
This study describes a web-based survey of Odum Institute research customers at the University of North Carolina at Chapel Hill. The survey was conducted to determine the applicability of digital repository software called the Dataverse Network in social science research workflows.

The software allows researchers greater personal control over ingest processes, bridging the gap between researchers and archives and possibly increasing submission rates of valuable data to archives. This archival technology provides tools that manage automated ingest, data cataloging, advanced search, as well as rights management issues. Archival tools also provide proper citation, creation of persistent identifiers, automatic creation of preservation formats, format migration, and statistical analysis of data. Customized branding and citation management can provide investigators collecting these data with a tool that will ensure that they get the credit they deserve. By working together with data producers, archivists can have more control of the digital curation process.

Headings:

Archives—Technological Evaluation

Archives—Digital Archive

Special collections—Technological innovations

Surveys—Internet

Social Science
Odum Institute Data Archive User Study: Exploring the Applicability of the Dataverse Network

by
Jonathan D. Crabtree

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

Chapel Hill, North Carolina
April 2011

Approved by

Helen R. Tibbo
Introduction

The research community is faced with expanding burgeoning collections of digital data. As researchers and scientists struggle to deal with this vast amount of information, they still have to continue their primary scientific work and would like assistance in this process (Science, 2011). A recent poll of Science’s peer reviewers shows that 20% of those asked were creating data sets larger than 100 gigabytes and 7% used data sets greater than 1 terabyte (Science, 2011). When asked where the respondents archive the data created by their research, over 50% claimed they stored the data in their labs (Science, 2011). Additionally, 38.5% reported that they archived their data on university servers while only 7.6% used community repositories (Science, 2011). This leaves most data to reside outside of archival repositories and beyond the care of data curators. This lack of stewardship places much data at risk and raises the questions of “what roles can digital archives play in the preservation process and when should they become involved in the data lifecycle”. Archivists and researchers hold varying views on how and when archivists should become involved in the research process.

To frame these views we can look to Ranganathan’s five laws of library science (Ranganathan, 1931) that remain foundational principles. Abrams et al. (2009) makes the loose analogy from these principles that underlie modern library practice to digital data
stewardship. Here the focus is on use and reuse of digital resources and preservation in service of use and future users.

… digital assets are preserved in order to be used. Furthermore, use entails that these assets be both discoverable and utile, that is, users can find assets of interest and the information content encapsulated into those assets are meaningfully exposed to their users. (Abrams et al., 2009)

To fulfill this vision the digital assets to be preserved will have to undergo many preservation operations. These operations can be subdivided into many individual tasks or services (Abrams et al., 2009). These curation services form the infrastructure needed to preserve data while adhering to the founding principles of library science (Raganathan, 1931). These services can be divided into categories that include identity, storage, characterization, catalog, annotation, fixity, replication, transformation, ingest and access services (Abrams et al., 2009). While a detailed exploration of the routines within each of these categories is beyond the scope of this paper, they all can be collectively defined as processes of digital curation.

The Digital Curation Center (DCC) of the UK states that digital curation “involves maintaining, preserving and adding value to digital research data throughout its lifecycle.” The DCC goes on to argue “The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence.” (DCC)

The term “digital curation” was first used in 2001 during a conference held by the Digital Preservation Coalition (Beagrie, 2006, p4) and was seen by attendees as an
essential beginning dialogue between archivists, libraries, information specialists and data managers (Beagrie, 2006). The application of these digital curation services can be seen as part of a lifecycle model that encompasses the entire research lifecycle (Higgins, 2008). Other commentators argue that the line between these services and who provides them could be blurred further into a digital curation continuum (Treloar et al., 2007).

The post-custodial archiving view described by Cook (1994) would have the digital objects remain in the control of their creator. Advocates of post custodialism argue that the content creators have the software to maintain the digital objects and deep knowledge to contextualize them. This is often the de facto approach taken when there is no centralized digital archive or when data production is distributed and no archiving mandate exists. Despite the pragmatic claims, post custodialism has not proven to provide uniform trustworthy archiving environments. The government of Australia used this approach for a number of years when they lacked a national digital archive but moved away from efforts to preserve content for the long-term through post custodial archiving (ANDS, 2007).

A majority of digital social science data continue to reside with their content creators, and thus in a post custodial and often risky environment. Historically, few social science data archives have played a role in data preservation until projects are over and researchers have released their data. This decreases the likelihood that important data will get from researchers to the archive and that archivists will work with producers to manage that data throughout the continuum. The concept of digital curation throughout
the lifecycle of data has been shown to be an effective approach (DCC) and could be advantageous when applied to the social science discipline. By applying digital curation processes throughout the lifecycle of data, the likelihood of long-term preservation and re-use of social science data may increase.

Many of these managed activities required for preservation can be made simpler and easier to accomplish if the process is begun earlier in the life of the research data. The goal of digital curation is to work with data creators all through the research process and provide the many curation services (Abrams et al., 2009) where they are needed to simplify the curation process. Researchers are interested in assistance with the curation and storage of their research data earlier in the research lifecycle but they need digital curation tools that offer them control over their important data (Feijen, 2011).

The Institute for Quantitative Social Sciences at Harvard (IQSS) is a founding member of the Data-PASS organization and as part of their consortial efforts they have created an open source digital data archive solution called the Dataverse Network (King, 2007). This application provides social science archivists and researchers with tools that can help in the digital curation process by performing many of the functions defined above. The Dataverse Network provides functionality that social science researchers can use early in the research data lifecycle and allows them more control in the process while at the same time allow archivist valuable input as well. This technology blurs the life cycle stages in research data development (Cook, 1994) while at the same time securing the data within the preservation and documentation systems of a trusted social science
digital archive. But would digital data archiving technologies like the Dataverse Network (King, 2007) be helpful to social science researchers? Would the researchers use these new tools? The goal of this paper is to investigate the current uses of these tools and the potential expanded use within the social science research community. Needs and attitudes surrounding these issues will be investigated by way of an archive user survey performed at the Odum Institute for Research in Social Science (Crabtree, 2010).

Social Science Background

We all live in a world surrounded by data, much of it generated through social science research and polling. In our everyday lives we hear news reports that reference data supporting the reporters’ claims. We read newspapers and browse Web sites that highlight public opinion surveys. We depend on data to make financial, health, and business decisions. Alarmingly, recent estimates show that the world is creating far more data than we are able to store (Collins, 2011). This raises the questions of “how will our society keep track and manage all this data?” and “how will we know what to save and preserve for the long-term?”

Access to empirical social science data is fundamental to successful social science policy development, research, and education. Students and teachers who wish to gain a deeper understanding of findings in economics, psychology, political science, sociology, educational research, and other social sciences must be able to discover and access the data that constitute these studies. Teachers and students in the natural sciences routinely
encounter the products of empirical social science in surveys, newspaper editorials, magazine articles, and other academic research products.

These sources are the results of researchers and scientists collecting vast amounts of data worldwide. Social science researchers often times collect their own data, but in many cases the data they use have been repurposed from an earlier study. For quality data to be available for reuse in later studies they must be documented and properly stored after the data has been collected. Nevertheless, a study at the Interuniversity Consortium for Political and Social Research has shown that only 20% of the data collected for projects funded by the National Science Foundation and the National Institutes of Health has been archived (Pienta et al, 2008). These results are disturbing given that both the NSF and NIH now require researchers to provide a data management plan that should include long term access (NSF, 2011). Data archives, which play an important role in the preservation of these data, must make efforts to fill these gaps. A group of United States based social science data archives with initial funding from the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) have formed the Data Preservation Alliance for the Social Sciences (Data-PASS) to help work in collaboration to preserve a greater amount of social science data (Data-PASS, 2011). A number of scholars in their respective disciplines are champions of documenting and archiving their data, but these are difficult tasks that seldom receive adequate funding and are often the first items cut from budgets.

The research community will continue to experience this issue until the tools for
Curating data are integrated into the data lifecycle (Green & Gutmann, 2007). The virtual archiving technology required to bridge the gap between the data producers and archives already exists and is in use at a few social science data archives (Odum & IQSS). The next step is enhanced awareness of the accessibility of this technology and its value to both researchers and the scholarly community. Data-producing organizations and scholars are typically supportive of archiving the materials produced, but economics and workflow issues tend to inhibit attempts at comprehensive archiving. Researchers must often move from one project to the next with minimal downtime in order to maintain the economic feasibility of their data collection organizations, yet this frenetic pace often undermines attempts to adequately archive valuable data. The costs associated with rehiring qualified staff when new projects arise — or of keeping staff on the payroll with no outside financial support — are considerable. Organizations and researchers must maintain a queue of new projects and opportunities for implementation in order to justify their organizations’ continued existence, a strategy that reserves little time and resources for adequately archiving data.

Collecting social science research data, particularly public opinion data, is ensnared in this problematic process, resulting in the loss of numerous valuable datasets (Parry et al, 2006). The data that underpin many social science research studies, policies, and discoveries have not been consistently archived despite mandates from funding agencies such as the National Institutes of Health and National Science Foundation. This is due primarily to the enormous degree of post-project effort required to prepare data for archiving. Funding agencies are moving toward requiring detailed data management
plans in efforts to stress the importance of data archiving and reuse (NSF, 2011). But without integrated tools and simple workflows this will become a burden to researchers. To help ease this weight, data archive managers and data producers must work cooperatively. The virtual archives in the Dataverse Network may be a key factor in engendering this cooperation (King, 2007).

**Dataverse Network Tools and its Digital Virtual Archive**

The Dataverse Network is an open source data repository that allows social science researchers and archivists tools to store, access, analyze, and document digital data. The application allows researchers to search and discover a vast amount of social science research data. The system provides a federated approach to digital archives and uses the Open Archives Initiative Protocol for Metadata Handling (OAI-PMH) to distribute and synchronize the holdings of social science data archives using the system. Researchers can access all the data in the network from a single archive website. Once the data has been located social science researchers can download, subset and analyze the study materials. Properly formatted citations are provided so users can reference the online results in published works. Persistent identifiers are assigned to each study ensuring the authenticity of the acquired data. To track any data corruptions Unique Numeric Fingerprints or UNFs are assigned to digital statistical formats. These UNFs are unique statistical computations that provide unmistakable identity for statistical data in a checksum format (Altman, 2008). To properly ingest research data into the application, web based metadata entry forms allow users to document data following onscreen
instructions to ensure the proper format is followed. The metadata is recorded using the Data Documentation Initiative (DDI) metadata standards and is encoded in XML. Unique to the repository are the automated metadata creation tools with the ability to read modern statistical formats and create variable level metadata without need for extensive hand coding by the social science researcher.

A powerful feature of the Dataverse Network is its ability to manage a large number of different archival collections within a single instance of the Dataverse server. Archivists as well as researchers have the ability to create individual or project based Dataverses within a single Dataverse Network installation. As the number of individual Dataverses grow, the entire Dataverse Network has access to them. These individual Dataverses can be created by local archivists or individual researchers, or harvested from other servers via the built-in OAI-PMH metadata exchange facilities. Along with bringing content under archival stewardship earlier in its life cycle, the system’s ability to manage multiple archival collections with different access rights and policies has the potential for significantly impacting traditional digital data archival workflows.

Because each Dataverse within the larger Dataverse Network can be customized to the needs of the researcher it allows great flexibility. Researchers can provide customized policy statements or embargo dates to satisfy their needs. In addition, the user interface can be customized to look exactly like the researcher’s personal homepage or research project site. Archivists and researchers can create these customized templates for individual Dataverses that allow the creation of “virtual archives” within the larger
Dataverse Network managed by social science data archives. While these virtual archives look and feel just like part of the researcher’s personal homepage or project website, they physically exist within the professional archive and curators have the ability to review documentation and assist in the creation of proper metadata records. It is the ability of the virtual archive to provide custom project requirements and customized interfaces that has the potential for impacting the future of social science archive workflows.

In traditional digital data archival workflows, the physical transfer of the material from the researcher to the archivist places the data in the custody of the archive after the project is complete. The research teams have to struggle to find time to assemble the data, materials, and any existing documentation and forward these to the archive for ingest processing. This process occurs after the funded project is over. Ingesting involves preparing data for archiving, de-identifying personal and confidential information, creating standard file formats, building any necessary metadata, and documenting this process. The depth and quality of the ingest process varies greatly, and the effort required to assemble the components often limits the amount of materials archived. In many cases it is clear that researchers have already moved to new projects and do not have the time to follow through with archiving previous datasets.

Virtual archiving can provide digital curation tools to researchers while at the same time allowing professional archivists input during the process. With virtual archiving, researchers begin using simple archival tools within the Dataverse Network earlier in the research process as the data is created. These simple Web-based tools allow
researchers and their staff to manage their data and its documentation throughout the lifecycle of the project. The virtual archives that result recreate the look-and-feel of the home-institutions’ Web sites and provide users the same ease they had when placing data on their local sites. The simple ingest procedures provide metadata validation routines that assist in documentation and even prompt researchers to enhance their metadata. When quantitative data is ingested, automated routines create detailed variable-level metadata without requiring costly manual procedures.

As these research teams begin placing their datasets into these customized virtual archives, ingest tools collect and verify much of the required metadata. Persistent identifiers are created along with the creation of Unique Numeric Fingerprints to ensure data authentication (Altman, 2008). Documenting the data becomes a simpler task when it is performed during the data collection phase of a project. When the research team releases the dataset by setting appropriate permissions, the archival submission process is complete. Though the whole process seems to the research team to be local on their Web site, the data is in fact stored at the remote archive site. Trained archivists manage the process and ensure that documentation and archival formats are created. After datasets are archived, users will be able to search for the data from the producer’s local Web sites and have access to online analysis tools for quantitative data. Credit and acknowledgement for the data will remain with the research teams themselves since their customers and users will recognize the customized branding. Unless the users notice the hyperlink locations they will never know they were actually accessing data on the larger Dataverse Network maintained by the social science data archive.
All of these features provide an incentive for researchers and project staff to manage research data within their own customized Dataverse. Placing data in the system is just as simple as uploading data to their local website or file server and the automated templates and metadata creation tools assist with metadata creation. Since the research project is still in progress, project staff are available to assist in this documentation and since the project is current it is easier to document. If integrated into researcher’s and data producer’s workflows, these tools may have a significant effect on the workflows within the social science data archives. Archivist will have input and control over documentation of research data much earlier in the lifecycle. This process can seem like the post-custodial blurring of the data lifecycle concept with archivist assisting data producers (Cook, 1994), but with a major exception. The data will not physically reside with the creator. It will be preserved and curated during the entire process by professional archivists. The concepts of digital curation throughout the entire lifecycle of the data will be utilized to help ensure the protection of the research data. Before these tools can become part of the research workflow they need to be acknowledged by the social science research community as helpful in the research process. Some archivists believe these virtual archives can be helpful but would social science researchers and their project staffs embrace these new additions to their workflows?

Archive User Survey to Connect with Researchers

The successful application of virtual archives by the Institute for Quantitative Social Science at Harvard University (IQSS) has shown that institutes, centers, journals,
research projects, and individual scholars can benefit by integrating these services into the lifecycle of research data (King, 2007). As a part of this masters research project, I conducted a survey of Odum customers investigating social science researchers’ awareness of recent NSF requirements for data management plans (NSF, 2011) and the knowledge of current archival tools available to help with digital curation. The survey was designed to provide indications of researchers’ needs in data management tools, willingness to collaborate, and services they expect of such new archival tools. Could the needs identified in this survey be filled by the use of the virtual archives within the Dataverse Network application? Social science data archives are well positioned to assist in this effort but do researchers need the help and are they aware of the potential services offered by archives?

It has been demonstrated that building partnerships is critical in the process of moving data from the researcher’s environment to that of the data archive (Crabtree & Donakowski, 2006). It is these personal relationships that grease the wheels of data donations. Technology alone will not solve data management problems. Barreau’s 2008 work notes that even with new tools, managing electronic data could be challenging (Barreau, 2008). This is especially true of social science research. Social science data archives will have to build partnerships and work closely with data producers to solve these issues. The methodology used in the survey keeps this in mind and builds on existing relationships within the local social science community at the University of North Carolina at Chapel Hill.
**Survey Methodology**

In this spirit of building relationships, this survey was designed to focus on local researchers and built upon already established connections between the survey respondents and the Odum Institute. The Odum Institute is well recognized among this population as a partner in social science research and data archiving. This study used the population of the social science community at the University of North Carolina at Chapel Hill as a sampling frame. More specifically, the up-to-date contact list of researchers maintained by the institute was the sampling list used. The contact list is updated automatically each semester through use of programs that poll the UNC Campus Light Weight Directory Application Protocol server. The total number of e-mails used in the survey distribution was 7,763 addresses. These addresses represent a wide range of researchers, students and staff in the social science community at the University of North Carolina at Chapel Hill. Many of these are researchers and staff members at the institutes and centers doing social science research at the University. This sample of potential Odum customers represents a large collection of social science faculty, students and staff members of the university community. Demographic questions were added to determine a more fine-grained breakdown of each respondent’s academic roles. After the invitation e-mail (see Appendix 2) was delivered, I had 59 users request to be removed from the sample. The removal of this group from the sample gave 7,704 as a total number of participants. The study had 1,174 users attempt to complete the survey, for an overall response rate of a little over 15% (Crabtree, 2010).
Social science faculty is a critical customer base for the Odum Institute. It is important for the archive to understand the needs of this faculty, as they are the driving force behind much of the social science research done at the university. Contained within the sample were 961 e-mail addresses in the faculty category; the data included 279 faculty responses. This response rate of over 29% provides substantial support for positive inferences across the general population of social science faculty members (Crabtree, 2010).

The survey instrument was administered as a Web-based survey using Qualtrics survey technology. A survey is appropriate for this population given that the respondents often use surveys in day-to-day research. E-mail invitations (see Appendix 2) were delivered to each participant that contained an Internet link to the survey instrument (see Appendix 1). Reminder e-mails were sent to participants who had not responded at the end of the first and second weeks the survey was in the field. E-mail correspondence with the participants was planned so the customers that had questions about the survey could get answers in a timely fashion. Participants who wished to be removed from the reminder list were removed immediately to prevent any negative backlash from the study.

The design of good survey questions is critical in this study. It is not surprising that the most challenging part of this study was the design of the survey instrument. Admitting a pronounced bias toward the use of these modern archiving tools is the first step toward controlling that bias. To further assist in this mitigation of bias risk, I consulted with the Odum Institute survey methodology group to advise in the design of
the survey instrument. The survey questions were formed with all due diligence so they would not purposely lead the respondent to the answer and truly gather the opinion of the researcher. Simple, clear and direct questions were used to provide quality variables for final analysis. Contingency questions were utilized in an effort to focus the target group and reduce the survey time for respondents that are not collecting social science data. I did not want to force respondents who are not collectors of original social science data to answer questions they may not feel comfortable answering or for which they do not have knowledge.

Questions to track independent variables such as academic status and recent research activity were designed to aid in final analysis. A primarily quantitative approach was used to analyze the data. The questions were designed as a mix of single response questions and questions using a Likert Scale that allowed the respondent to rate their answer. The survey was designed in a streamlined fashion to be as short in duration as possible. The goal was for participants to finish the complete survey in less than ten minutes. The final mean for the survey duration was eight minutes and nineteen seconds, falling well within the ten-minute goal. The goal of this analysis is to provide indications of researchers’ needs for data management tools, willingness to collaborate with digital social science archives, and services they expect of such tools.

**User Survey Analysis**
One of the critical functions of a research university is the ability of its faculty to secure grant funding. The group of faculty that administrates these grants is an important subset of the social science community at the university. To help define this group the participants in the survey were asked their primary university or employment status (Fig 1).

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assistant Professor, Tenure-Track</td>
<td>36</td>
<td>3%</td>
</tr>
<tr>
<td>2</td>
<td>Assistant Professor, Fixed-Term</td>
<td>29</td>
<td>3%</td>
</tr>
<tr>
<td>3</td>
<td>Associate Professor with Tenure</td>
<td>48</td>
<td>4%</td>
</tr>
<tr>
<td>4</td>
<td>Associate Professor, Fixed-Term</td>
<td>16</td>
<td>1%</td>
</tr>
<tr>
<td>5</td>
<td>Professor with Tenure</td>
<td>70</td>
<td>6%</td>
</tr>
<tr>
<td>6</td>
<td>Professor, Fixed-Term</td>
<td>6</td>
<td>1%</td>
</tr>
<tr>
<td>7</td>
<td>Lecturer/Instructor/Adjunct Professor</td>
<td>38</td>
<td>3%</td>
</tr>
<tr>
<td>8</td>
<td>Professor of Practice</td>
<td>3</td>
<td>0%</td>
</tr>
<tr>
<td>9</td>
<td>Post Doctoral Researcher</td>
<td>33</td>
<td>3%</td>
</tr>
<tr>
<td>10</td>
<td>Doctoral Student</td>
<td>274</td>
<td>25%</td>
</tr>
<tr>
<td>11</td>
<td>Master Student</td>
<td>135</td>
<td>12%</td>
</tr>
<tr>
<td>12</td>
<td>Undergraduate Student</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>13</td>
<td>University Staff</td>
<td>250</td>
<td>23%</td>
</tr>
<tr>
<td>14</td>
<td>State Agency Staff</td>
<td>21</td>
<td>2%</td>
</tr>
<tr>
<td>15</td>
<td>Private Agency Staff</td>
<td>13</td>
<td>1%</td>
</tr>
<tr>
<td>16</td>
<td>Other (Specify):</td>
<td>116</td>
<td>11%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,089</td>
<td>100%</td>
</tr>
</tbody>
</table>

Fig. 1 Question: What is your primary university or employment status?

These primary appointments were processed during the recoding of the data to represent four populations of interest. Group 1 was defined to represent the participants who hold faculty appointments or professorships. These participants are recorded using selections
one thru nine in Figure 1. Group 2 consisted of undergraduate, masters, and doctoral students who responded to the survey (Fig. 1). Staff members and the “other” category make up Groups 3 and 4, respectively. The “other” category also contains alumni and professors with emeritus status who no longer work in the field.

Additional questions were asked to determine recent grant recipients within each group. In the past 5 years 437 of the respondents have received 607 total grants or fellowships. Many of these grants will either collect original data or re-use existing data. When analyzing the results, 63% of the participants in this survey engaged in research that collects original data as part of the grant (Fig. 2).

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>663</td>
<td>63%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>383</td>
<td>37%</td>
</tr>
<tr>
<td>Total</td>
<td>1,046</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Fig 2 Question: Does your research involve the collection of original data? (Assuming a 95% Confidence Interval of 3%)

The Institute is interested in researchers even if they do not collect original data. Often researchers who do not collect original data tend to be great users of secondary data. One of the primary missions of the Institute is to provide data and services to this research population as well. Many of the services offered by the Dataverse Network are designed to help these users. When respondents were asked if they used secondary data in their research, 524 say they do.
Beginning in February, 2011 the National Science Foundation will require all grant submissions to include a two-page data management plan (NSF, 2011). The survey shows that 47% of the Odum Institute user community is not aware of this new requirement (Fig. 3).

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>504</td>
<td>53%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>445</td>
<td>47%</td>
</tr>
<tr>
<td>Total</td>
<td>949</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 3](image-url) Question: Are you aware of the current trends requiring grant submissions to include data management plans for data access and preservation? (Assuming a 95% Confidence Interval of 3%)

An educational campaign is clearly needed. The Institute teaches short-courses on data management and tools within the Dataverse Network that can help with these new requirements. The Institute would like to see a greater number of these researchers knowledgeable in this area.

For the respondents who did collect original data they were asked if they archive the data once their projects are completed; most think they do. But further responses reveal that many of them have varying ideas about what “archiving” data means. In fact, only 18% of those surveyed submit data to an official archive as part of their research practices, and 56% thought placing data on a disk, USB drive, or hard drive were proper archival methods (Fig. 4). These results are similar to the study published in Science (Science, 2011).
Researchers need to have a better understanding as to what proper archiving entails. The use of the word “archive” is often confused with process of making a digital backup. O’Toole describes what makes records archival is “neither age nor appearance, but rather content, meaning, and usefulness.” (O’Toole, 1990) Researchers’ backup copies of data stored on USB drives or on websites often lack the documentation needed to be useful for researchers in the future. In addition, these data will require format migration over the years as statistical formats become obsolete. Professional archives create preservation capable formats, then monitor changes in those formats and migrate archived data to new formats when necessary. Social science digital archives offer many such curation services ranging from preservation redundancy to metadata creation in efforts to provide
access to and usability of these data for generations of future researchers. The Dataverse Network has the potential to allow researchers who currently place data on their websites to just as easily place that data within a proper archive environment that provides a full range of digital curation services (Abrams et al., 2009). Proper archives providing these services should include appraisal, ingest, arrangement, description, and reference as well as user access (O’Toole, 1990). Many of these concepts originated in traditional archives but still hold value for their digital counterparts. Results of this study elucidate the need for greater education in the social science user community in the area of archiving and data management. The promotion of the Dataverse Network for use in the social science community could help this effort.

Despite this misconception and apparent lack of detailed knowledge of proper archival practices, researchers also indicated interest in learning about tools to simplify and assist in the archival process (Fig. 5).

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>469</td>
<td>81%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>108</td>
<td>19%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>577</td>
<td>100%</td>
</tr>
</tbody>
</table>

Fig. 5 Question: Would tools and services that simplify the process of archiving data increase your likelihood of consistently archiving your collected data? (Assuming a 95% Confidence Interval of 3%)

High on the researchers’ priority list are services and tools related to data management and data management plans. When asked what other archive-related services or programs they would like to see, many respondents indicated interest in learning more
about the archive services currently available. This evident eagerness should be seen by archivists as an invaluable opportunity to educate researchers as well as increase future submission rates.

As noted previously, a large number of respondents were not aware of the new National Science Foundations requirement for data management plan in grants (NSF, 2011). One would predict that those respondents who are active in grant work should be more aware of the requirement. A cross tabulation of those participants who collect original data and know about the new NSF requirement show that 58.6% are somewhat more likely to know about the policy using a Chi-Square value of 14.67 where P< .0001. When comparing the results across the user groups, faculty members have the highest likelihood of knowing about the policy at 65% with a Chi-Square value of 38.56 where P< .0001 as well. These results are to be expected but the number of participants who do not know about the policy still suggests that social science data archives need to provide more education on the issue. A current initiative by the Data Preservation Alliance for the Social Sciences (Data-PASS) organization seeks to provide data management plan templates as well as educational information to help bridge this gap.

When asked if once their research projects are complete do they archive the data most researchers think they do as noted above. From a closer look at the answers for question seven (see Appendix 1) one sees that in some cases the individuals who do not archive the data typically feel they must destroy the data due to IRB stipulations or other privacy related reasons. In many cases researchers simply either did not know how or
what it means to archive their data. In many cases it may be that social science researchers who collect original data are just not aware of the many options for archiving available today. This lack of knowledge in archiving is further shown within the data that greater than 90% of Odum Institute users are not familiar with the Odum Institute archive services (Fig. 6).

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>43</td>
<td>9%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>415</td>
<td>91%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>458</td>
<td>100%</td>
</tr>
</tbody>
</table>

Fig. 6 Question: Are you familiar with the Odum Institute Archive Services? (Assuming a 95% Confidence Interval of 3%)

Of the respondents that were familiar with the services offered by Odum only a very few had visited the website in the past 12 months and almost none have archived their data recently. Since 72% of the respondents claim they archive their data but very few utilize the Odum Archive it is important we figure out what they are doing. A qualitative look at responses to question 12 (see Appendix 1) reveals a concerning problem. Only 18% of the places used by respondents to archive data remotely resembled services of a true archival process as describe by O’Toole (1990). Clearly many researchers lack knowledge of how to optimally management and archive research data although we have seen that researchers are interested in learning about tools to simplify and assist in the archival process (Fig. 5). While we have seen interest in learning new data management tools (Fig 5), an even more detailed look shows that the faculty and student groups display a very high likelihood that they would be interested in these new tools with affirmative answers of 82.2% and 85.1% respectively with a Chi-Square of
6.91 and P of .075. When asked to further describe these areas of interest, a large number of respondents placed high priority on services and tools related to data management and data management plans (Fig. 7).

<table>
<thead>
<tr>
<th>Question</th>
<th>Extremely Interested</th>
<th>Very Interested</th>
<th>Moderately Interested</th>
<th>Slightly Interested</th>
<th>Not at all Interested</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example Data Management Plans</td>
<td>29%</td>
<td>30%</td>
<td>23%</td>
<td>10%</td>
<td>8%</td>
<td>571</td>
</tr>
<tr>
<td>Data Management Plan templates</td>
<td>31%</td>
<td>32%</td>
<td>21%</td>
<td>9%</td>
<td>7%</td>
<td>570</td>
</tr>
<tr>
<td>Organizational/Researcher commitment to archiving project data</td>
<td>17%</td>
<td>21%</td>
<td>34%</td>
<td>15%</td>
<td>13%</td>
<td>558</td>
</tr>
<tr>
<td>Consultation on data management issues</td>
<td>23%</td>
<td>27%</td>
<td>28%</td>
<td>13%</td>
<td>9%</td>
<td>568</td>
</tr>
<tr>
<td>Training on data management program implementation</td>
<td>23%</td>
<td>27%</td>
<td>28%</td>
<td>13%</td>
<td>9%</td>
<td>568</td>
</tr>
</tbody>
</table>

Fig. 7 Question: For each of the following archive training, resources, and functions, please indicate the degree to which they are of interest to you and your research. For proposal development:

In addition, the quest for tools that assist with data dissemination and authenticity are also in high demand by researchers as most are extremely to moderately interested in these tools (Fig. 8).
<table>
<thead>
<tr>
<th>Question</th>
<th>Extremely Interested</th>
<th>Very Interested</th>
<th>Moderately Interested</th>
<th>Slightly Interested</th>
<th>Not at all Interested</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide data citation for users</td>
<td>49%</td>
<td>22%</td>
<td>21%</td>
<td>5%</td>
<td>3%</td>
<td>166</td>
</tr>
<tr>
<td>Customized website for archive access that retains recognition of the data producer</td>
<td>45%</td>
<td>27%</td>
<td>18%</td>
<td>6%</td>
<td>4%</td>
<td>166</td>
</tr>
<tr>
<td>Access controls and user-based authentication</td>
<td>45%</td>
<td>27%</td>
<td>19%</td>
<td>6%</td>
<td>3%</td>
<td>165</td>
</tr>
<tr>
<td>National and/or international-based discovery network</td>
<td>27%</td>
<td>24%</td>
<td>31%</td>
<td>12%</td>
<td>6%</td>
<td>163</td>
</tr>
<tr>
<td>Online analysis tools for users</td>
<td>40%</td>
<td>21%</td>
<td>21%</td>
<td>12%</td>
<td>6%</td>
<td>166</td>
</tr>
<tr>
<td>User-based permissions for project data access</td>
<td>38%</td>
<td>30%</td>
<td>22%</td>
<td>7%</td>
<td>3%</td>
<td>166</td>
</tr>
</tbody>
</table>

Fig. 8  Question: For each of the following archive training, resources, and functions, please indicate the degree to which they are of interest to you and your research. For data dissemination and authenticity:

All of these features are offered inside the Dataverse Network. If researchers would use these existing tools some of these concerns could be met. The final question of the survey really sums up the outcomes. When asked what other archive-related services or programs would they like to see many respondents were interested in learning more about the archive services available (see Appendix 1). By looking thru the entire open-ended questions with a qualitative eye it seems like a strong possibility exists that the respondents of this survey are thirsty for more information about Odum’s archive services.

Current Projects Exploring Virtual Archives

The successful deployment of the Dataverse Network tools at IQSS has prompted ongoing demonstration projects at Odum that are examining the benefits and issues surrounding deployment of virtual archives with existing Odum customers. Funding from
the Institute for Museums and Library Services has allowed the initial development of virtual archives for five members of the National Network of State Polls (IMLS). The goal is to examine the workflows of public opinion survey centers and design systems and procedures to integrate the use of virtual archives earlier in the data collection lifecycle. Though this program is only in its first year, it is offering encouraging signs. Early results from phone interviews seem to agree with the results of the Odum user survey (see Appendix 1). Researchers and data producers do want assistance in data management. In these preliminary interviews researchers have expressed the willingness to embrace the idea of digital curation earlier in the life cycle of the survey data.

The public opinion research community has expressed interest in documenting survey methodology much earlier in the survey research process. The American Association for Public Opinion Research (AAPOR, 2011) has been working with its membership to develop a “Transparency Initiative” that will publish the methodology of public opinion polls at the time they are reported on in the news or in publications. This is typically long before the data has reached the social science data archives and metadata has been created. Often many interim releases of a poll will be reported as the dynamics of an election or public issue unfold.

With a typical archival workflow the metadata record would not exist at this time so it will require a change in procedure. By using virtual archives the polling agency will have access to the archive workflow earlier in the process making this initiative possible. Polling agency staff can create and update metadata records for the study long before the
data are ready to be deposited. With the assistance of trained archivist the metadata can also be recorded and checked prior to the release of any publications that use references to that polling data. The polling agencies will have full control of the datasets and their release dates. Since polling agencies are already accustomed to placing preliminary data on their websites it will not be a radical change in workflow to use these new virtual archives. Yet at the same time, archivist will have early access and influence in the ongoing digital curation of the data.

An example of these influences can be seen in the AAPOR Transparency Initiative project. In the AAPOR initiative the existence of the Dataverse and its virtual archives helps researchers avoid creating a custom set of metadata requirements when useful standards already exist. The built-in entry template in the virtual archives uses the Data Documentation Initiative (DDI) standard that is widely accepted by the social science community as their unified standard. Collecting the new metadata and recording it in this standard allows the exchange of catalog records among social science archives. The AAPOR Transparency Initiative (AAPOR) also requires monitoring of this metadata for its proper use and completion. Without using a published standard this reporting and auditing of compliance would have been very time consuming and expensive.

Current initiatives and projects look favorably toward the adoption of virtual archives. The ability for the social science data archivist to assist early on in the research life cycle of data is powerful. Current projects have shown that many advantages can be gained. In agreement with the Odum Archive user survey it seems researchers have been
willing early on to embrace the use of virtual archives and allow archivist early digital curation input.

**Final Thoughts**

This research effort focused on understanding the needs of Odum data archive users and determining if services in data management that the archive has the ability to offer would be used. Social science researchers can benefit from the tools offered by the Dataverse Network. The Dataverse application has been in use at Harvard University for several years and has been well received in that environment. For the application to be adopted in a more wide ranging fashion the workflow of both the researcher and the archive will need to be modified.

The changes will not be simple to enact. Changing the workflows and procedures used by social science researchers is an undertaking that will take time to implement. Archives must show the willingness to work with researchers to prove that new tools for virtual archiving will save time and provide compliance in the areas of sponsor-mandated data management. An educational effort will be required to inform the research community of the advantages. If done successfully the archives can provide a valuable service to the researchers and increase submission rates, thereby providing additional data for secondary research analysis. As submissions increase this in turn will increase the number of studies available and result in an increased usage of the archive materials.
This will not happen overnight; archives must work to build partnerships with data collecting organizations and individual researchers to make this happen. The study presented here received a strong response rate from our UNC faculty. Hence, we can make inferences across the population of potential Odum Institute archive users. In many cases the respondents seem to be receptive to learning about new tools that may offer assistance. Many of the survey questions pointed to a lack of knowledge in archival policy. This was somewhat expected but it is astonishing to see it on paper. Indicators also show that there is a willingness to learn about this area. Users are somewhat familiar with the new initiative by funding agencies to require data management plans and would like some help in that area.

The Dataverse Network offers a powerful digital curation tool for archives and researchers. As users understand how the customized Dataverses can be managed and developed they will learn to apply these to their projects. Researchers may be receptive for assistance in data management but they still want control over their valuable data (Feijen, 2011). The ability to control policies for access to research data while at the same time modify the look and feel of the user interface is a strong combination. The use of virtual archives can be a useful tool in this regard. Once researchers become accustomed to placing their data in customized Dataverses they will receive the valuable academic credit for the creation and development of the data. While under the covers archivist are assisting with documentation and organization by providing tools for metadata creation as well as providing preservation services for all the data. The injection of archival assistance earlier in the research data lifecycle will likely increase archive submission
rates (Green & Gutmann, 2007) and has the potential for ensuring a higher quality archival object. Tools such as Unique Numeric Fingerprints will be used in a more widespread manner, thereby ensuring the integrity of repurposed data (Altman, 2008). Use of virtual archiving also means that researchers would spend more time focusing on new research projects and less time reformatting data for public distribution. If virtual archiving becomes pervasive in research workflows, a much larger percentage of publicly funded data collections will be archived for future use. Social science data archives must strive to increase the submission rate of research data. The Dataverse Network and its virtual archiving can be a critical tool in that effort.
Acknowledgements

I would like to convey a world of gratitude to the Odum Institute for Research in Social Science for the freedom to work on projects advancing digital archival research. I would also like to thank all our Data-PASS partners for assistance in the common goal of better social science data preservation and access. In addition, I would like to thank the Institute for Museums and Library Services for the funding of the deployment of virtual archives in a demonstration project for public opinion data collection centers. IMLS National Leadership Grant 2010: Award Number LG-07-10-0240-10
References

AAPOR, Transparency Initiative, Retrieved March 2011 from
http://www.aapor.org/Content/NavigationMenu/AboutAAPOR/TransparencyInitiative/default.htm


Crabtree J., (2010), Odum Archive User Study 2010,
http://hdl.handle.net/1902.29/10533

Data-PASS, Data Preservation Alliance for the Social Science,
http://www.icpsr.umich.edu/icpsrweb/DATAPASS/

Data-PASS, (2011), Data Preservation Alliance for the Social Science,
http://www.icpsr.umich.edu/icpsrweb/DATAPASS/about.jsp

DCC, Digital Curation Center United Kingdom,
http://www.dcc.ac.uk/digital-curation/what-digital-curation

DDI, Data Documentation Initiative, Retrieved March 2011 from
http://www.ddialliance.org/

DVN, Dataverse Network Repository Software,
www.thedata.org

Feijen M., What Researchers Want, Retrieved March 10, 2011 from


IQSS, Institute for Quantitative Social Science at Harvard University, http://www.iq.harvard.edu/


Odum, Howard W. Odum Institute for research in Social Science, www.odum.unc.edu

O’Toole, J.M. (1990), Understanding Archives and Manuscripts, Archival Fundamentals Series, Society of American Archivist


presented at the annual meeting of the American Sociological Association Annual
Meeting, Sheraton Boston and the Boston Marriott Copley Place, Boston, MA. Retrieved

Ranganathan, S.R. (1931), The Five Laws of Library Science, Madras

6018 pp. 692-693 DOI: 10.1126/science.331.6018.692, Retrieved March 26, 2011 from
http://www.sciencemag.org/content/331/6018/692.short

Treloar, A., Groenewegen, D., Harboe-Ree, C., The Data Curation Continuum: Managing
Data Objects in Institutional Repositories, *D-Lib Magazine*, Volume 13, Number 9/10,
Appendix 1

Odum Archive User Survey Instrument

INTRO: Thank you for participating in the Odum Archive User Survey. The data you provide will be used to enhance future archive services. This survey is also being conducted in partial fulfillment of School of Information and Library Science degree requirements.

Your participation is voluntary and your answers are completely confidential. You may skip any question you choose not to answer.

1. Please enter the name of the institution/organization and department with which you are primarily affiliated.
   [OPEN TEXT]

2. What is your primary university or employment status? Select one.
   1. Assistant Professor, Tenure-Track
   2. Assistant Professor, Fixed-Term
   3. Associate Professor with Tenure
   4. Associate Professor, Fixed-Term
   5. Professor with Tenure
   6. Professor, Fixed-Term
   7. Lecturer/Instructor/Adjunct Professor
   8. Professor of Practice
   9. Post Doctoral Researcher
   10. Doctoral Student
   11. Master Student
   12. Undergraduate Student
   13. University Staff
   14. State Agency Staff
   15. Private Agency Staff
   16. Other (Specify): 

3. In the past 5 years, which of the following grants or fellowships have you been awarded? Select all that apply.
   1. National Endowment for the Humanities (NEH) grants or fellowships
   2. Fulbright Grants for Scholars or Professionals
   3. Agency for Healthcare Research and Quality (AHRQ) grants or fellowships
   4. National Institutes of Health (NIH) grants or fellowships
   5. Guggenheim Fellowship
   6. National Science Foundation (NSF) grants or fellowships
   7. Centers for Disease Control and Prevention (CDC) grants, fellowships, or cooperative agreements
   8. Institute for Museum and Library Service Grant
   9. Other Federal Grants (Specify):
10. Other Private Grants (Specify):
11. I have not been awarded any grants or fellowships in the past 5 years.

4. Are you aware of the current trend requiring grant submissions to include data management plans for data access and preservation?
   1. Yes
   2. No

5. Does your research involve the collection of original data?
   1. Yes
   2. No [SKIP TO Q15]

6. When research projects are complete, do you archive the data collected?
   1. Yes [SKIP TO Q8]
   2. No

7. Why do you not archive the data collected?
   [OPEN TEXT] [SKIP TO Q13]

8. Which of the following practices have you or your organization implemented to archive research data? Select all that apply.
   1. Convert datasets to SPSS portable format (.por)
   2. Make copies of datasets on disks, USB drives, or computer hard drives
   3. Store copies of datasets on servers
   4. Store datasets in a secure location
   5. Restrict access to datasets
   6. Upload data to project website
   7. Create metadata
   8. Submit data to an archive
   9. Other (Specify):

9. Are you familiar with the Odum Institute Archive Services?
   1. Yes
   2. No [SKIP TO Q12]

10. How many times have you visited the Odum Archive Services website in the past 12 months?
    1. 1
    2. 2
    3. 3
    4. 4
    5. 5
    6. More than 5 times
    7. I have not visited the Odum Archive Services website in the past 12 months.

11. Have you ever archived your data at the Odum Institute?
1. Yes
2. No

12. Where do you typically archive your data?
   [OPEN TEXT]

13. Would tools and services that simplify the process of archiving data increase your likelihood of consistently archiving your collected data?
1. Yes
2. No

TEXT: The Odum Institute Social Science Data Archive would like to help decrease the burden of data management and the archive process. Below are potential areas that we would like to address. The goal is to simplify the process for researchers and to increase archive submissions.

14. For each of the following archive training, resources, and functions, please indicate the degree to which they are of interest to you and your research.
For proposal development:
a. Example Data Management Plans
b. Data Management Plan templates
c. Organizational/Researcher commitment to archive project data
d. Consultation on data management Issues
e. Training on data management program implementation
   For data processing and archive submission:
f. Tools to assist in the creation of metadata
g. Web-based access and/or data uploads to the archive
h. User based access permissions for project data
   For data dissemination and authenticity:
i. Provide data citations for users
j. Customized website for archive access that retains recognition for the data producer
k. Access controls and/or user-based authentication
l. National and/or international-based discovery network
m. Online analysis tools for users
n. User-based permissions for project data access
1. Extremely interested
2. Very interested
3. Moderately interested
4. Slightly interested
5. Not at all interested

15. Does your research involve the re-use of social science data (i.e. secondary data analysis)?
1. Yes
2. No
1. Yes, to archive original data
2. Yes, to obtain data for secondary data analysis
3. Yes, to discover data for reference purposes
4. Yes, for other purposes (Specify):
5. No [SKIP TO Q18]

17. For each of the Odum Dataverse Network features listed below, please indicate how useful they are in your research?

a. Federated search across member archives
b. Multiple format data downloads
c. Detailed metadata and citation information
d. Basic descriptive statistics
e. Advanced data analysis tools
f. Ability to create your own archive and/or upload your project data
g. Advanced question-level survey search

1. Extremely useful
2. Very useful
3. Moderately useful
4. Slightly useful
5. Not at all useful

18. What other services or programs would you like to see the Odum Institute Archive provide? [OPEN TEXT]

Appendix 2

Initial Invitation Email

Subject: Share your views about the Odum Institute Data Archive Services

Dear [Name],

This email is to ask for your participation in a brief survey that is being conducted by the Odum Institute Archive Services at the University of North Carolina at Chapel Hill. We are interested in learning more about data management practices in your organization and your interest in future Odum Archive training courses and services. Your input will help guide the content of these new programs. The survey is also being conducted in partial fulfillment of School of Information and Library Science (SILS) degree requirements.

Your participation is completely voluntary, and the information you provide will be kept confidential. Results will be reported only in aggregate form; your name will never be associated with your data.

The survey should take less than 10 minutes of your time.

Follow this link to the survey: ![SurveyLink]
Or, copy and paste the URL below into your Internet browser:
${l://SurveyLink}

We thank you in advance for your cooperation.

If you have any questions about the survey, please contact Jonathan Crabtree at jonathan_crabtree@unc.edu or Dr. Helen Tibbo, SILS advisor, at tibbo@email.unc.edu. If you have any questions about your rights as a research participant, you may contact the University of North Carolina Institutional Review Board at IRB_Subjects@unc.edu and mention study number 10-2108.

Sincerely,

Jonathan Crabtree
Assistant Director of Archive and Information Technology
H. W. Odum Institute for Research in Social Science
CB# 3355
University of North Carolina at Chapel Hill
Chapel Hill, NC  27599-3355

Web Survey Intro Screen Template when Respondents Have Received Email or Postal Mail Invitation Containing Elements of Informed Consent

Thank you for participating in the Odum Archive User Survey. The data you provide will be used to enhance future archive services. This survey is also being conducted in partial fulfillment of School of Information and Library Science degree requirements. Your participation is voluntary and your answers are completely confidential. You may skip any question you choose not to answer. Please click the [>>] button below to begin the survey.
Reminder Email to Target Only Non-respondents

Subject: The Odum Institute Data Archive needs your help

Dear [Name]:

About a week ago we invited you to complete a survey on Odum Archive Services. As of today, your survey has not been completed. In order for our results to represent all groups, we really need your participation. The survey is also being conducted in partial fulfillment of School of Information and Library Science (SILS) degree requirements. We hope you will take a few moments now to click the link below and complete the survey.

Follow this link to the survey: ${l://SurveyLink}$
Or, copy and paste the URL below into your Internet browser:
${l://SurveyLink}$

The survey will take approximately 10 minutes to complete. Your participation is completely voluntary, and the information you provide will be kept confidential.

If you have any questions about the research project or the survey itself, please contact Jonathan Crabtree at Jonathan_Crabtree@unc.edu or Dr. Helen Tibbo, SILS advisor, at tibbo@unc.edu. If you have any questions about your rights as a research participant, you may contact the University of North Carolina Institutional Review Board at IRB_Subjects@unc.edu and mention study number 10-2108.

Sincerely,

Jonathan Crabtree
Assistant Director of Archive and Information Technology
H. W. Odum Institute for Research in Social Science
CB# 3355
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3355