.

I must also give a very hearty thanks to all of my willing and brilliant interviewees. I have learned so much from all of you! This was my first experience working with you and learning about your research practices and needs. I hope I continue to build on these experiences and skills and continue bringing researchers into the conversation when building new academic library tools.

I also want to thank the Data Practices team at UCLA for accepting me into your group and providing both amazing support and constructive criticism along the way—Jillian Wallis, Laura Wynholds, Ashley Sands, and especially Lizzy Rolando, one of my best friends and biggest supporters! The Data Girls rock! Last but certainly not least, I need to thank my amazing family for always being there for me. And more importantly, listening to me talk about data and data management… constantly. And of course, to my best friend and partner, Eric Hounshell—you inspire me every day to try my best and encourage me to always believe in myself.

Introduction

Advanced technologies have enhanced the ability to generate vast amounts of data and many forms of digital research output. As a result, the methods and practices of scholarship in many fields have undergone profound changes. Though these changes may be most apparent in the sciences, the scholarship practices of other academic disciplines have also been affected by the proliferation of digital technology. Scholars in the digital humanities are now focusing on the use of data and other digital materials to engage with new methods of constructing scholarly arguments and disseminating research. This focus on data as its own intellectual entity across academic fields, especially in areas that traditionally did not consider their work to involve 'data,' demonstrates the need for information professionals to develop new ways to harness the data deluge.

Data management is therefore necessary, and yet remains a complex problem. There are many different kinds of tools aimed at supporting scholarly output. Certain disciplines, especially in the sciences and social sciences, have well-established publishing standards as well as the requisite tools such as data repositories, which are willing to take on new data and provide the necessary stewardship. However, many fields especially in the humanities still lack the support necessary for managing the data they are producing. In light of this need, there are other initiatives towards assisting those interested in managing their data have developed. Data registries acknowledge that data may live in multiple silos and instead physically storing the data, they bring together surrogate records of the data. Data registries along with other tools such as institutional repositories and researcher identification systems are all examples of a possible solution to making data discoverable. Yet diverging definitions of data across academic fields make it difficult to create a single system in support of all research. Therefore, information

professionals must engage with the wide range of research practices and consider the actual needs of the community.

Background of UCLA Data Registry Project

In July 2011, Professor of Information Studies, Christine Borgman and Todd Grappone of the UCLA Library were awarded a grant from the Institute for Digital Research and Education (IDRE) to carry out the UCLA Data Registry project. The concept for this endeavor was first envisioned in 2009, as the Center for Embedded Network Sensing (CENS) established a pilot effort to register research data to make them more publicly accessible. With CENS's official closing date set as July 31, 2012, barely a year was left to capture the legacy of data and publications from UCLA's first NSF-funded Science and Technology Center. The UCLA library was therefore in a unique position to migrate the small data registry that CENS has created, but also to use CENS's content and experiences as the basis for a sustainable data registry that could be expanded to include the greater UCLA research community (Borgman et.al., 2011).

Once the registry is developed into a working prototype, user testing will occur to determine if any changes need to be implemented before the tool is rolled out as one of the suite of services offered by the UCLA Library. The library will maintain the registry and continue to develop future phases as the tool is integrated into the UCLA research community and the larger infrastructure in support of scholarship.

Problem Statement and Goal of this Research

Given the amount of data and digital research output being produced across all academic disciplines, the first goal of this thesis is to understand the current research practices and data

needs of UCLA researchers and then to determine what tools they require to manage, curate and disseminate their work. If researchers are unable to discover each other's work, the chance for data sharing and reuse between researchers diminishes. However, discoverability of the data may mean different things to different kinds of researchers. Therefore, discussing data and research practices with scholars from a wide range of academic disciplines will help to determine what researchers consider their data to be as well as attitudes and behaviors that researchers have towards the discoverability of this data. Based on the results of these discussions, the project will also determine the implications for the design of the imminent UCLA Data Registry, as well as other kinds of tools that can support research.

Review of the Literature

Diverging Research Practices and Definitions of Data

Though many forms of data and research output are increasingly digital, all bits are not created equal and all researchers do not share a common definition of data. What a researcher considers her data to be depends on the research practices in her field. Furthermore, a distinction between data and research output must be drawn during this discussion. Across academic fields, different outputs of research are commonly created and disseminated. In the sciences and social sciences, the most common form of research output has been the journal article publication. In the traditional humanities, research output has been the monograph, but in the digital humanities, the output is taking new forms such as three-dimensional reconstructions and models, as well as social network graphs.

In the sciences and social sciences, access services have shifted from only considering journal publications to also including data, recognizing that they may be a valuable research

output (Arzberger, et. al., 2004). This relatively new focus on data as its own intellectual entity has not occurred in the humanities to the same extent. Although the output of research is also shifting in the digital humanities, the shift is not necessarily towards the focus on data in its own right. To understand the relationship between research practices and the tool required for managing data, it is necessary to understand how scholars in each of the three general areas of research (science, social science and humanities) define, disseminate, and use data in their research.

In the sciences, large and small datasets are generated by computers or collected as results from an experiment. Scientific data can also include images such as brain scans and X-rays (Borgman, 2007). According to Jim Gray, the advance of digital technology has caused a scientific data deluge, which has shifted our relationship with the ways we produce knowledge. He argues that in some scientific fields, a new age of scientific production and discovery is upon us. The 4th Paradigm, or data-intensive science affords researchers a new set of methods, beyond empiricism, theory, and simulation (Hey, Tansley & Tolle, 2009). Chris Anderson, Editor-in-Chief of Wired echoes this notion, arguing that given our ability to generate such vast amounts of data, we have moved beyond theory, such that the "data speak for themselves" (Anderson, 2008).

This extreme claim that theory no longer governs knowledge production is no doubt contentious. Many scholars do agree that both the amount of data produced and cyberinfrastructure, or "the distributed computer, information, and communication technology" required for a "knowledge economy," (Atkins et al., 2003, p. 5), have indeed altered the methods of scientific research and have created new opportunities for collaboration and sharing data. However, to argue that the 'data speak for themselves,' has broad sweeping implications

regarding our relationship to information and world around us—access to raw data does not mean one has direct access to raw knowledge (Boyd & Crawford, 2011). Yet sharing raw data may open up new opportunities for combining and reusing data in ways that can advance science, and provide a "far greater return on investment in research" (Buckland, 2011).

Determining what constitutes data may be more difficult in the social sciences than in the sciences due to the many possible data sources, including data collected by researchers through experiments, as well as data collected by private agencies or the government (Borgman, 2007). Data collected by researchers in the social sciences include qualitative interview or ethnographic data as well as quantitative survey data. One defining quality of the kinds of data being produced in social science is the issue of privacy, as much of the information collected involves personal and identifiable information. Therefore, an important step in the social science data practices is the data-cleaning phase, which includes making data anonymous. Like the sciences, social scientists have faced their own data deluge (Borgman, 2007). However, an important difference between the sciences and the social sciences is that a good portion of the data used by social science researchers are not generated by the researchers themselves, which may create a more open environment for sharing and depositing into shared repositories.

The emerging discipline known as the digital humanities not only utilizes increasingly more digitized source material, but scholars are also exploring new methods and ways of using technology to ask fundamentally humanist questions. Defined by their gradual break from the traditional methods of conducting humanities research, the digital humanities was born out of a literary community, who turned to "statistical analysis of a text's linguistic features, for example, or author-attribution studies or studies that rely on data mining"(Fitzpatrick, 2011, n.p.). However, this area has expanded to include "scholars in history, musicology, performance

studies, media studies, and other fields that can benefit from bringing computing technologies to bear on traditional humanities materials" (Fitzpatrick, 2011, n.p.). The task of defining data is very difficult in the humanities, as humanists are able to use publications, archival records, physical artifacts and other documents as sources of data (Borgman, 2007). Further, with the rapid introduction of digitized material, the notion between digital and data is often conflated and this complicates the issue of defining data even more.

This shift towards the use of digital technology in traditionally non data-driven fields is not entirely new. Tara McPherson argues that a gradual transition occurred as humanists adopted more digital technology into their practices (McPherson, 2008). However, the digital humanities are still currently at an interesting moment; they are experiencing a move towards becoming more of an established and recognized discipline. Scholars working within this framework are thus engaging in ongoing debates and struggles to define what it means to be a digital humanist. Fitzpatrick notes that it is not the digital that renders the digital humanities something new, but rather that scholars are asking traditional humanities questions using contemporary computational methods.

The methods used in this research lie at the heart of the debates about how to define the digital humanities. Many scholars concern themselves with "making their methodology accessible to a broader humanities audience" (Gibbs and Owens, 2011). There is now a focus on documenting and understanding the processes and methods used in the digital humanities. The goal of certain projects is create a reusable method that can be repurposed for different projects and questions.

Another cause of tension between scholars is the production and use of tools. According to Fitzpatrick, the tensions center around whether the field should be focused on producing and

creating tools for other scholars to use such as the Text Encoding Initiative and the William Blake Archive, or should scholars instead continue to work on interpreting and engaging with more theoretical endeavors, once essential to the humanities. Fitzpatrick likens this schism to the age-old 'theory vs. practice' divide, which has already existed within the humanities before the digital was even introduced. However, Fitzpatrick (2010) argues that in recognizing that "boundaries between the critical and the creative are arbitrary," these tensions and debates are helping to not just define but also to expand, evolve and develop the digital humanities (n.p.). She reiterates a point previously made by Neil Fraistat, director of the Maryland Institute for Technology in the Humanities stating, "these debates can be most productive if we understand them as a means of opening ourselves to the kinds of conversations that true interdisciplinarity can support" (2011, n.p.).

Across all fields, the question of what conditions and incentives for researchers to make their research outputs available to others is still unresolved (Borgman, 2012). Both scientists and social scientists self-archive papers on websites and repositories as well as contribute to disciplinary and institutional repositories (Borgman, 2007). "Scientific publication practices, which are oriented toward journal articles and conference papers rather than books, have reflected a steady shift from paper to electronic forms" (Borgman, 2007, p. 181). In the humanities, the monograph is now widely available in digital form and the digital humanities have taken leave from this "gold standard," instead gravitating towards new methods of disseminating data such as producing three-dimensional models and social network graphs.

Questions regarding data ownership also affect researchers' willingness to share data. More than a decade ago, Phil Agre argued that data ownership was an important and often absent element of data are concerns. Though rules and understandings of where data came from and

what others are allowed to do with them can be embedded in data and software, once data begins to migrate, and get merged with other data, the lines between disparate data sources are blurred, and the rules no longer govern as they had before (1994). This uncertainty of what can happen to data once released, fear of data being misused, or not receiving attribution for shared data may affect researchers' willingness to release data.

Funding also weighs heavily in the ongoing conversation about data ownership and willingness to share. Many argue that "open access to research data from public funding should be easy, timely, user friendly and preferably Internet-based" (OECD, 2007). Certain funding agencies that support research require open sharing of the data produced during research (National Academy of Sciences, 2009). Of course these types of mandates rest on assumptions about the usability and understandability of research data, and questions about what types of research receive public funding. And ensuring research data are easily accessible and able to be used widely is a "matter of sound stewardship and public resources" (Arzberger et. al., 2004). Thus, it is likely that unfunded research, or privately funded research may result in a situation where a researcher is not willing to make their data accessible to others. Researchers are also often only willing to share their data once they have had the opportunity to publish on it first, from fear of having their work stolen by those competing for funding, jobs, and prestige (Hilgartner & Brandt-Rauf, 1994). This practice of releasing data only after publication has become common practice, but must not be forgotten in developing tools for data management.

Data Discovery Tools

Different kinds of tools have been developed in the service of scholarship and aimed at making digital data and other forms of research output discoverable by other researchers. Though

instantiations of the many different tools that can assist researchers are custom built and deployed on individual research university campuses, the greater research community defines the general requirements and functions of these systems. The design of these scholarly tools are often best informed by communicating with researchers to figure out how a specific tool might support researchers and their research output. In many cases, the most effective tools are not stand-alone services, but instead work in conjunction with other tools to function as a suite of services that collectively form an information infrastructure, which forms a "value chain of scholarship" (Borgman, 2007). Links between journal publications and the datasets on which the article is based, allow an interested user to discover either the dataset or the publication first, as the link provides the user with easy access to other components in the chain. Therefore, one is able to enter the chain at any point and continue to "follow the relationships" (Borgman, 2009). Since each tool provides different services and functions, they become more effective when they are connected to one another. The following tools are all in the service of discovering digital data and research output, yet each performs a distinct function.

Journal Publications and Institutional Repositories

Based largely on the academic reward system, which still places heavy emphasis on the journal publication, scholars have strong incentives to publish their work in this form, which of course has contributed to the notion that publications are the most valuable kind of research output (Borgman, 2007). Depending on the field of the publication, there are different community standards surrounding the inclusion or exclusion of the data on which a publication is based. Therefore, discovering raw data from a journal publication is less likely. However, summarized versions of the data or visualizations of the data are often included. In the sciences

and social sciences, these journal articles have made a relatively easy transition from paper-based over to electronic forms (Borgman, 2007). The humanities, whose standard of publication was the monograph, for the most part, have also gradually shifted over to electronic copies of published work. In addition to housing the publications produced by its own faculty, the university research library attempts to provide access to as many of these e-resources and serials to its patrons. However, more recently, many universities have created institutional repositories, which represent a more focused effort towards accumulating, preserving and providing access to the intellectual assets produced by its own faculty and affiliated researchers.

According to Clifford Lynch (2003) of the Coalition for Networked Information, a university-based institutional repository is a "set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" (n.p.). Besides electronic journal publications or digitized articles, some institutional repositories have also attempted to embrace the stewardship of research data. Essential to this definition is the organizational commitment to the stewardship of the material, including the long-term preservation and maintained access. The Association of Research Libraries adds, "Repository services are built upon a foundation of content, context and access" (ARL, 2009). Each of these three components is vital, and each presents its own set of problems.

Many institutional repositories have struggled to recruit content because as Sayeed Choudhury points out, "technology alone cannot engender transformation" (Choudhury, 2008, p. 212). By this he means that in addition to the technical ability to provide repository services, the users must also be interested in using the repository and contributing to it, otherwise the service does not serve its intended purpose. Some of the problems that Choudhury identified with

current repository efforts include figuring out how to motivate researchers to contribute to the repository, developing the repository in such a way as to support the varying needs of documents and data simultaneously, and also allow for multi-institutional projects (Choudhury, 2008). The Association of Research Libraries acknowledged that on one hand, many repositories did struggle to acquire content, but on the other, repositories which have been more successful in acquiring content from digitization have tended to struggle with "context issues such as metadata creation" (ARL, 2009).

In recognizing that the repository issues may be as diverse as the research materials that they are trying to care for, an alternative approach to preserving material for the long-term and providing access to future users is the micro-services approach that the California Digital Library (CDL) has adopted. CDL's ideology stems from the recognition that the architecture of a centralized repository, "often leads to large, cumbersome systems that are expensive to deploy and support (Abrams, Cruse & Kunze, 2009). Instead, they argue that preservation is not a place in which content is contributed, but rather a process (Abrams, Cruse & Kunze, 2009). This approach employs a set of ten discrete micro-services, with each service performing a single function.

Data Repositories

While institutional repositories are aimed at storing, preserving and maintaining access to a wide range of digital research output including journal publications, electronic books, and in some cases even datasets, certain disciplines have domain-specific data repositories dedicated to caring for the datasets produced in those fields. In the biological sciences, the GenBank or the Protein Data Bank or the Inter-University Consortium for Political and Social Research (ICPSR)

for social science data offers researchers a stable environment for the data stored there.

Data repositories like the GenBank take on the responsibility of providing long-term care and stewardship of the data. According to the National Science Board's Long-Lived Data Report, a long-lived digital data collection is something that satisfies four criteria: 1) that the collection is a collection of data—meaning anything in digital form. 2) In addition to stored data, the collection is comprised of infrastructure, organizations and the individuals necessary to preserve access to the data. 3) These collections can be accessed electronically, via the Internet and 4) That the collection is long-lived (2005). The concept of long-lived data actually refers to the definition provided by the OAIS model, which holds that a collected is saved for the long term if it must be concerned with the impacts of changing technologies and even changing user communities (OAIS, 2002). Data deposited in a repository that fits these descriptions are maintained and cared for into the future. While data repositories may be one of the premier tools for actually discovering data, they are usually only geared towards a certain kind of data, or data produced by a single discipline. Repositories can exist on campuses or elsewhere, if they are funded by government agencies or other organizations.

Data Registries

Data Registries are just starting to become internationally recognized as another possible solution for making data discoverable. Humanities fields lack the support necessary for managing the data they are producing. In light of this need, there are other initiatives towards assisting those interested in managing their data. Data registries maintain a record of the data and a description of the datasets that a researcher is willing to make available. A successful data registry will assist two groups of users, data producers and data seekers. The data producer is a

campus researcher, faculty member or graduate student whom generates data or collects any form of data or research output. The data seekers are researchers, faculty or students interested in obtaining information about how to find primary research datasets, descriptions about research projects being conducted on campus, and also contact information of those researchers in charge. These categories of users are also fluid—a data producer can also be a data seeker. All users will be coming from disparate research interests and subject areas. The Data Registry should thus be both general enough to accommodate for multi-disciplinary data and yet specific enough to contain enough detail about the registered data in order to provide data seekers with adequate information about the data they are hoping to acquire.

Unlike a repository, data registries never actively store the data. Instead, a registry is as Chaven and Ingwersen describe a "data portal" which provides access to surrogate records of primary data and contact information of the institution or researcher in charge of a given dataset (2009, p. 2). Using the registry, someone who is interested in the data can contact the researcher in order to discuss a share between them. Additionally, a data registry can also provide more exposure to data safely stored in a discipline-specific or some other kind of repository.

The German National Library developed an example of this type of architecture. For this registry, they utilized 20 metadata elements total, and whenever possible they adopted elements from the Dublin Core schema in order to describe each of their primary datasets (Brase, 2004). Another example is a system designed by the Australian National Data Service (ANDS), called "Research Data Australia" where a "mesh of highly findable web pages describing (and where possible linking to) Australian research data collections" (ANDS, 2011) assist a wide variety of researchers and organizations and by providing a stable environment for the discovery of their data.

These examples of data registries provide excellent models for the UCLA Data Registry, however the real point of departure for this project is the CENS Data Registry, created by Matt Mayernik and Jillian Wallis. In working with CENS researchers since CENS′ inception in 2002 to develop data management plan services that would facilitate reporting of data to the NSF, the team worked through a few iterations of the service, which would ultimately become the CENS Data Registry. In gathering data about the research practices of the scientists, they realized that the system they wanted to implement needed to be dynamic and account for heterogeneous data and inconsistent data practices. After departing from their first idea of building a data repository called Sensor Base, which would capture the actual data being produced at CENS, they crafted the idea of a metadata repository (Wallis, Mayernik, Borgman, Pepe, 2010). The goal was to build a system that can "enable potential data users to discover what CENS data exist, to determine whether those data may be useful, and to learn how to acquire data of interest" (Wallis, Mayernik, Borgman, Pepe 2010, p. 336). This approach allows the data to be discovered via a surrogate of the data or a record of metadata, rather than the data themselves.

One method of creating motivation for researchers to contribute their data to shared platforms such as data repositories and data registries is the ability to create a citation for the data. Creating a method of citing data would allow the producer to receive academic credit for the data they are willing to make available. There are several prominent organizations supporting the practice of providing citation standards for datasets. The Committee on Data for Science and Technology (CODATA) is an interdisciplinary scientific committee of the International Council for Science, which was established 40 years ago. This group advocates for the improvement of the quality and reliability of management of scientific and technological data. They point to the various issues requiring attention if a standard for citing data is to be widely adopted. Among

them are the technical and infrastructural issues, the financial and institutional support needed, the question of persistent identification and also the socio-cultural and community norms of giving and receiving credit for data citations.

Researcher identification system

Another kind of system that can help to expose scholarly output and encourage data sharing is a researcher identification and discovery system such as VIVO developed at Cornell University. Whereas most of the systems discussed so far have been developed in the service of the research output itself, this one offers a different perspective on research. Using the VIVO system, faculty, students, and other campus staff can discover information about the researchers, departments, events, courses, publications and grants on their campus. This system encourages research output discovery, collaborations among researchers, by bringing together "publicly available information on the people, departments, graduate fields, facilities and other resources that collectively make up the research and scholarship environment in all disciplines at Cornell" (2012). This system supports browsing by categories such as people, which are sub-divided further into categories such as faculty members, librarians, and non-academic. Users can also browse by other categories such as organizations on campus, events, and more academic categories such as research, which is sub-divided into smaller categories such as journal, digital resource, and digital collections software. Cornell supports this system by employing a staff of curators and subject experts in the library, who are also part of the VIVO project team.

All of these tools offer different services and perspectives on data management and data discoverability. The goal of this thesis is to extrapolate from researchers' data practices the implications for the design of tools that can support research, and determine if the Data Registry

is the right tool to build. However, given that a grant was received for the specific purpose of building the UCLA Data Registry, another objective of this thesis is to determine how the UCLA Data Registry should be built if it is to effectively meet the research community's needs.

Research Questions

To accomplish the goals of this project, I formulated the following three research questions. These questions guided the recommendations for the design of the UCLA Data Registry, as well as other kinds of tools that UCLA researchers need to support their data and their current research practices.

1. How can the UCLA Data Registry benefit UCLA researchers?
2. What incentives and disincentives do UCLA researchers have to register their data?
3. What are the challenges or obstacles that need to be overcome before services like the Data Registry can be built and utilized by researchers?

Research Methods:

Building the Data Registry requires an in-depth understanding of a wide range of research practices. The best method to access to this type of detailed and quite often very idiosyncratic information is through qualitative data collection and analysis. Therefore, one-on-one interviews with a representative set of researchers were conducted. The sample size of 20 interviews was determined based on the need to obtain nearly equal sample sizes of five to seven researchers from each of three general academic areas: science, social science and the humanities. Each interviewee was asked to self-identify as belonging to any, all or none of these general

categories. Thus, these categories are amorphous rather than strict.

While most researchers were able to place themselves in a category, some were reluctant to do so, and many noted the category they fit into actually depends on the project they are working on and whom they are working with. Methods often overlap, and so most researchers in the humanities and social sciences felt that they belonged in either at any given time. One researcher argued that she belonged to all three categories.

The initial recruitment for interviews were drawn from the pool of IDRE awardees, as the grant placed a heavy emphases on collaboration with other awardees whenever possible. This created an excellent opportunity for "getting in," and building camaraderie with potential interviewees (Lofland, et. al., 2006). In using the shared experience of receiving grants for working towards improving infrastructure for data, establishing this connection will hopefully lead to future opportunities to conduct user testing once the system is in its prototyping stages. The initial group of six grant winners was comprised of four scientists, one humanities scholar and one archaeologist, who considered herself to be part of all three general areas of research, however I placed her in the humanities group since she is heavily involved with the Center for Digital Humanities. From each of the researchers within this initial pool of participants, two more referrals were elicited.

Although the initial goal was to obtain a nearly equal number of researchers in each of the three main categories, the final breakdown below obviously shows a heavy skew towards the humanities. Many of the interviewees in the initial pool acknowledged that the Data Registry project might be of interest to those working in the digital humanities and often times offered names of researchers working within that framework. Therefore the resultant focus on the digital humanities was by accident rather than design. The results and discussion to follow will offer

explanations as to why the digital humanists became a focus of this project. The final breakdown of participation by discipline was as follows:

- Science: 5
- Social-science: 5
- Humanities: 10

All 20 interviews were conducted between January and April of 2012 and each lasted between 30 minutes and 1 hour. Each interviewee was asked a series of 16 questions aimed at being open ended, in order to give the interviewee a chance to expand and reflect on his or her work. All of the interviews were audio recorded. The data were analyzed according to the general advice and principles from Lofland, et. al.'s, *Analyzing Social Settings*, which hold that qualitative data analysis can arise out of the data themselves, by induction rather than deduction. This method is often referred to as grounded theory, because, "when empirical or theoretical observations emerge inductively, they are often said to be 'grounded' in the sense of emerging form the group up rather than being called forth by prior theoretical constructs" (Lofland, et. al.,2006, p. 195). However, instead of establishing formal codes from this method of induction, general concepts, themes or categories were established. With these themes, other strategies such as memoing and concept mapping were implemented to analyze the data (Babbie, 2007). Memos were taken both directly after an interview and throughout the analyzing process, as themes both emerged and were rejected. The concept mapping, or the process of physically or graphically formatting data in virtual or physical space to determine how certain categories related to each other, proved to be an extremely useful exercise in the analysis of the data.

Results of the Research

The results of the interviews are organized by the three research questions. The first two

questions are aimed at figuring out how the UCLA Data Registry will support research at UCLA, while the third question was aimed at a broader understanding of research practices and concepts of data or digital research output in each of the three areas of research—science, social science, and the humanities—as well as the changes in scholarly practices that have occurred. In developing the guidelines for building any kind of information technology in support of scholarship such as the Data Registry, a deeper understanding of how researchers do academic work is needed.

Research Question 1: How will the UCLA Data Registry benefit UCLA researchers?

The major benefit that all three groups of interviews acknowledged that the Data Registry would provide is additional exposure to their work. An adjunct faculty member noted, "A lot of work in academia and publishing is trying to get other people to notice your work" (Archaeologist, Adjunct Faculty). For this researcher, discovery of her work is very important. She explained that she is still trying to secure a full time faculty position. She stated that "the more you get your work out there, the better off you are" (Archaeologist, Adjunct Faculty). Another archaeologist mentioned that he would like to register all of his data because journal publications "only reflect a sheer fraction of the data we generate" (Archaeologist, Faculty). The Data Registry would provide him with an additional method of exposing the data he has collected that may not be included the publication. He acknowledged other sources of dissemination, such as eScholarship, but he still saw the Data Registry as providing additional avenues of making his data available to others.

Many researchers also saw the Registry as an alternative venue for first publishing data. A researcher working in the sociology department explained that she needed to find additional

methods for disseminating her data outside of traditional publications, as her current research projects combine methods from sociology and anthropology and focuses on a digital community. She stated that finding a journal to publish this type of multidisciplinary study has become increasingly difficult (Sociologist, Faculty). The registry would give her a chance to make her data and her methods available to the rest of the community without a formal publication.

Researchers also noted that the registry could be beneficial beyond the services that it might provide for data. One interviewee stated that there is a real need for the Data Registry because there is no other "clearinghouse of information about research projects on campus" (Architectural history, Faculty). She described that the registered projects could serve as a 'research profile' with links to publications, CVs, etc. She supported her assertion that the registry would serve as a better tool that can be used to find people and their interests rather than data because the registry would only be able to provide the very basic descriptions applicable to all research projects.

Research Question 2: What incentives and disincentives do researchers have to register their data?

The interviewees in each of the three general categories of research were able to identify motivating factors that would encourage them to contribute to the Data Registry and also factors that might deter from contributing. These incentives depended heavily on the individual researcher and the type of work they were doing, rather than the general area of research that they identified with. The following section is organized by the incentives to contribute, followed by the disincentives to contribute.

Incentives to contribute:

Those researchers interested in registering their data described incentives such as curiosity in what others within the same research area are doing, additional methods of managing their data, and the feeling of obligation to register due to receiving public funding. One learning scientist noted that she would be interested in other projects being conducted in a similar field, because she would be interested in establishing possible collaborations for future projects (Learning sciences, Faculty). Another motivation to which researchers pointed was an additional method of data preservation. In archaeology, where data loss is a big concern since an excavation can never be reproduced, providing multiple forms of preservation of even the digital content is advantageous. One researcher stated, "Data preservation is key because it's the only thing that's left" (Egyptologist, Faculty).

Another interviewee working on creating a searchable database for ancient magical artifacts said that his motivation came from his feeling of personal obligation to make his work available to others. He stated, "I truly believe that research paid for with public money must be publically and freely available"(Near Eastern Studies, Faculty). Likewise, a researcher in bioinformatics acknowledged that contributing to the Data Registry would fulfill funding agency requirements. Additionally, he pointed out that he felt personally motivated to both register and seek data because the Data Registry could be a place to find more data. "The data is difficult. We always struggle with getting enough data" (Bioinformatics, Faculty).

Disincentives to contribute:

A computational scientist stated that he does not see an incentive to register before he had a chance to publish on his results, because others might try to compete with him by stealing his

results. He explained that his trepidation stems from the fact that his data are generated by a shared super computer, which can take from one up to seven days to compute. This query is very expensive and so there is a high level of concern when it comes to sharing data with others. Similarly, a graduate student in biochemistry recalled an occasion when his research ideas had been stolen from him without attribution. Therefore he too is wary of sharing ideas and data before he has published his results.

Another huge factor that contributes to a researcher's unwillingness to make data available is the incredible amount of effort that goes into creating metadata that renders the data useable to anyone else besides the data producer. The biochemist acknowledged that metadata is rarely used in his lab and that someone else viewing the data would see it as "a bunch of files"(Biochemist, graduate student). However, he has contributed to a shared repository such as the Protein Data Bank, which has very strict metadata standards for submission. So when he contributes to repositories with established standards, he conforms to the metadata requirements, however on a daily basis he does not find it necessary to add metadata.

For other kinds of projects, specifically those in the digital humanities that create scholarly arguments by combining disparate files into visualizations rather than text, the Data Registry might not be able to capture the true essence of the research. One faculty member questioned, "what does it mean to save the data? Are you going to save the files? The XML files and image files? You can't display them" (Digital humanities, Faculty). She argued that unless you can also save the environment where the project as a whole works, "its like taking all the paragraphs of your book out of their chapters" (Digital humanities, Faculty). She would only be interested in registering her entire project, and so unless it lived on a server or was web-based, she would not be interested in creating a registration page for her research.

Research Question 3: What are the challenges or obstacles that need to be overcome before services like the Data Registry can be built and utilized by researchers?

Challenges in the Sciences

In the sciences, researchers continue to struggle with issues surrounding metadata practices. Whereas some scientists are working towards developing better metadata practices and metadata standards, others are still resistant. The researchers working in biomedical informatics and neuroscience are on one end of the metadata spectrum. The entire medical community is experiencing great advances towards metadata standardization, as the new mandate to make all medical records electronic, vast improvements in the standardization of data and metadata is imminent. The entire community is hopeful for what this means for advancing and improving their work. "We are on the cusp of something. Because the hospital is changing over to this new system, there is a lot of potential if its done right, to make research that much easier because you have one centralized system that you are going to be accessing"(Biomedical informatics, Faculty).

Imposed standardization is not the only driving force in this area of research. The researcher working in neuroscience stated that he attempts to apply as much metadata as possible so that others can reuse them. His reasoning was, "You don't know what people are going to want to do with that data. Cognitive neuroscientists might want to know about the task paradigm, but the scanner parameters… they might not care." He makes a point that if the researcher producing the data only chooses to apply metadata describing certain things, then the uses of the data are limited (Neurology, Faculty).

However, this trend has not quite spread to other areas of science. In the computational

science field, they are able to generate big data, yet the lack of metadata standards makes sharing this data virtually impossible. The field is however aware that this is a huge problem. Due to the cost incurred by the query alone, sharing data would be beneficial. The metadata itself plays a very large role in the research. It essentially translates the computer-generated output into a human-readable visualization. Without the metadata schema, the computer's output is completely unintelligible. There have been many attempts in the field to create sharing mechanisms such as APIs which would share the metadata along with the data. However, the complexity of the creating this API becomes quickly overwhelming quickly. The faculty member in computational science stated, "We are kind of dropping the idea of making something really universal because, universal sounds good and standard, but if you try to cover everything it becomes really complicated" (Computational scientist, Faculty).

Bleaker yet, in some fields, the lack of technical standards is not the biggest problem preventing opportunities for data sharing. The cultural practices of a given lab weigh heavily on the metadata practices. When asked about the kind of metadata practices that are used in his lab, the biochemist answered, "We don't normally…we don't really have any ontologies or taxonomies. We really don't have any good system, where we have data that describes our data. Its just a bunch of data in folders" (Bio-chemist, Graduate Student). Thus, the absence of community standards of providing any metadata whatsoever definitely limits a researcher's willingness to make his or her data available.

Additionally, the sciences continue to struggle with developing proper storage for the data they are generated. Just as the researcher at LONI spoke about how the five terabytes of data that they have are actually not really enough, the computational scientist argued that a more pressing issue is the question of where to house these large datasets rather than the issue of

standardizing metadata.

Challenges in the Social Sciences

Based on the interviews collected for this thesis, one possible problem facing the social scientists is the amount and diversity of data being produced. Social scientists have definitely experienced their own data deluge, as they have begun to gravitate towards more data-driven and rigorous methods of conducting their research. However, the amount of data being produced probably would not prevent them from utilizing tools like the UCLA Data Registry. In fact, many researchers in the social sciences were very interested in learning about new tools and methods to manage, preserve, and disseminate the data they are producing.

Challenges in the Digital Humanities

The interviews with humanists conducted for this thesis were largely with researchers working in the digital humanities. The challenges faced by digital humanists stem from the many changes that are currently taking place, as well as shift from being mostly project based to becoming more of an established discipline. In moving away from traditional humanities scholarship, researchers in the digital humanities begin to adopt the language of data and metadata, move from two-dimensional text based arguments to visual three-dimensional space-based arguments, and engage with new forms of scholarly communication. These changes have made communication between digital humanists and information professionals more difficult, as their needs vary greatly from that of scientists and social scientists.

One of the major changes that has occurred in the digital humanities is the gradual adoption of the language of data and metadata, however, the understanding of what these terms signify may be quite different from the concept of data in the sciences and social sciences.

Another major change that has occurred within the digital humanities is the move from creating arguments in texts to utilizing technological tools to create arguments in an entirely new medium. The new arguments often go through a process of combining files from disparate sources into a visual representation of the argument. Many of the interviewees working on these types of visualization were three-dimensional, so in addition to using digital files, these researchers utilize complex algorithms, GIS data, or other sophisticated software to create their projects. A researcher working in digital cultural heritage said that his three-dimensional model of the Roman Forum explores questions of space in social history, which had previously only been explored in text, but "couldn't be fully articulated" (Digital Cultural Heritage, Faculty). Another researcher working in architectural history is developing a three-dimensional model of an exact replica of a building from a World's Fair. This project has been underway for the last 10 years, as she continuously gathers more and more information about the building she is modeling. Though both of these projects are working with three-dimensional tools to reconstruct sites of the past, the goals and methods used to create each project are quite different. Whereas the architectural historian is attempting to develop the precise copy, the historian building the Roman forum is more interested in the spatial relationships that the particular monuments have to each other and how this affects an understanding of the social history of the Romans at that time. In order to carry out each of these goals, the methods used to generate their models are very different as well.

A third major shift in the digital humanities is that alternative forms of publications are emerging and to an extent replacing more traditional forms of publications such as the journal article. A professor in the Film and Television program at UCLA who works with iPad applications to teach his students how to conduct visual analysis of films, argues, "This is

publication. It's an alternative form of publication. That's part of the whole turn-around that I think it happening now" (Film and Television, Faculty). Again, every researcher embraces this change to fit the needs of his or her research. Another researcher working on developing a database that will contain ancient magical artifacts, described the database as a "specialized academic tool" (Near Eastern Languages and Cultures, Faculty). He doesn't necessarily see his database as a publication at all, but rather as a tool that he can use to interpret his data. He explained that, "Its only within this larger framework, that I can really come to an understanding of what motivated people, why they did certain things and why they made certain choices"(Near Eastern Languages and Cultures, Faculty). Only after he was able to bring together the sources of information together into a whole was he able to understand his research in a new and ultimately more valuable way.

These changes occurring in the digital humanities have introduced different concepts of data and the role that data plays in research and research outputs. As a result, the needs of researchers are more difficult to address as well. If information professionals are going to build tools aimed at providing services to these researchers, then it is important to first understand the varying definitions of data, how the data fit into the larger research project, and the different attitudes towards the discoverability of the data.

Discussion of the Results

The discussion of the results is organized into three main sections. The first section discusses the main finding of this research—that varying concepts of data and data practices across academic disciplines directly affect researchers' attitudes and behaviors towards data discoverability. In taking the different notions of data and the researchers' attitudes towards data discoverability into account, the second section discusses how to implement this research into the

design of the UCLA Data Registry. However, the Data Registry will not be able to provide all of the support that researchers need for their research, as indicated by the kinds of benefits, incentives and disincentives, as well as the challenges they are currently facing. The last section provides an explanation for the other tools that could be implemented in conjunction with the Data Registry to provide service for the many different kinds of research being conducted on campus.

The Spectrum of Data Discoverability

During the course of conducting the interviews with researchers it became apparent that the way researchers characterize their data, and how data are used in their work, varies among the different disciplines, and therefore their attitudes towards what data should be shared or discovered by others varies as well. The limited size of the sample makes drawing conclusions based on trends very difficult. However, certain similarities in how researchers within the same general research area characterize their data and research output can be identified.

The distinction between data and research output can be helpful when considering the spectrum of definitions of scholarly output across the academic disciplines. When the scientists in my sample discuss their research process, they describe the series of standardized steps they take to collect, clean, use data. Data are generated with the goal of producing a journal article, which in some cases is also a very formulaic endeavor. The data play different roles in the publication for different fields; for example, the raw data is never included in the publication for computational sciences, but rather just the processed data in the form of a visualization. In fact, in this field, the raw data that is generated by a super computer is not the end product or the research output at all, but rather an intermediary stage, or even the input (Arzberger et. al, 2004)

of the research. Once the visualization has been created, the raw data is no longer as important as the processed data. Though they are concerned about having enough storage for all of the data generated, the real concern is for making the processed data available in the future, rather than the raw data.

In neuroscience, the data they collect are three-dimensional brain scans, and even though the original Dycon are not included in the publication, and derivative images in JPEG or TIF formats are included in the publication, Dycon images are still very important to the researcher. Once the data have gone through the cleaning process and being "defaced" or stripped of any identifiable features, it is stored for future use. The neuroscientist said that data reuse is common, as other scientists asking different questions can use the image in different ways. Towards this end, robust metadata is applied to the data throughout its lifecycle. In recent years the ability to apply metadata describing the processes that the data have gone through has been made more automatic and easier with a tool called the LONI pipeline. This tool effectively tracks the provenance, or the ways that the data has been altered or changed from its original state (Gil, 2010).

In neuroscience along with bioinformatics, data sharing is just starting to become more commonly practiced. With the implementation of the electronic medical record, new opportunities for exchanging data will arise due to the availability of data in standardized forms. Therefore, the data are being valued as an important research output that should be stored for future use. As the neuroscientist at LONI argued, it is important to capture as much metadata as possible due to the different uses of the same images.

However, in other science fields, the culture of sharing is much more limited, as is the case in biochemistry. Even though the Protein Data Bank has strict metadata standards for data

deposit, the interviewee in biochemistry explained that his lab does not use metadata on a daily basis. Essentially, he works to meet requirements of the funding agencies, but isn't interested in providing additional avenues of discovery for his data. This researcher also explained that unfinished projects by graduate students are never completed. The data generated are forgotten on a server. The most important research output in this community is still the journal publication. This is somewhat surprising due to the domain specific data repository dedicated to preserving this type of data. Therefore, the culture of a particular lab may be more effective in engendering certain attitudes towards data sharing and data discoverability.

The social scientists in the sample collect both qualitative and quantitative information. Though social science research can be considered data-driven, as the conclusions are drawn from the data collected in the field as well as standardized, quantitative data, the data collected are only an intermediary step in the project as a whole. The main goal in most of the projects described by researchers in my sample, is to create a workable theory, which uses the data as evidence, or to create a useable tool, such as the researcher in GIS who collects geographic coordinates and for interpreting radiation levels in a given geographic area. In the learning sciences, the data are entirely qualitative, ethnographic information collected in the field. The process of collecting the data is very time-intensive, as the researcher needs to gradually discover emerging trends in the data about her research subjects. From the data collected for just one project, she explained that she would write up to around 15 journal publications on her findings. The raw data itself can never be released in its original form, as the personally identifiable information about her research subjects would violate her Institutional Review Board (IRB) agreement. However, she is able to release summaries of the anonymous data. Once again, it is common to consider the journal publication or a complex visualization tool as one of the most

important outputs of the research, rather than the data collected. However, many of the social scientists in my sample indicated that finding a journal to publish work has become increasingly difficult, especially when adopting an interdisciplinary or mixed-methods approach to conducting research. Therefore, they are open to adopting new tools, which would help them to disseminate their research.

Though most of the interviewees in the digital humanities have begun to adopt the language of 'data' and 'metadata' when describing their research, this alone does not actually reveal much in particular about the work that they do, or their attitudes towards the discovery of their research. In fact, during the interviews there was a wide range of success with regards to the appropriateness of the interview questions in terms of how well they related to any given researcher's work. Often times, questions about data and metadata would require re-phrasing to relate it better to the work that these scholars were doing. These moments would result in the necessity to engage in a more in-depth conversation about the project including the background to the project, the kinds of research questions that are being asked, and the methods used. Essentially, the project as a whole needed to be further described and fleshed out to understand the various components such as the files or the 'data' and how they fit together. However, these experiences often yielded a deeper understanding about a research practices and the research project.

Their work may involve data, but their final product is not a dataset, necessarily, but rather a whole project, often involving many components. Often times, it seemed like there was conflation of the terms data an digital, where digital materials were used, but these materials were not actually data. The output of digital humanities research can include data in addition to other components used to create the project such as software, visualization platforms, and analog

archival material. In many different fields, the data are always **"**difficult to separate from the software, equipment, documentation and knowledge required to use them" (Borgman, 2007, p. 183). But a point of diversion and comparison between the digital humanities and the sciences and social sciences, is that the actual research output is no longer text-based, but rather visual representation of the argument, which may further complicate the discoverability of these digital objects and the data used in their production. In order to exemplify the concept of data and its role in the digital humanities, it will be helpful to use three specific digital humanities projects as representative of the kinds of work being produced in the digital humanities.

The first example of a kind of digital humanities project, are the types of projects where the creation of the project is itself the research output. These bespoke digital objects are created with two main goals in mind, as a technical and even artistic feat and also as a usable object, however these objects are not necessarily reusable tools. The World's Fair replica being created by the architectural historian is an example of this type of digital humanities project. The creator hopes to use her model in classroom instruction which can guide a lesson on the architecture used in the fair. She used various data sources in her project, but her research output is much more than the entire list of the files she used, as she weaved these sources together to create her object. The software used to create the model is also a necessary element in her project. For projects of this nature, the discoverability of this object may not be a realistic goal. Since the object is not reusable in a sense, its creation is not towards creating standards that can be representative of the methods one takes to create these types of projects. Instead, it is created as an end in itself to convey an argument.

Some projects in the digital humanities do create reusable tools, in which case the discoverability of these tools may be more important. Though the researcher working in digital

cultural heritage is creating a model of the Roman forum, the goal of his research is not to create a perfect replica, but rather to create a virtual world that can then be used to explore space and spatial relationships. This project involves a lot of data including GIS data and images. But the research output is the combination of Google Earth, the text-arguments supporting his decisions to build the forum in the way that he did, the images used as evidence such as coins or other artifacts, and the computer programs and algorithms he used to create the buildings and the avatar. Part of his goal is to develop is own argument about the way that this particular world was spatially organized, however it can also be used a as a tool to explore processions and other social and cultural events.

The third example of the kind of projects being created in the digital humanities are those whose major contributions are geared towards developing standardized methods for doing certain kinds of research or work in the digital humanities. The actual projects may vary greatly, but what they all share is the interest in developing a reusable method for doing a kind of project in the digital humanities. For example, the researcher working towards creating an archive for Twitter feeds is trying to develop a standard method for how to aggregate this type of data based on certain parameters. Furthermore, many digital humanists now employ social network graphs and text analysis tools to conduct their research. Since the software used to generate these types of projects have become easy to access and very intuitive to use, the valuable technical know-how to be gained from these projects instead lies in making explicit the decisions about how to analyze a text, and the processes taken to do so.

The discovery of digital humanities projects can be more complicated than other kinds of research. In my sample of interviewees, their research output often involved a software component such as a visualization platform, both digital and analog data sources, and both

explicit and implicit methodologies that are required to understand, use or reuse the project. Tara McPherson explicitly states that the visualization is afforded by the technology that digital humanities deliberately turn to when conducing their research (2008), which means that the visualization is essential to the project. The visualization needs to be considered when developing data management plans or services, including data registration. The OAIS model recognizes that "as digital technology evolves, multimedia in technology and the dependency on complex interplay between the data and presentation technologies will lead some organization to require that the look and feel of the original presentation of the information be preserved" (OAIS, 2002).

The question posed by a faculty member asking 'if saving the data meant saving the files,' really gets to the heart of the problem that essentially every project utilizes different concepts of data and how the data fits in with the project as whole. Even in certain cases, when the answer to the question seems like it should be 'yes', truly capturing the entirety of the project requires much more than having 'just the files'. In the work of the Film and Television professor, his data really are just XML files that are then linked to a film in an iPad application. However, the code that he wrote for the iPad application, which actually connects the XML files to the film he wishes to analyze could also be considered his data. The application is more than just a connection between the XML and the film, it exemplifies the technical skill and ability of the professor, creates an environment that allows the user to move fluidly through the film, which helps to drive home the arguments made about the film.

Implications for the UCLA Data Registry

Within the bounds of the definition of a data registry, certain kinds of research projects

would benefit from the UCLA Data registry. Many of the interviewees working in the sciences have data that are currently stored on UCLA servers, such as those working at the LONI neuroscience lab. These data have a permanent home, and given the researcher's interest in sharing data, a registration page would provide additional exposure to his data. For those social scientists interested in exposing their work but are having difficulty finding venues for disseminating their work, the UCLA Data Registry might be a useful method of providing access to one's research until a more formal article can be published. However, many scholars are still in need of storage space for their data. And for these scholars, the Data Registry would not be able to provide the support that they need.

The ideal UCLA Data Registry would function within a larger infrastructure of tools. In addition to the large infrastructural requirements such as additional storage options, the UCLA library should consider linking up current data management related library tools to the Data Registry, such as the Data Management Planning Tool, which helps researchers to generate Data Management Plans for different funding agencies. If a researcher is creating a data management plan, they could then be easily directed to the Data Registry to create a registration page, with much of the same information, or even harvested information if the two systems could become interoperable. The Data Registry could also point to datasets stored in California Digital Library's Merritt repository, or those projects residing in eScholarship, UCLA's institutional repository.

In taking all of this information into account, the wireframes in Appendix C present the first phase of the UCLA Data Registry. It currently utilizes a core set of generic metadata elements, which can be used to by any community, which an be seen on the "Register Data Page." The metadata elements used in the Data Registry was borrowed from the data registry

developed by Matthew Mayernik for the CENS registry, which was instantiated in the Dublin Core metadata standards. This page displays the fields that researchers will contribute information to in order to generate the surrogate record of the data they are registering. The starred fields are required fields, while the others will be optional. Researchers are also able to contribute links to actual datasets, if a stable URL or DOI already exists as indicated by the 'link to data' field.

The submit button on the bottom-right also has an additional component, the "EZID." The EZID service is a service developed by the California Digital Library, in which researchers can generate unique identifiers for datasets including DOIs or ARKs (CDL, 2012). While, the researcher may or may not have an actual dataset to register and link to with the Data Registry, each registration page will received a DOI, which will facilitate easy retrieval, and can be used to cite in publication as a location of the surrogate record of the data. The ability to generate DOIs may be among the most important incentives to contribute to the Data Registry. The library can also use these DOIs to track the usage of the Registry. The "Data Paper Page," is the surrogate record of the data, which is populated with the fields that had been previously contributed by the researcher. This information can be updated via the "Update Registry" option, and these records can be searched via keywords. When a keyword is successful it will populate the "Look up data" page with the records that contain that key word.

What Tools do UCLA Researchers Need?

Based on the discussion of research practices and attitudes towards data discovery, it's clear that other tools, which can work in conjunction with the Data Registry are also needed to support the wide range of research being produced by UCLA researchers. When asked about the

possible benefits and incentives to register with a system like the proposed Data Registry, researchers identified benefits such as finding information about research projects on campus, a tool that would encourage future collaborations, and the ability to gain additional knowledge about the type of work being conducted. This set of requirements indicates that one kind of tool that the UCLA researchers would benefit from is a researcher identification system like the VIVO system at Cornell University. Researchers across all fields interviewed for this study indicated that knowing what kinds of projects were being conducted on campus would be a valuable information resource. For some researchers, this type of system would help them to identify and initiate collaborations, and more importantly create a research profile. This system can incorporate links to publications, grants, CVs and serve as a central location to find researchers' personal website.

Another set of requirements that researchers identified were that of data preservation, fulfilling funding agencies' data management plan requirements, and the necessity for more storage of the data being generated. Once again, these requirements will not be fulfilled by the Data Registry, but instead signify that a data repository is needed for all the various types of data that are being produced on campus. There are some options for these researchers already on campus, for certain kinds of data. However most of the researchers interviewed currently have their data stored on a UCLA server, rather than a repository whose aim to provide long term preservation and storage needs. In addition to the physical hardware needed for data storage, further technical standards need to arise which guide the selection and appraisal, a concept which originated in the archival community, which holds that constant reevaluation of a collection of information—records or in this case data—to make sure it is necessary to continue storing (Stewart, 1976). While the data at the LONI lab takes up a lot of storage space, there is a culture

of sharing, which might mean that the data will serve multiple uses for multiple researchers, making it a valuable asset to store into the future. However, this raises questions and concerns about the data generated in other fields such as computational sciences. The data is not easily sharable, or useable by researchers other than the data producer. And while, the interviewee stated that storage is a big concern, the question that seems to follow is, what sort of guidelines also need to be developed alongside the storage system that can evaluate the data and make sure that it is worth storing?

The long terms needs of digital humanities projects require a system that will allow the entire project to be useable in the future, which includes software, applications and the storage of all the constituent parts of the project. This requirement makes the discovery of their data very difficult, as their data is not standardized across the discipline, nor are the tools that they use to generate their projects. They need a kind of repository that allows for mixed media formats and a relatively open structure.

One possible solution to the multi-media storage needs of many of the researchers at UCLA is Harvard's Dataverse Network. This open source environment allows researchers to "publish, share, reference, extract and analyze research data" (Dataverse, 2012). The Dataverse Network supports multiple dataverses. Each dataverse contains collections of actual data files, text files or other forms of research output as well as a description about the data. In addition to providing the requisite space requirements, the Dataverse Network provides researchers with a standard for receiving citations for their work, which may encourage data sharing. Though the Dataverse Network began at Harvard, it is not only affiliated with that university. Since it is open-source, a UCLA researcher in need of the services provided by the Dataverse Network can create her own dataverse for her research. Once she has her dataverse, she can create a

registration page with the UCLA Data Registry in order to provide other UCLA researchers and students with easy access to her research.

Conclusion

When building any kind of information service in support of the wide range of research being conducted on a university campus, it is necessary to understand how researchers do academic work. The research conducted for this thesis provided the requisite background study for how researchers on the UCLA campus might utilize the UCLA Data Registry—a tool aimed at making surrogate records of data discoverable. In interviewing researchers from the sciences, social sciences and digital humanities, it became clear that a scholar's interest in making their data available using a system like the Data Registry is dependent on the particular conception of data and what is held as valuable research output in a given field. It also became clear that in addition to building the Data Registry, other tools should also be implemented in combination with the Data Registry. Connecting the Data Registry with other kinds of tools aimed at the discovery of research output, such as a researcher identification system, and personal storage space like Harvard's Dataverse Network might provide a more useful infrastructure for the researchers at UCLA.

As the UCLA Data Registry develops, the UCLA library must also continue working with faculty to test the system and make changes that are deemed necessary. As indicated by the interviews, many humanists were interested in the Data Registry, yet certain kinds of research may still be left out. The UCLA Data Registry project is an example of the type of data management service that academic libraries need to develop as the ever-growing amount of research output continues to be produced in all academic fields. As information professionals, we

must strike a balance between advocating for standards and tools that will assist us in the process of providing description, preservation and accessibility to research output and the actual needs of researchers and scholars.

Though the success or failure of the UCLA Data Registry is unknown at this time, what is known is that scholarship is changing. If the UCLA Data Registry is successful, then the library should continue supporting the registry, and potentially building off of its services in new ways. However, if this particular service is not actually of value to the greater research community, the lessons learned should be applied towards a new service. The university research library and digital libraries need to embrace the changes occurring in the sciences, social socials and especially need to continue working together and collaborating with digital humanists to implement standards when they can, but not impose standards when they are not be helpful.

Appendix A

Recruitment Letter

I used the following recruitment letter in my first email to every researcher that I interviewed:

Hello ___,

I am conducting research for my master's thesis in the department of Information Studies. My project focuses on gathering the functional requirements necessary for implementing a data registry, a new library tool for data management and discovery. My advisor, Christine Borgman and UCLA Library's Todd Grappone received a grant awarded by the Institute of Digital Research and Education and the Institute of Informatics to build a data a registry for the UCLA campus. We envision the registry as a web service where researchers contribute descriptions about the data they produce and also an environment where students and researchers can find data and learn about other research projects on campus.

My specific part in this project is to gather the requirements of the data registry by talking to researchers on campus about the data they produce, what methods they currently use for storing them, and how a tool like a data registry might benefit their work. I am particularly interested in how researchers describe their data, and the differences in data description among researchers.

I have contacted you because a discussion about your research and data practices will guide me in figuring out how to create this tool for researchers and students. Are you the person on your research team that primarily collects, processes, and analyzes the data? If you are interested, I would like to set up a time for a quick, informal interview. If you would like to hear more about my project and research before we set up an interview, please don''t hesitate to email me any questions or concerns you might have. I am looking forward to meeting with you!

Thanks so much,

Rachel Mandell

Appendix B

Interview Instrument

The interview instrument I used to conduct interviews with UCLA researchers was as follows. First I would provide a brief introduction to my research, including the purposes and goals of the research. Though this part of the conversation was usually completely ad hoc, I would make sure to say something like the following:

Hello, as explained in the recruitment letter, I am working on figuring out how to build a data registry, which will be a new tool maintained by the UCLA library. This web service is aimed at assisting researchers in making their data more accessible to others and also at facilitating data discovery by those students and faculty interested in other campus research efforts. In order to know what sorts of components the registry should include, I need to understand what types of data you produce, if you would be interested in finding data you didn't produce and if so what kind of data you would want to find in a system like the registry, and also how you describe your data. The ways you describe, or don't describe your data will help me to know what sorts of metadata fields will help researchers both register and seek data. My questions will help to guide our discussion, but I want this to also be an informal meeting, where I just try to understand your daily tasks as a researcher, so of course there are no right or wrong answers here.

The following list of 15 main questions and subquestions were what I used to guide the conversations I had with UCLA researchers.

1. Briefly describe your current research and the data you use, generate or collect?
   - What format(s) do they come in?
   - What constitutes a dataset?
   - How much data do you produce?
2. Can you walk me through your data collection process, from start to finish?
   - To what degree is your data collection process ad hoc or standardized?
3. Where are your data currently stored? How are they stored?
4. What are you current data management methods?
   - Where are your data management pain points?
5. How do you describe your data or apply metadata? What language/terms do you use?
   - Are there metadata standards that you use?
   - If so, where did they come from?
   - Are there standard taxonomies that you use?
6. What is the purpose of describing your data? For yourself for future use? For others on your team? For potential deposit in a repository?
7. Is your data included in the publication? Is it essential to the published paper?
8. Have your data, the collection process, or metadata practices changed in recent years?
9. Have you submitted data and/or metadata to a shared community repository?
   - If so, what data did you contribute?
   - If not, why not? And hypothetically, which of your data would you contribute?
10. What would you have to do differently in terms of annotating, describing your data if you
11. knew that there was a possibility that you were going to make them available to others?

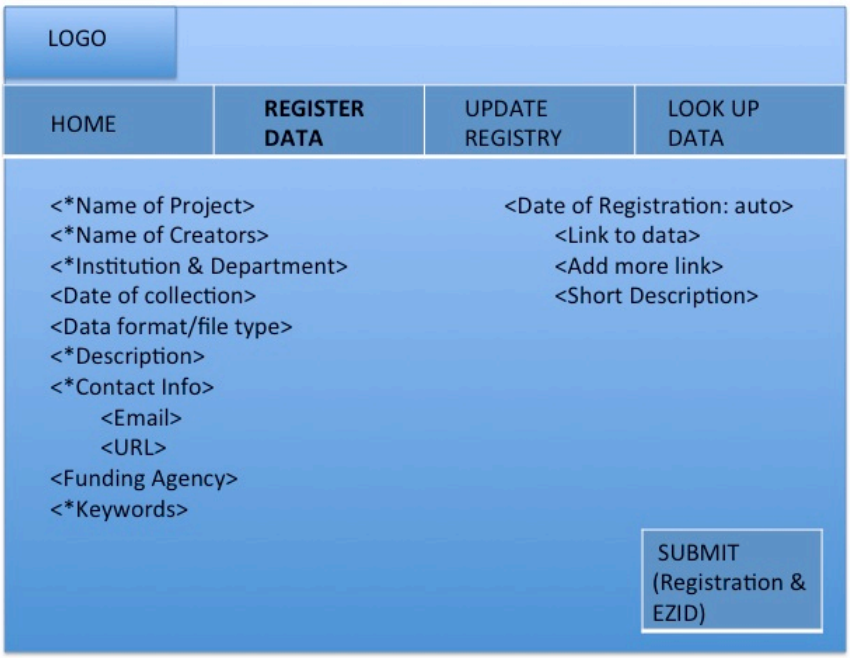
12. Is there a strong ethic of sharing in your research community?
13. Under what conditions would you release data to others?
  - What data would you make available?
14. Are there any barriers to sharing your data that are out of your control?
15. What would your motivations be for registering your data? (Some possibilities might be: To make your data available? To fulfill a mandate by a funding agency? To receive credit for research output?)
16. Where do you receive most of your funding?
  - Are there any data management requirements?

Appendix C

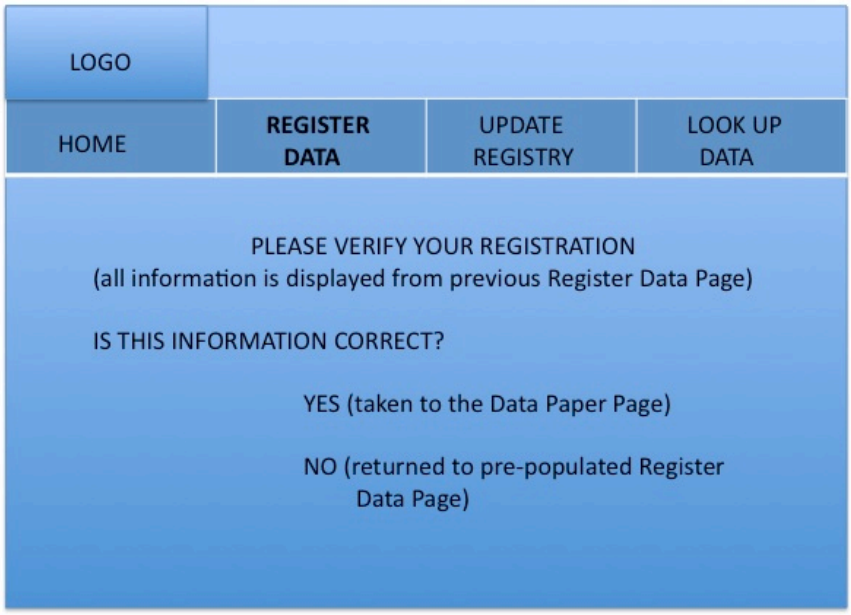
Wireframes for the UCLA Data Registry:

HOME PAGE

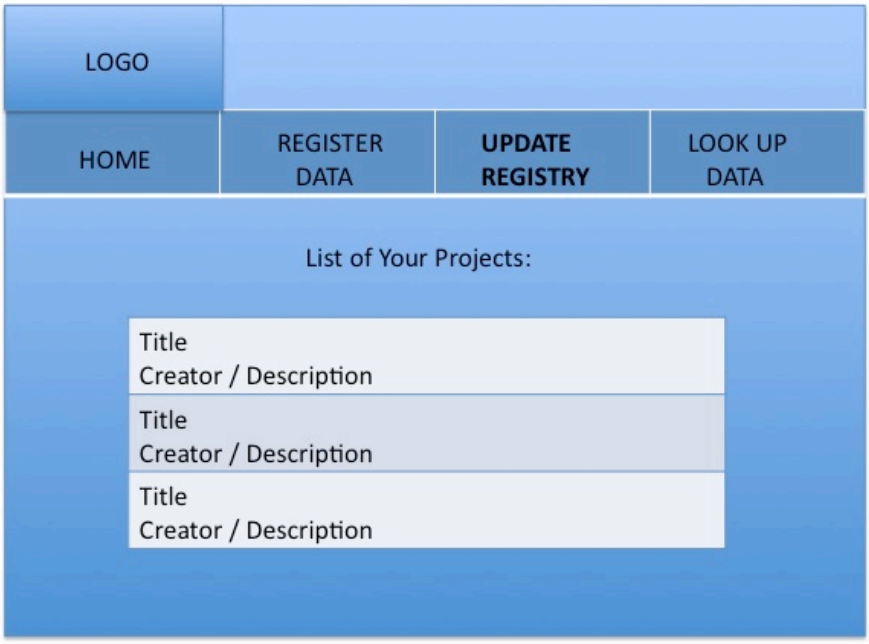
REGISTER DATA PAGE: * = required field, and the star should be visible to the end user



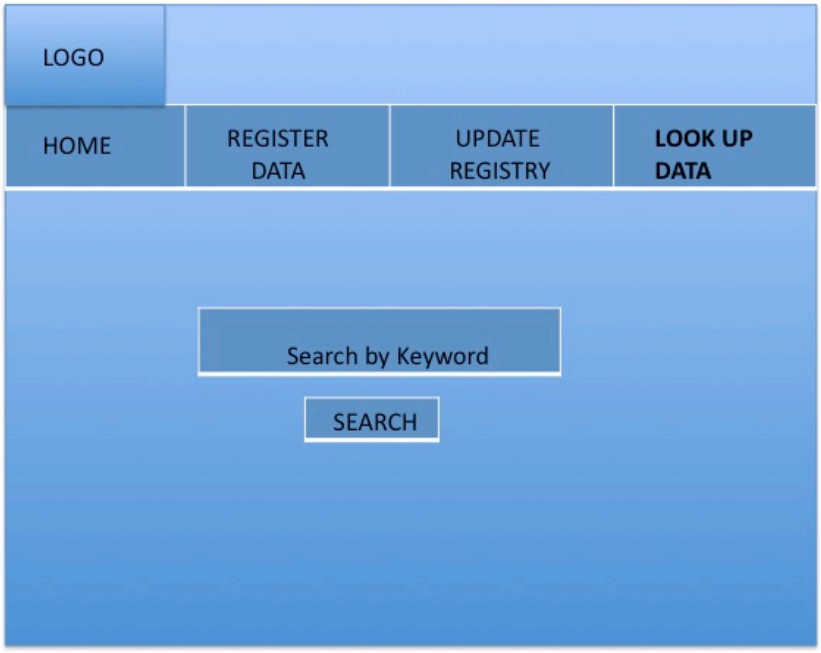
VERIFICATION PAGE: Data from the previous Register Data Page. All information in parentheses is not displayed on the page.



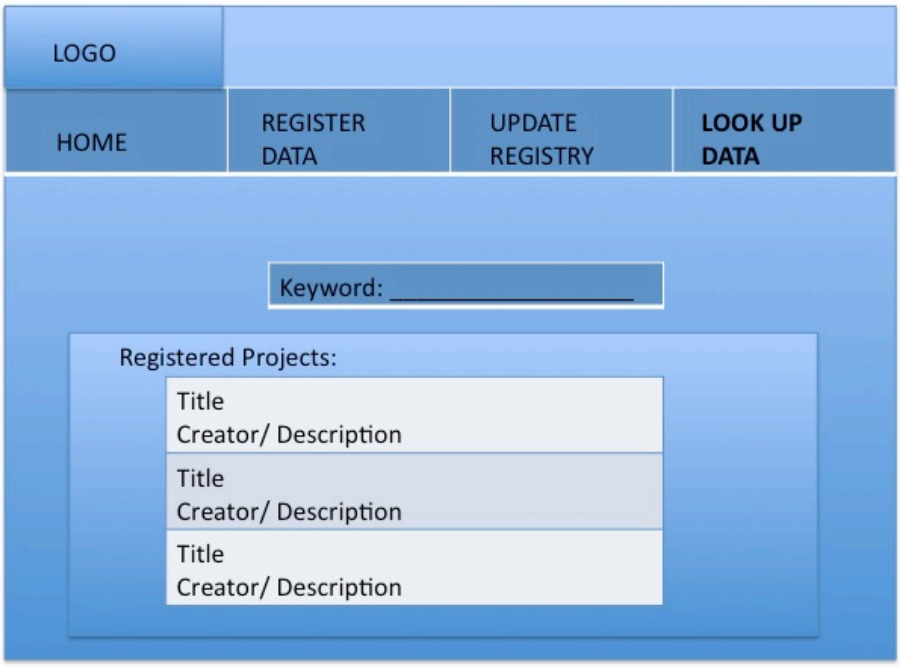
UPDATE REGISTRY PAGE: from the home page. When 'UPDATE REGISTRY' is selected, the user will be prompted to log-in and the list of projects registered to that user will be displayed. The description will be the first sentence of the description registered.



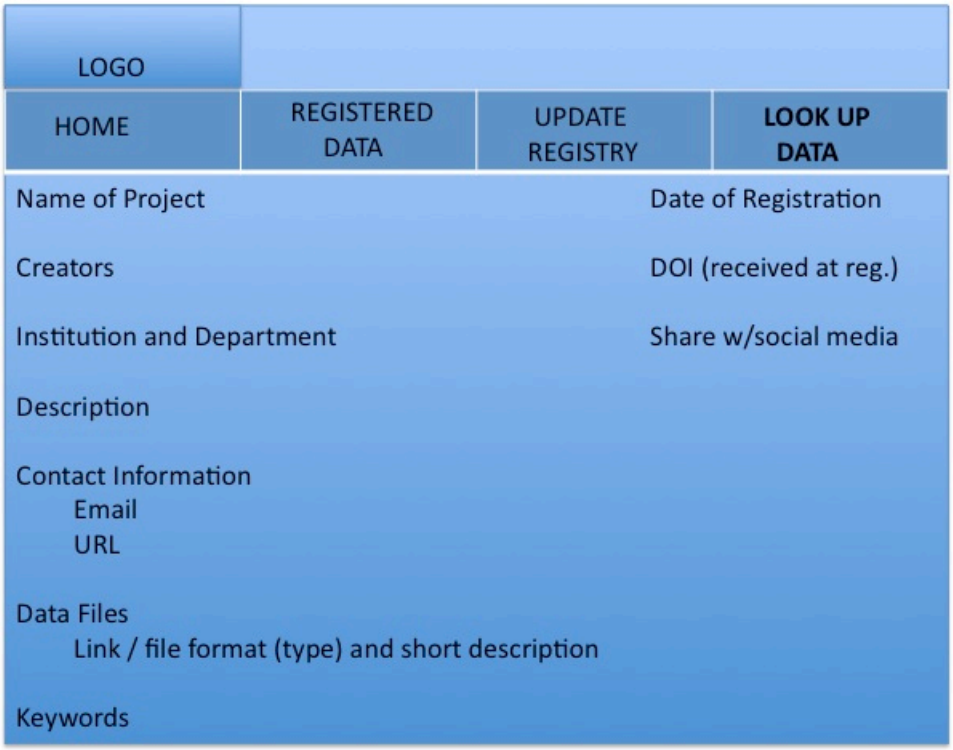
LOOK UP DATA PAGE: from Home page, 'LOOK UP DATA' via keyword searching



SELECT REGISTERED PROJECT PAGE: from 'search by keyword' page.
The keyword search returns a list of registered projects that contain the keyword. The 'description' shows the the sentence in the registered description that contains the keyword



DATA PAPER PAGE: from the 'registered projects' page which was populated based on the keyword.



References
Note all URLs and DOIs last visited on June 2, 2012.

Abrams, S., Cruse, P., & Kunze, J. (2009). Preservation is not a place. *International Journal of Digital Curation*, 4(1). Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/98/73

Agre, P. (1994). Living data. Wired, 2(11). Retrieved from http://www.wired.com/wired/archive/2.11/agre.if.html

Arzberger, P., Shroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D, et. al. (2004). Promoting access to public research data for scientific, economic, and social development. 3(29). Retrieved from https://www.jstage.jst.go.jp/article/dsj/3/0/3_0_135/_pdf

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 16(07). Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Association of Research Libraries. (2009). The research library's role in digital repository services. Retrieved from http://www.arl.org/news/pr/repositories-3feb09.shtml

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Messerschmit, D. G., Messina, P., Ostriker, J. P., et al. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

Australian National Data Service. (2011). About ANDS. Retrieved from http://www.ands.org.au/index.html on

Babbie, Earl. (2007). *The practice of social research* (11th ed.). Belmont, CA: Thomson.

Berman, F., Lavoie, B., Ayris, P., Cohen, E., Courant, P., Dirks, L., Friedlander, A., et al. (2010). Sustainable economics for a digital planet: Ensuring long-term access to digital information: Final report of the blue ribbon task force on sustainable digital preservation and access. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Borgman, C.L., Grappone, T., Strong, G., Goldman, J., & Wallis, J. (2011). Institute for digital research and education proposal: UCLA Data Registry System. Retreived from http://works.bepress.com/borgman/259

Borgman, C. L. (2007). *Scholarship in the digital age : information, infrastructure, and the Internet*. Cambridge MA: MIT Press

Borgman, C.L. (2009). The future is now: a call to action for the humanities. *Digital Humanities Quarterly*, 3(4). Retrieved from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

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

Boyd, D. & Crawford, K. (2011). Six provocations for big data. Proceedings from Oxford Internet Institute's: A Decade in Internet Time: Symposium on the dynamics of the internet and society. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Brase, J. (2004). Using digital techniques: Registration of scientific primary data. In Heery, R., Lyon, L. (Eds.) , *Lecture Notes in Computer Science*, 3232, 488-494. Heidelberg, Springer. DOI: 10.1007/978-3-540-30230-8_44

Brase, J. (2009). DataCite-A global registration agency for research data. Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (pp. 257–261).

Buckland, M. (2011). Data management as bibliography. *Bulletin of the American Society for Information Science and Technology*, 37(6), 34–37.

California Digital Library. (2012). EZID. Retrieved from http://www.cdlib.org/uc3/ezid/

Chaven, V., and Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10(Suppl 14). Doi: 10.1186/1471-2105-10-S14-S2

Choudhury, G.S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, 57(2): 211-220.

CODATA. (2011). Data citation standards and practices. Retrieved from http://www.codata.org/taskgroups/TGdatacitation/index.html

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

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)*. Retrieved from http://public.ccsds.org/publications/RefModel.aspx

DataCite. (2011). Why cite data? Retrieved from http://datacite.org/whycitedata

Dataverse Network. (2012). Retrieved from http://thedata.org/

Digital Curation Centre. (nd.d) What is digital curation? Retrieved from http://www.dcc.ac.uk/digital-curation/what-digital-curation.

Fitzpatrick, K. (2011). The humanities, done digitally. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/The-Humanities-Done-Digitally/127382/

Freeman, G. (2005). Library as place: Changes in learning patterns collections, technology, and use. In Library as place: Rethinking roles, rethinking space (1-9). Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub129/pub129.pdf

Gibbs, F. and Owens, T. (2011). The hermeneutics of data and historical writing. In Jack Dougherty and Kristen Nawrotzki (Eds.), Writing History in the Digital Age. Retrieved from http://writinghistory.trincoll.edu/data/hermeneutics-of-data-and-historical-writing-gibbs-owens/

Gil, Y. (2010). Provenance XG final report. Retrieved from http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle (Eds.), The fourth paradigm: Data-intensive scientific discovery (p. xix–xxxiii). Redmond, WA: Microsoft.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership and control: Toward empirical studies of access practices. *Knowledge*, 15, 355–372. Retrieved from http://scx.sagepub.com/content/15/4/355.full.pdf+html

Lofland, J., Snow, D., Anderson, L., & Lofland, L.H. (2006). *Analyzing social settings: a guide to qualitative observation and analysis*. Belmont, CA: Wadsworth/Thomson Learning.

Lynch, Clifford A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *Association of Research Libraries Bimonthly Report* (226): 1-7. Retrieved from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml.

Mayernik, M. S. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators. (Doctoral Dissertation). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042653

McPherson, T. (2008). Dynamic vernaculars: Emergent digital forms in contemporary scholarship. Lecture presented to HUMLab Seminar, Umeå University. Retrieved from http://stream.humlab.umu.se/index.php?streamName=dynamicVernaculars.

National Science Board Members. (2005). *Long-lived digital data collections: Enabling research and education in the 21st Century*. Washington, D.C.: National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

Organization For Economic Co-Operation And Development. (2007). *OECD Principles and guidelines for access to research data from public funding*. Paris: Organization for Economic Cooperation and Development. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Stewart, V. R. (1976). A primer on manuscript field work. *Midwestern Archivist*, 1(2), 3–20. Retrieved from http://minds.wisconsin.edu/handle/1793/44053.

Svensson, P. (2012). Beyond the big tent. In M. K. Gold (Ed.), *Debates in the Digital Humanities,* 36–49. University of Minnesota Press.

The William Blake Archive. (2010). http://www.blakearchive.org/blake/

Wallis, J. C., Mayernik, M. S., Borgman, C. L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: From vision to practical reality. Proceedings of the 10th annual joint conference on Digital libraries. Doi:10.1145/1816123.1816173.