A New Tool for Managing and Discovering Research Data: Creating the UCLA Data Registry

Rachel Mandell
UCLA Department of Information Studies, GSE&IS Building, Box 951520, Los Angeles, CA 90095.
Email: ramandell@gmail.com

The UCLA Data Registry will be 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 more researchers are interested in registering data than seeking data; that the Data Registry will be used mainly for gaining additional exposure to research output, and also discovering new collaborations and projects are taking place on campus. 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 new types of scholarship.

Introduction
Advanced digital technologies have enhanced the ability to generate vast amounts of data, and as a result the methods and practices of scholarship in many fields have undergone profound changes. Though these changes may appear to be most apparent in the sciences, they are not the only academic disciplines where scholarship practices have been affected. Scholars in the digital humanities and archaeologists working on virtual cultural heritage are now focusing on data use and production. 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. Whereas certain fields, especially in the sciences and social sciences, have well-established data repositories poised to take on new data, many fields, lack the support necessary for managing the data they are producing. Data registries are just starting to become internationally recognized as the possible solution to make orphan data discoverable.

Background of the UCLA Data Registry Project
According to Jim Gray, the data deluge has shifted our relationship to the production of 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 that with our ability to generate the amount of data that we do, we have moved beyond theory, such that the “data speak for themselves” (Anderson, 2008).

This extreme notion that theory no longer governs our knowledge claims is no doubt contentious. However, many scholars agree that along with the amount of data being produced, cyberinfrastructure, or “the distributed computer, information, and communication technology” required for a “knowledge economy,” (Atkins et al., 2003), has indeed created new opportunities for sharing data and developing collaborations. But sharing data still remains a complex problem, or as Christine Borgman (2012) argues, a “conundrum” due to the many factors that contribute to a researcher’s willingness or unwillingness to release data to others.

In addition to the social sciences’ adoption of digital information such as survey data and online maps, an emerging field known as the Digital Humanities are utilizing digitized source material, and exploring new methods of using technology to ask fundamentally humanist questions. In Kathleen Fitzpatrick’s (2011) article from the Chronicle of Higher Education’s Profhacker blog, what was soon to become 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”(n.p.). As this area started gaining momentum, it 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.).

Given the diversity of the current research data landscape and the need for enhanced data management services, university libraries have embraced new opportunities to get involved in early stages of the research process to help ensure the long-term preservation of valuable data (Gold, 2007). In addition to building digital services such as repurposing institutional repositories for digital data (Choudhury, 2008), Patrik Svensson (2010) argues that the library can also reclaim its position as a physical space that encourages collaborative learning. In a recent interview conducted at the Coalition for Networked Information (CNI) 2010 Fall Member Meeting, Gary Strong, University Librarian at UCLA, discussed how the library can provide the kinds of spaces and learning environments which are required to produce the new forms of knowledge that are emerging in areas such as the digital humanities.

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 publically accessible. With CENS’ official closing date of 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 will be expanded to include the greater UCLA research community (Borgman, Grappone & Wallis, 2011).

Research Questions
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 who generates or collects any form of data. The data seekers are researchers, faculty or students interested in obtaining primary research datasets, descriptions about research projects being conducted on campus, and contact information of those researchers. These categories of users are also fluid—a data producer can also be a data seeker. Users will come from disparate research interests and subject areas. A data registry should thus be both general enough to accommodate multi-disciplinary data and yet specific enough to contain detailed information about the data registered.

The UCLA Data Registry will be designed to serve the greater UCLA research community by collecting and making available surrogate records of research datasets produced by a broad range of researchers on campus. This joint venture between my advisor, Christine Borgman and the UCLA Research Library promises to facilitate the discovery of valuable research output. The goal of this project was to utilize my research in collaboration with a systems developer to design a tool that will effectively meet the research community’s needs. To this end, I formulated the following four research questions, which were used to guide the development of the Data Registry:

1. What data will be registered and what data do users hope to discover in the UCLA Data Registry?
2. How can the UCLA Data Registry benefit the work of researchers?
3. What factors motivate researchers to contribute or not contribute to the UCLA Data Registry?

Research Methods
Answering these questions and subsequently building the Data Registry requires an in-depth understanding of a wide range of research practices and data description. The best method used for gaining access to this type of detailed 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 identify with one, 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 or even whom they are working with. One researcher argued that she belonged to all three categories.

The initial recruitment for interviews were drawn from the pool of fellow IDRE awardees, as the grant placed a heavy emphasis on collaboration with other awardees whenever possible. A group of six of these grant winners were comprised of four scientists, one humanities scholar and one archaeologist, who considered herself to be part of all three general areas of research. From each of these participants, two more referrals were elicited. Although the goal was to obtain a nearly equal number of researchers in each of the three main categories, 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 offered names of researchers working within that framework. Therefore the resulting focus on the
humanities was by accident rather than design. The final breakdown was:

- Science: 5
- Social-science: 5
- Humanities: 9
- All three categories: 1

All interviews were conducted between January and April of 2012 and each lasted between 30 minutes and 1 hour. Each interviewee was asked the same 15 questions. These were open ended to give the interviewee a chance to expand and reflect on his or her work. All of the interviews were audio recorded. Given the scope and time frame of the project, formal coding was not used to conduct the analysis. Instead the data was analyzed according to the general principles held by grounded theory. The themes and categories emerged from the data rather than being imposed on the data as a system of a priori labels. This method of what Lofland, et. al., (2006) refer to as “focused coding” concentrates on understanding the corpus that accumulates as one goes through the data and looks for what can be discovered within the material.

**Results**

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

*Research Question 1: What data will be registered and sought in the UCLA Data Registry?* Of the 20 researchers interviewed, a total of 15 were interested in registering their data with the UCLA Data Registry. The largest group of six researchers belonged to the humanities followed by a group of five social scientists. The smallest group of three was the scientists. The researcher in all three categories was also interested in registering and the five remaining researchers either indicated that they were not interested or did not explicitly state whether or not they would be willing to contribute. The types of data that researchers would want to register are:

- 3D brain image scans
- 3D architectural models
- 3D archeological dig site models
- Analysis of art/museum objects restoration
- Images of tumors and clinical information
- Archeological data: descriptions of artifacts, sketches, scans, notes
- Visualization platforms
- Summarized statistics of social interactions among school children
- Database of ancient magical artifacts
- Twitter feeds

These types of data come in multiple formats, many of which are standard formats; some are open formats and others are proprietary. Some of the projects are constructed using highly specialized software, such as the 3D models.

Fewer researchers indicated that they would be interested in using the UCLA Data Registry to potentially find registered datasets. However, there are also fewer total researchers represented in the breakdown of interview data collected, as many researchers were not explicit about whether they would seek data in a system like the Data Registry or what they would be specifically interested in finding. In this case, three scientists were the group most interested in using the registry to find data, while only two humanists and two social scientists were interested in seeking data using this system. Nearly all researchers that indicated they would be interested in seeking data stated that they would want to find data that are similar to their own or to research projects in their fields.

*Research Question 2: How will the UCLA Data Registry benefit researchers?* The three main benefits that researchers identified were providing additional exposure to their work, providing additional methods and venues of disseminating their work, and the potential of using the registry as a UCLA research directory for information about the researchers on campus.

Additional exposure to one’s research was the most common benefit that researchers saw in a system like the Data Registry. 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). Therefore, this researcher is interested in registering solely to help her work gain additional avenues of exposure. Many researchers adopted a “why not?” attitude when asked to articulate possible benefits of the Data Registry.

In addition to the enhanced exposure of one’s work, researchers also saw the Data Registry as an environment that can be leveraged to disseminate data, as an alternative to the traditional, dominant method of data dissemination, which in many fields is still the journal article publication. One faculty member acknowledged that the journal article or even the monograph, “only reflects a sheer fraction of the data we generate” (Archaeologist, Faculty). Since the traditional publication is somewhat limiting, this researcher would be interested in registering data in order to disseminate more of his data.

One researcher in the sociology department echoed this point about needing to find additional outlets of disseminating data outside of traditional discipline-specific journals. However, her needs were somewhat different. She saw the Data Registry benefiting her work because her current research project combines multiple methods of inquiry, borrowing from anthropology and sociology. She stated that finding a journal to publish this type of...
multidisciplinary study has become increasingly difficult. Therefore the Data Registry would give her a chance to make her data and methods available to the rest of the community (Sociologist, Faculty).

The third benefit that researchers imagined the Data Registry providing was an active and up-to-date research profile and directory about the research projects being conducted on the UCLA campus. An architectural historian noted that there is “no clearinghouse of information about research projects on campus” (Architectural Historian, Faculty). While this may not have been one of the intended purposes of the Data Registry, the need has been articulated and is therefore valuable information for future development of the Data Registry.

Research Question 3: What are the motivations to contribute or not contribute to the UCLA Registry? Those researchers interested in registering their data gave multiple reasons for being motivated to do so. 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). A researcher in bioinformatics acknowledged that contributing to the Data Registry would be “huge step forward and huge asset,” he noted that in addition to fulfilling funding agency requirements, he could use the Data Registry as a place to find more data. “The data is difficult. We always struggle with getting enough data” (Bioinformatics, 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). One researcher 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).

However, not all interviewees shared this sentiment. Other factors contributed to some researchers’ motivations not to contribute to the Data Registry. A computational scientist stated that he would not be motivated 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 is 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. A graduate student in biochemistry also noted that stealing data can unfortunately be somewhat common in his community and 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 sometimes 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).

Discussion

The interviewees’ responses demonstrate that more researchers were interested in registering their data than actually seeking data. This may be somewhat surprising due to the ongoing struggle of advocating for the importance and value in sharing data (Borgman et. al., 2010; Buckland, 2011). Based on my data collected and supporting literature, one way to interpret these results is to see the Data Registry as more of a personal tool to be used by researchers to manage their digital content and research projects, rather than as truly a place to find reusable primary datasets.

Towards this end of remaining flexible and open, the Data Registry should be available for three different potential scenarios. The first scenario describes those researchers who already have stable environment for their data, such as a server or a repository. In this situation the registry would provide an excellent method of exposing their work. The second scenario allows those who are unwilling to share the actual data to generate a registration page with a description of their research. Lastly, for those willing to share data after they have published their work, they can opt for an embargo period on their data.

If the Data Registry is implemented as a personal tool, it needs describe the registered data using an open schema, allowing for the ability to provide free text descriptions of
data and very few required fields. While some standard must be chosen, it is important to avoid a ‘one-size-fits-all’ mentality. Given the diversity of research output being generated on the UCLA campus, the registry should not impose a standard that may limit the scope of the registered data. As Sayeed Choudhury points out, “technology alone does not engender transformation”(2008, p.212), and so the tools that do not fit a particular community’s needs, will not be utilized by that community. Therefore the Data Registry needs to accommodate for many different kinds of data and data needs.

One of the most profound implications that this research has revealed is that the culture of publishing and scholarly communication is starting to take a new form in multiple areas of research. Scholars working in the digital humanities are now utilizing advanced digital technologies to develop their arguments in formats and frameworks that analog, two-dimensional text could not support. An interviewee working in digital cultural heritage explained that he turned to the three-dimensional tools to study the Roman forum and questions that had been previously only explored in text form. He argues that these questions about space and social history couldn’t be fully articulated until they were in the 3D world.

The methods of publishing their work are also undergoing changes. Whereas the humanist tradition of publishing involved the tedious and arduous task of working at producing the single work or the career-defining masterpiece, the digital humanities are currently experiencing the more rapid creation and development of knowledge in episodic spurts (Strong, 2011). Along with this new form of publication, comes a new form of knowledge—the prototype as something to be tinkered with, discussed, and worked out collaboratively.

The requisite infrastructure may not yet be up to speed to truly support this kind of scholarship. Data management still remains a complex problem, as the researcher in the digital humanities pointed out that saving the visualization platform is vital to certain projects. Additionally, there must also emerge a new standard for citing and receiving academic credit for this type of work. Though the digital humanists are not the only group in need of a new form of standard citation. Groups such as CODATA and DataCite have been advocating for an adopted standard for citing research datasets for quite some time. Though their needs are very different, both areas of research are changing the face of scholarship and scholarly communication.

Conclusion
The UCLA Data Registry promises to provide a stable environment for research data produced on campus. Based on the interview data collected, more researchers are interested in registering data than seeking data. This indicates that researchers are in need of more tools that will help them to manage and disseminate their data, which may otherwise be ignored. Researchers also acknowledged that the Data Registry could function as a campus-wide research directory. This type of system can provide valuable information about the types of research being conducted on campus. Faculty would then utilize the Registry as a means of potentially finding new collaborations.

As of April 2012, the UCLA Data Registry is an early prototype. As it develops, we must also continue working with faculty to create a system that fits the needs of the community. Further consideration for how to manage, preserve and maintain access to new types of scholarship is still needed. 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 data continues to be produced in all academic fields. However, it still remains vitally important to consider the actual needs of the community, and in developing a new system, we aren’t also developing phantom uses of the tools we are creating.

ACKNOWLEDGMENTS
I would like to thank Professor Borgman, my advisor and mentor, the Data Practices team at UCLA including Jillian Walls, Laura Wynholds, Ashley Sands and Elizabeth Rolando, and of course, my family.

REFERENCES
*Note all DOIs and URLs last visited on May 10, 2012


Borgman, Grappone, & Walls. (2011). CENS library proposal to the Institute for Digital Research and Education


