Rebecca L. Greenstein. Fixing a Hole: Discerning Usage Patterns of Datasets in an Open Access Data Repository. A Master's Paper for the M.S. in L.S. degree. April, 2017. 54 pages. Advisor: Denise Anthony

In recent years, the federal government has mandated that data produced using federal funds be made available to the public. This, and the recent surge in the amount of data produced and the size of datasets, have made the pressure to share data ever the more urgent. Data can be shared using open access repositories, which can be institutional or domain-specific. In the social sciences in particular, data sharing is unique because of the various sources and types of data produced. This paper examined the usage patterns of the datasets in one social science repository based on production date. It found that the average number of download statistics for each year was remarkably consistent, but the data were extremely skewed. Further analyses could look at usage patterns based on topic/keyword, non-use of datasets, or time of usage of particular datasets.

Headings:

Data analysis

Social science archives

Repository libraries

Open access publishing

Metadata

FIXING A HOLE: DISCERNING USAGE PATTERNS OF DATASETS IN AN OPEN ACCESS DATA REPOSITORY

by
Rebecca L. Greenstein

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 Library Science.

Chapel Hill, North Carolina
April 2017

Approved by

_____

Denise Anthony

1. Introduction

In 2013, the Executive Office of the President's Office of Science and Technology Policy (OSTP) issued a memorandum mandating that the results of federally funded research be made available to the public, industry, and scientific community for free. The memorandum, addressed to the heads of executive departments and agencies, stated that sharing these data would boost the economy and help companies focus research in order to use others' discoveries for their best interest. The data shared would include both peer-reviewed publications and digital data. The memorandum also stipulated that federal agencies with over $100 million in annual conduct of research needed to develop a plan to support public access to data, the format of which should ensure interoperability and no technological problems (Holdren, 2013).

This mandate was not without cause. In recent years, the concept of data has become more nuanced as science and technology have grown and data-driven research has become more prominent. This has led to the production of larger datasets and a need to manage them. For example, the federally funded Human Genome Project resulted in a database with enough information to fill 2,000 computer diskettes (Zimmerman, 2003). A sequenced genome can provide valuable information for the layman, as information about susceptibility to heritable diseases is written into one's genes and can provide life-changing facts. Open access to this type of data is of the utmost importance, for both the scientific community to further the research and the layman to be informed about their hereditary make-up. Bodies other than the OSTP, such

as funding agencies and journals, are requiring both data management plans and deposition of published data – the time is now to start managing and sharing data wisely (Borgman, 2012).

In theory, the OSTP memorandum seems like a good idea. If the public's tax dollars are contributing to scientific research, shouldn't the public be able to view the data underlying the results of that research? Internet-based forums to view and comment on others' data would allow for an open access conversation on methods and best practices (Nielsen, 2011). In reality, however, it is a little more complicated. Data sharing practices vary between people, countries, and fields. Data rarely stand alone: they must be accompanied by procedures, software, and/or lab and field conditions. They are rarely self-describing, as a fair amount of documentation is needed for interpretation outside of the original context. An important question is how to cite data: which unit should be used for citation? The question of who should fund and manage data sharing is also unanswered at the moment (Borgman, 2015).

Even with these lingering questions, modern-day researchers do have multiple options for sharing their research data with their peers and with the public, thus making them available via open access. They can use repositories, which can be national, domain-specific, or institutional; there are currently 267 Social and Behavioral Sciences data repositories indexed by re3data (Re3Data, 2016). Back in the 1950s, researchers stored results from large survey interviews on IBM cards so that they could use the data gathered after the original study was over. National repositories were also common sites for researchers to store machine-readable data. By the late 1980s, the two types had

converged: all archives had individual files from sample surveys and files from national databases (Tannenbaum & Taylor, 1991).

By 2003, data sharing was not a new concept; what was new was the government's pressure on individual researchers to share data. Pressure to share data also came from journals, institutions, and disciplines. Disciplines such as agriculture, astronomy, genetics, and meteorology had well-established data repositories, but for other disciplines such as the social sciences, there was less of an infrastructure for open access data sharing (Zimmerman, 2003); the landscape has changed significantly since then. The benefit of depositing data in open access repositories has been shown by multiple groups: papers with data available by open access receive more citations than those without publicly available datasets (Pienta, Alter, & Lyle, 2010; Piwowar, Day, & Fridsma, 2007; Piwowar & Vision, 2013; Xia & Liu, 2013).

The social sciences are a broad field of research with many data types, and social scientists across disciplines will look at data sharing from different perspectives. The types of data used in the social sciences vary greatly, from maps (archaeology) to sounds (linguistics) to oral interviews (sociology) (Gómez, Méndez, & Hernández-Pérez, 2016). Economists tend to reuse existing government data and don't produce very much original data, whereas political scientists collect a fair amount of original data (King, 1995). It thus follows that economists would take a more liberal attitude toward data sharing than political scientists would. Historically, sociologists have been at the front of data sharing advocacy, but as time has passed economists have taken their place (Freese, 2007). A look at social sciences data repositories specifically is thus warranted, as data sharing practices are ever-changing and data take many forms in this broad field.

Social scientists choose to share, or not to share, data for a variety of reasons. Of 204 social scientists polled about their data sharing attitudes from 2009-2010, 39.8% think that they can integrate data from disparate sources, and 23.2% agree that others can easily access their data, a statistic slightly lower than that of the group of respondents as a whole (Tenopir et al., 2011). Another factor in social scientists' willingness to share data is privacy. Social science studies frequently collect information about human subjects, and sharing personally identifiable information with the general public can be a breach of confidentiality. Repositories' various methods of handling personally identifiable information are discussed in the next chapter.

In order for scientists to be willing to re-use data, a few factors need to be in place. Users need to be able to trust the repository, in that its data are valid, the staff are trustworthy, and its documentation is clear (Yoon, 2014). Scientists also must be satisfied with the data quality of a repository in order to re-use its data; data relevancy, completeness, accessibility, ease of operation, and credibility are positively associated with the satisfaction of data re-users (Faniel, Kriesberg, & Yakel, 2016). When scientists are willing to re-use data, the re-used data can serve multiple purposes in various papers. They can give credit for related work, evaluate analysis, serve as a meta-analysis for summary data, evaluate an analysis method, support data for new studies, or act as raw data (He & Nahar, 2016).

There are three steps involved in making use of data that have been deposited in an open access data repository: the initial sharing of data, use of those data, and re-use/citation of data. A gap exists between the abundance of research about the first and third steps and a paucity of research regarding the second step. Knowing the usage

patterns of a repository, something this paper will seek to parse out, is vital because these indicate whether a repository is beneficial to its users. A repository could also try to gain access to other datasets that the statistics show are used more frequently than others. For example, if datasets made public in 2012 are used more than those made public in 2007, efforts could be made to gain access to datasets that have been created in recent years. To that end, this paper will analyze scholars' use of one particular social science repository, the Howard W. Odum Institute at UNC-Chapel Hill.

2. Literature Review

This literature review will analyze some of the research conducted about data sharing in general: definitions surrounding it, benefits and challenges, and various pressures on researchers to share data. It will then address the data sharing practices and issues facing social scientists specifically before moving into data repositories, examining types of data repositories, facets of trustworthiness, and good practices of data repositories. Next, it will look at social science data repositories specifically before describing the social science data repository under scrutiny in this study in order to frame the research question.

2.1 Data Sharing

Definitions

In order to understand the minutia of data sharing, we must first examine some related definitions, including those of data, dataset, database, and data sharing. In a nod to their significance, the National Research Council (1985) described data as "the building blocks of empirical research" (p. 3) and Nicholson and Bennett (2011) said they are the "outputs of research, inputs to scholarly publications, and inputs to subsequent research and learning…the foundation of scholarship" (p. 505). The National Academy of Sciences (2009) defined data as "information used in scientific, engineering, and medical research as inputs to generate research conclusions," (p. 22) which encompasses many different types of data. In 1999, the National Research Council clarified their description to "facts, numbers, letters, and symbols that describe an object, idea,

condition, situation, or other factors" (p. 15), although Borgman (2010) cautions that data vary both between collaborators and by discipline when citing this definition. This paper will use the 1999 definition from the National Research Council because the 1985 National Research Council and Nicholson and Bennett descriptions are too broad – when dealing with actual data downloads, specificity is vital. The National Academy of Sciences definition is too focused on hard science, while this study focuses on social science.

Relatedly, data sharing, dataset, and database must also be defined. Borgman (2012) describes data sharing as "the release of research data for use by others" (p. 1060); methods of sharing can vary from email to deposit in a repository; documentation methods vary as well. Similarly, Kim and Adler (2015) define data sharing as "providing the raw data of your published articles to other researchers outside your research group(s) by making it [sic] accessible through data repositories/public web spaces/supplementary materials or by sending the data via personal communication methods upon request" (p. 409). This definition is too narrow, however, as data sharing can take place inside one's research group, and cleaned-up data can be shared as well as raw data; this paper will use Borgman's definition. According to the National Academy of Sciences (2009), datasets are "collections of similar or related data" that are stored in a database, which is "a collection of data that is organized to permit search, retrieval, processing, and reorganization of stored information" (p. 29).

Advantages

There are many advantages to data sharing, and a fair number relate to improving the data or methods themselves. New research using existing data can be promoted and

used to make new discoveries. Data sharing encourages appropriate data use, in that the data could be used incorrectly the first time around; reuse seeks to prevent this from reoccurring. Measurement and data collection methods can be improved after data have been shared. The development of theoretical knowledge and analysis of analytic techniques are promoted by sharing data. Data sharing encourages multiple perspectives on a single dataset and protects against faulty data (National Research Council, 1985). It also advances the state of science by letting data function as a resource rather than a method (Borgman, 2010).

Other advantages to data sharing relate to the people involved. Data sharing helps scientists verify, refute, and refine original results. It can also provide training resources for students (National Research Council, 1985). Another incentive for data sharing is reciprocity, as depositing data could be a way to gain access to both others' data and to useful tools for analysis and management (Borgman, 2010). Researchers from many backgrounds can benefit, as they see the same dataset from multiple angles and could provide fresh methods on how to use it (Kim & Adler, 2015). Also, it has been shown multiple times that datasets deposited in open access repositories are cited more than those that aren't (Pienta et al., 2010; Piwowar et al., 2007; Piwowar & Vision, 2013; Xia & Liu, 2013); data sharing increases the citation counts of researchers' articles, which have the potential to lead to promotion or tenure.

The last group of incentives relates to scientific research as a whole. Open access (scientific work being available to the public at no cost) and peer review (scientists reviewing each other's work) support the idea of open scientific inquiry (Borgman, 2010). The idea of science as a social contract, that scientists publish their work such that

it can be confirmed or refuted, is also at work here (Vision, 2010). Social and hard scientists think that data sharing is a norm of science (Ceci, 1988). Data sharing is also prompted by coercion, which often comes in the form of funding agency requirements (Borgman, 2010); the NSF, NIH, and IMLS now require data sharing and management plans as part of grant applications (Kim & Adler, 2015).

Disadvantages

There are also disincentives to sharing data, some of which relate to the various stakeholders involved. The two most obvious disadvantages to sharing data are a lack of both time and funding on the part of the researchers involved (Tenopir et al., 2011). Some researchers lack knowledge regarding when and how to share their data (Gómez et al., 2016). Researchers also might be hesitant to share data because they are worried about privacy concerns of the subjects involved; datasets could contain personally identifiable information that the subjects would probably be reluctant to share with the public (see below for an in-depth explanation of this aspect of data sharing).

Factors related to the data themselves can be a barrier to data sharing. It is difficult to manage research data, but it is even harder and more labor-intensive to make those data available and understandable to others (Borgman et al., 2015). Furthermore, the documentation methods to make one's data understandable to researchers outside of one's research team (Borgman, 2010), and the protocols for deposit in a repository, are time-intensive; the latter can cost money, depending on where the deposit occurs. Also, software programs are constantly changing as time goes by, and may differ between research groups, creating another obstacle for sharing data.

Disincentives are also related to the state of science in our society. For example, publishing papers carries more weight than collaboratively sharing data on a wiki and keeping science open (Nielsen, 2011). Rewards come from publication, rather than data management. Specifically, publishing in high impact journals rather than depositing data in open access repositories, among other things, leads to faculty receiving tenure. Data sharing comes with concerns about intellectual property rights, which differ in various countries and might require a license to share a dataset (Borgman, 2010).

Pressures

Pressures to share data, as alluded to above, come from a variety of sources. Journals can exert regulative pressure on scientists to share their data by requiring deposit of raw data at the time of publication (Kim & Stanton, 2016). Higher impact journals tend to have more stringent policies for data sharing, but these can be vague in that only a few provide specific examples of repositories that they suggest the authors deposit data into (Sturges et al., 2015). Funding agencies can also provide regulative pressure by requiring data management plans, or DMPs (see below for further discussion of DMPs). Communities can put normative pressure on scientists to share data if many scientists in a community commonly share their data (Kim & Stanton, 2016). For example, out of 25 scientists interviewed in one study, 13 felt that data sharing was part of their professional mission to further the development of science, and seven felt pressures from colleagues to share data (Kim & Stanton, 2012). Scientists can also put pressure on themselves to share data if they think it will be self-beneficial; on the other hand, scientists might be dissuaded from sharing data if they think they will get scooped. Self-pressure also comes in the form of perceived effort and scholarly altruism, or wanting to assist one's

colleagues. The availability of data repositories at a discipline level is another factor that can pressure scientists to share their data with fellow researchers: if it exists they feel they must contribute (Kim & Stanton, 2016).

In a multilevel analysis of data sharing pressures, Kim and Stanton found that pressure from journals, normative pressure, perceived career benefit, and scholarly altruism had positive relationships with behaviors in data sharing. Perceived effort had a negative relationship on data sharing activities. Pressure from funding agencies, perceived career risk, and availability of data repositories did not have a significant relationship with data sharing behaviors (Kim & Stanton, 2016). This last finding is troubling, particularly because a study by the same authors from 2012 found the exact opposite: that a lack of data repositories is a barrier to sharing in multiple disciplines (Kim & Stanton, 2012). The more recent finding was described as "unexpected" by the authors: a few possible reasons for the oddity is that data repositories can be hard to submit to, and they might not support all types of data generated in a particular discipline (Kim & Stanton, 2016).

2.2 Social Scientists and Data Sharing

General Practices

The social sciences is a large and diverse field, and it necessarily contains many different data types. Not all social sciences data are collected digitally: interview and survey data, which can be digitized, are common in sociology, but in archaeology, observational data are linked to geographical coordinates, samples, and drawings, which are in paper or other physical forms (Gómez et al., 2016). New data types in the social sciences include Facebook posts and tweets, which can be used as measures of public opinion and/or to predict upcoming election results. In a process called georeferencing,

spatial data can be merged with survey data to provide more potentials for analysis, such as finding possible correlations between living environment and behaviors or opinions (Recker, Müller, Trixa, & Schumann, 2015).

Another consideration when examining how social scientists share data is differences in data source. Many social scientists produce their own data through interviews or surveys, but public records can also be used as data sources (Borgman, 2012): sometimes investigations can be based on data that weren't originally produced for that purpose (e.g. analyzing government data or tweets to get new data). Linguistics data contain sounds, and historical data can be in the form of maps, journals, or photographs (Gómez et al., 2016). Sources also vary by discipline: political scientists tend to collect original data, whereas economists often analyze existing governmental data (King, 1995).

Data sharing norms also differ among disciplines in the social sciences: social science disciplines such as political science, sociology, and economics were among the first disciplines to coordinate data sharing efforts (Pienta et al., 2010). In a slight departure from this finding, Freese (2007) describes how historically, sociologists were close to the forefront of data sharing advocacy, but economists have since emerged as chief data sharing advocates. Discipline-specific norms also stem from the norm produced by the discipline's journals and whether or not they require data sharing (Freese, 2007). Another factor is that some repositories don't support all types of data generated in a discipline (Kim & Stanton, 2016); a researcher would probably be less likely to contribute to a repository if it didn't support their data type.

Given the heterogeneity of their data types and sources, one wonders how social scientists share data. In a survey of 724 NSF grant awardees across various disciplines, Kowalczyk found that CDs, DVDs, and removable hard drives are the most popular storage devices across disciplines. Repositories were used by fewer than half of surveyed researchers. Social scientists are more likely to use a domain-specific archive and have personal reasons for submitting to an archive. The study found that if a professional data manager were available, researchers were more likely to use sustainable technologies, which social scientists are less likely to have than those in other fields (Kowalczyk, 2014). Only 10.8% of social scientists polled in a meta-analysis of data sharing practices store data in repositories (Tenopir et al., 2015), which is less than the 46% figure from a paper published a year later (Gómez et al., 2016), illustrating a difference in sampling or methodology between the two groups.

Perhaps as a result of their non-sustainable sharing practices, social scientists often do not fulfill others' requests for data. In one study, 141 psychologists who had said they would comply with the American Psychological Association's ethical principles, which include sharing data for independent verification, were asked to provide published data. Only 25% of them actually shared these data, however, even after six months of continued requests from the researchers (Vision, 2010).

Data Sharing Attitudes

In two large-scale studies about the data sharing practices of researchers across disciplines, Tenopir and colleagues surveyed more than 1,000 researchers in 2009-2010, and then again in 2013-2014. They were looking for differences across discipline, age, and geographic location, and to see if changes occurred over time. In the first study, they

found that 74.9% of all respondents share their data with others, and 50.1% of the researchers think that a lack of access has restricted their ability to answer scientific questions. 67.2% of all respondents think that lack of access to data generated by researchers at other institutions is a barrier to scientific progress, while 36.2% think others can access their data easily, an example of researchers not practicing what they preach. For the social sciences specifically, 39.8% think that they can integrate data from disparate sources, and 23.2% agree that others can easily access their data, a statistic slightly lower than the general population's. 80.2% of scientists, however, believe that their data will be used in ways other than intended (Tenopir et al., 2011).

The follow-up study contained fewer statistics and more broad claims about data sharing practices among those polled. This study contained a different survey and more international respondents. The researchers polled in this iteration felt less satisfaction with long-term data storage and tools and more of a desire for data management best practices than those polled previously. A pattern emerged across disciplines based on whether or not the researchers used human subjects and if they needed to consider their subjects' privacy. Hard and social scientists had less of a strong feeling about the lack of access to data as a barrier to research, and were more likely to think that scientists should not share data if they work with human subjects. As in the 2011 study, scientists are dissatisfied with their ability to integrate data from disparate sources (Tenopir et al., 2015).

Trust and Satisfaction

A major factor in whether social scientists choose to both share their data and re-use others' data is whether or not they trust the repository. Trust as a concept is hard to

define. In 2014, Yoon interviewed researchers who re-used datasets in social science repositories. They defined trust as data validity and data integrity. Factors affecting trust identified in this study included user communities, organizational attributes, past experiences, repository process, and a perception of the roles of the repository. Half of the study participants believed in the integrity of their specific repository, and knowledge of the staff builds trust. Repository reputation is important, as is users' experience. A repository with clear documentation about its practices is also more likely to be trusted than one without this resource (Yoon, 2014).

Other factors affecting repository trust, as mentioned by interviewed social scientists (both depositors and re-users), include identification, benevolence, integrity, and transparency. Identification is the acceptance of stakeholder interests, and relates to how a repository understands its community's needs. Benevolence is the view by customers that the object of trust (repository) shows good will towards the customer (repository user), and integrity is the idea that the organization is honest and respects its stakeholders. Transparency is the willingness to share information with stakeholders. Specifically, depositors cited metadata standards, data quality, and migration as factors influencing their trust in a repository (Yakel, Faniel, Kriesberg, & Yoon, 2013).

For researchers who re-use data in repositories, a factor in their repository choice is satisfaction with that repository. Authors who cited datasets from the Inter-University Consortium for Political and Social Research (ICPSR) were polled by Faniel et al. Attributes that influenced their satisfaction with data quality were relevancy, completeness, accessibility, ease of operation, and credibility, which all had significantly

positive associations with the satisfaction of data re-users. 60% of the respondents felt that adequate data are available for reuse in their fields (Faniel et al., 2016).

Data Sharing Timing

In an effort to guide researchers in data management, many journals and funding agencies now require data management plans, or DMPs. The content of DMPs can vary depending on the body requiring them. As an example, the National Science Foundation's Social, Behavioral and Economic Sciences template delineates who in the planned study will be responsible for data management, what forms the data are expected to take, how long the data need to be retained, modes of data format and dissemination, and how the data will be stored and preserved ("DMP Preview: DMPTool," n.d.).

Data repositories can assist researchers with figuring out when to think about data sharing and archiving. Specifically, in theory, researchers can start talking to repositories in the planning and discovery phases. They can then begin to collect their data. When they prepare and analyze their data, they should give the repository one copy for archival purposes. When they publish and share their data, a final copy should be shared with the repository to ensure long-term management (Green & Gutmann, 2007). The data life cycle and the research life cycle thus cannot be considered independently of each other (Tenopir et al., 2011).

Guss' study of social scientists showed that this model is not always the case in practice, however. Respondents in this study reported that the content of DMPs varied drastically: 3.8% had a clear statement of data retention or archiving, 13.2% had equivocal language about data retention, 34% contained equivocal language that might disqualify the data from being archived, 18.9% contained a statement about destroying

data, and 30.2% did not mention the future of data (Guss, 2009). Clearly, it is important to think about data management in the planning stages of a research project.

2.3 Data Repositories

Data repositories generally fall into two different categories: institutional and domain-specific. As the name suggests, institutional repositories are based at one institution (see https://cdr.lib.unc.edu/ for an example) and contain a stockpile of electronic versions of articles or datasets produced by researchers from a variety of disciplines at that particular institution. Domain-specific repositories are for one specific discipline (see https://www.icpsr.umich.edu/icpsrweb/ for an example) and seek to supply a common infrastructure that makes the boundaries between institutions, researchers, and locations disappear (Green & Gutmann, 2007). Use depends on a given researcher's needs for a specific article or dataset; a researcher could use different repositories for different functions.

Goodman et al. (2014) delineate a list of ten rules for caring for and feeding scientific data and argue that scientists should foster and use data repositories, whether institutional, domain-specific, or both. Researchers are encouraged to talk to librarians or data scientists if they aren't sure how to use data repositories to best suit their needs. Good repositories give their datasets persistent identifiers for citation purposes: these also can be called digital object identifiers (DOI) or handles (hdl) (Goodman et al., 2014). Since URLs are rapidly decaying - "404 not found" errors when a URL is no longer active are commonplace - persistent identifiers allow for long-term access to datasets or articles (Klump, 2011).

Other practices of good repositories are asking researchers to provide adequate documentation for their data and curating data post-deposit (Goodman et al., 2014).

Repositories should have clearly identifiable files and accessible and understandable datasets (Peer, Green, & Stephenson, 2014). There are many positive outcomes when repositories curate data, similar to the advantages for sharing data in the first place. Curating data enables reuse, allows for retention of unique data, makes more data available for future projects, increases the ability to validate research results, promotes data use in teaching, and should be done for the public good (Yoon & Tibbo, 2011).

Repositories are, in general, more trustworthy than sharing mechanisms such as Dropbox, floppy disks, or external hard drives. Technology such as floppy disks or external hard drives can become obsolete in just a few years, while repositories are designed with the intention of maintaining their enduring value. One group of researchers described a Dropbox system glitch where all of their files were bizarrely deleted; clearly their backup system was not performing as planned (Cliggett, 2013).

2.4 Social Science Data Repositories

Background and an Example

As Tannenbaum's and Taylor's (1991) article about the history of social science repositories asserts, "[s]ocial science data archives are among the most enduring products of the 1960's" (p. 225). They have evolved from the International Business Machines (IBM) cards used to store the results of large-scale survey interviews in the 1950s (Tannenbaum & Taylor, 1991) to vast quantities of linked data across multiple institutions all over the world. The sheer number of social science data repositories available demonstrates their lasting value: There are 267 Social and Behavioral Sciences repositories listed on re3data.org (Re3Data, 2016), the largest and most comprehensive web-based index of data repositories in existence (He & Nahar, 2016).

One example of a social science data repository is Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS), or German Social Science Infrastructure Services, which was founded in 1960. Based in Germany, the Data Archive for the Social Sciences section of GESIS both carries out its own research and assists researchers through their own research cycles. It, and all data repositories, needs to ensure that no copyright infringement has occurred before accepting data. As an international data repository, it has to deal with different criteria for advocating and licensing research data that are in place in Germany, the Netherlands, the UK, and Denmark. It is also trying to determine the best incentives for data archiving (such as academic credit) and how to deal with the "long tail" of data: individual researchers are doing smaller studies and quality control is harder than before. Another challenge is new data types, such as Facebook posts and tweets. The sheer number of data repositories, as described in the last paragraph, creates a heterogeneous landscape for social science data sharing, where researchers can't be sure that their data are being consistently curated and maintained (Recker et al., 2015).

Privacy Issues

Social science data often contain sensitive and identifiable information, and privacy is a challenge that social science data repositories manage in different ways. Researchers think that the risk of disclosure should lie with the repository rather than the researcher, which is sometimes the case; repositories have devised methods for carrying these out. As one example of a way to manage sensitive data, ICPSR allows patrons to use but not download sensitive data (Frank, Kriesberg, Yakel, & Faniel, 2015). Some repositories let users freely download data about human subjects, while others require a

lengthy process before approval to use sensitive data is granted to researchers (Nelson, 2009). Researchers can be dissuaded from sharing their data because of the liabilities associated with a breach of privacy, and the process of de-identifying data can be very time-consuming (Sayogo & Pardo, 2013).

In a study about social science researchers and their attitudes towards sharing sensitive data, Guss found that only 35.9% would feel comfortable archiving de-identified data in a repository, 48.4% would not, and 15.6% don't know. For those who said no (not comfortable sharing), concerns existed about the continued anonymity of their data, that the data wouldn't make sense to other researchers, and that they hadn't informed their participants about the possibility of archiving and that consent had not been given. The ones who said yes (were comfortable sharing) said they did not have any personally identifiable data and reiterated their beliefs in open access to data. The respondents who said they didn't know were unsure about ethics and data control (Guss, 2009).

Differences in the type of data being shared can lead to varying attitudes on privacy. For quantitative social sciences, data confidentiality risks are for living individuals, and the community has come to more of a consensus than the qualitative social sciences about keeping data confidential. For archaeology and other qualitative social sciences, data can come in the form of location data, and indigenous communities can be endangered because of a breach of privacy (Cliggett, 2013; Frank et al., 2015). There is a risk of historic sites being located and looted, although among the archaeologists surveyed, there was disagreement regarding whether or not to make historical site information available (Frank et al., 2015).

2.5 The Odum Institute

Founded in 1924, the Howard W. Odum Institute at UNC-Chapel Hill (henceforth Odum) is an example of a domain-specific repository, specializing in social sciences data storage and curation. It provides access to self-curated data collections as well as those curated by other institutions (Odum Institute, 2016). It offers three tiers of support for researchers: self-service ($0), guided service ($3000 base plus per dataset fee, which depends on the type and scale of the project), and lifecycle service ($5000 base plus per dataset fee). The self-service relies on the client to prepare data files, create metadata, and upload data into the Dataverse. Patrons who use the guided service get help from Odum staff members in data file organization, controlled vocabularies for metadata, and file access levels. Lifecycle services clients can expect help at all stages of the research process, including a review service for DMPs ("Odum Institute: Data Management Services," n.d.).

Odum has a rich history in data archiving and preservation. In the 1970's, its staff members were influential in determining how to catalog machine-readable data files, which helped with the discoverability, accessibility, and usability of large amounts of social science data. More recently, Odum has been part of the movement to archive and preserve at-risk social science data as part of the Data Preservation Alliance for the Social Sciences (Data-PASS) (Crabtree, 2013). Relationships like these are vital to successful data archives in order to achieve a common metadata standard like Data Documentation Initiative (DDI), but are time-consuming (Crabtree & Donakowski, 2007).

Fortunately, there are tools in place to assist with sharing and compatibility across multiple repositories. The Dataverse Network (DVN), which partners with Data-PASS and of which Odum is a member, is an open-source application for publishing,

referencing, extracting, and analyzing research data. Its goal is to incentivize sharing without hardware or software costs. Between its inception in 2006 and 2011, it hosted hundreds of virtual archives, 37,000 studies, and 600,000 files. The DVN contains multiple dataverses, all of which contain research data (Figure 1). Each dataverse contains research studies, which can grouped into collections of studies. Studies consist of cataloging information (metadata), text files, graph data and data set files (Crosas, 2011).

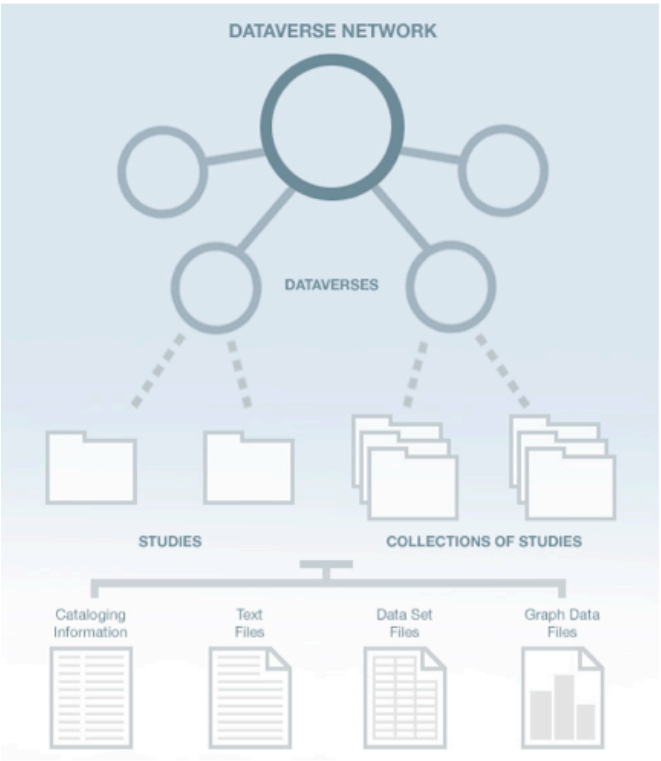

Figure 1. Dataverse structure (from http://www.dlib.org/dlib/january11/crosas/01crosas.html), used with the author's permission.

One characteristic of good repositories described in Section 2.3 is that they give their data persistent identifiers. The DVN passes this test because it uses handles, which serve as working URLs, as persistent identifiers in order to identify digital objects, in this case studies. An example of a handle is http://hdl.handle.net/1902.2/6635. Because each

study can have multiple datasets, each dataset has its own identifier called a universal numerical fingerprint (UNF), a string of alphanumeric characters. An example of a UNF is 3:aGYTy1ubiRXFTnPZBExcdA== (Crosas, 2011).

With these identifiers, Odum can easily track dataset downloads. In spite of this, no one has tracked usage patterns of their datasets. In addition, no one has examined usage patterns for any data repository in any field, whether institutional or domain-specific. Benefits of re-use of open access data have been shown by many groups; in other words, datasets that are made publicly available via open access receive more citations than those that are not (Pienta et al., 2010; Piwowar et al., 2007; Piwowar & Vision, 2013; Xia & Liu, 2013). Odum, however, does not currently have a mechanism in place to track reuse of its datasets. This study will use Odum as a case study to examine usage patterns of its datasets in the hope that this methodology and reasoning can be applied to other repositories. Specifically, it will try to determine the answer to the question, "What are the usage patterns of the Odum Institute datasets?" It will examine whether there are patterns of usage, which the Odum Institute could then use to inform future collection development practices. For example, if datasets made public in 2012 are used more than those made public in 2007, efforts could be made to gain access to datasets that have been created in recent years.

3. Methods

3.1 Design/Strategy

This study, the first of its kind, sought to examine the usage patterns of an open access data repository's datasets, using the Odum Institute at the University of North Carolina at Chapel Hill as a case study. It was a content analysis of the download data from the Odum Institute Dataverse that used statistical tests and analyses to attempt to discern a pattern of usage based on the production years of each dataset.

3.2 Sample

All download data for released datasets (not drafts or deaccessioned datasets) from the Odum Institute Dataverse Version 4 (hereafter DV4) until November 2016 were examined. Since staff members routinely download data for testing purposes rather than for their own analyses, the downloads by Odum Institute staff members were excluded from the 21,751 non-anonymous downloads in DV4. There are 25,137 datasets in the Odum Institute Dataverse, of which 3,452 (14%) have been downloaded. They are grouped into datafiles, of which 31,628 have been downloaded. In total, there have been 571,550 downloads.

3.3 Data Collection Methods

Data from DV4 were downloaded on December 25, 2016. They were generated using pgAdmin, the backend database of the Dataverse system, and by submitting SQL queries to generate the various tables. These logs are comprised of multiple users' downloads. Once the single Excel spreadsheet was downloaded from pgAdmin, the seven tables/tabs were split into seven Excel files: DV4_dataset, DV4_datasetfield, DV4_datasetfieldtype, DV4_datasetfieldvalue, DV4_datasetversion, DV4_dvobject, and

DV4_guestbookresponse. Five of the tables, DV4_dataset, DV4_datasetfield, DV4_datasetfieldvalue, DV4_datasetversion, and DV4_guestbookresponse comprise a relational database used to keep track of download statistics and dataset metadata.

Using the Excel file for DV4_guestbookresponse, the non-anonymous download were perused for downloads by Odum staff members, which were transferred to a new tab in the Excel file, in case they were needed later, and excluded from future analyses. Such downloads were found by looking for the phrases Dataverse, UNC-Chapel Hill, Odum, UNC, UNC-CH, CPSM, and CSSP in the email addresses or researcher affiliation fields. Each name associated with these downloads was searched on the Odum staff website (http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=16) and in the UNC directory. 144 such downloads were removed. Some staff members definitely downloaded the same dataset multiple times, but as this couldn't be accounted for in the anonymous downloads, no action was taken for these downloads in the non-anonymous downloads.

The tables DV4_datasetfieldtype and DV4_dvobject are not part of this relational database. DV4_datasetfieldtype is a glossary that describes the 154 metadata fields available to the inputters of metadata for the studies. Examples include title, name (of author), production date, summary, keywords, identifier, subject, grant agency, and deposit date. Only five of these fields (title, author name, email address, text/description of study, and subject) are required fields. DV4_dvobject contains information like "permission index time" that was not helpful for the purposes of this analysis.

The five tables in the relational database mentioned above contain more fields than was needed to create the database for analysis. Thus, only the necessary

columns/attributes were transferred from the original .xls files to new .csv files. Those attributes are diagrammed in the entity-relationship (ER) diagram in Figure 2. The included attributes are the primary and foreign keys for each table, the datafile_id (to indicate which datafiles have been downloaded), the versionstate (to indicate if the dataset is complete or not), and the fieldvalue (which contains the metadata for each dataset). Relationship diamonds are blank, since the database represents abstract concepts rather than physical objects. The relational database can also be expressed by the schema in Figure 3, which indicates primary/foreign key relationships and the attributes that correspond to each entity.

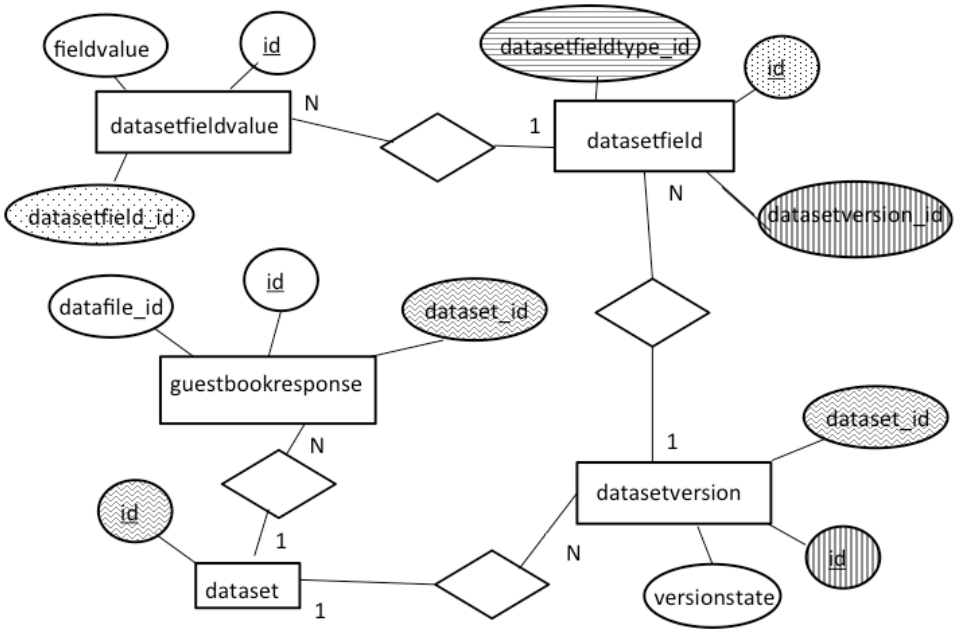

Figure 2. ER diagram for DV4. Entities are shown as rectangles, attributes as ovals, and relationships as diamonds. Cardinality for each entity group is indicated by 1s and Ns, and degree for each entity group is indicated by single lines. Primary keys are underlined, and primary/foreign key groups have the same background pattern (dots, horizontal lines, vertical lines, and squiggles).

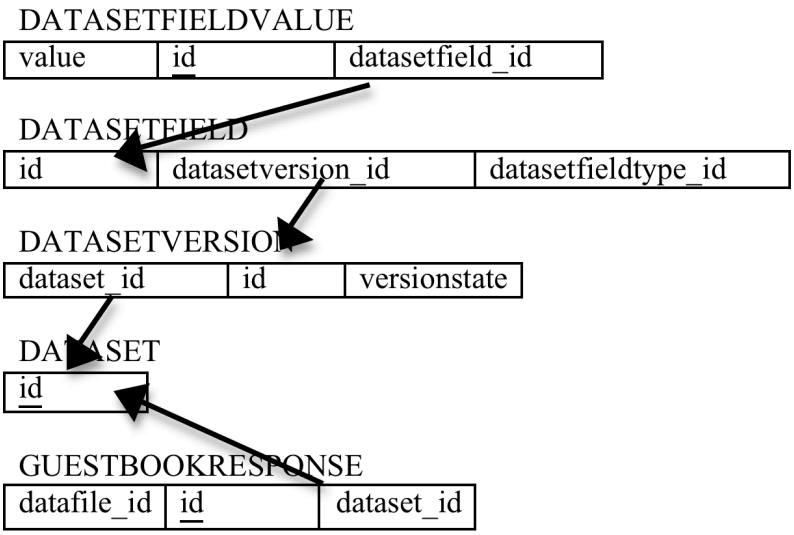

Figure 3. Schema for DV4. For each binary relationship, arrows point from each foreign key to each primary key.

The database was then created in the Mozilla Firefox extension SQLite by using the SQL create table statements listed in Appendix A, in the given order. As a few of the attributes were mistakenly left out of the original queries, the statements at the end of the appendix reflect the correction once the errors were discovered (one empty column was left out, and a second column including the values was left out). The contents of the first four tables were then transferred to TextWrangler, saved as .sql files, and inserted into the database after the regular expression statements shown in Appendix B were applied to all rows in the tables. These find the beginning of the line, the tab, and the end of the line and replace each one with the given values. The contents of the datasetversion table needed to be cleaned by deleting empty rows so they could be inserted into the database. As some of the values in the datasetfield table are null, the word NULL was inserted into the blank cells in the appropriate column in the Excel .csv file. Once the data were in TextWrangler, the apostrophes were removed using a special regular expression for that

table specifically to remove the apostrophes around each instance of the word NULL (Find 'NULL', Replace NULL).

The contents of the datasetfieldvalue table, which contains over 500,000 rows of metadata describing the studies, proved significantly more difficult than the others to insert into the database because the data were not formatted in a way that allowed easy insertion into the database. At first, the data were cleaned in TextWrangler, but as this method proved exceedingly slow, the data were ultimately cleaned in Excel. Since apostrophes, quotation marks, commas, and angle brackets are part of the SQL language, these were removed from the Excel file (found and replaced with a blank space), and a new Excel file was saved. The single large and clean Excel file was split into nine smaller Excel files of approximately 60,000 rows each to make it easier to manage the data. For each small Excel file, the rows were perused to find instances of not-three values in each row, as the datasetfieldvalue table accepts only three values. All rows without foreign keys were removed, as were any entries that took up multiple rows. The regular expressions find the ends of lines, so these would be placed into multiple rows in the database, even though they are only one entry.

Once the contents of each small file were cleaned, they were transferred to TextWrangler where the regular expression statements in Appendix B were applied. After that, attempts were made to transfer the data to the database, but once SQLite found an error, further cleaning steps in TextWrangler occurred. This was necessary because some unclean data were almost impossible to find in Excel (i.e. tabs in the middle of lines, which the regular expressions found and erroneously inserted commas and apostrophes into).

Summary statistics to determine how many times each datafile and each dataset were downloaded were produced using the queries in Appendix C. The query in Appendix D was then run to generate the data used for analysis. The output contains, for each datafile, the datafile_id, the number of downloads, and the year it was produced.

3.4 Data Analysis Methods

Study data were analyzed by comparing download frequency for each production year. In Excel, studies were separated based on production date metadata into groups; there were three possible date formats: year, month/day/year, and year-month in three sections in the data. The three columns (datafile_id, number of downloads, and production date) were always kept together. The data for the years 1932-1950 were placed in one tab, and each subsequent decade was put in a new tab. For the datafiles that had multiple dates entered, one was excluded using the following guidelines: If the dates were in the same year (i.e. 1988 and 1988-10), the more detailed date was excluded for ease of reading. If the dates were in different years (i.e. 1988 and 1986), the later date was excluded, as the user could have been looking for the data from the earlier form of the dataset. After the 108 duplicates were removed, 28,596 datafiles remained out of the original 28,704 (28,596/31,628 downloaded datasets (90%) of datafiles have production dates).

Data were grouped by year(s) of datafile production and, for the various analyses, by number of datafiles, total number of downloads, and average number of downloads along with the standard deviation. Two-tailed t-tests that assumed unequal variances to assess whether there is a difference in means between two groups were run. These tests were used to assess whether there is a difference in usage based on production date.

4. Results

4.1 Data Grouped by Year

After the data were sorted according to the procedure in the section above, graphs were made describing them. For the datafiles that have been downloaded, there are 71 years from 1930-2016 with production dates; most years have fewer than 1,000 datafiles in them (Figure 4). The number of datafiles in any given year varies considerably, from two (eight years) to 6,744 (1988). The number of downloads in any given year also varies considerably, from 25 (1952) to 112,901 (1988). Since it was hard to tell exactly how many datafiles were produced in the vast majority of the years under scrutiny, a new graph was made after the outliers (the three years with over 1,000 datafiles: 1988, 1989, and 1999) were removed (Figure 5). There were a fair number of datasets produced in the 1990s, but there is not a strong trend between the year and the number of datafiles: there is not a constant decrease or increase for any given stretch of time.

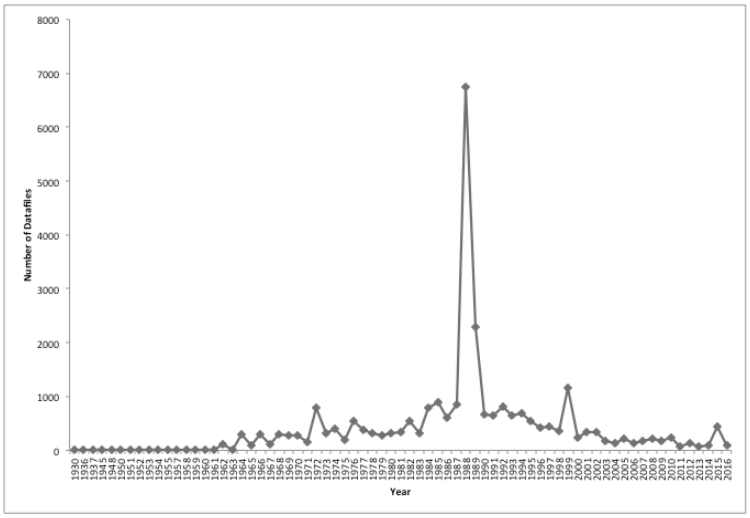

Figure 4. Year of production vs. the number of datafiles produced in each year.

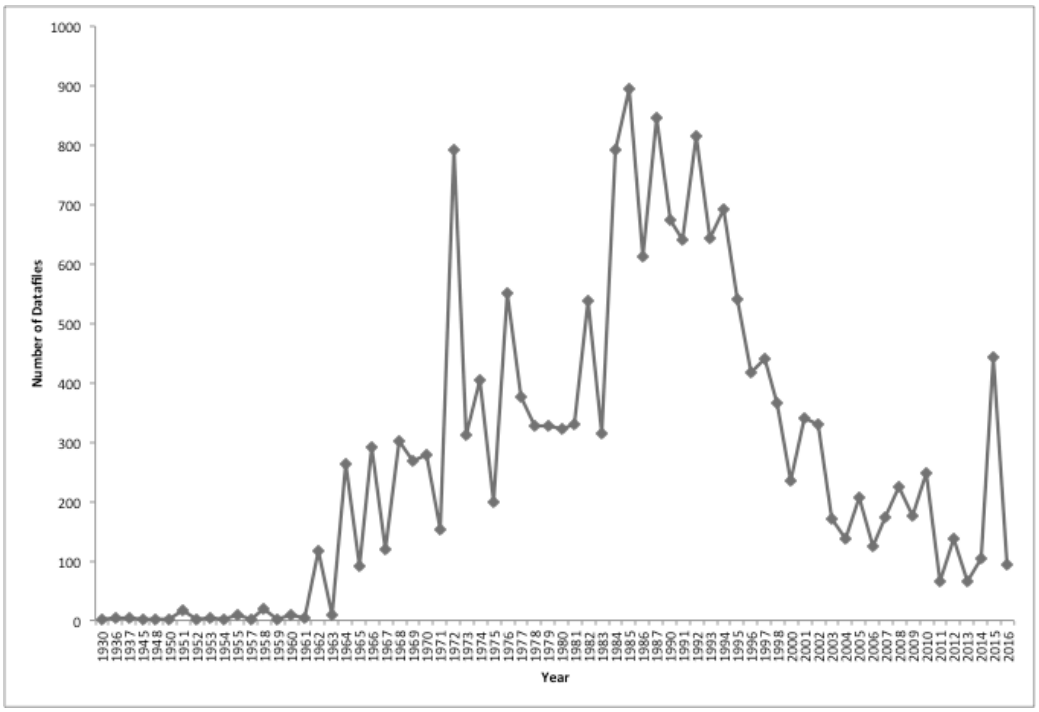

Figure 5. Year of production vs. the number of datafiles produced in each year, with outliers removed.

The year of production vs. the number of downloads was also graphed (Figure 6). Here, it is clear that four years, 1988, 1989, 1999, and 2015, have substantially more downloads than the rest of the years. They all had more than 30,000 downloads, and the next highest number of downloads/year is 18,995 (1992). Figure 7 shows the year of production vs. the number of downloads per year with these four years removed. Beyond the large number of downloads for datafiles produced in the late 1980s-early 1990s, there is no clear pattern in the number of downloads made for each production date year.

Next, the year of production was compared to the average number of downloads for each year (Figure 8). When these data were graphed, a trend emerged. The average number of downloads for each production year is remarkably consistent; it is between 10

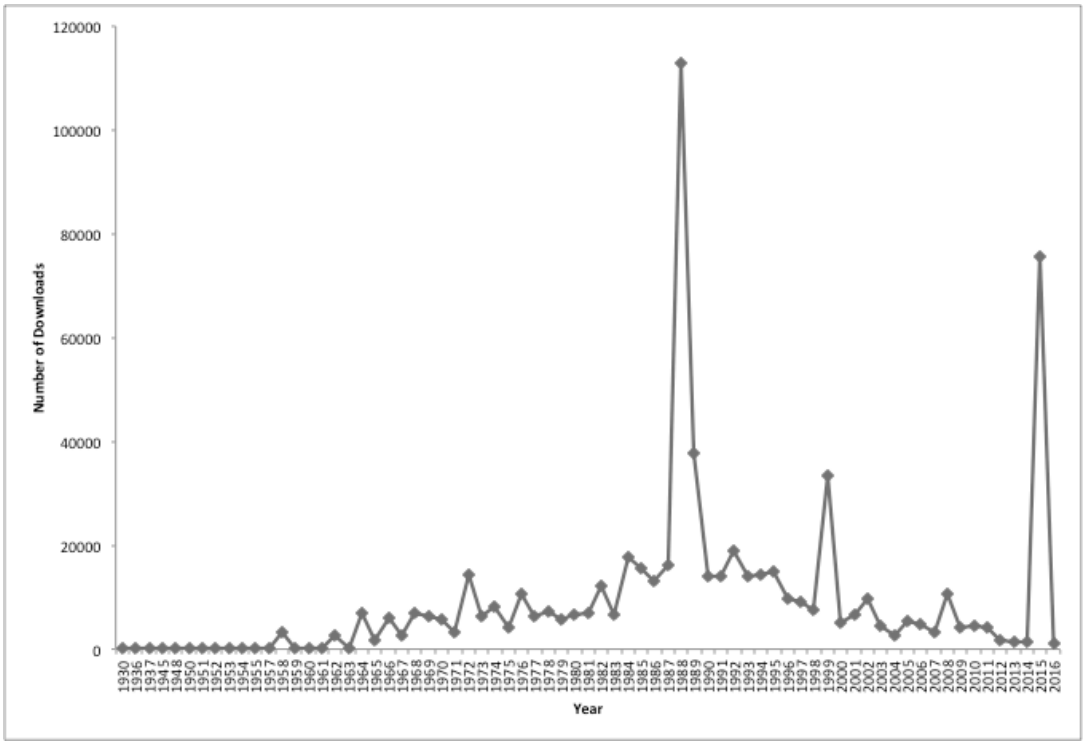

Figure 6. Year of production vs. the number of downloads for datafiles produced in each year.

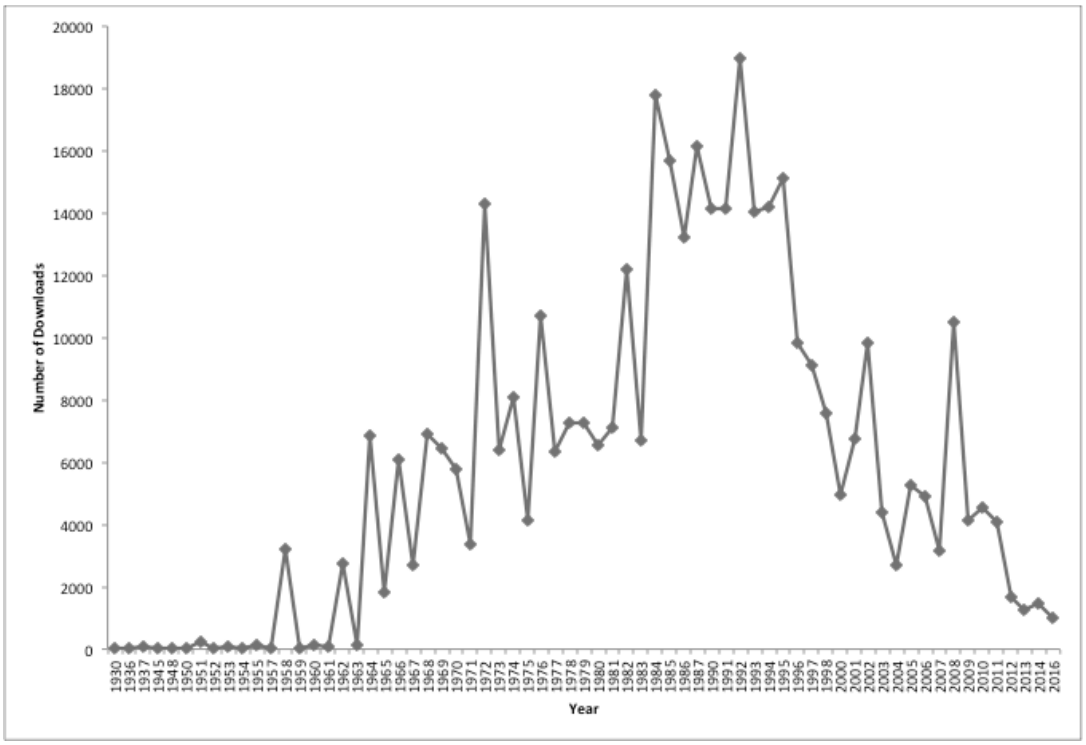

Figure 7. Year of production vs. the number of downloads for datafiles produced in each year, with outliers removed.

and 30 downloads for all but five years (1958, 2006, 2008, 2011, and 2013). Considering the range in the raw number of downloads, this is slightly surprising. The standard deviations are, on the whole, very high, with some even larger than the mean, indicating that the data seem to be quite skewed. For four of five of the outlier years, the standard deivation is larger than the mean, and when the number of downloads is viewed as a scatterplot for a single year (Figure 9), the data are, in fact, very negatively skewed. A number of t-tests were run that attempted to discern a difference in means between sets of two years. For the most part, the p-values were either close to 1 (not at all significant) or close to 0 (very significant), and it was thought that the skewness of the data in general, and the lack of any pattern, would render these t-tests unreliable. Also, running a t-test on two sets of data that both have sample sizes of 2 does not really allow the analyzer to come to any conclusions. The p-values are therefore not included.

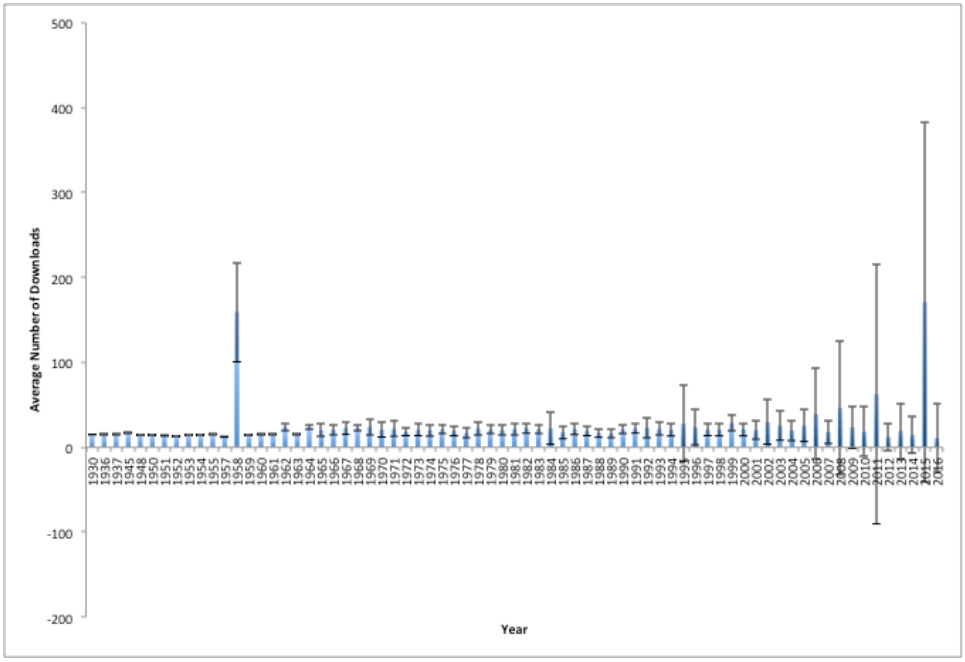

Figure 8. Year of production vs. the average number of downloads for datafiles produced in that year. Error bars are the standard deviation of the mean.

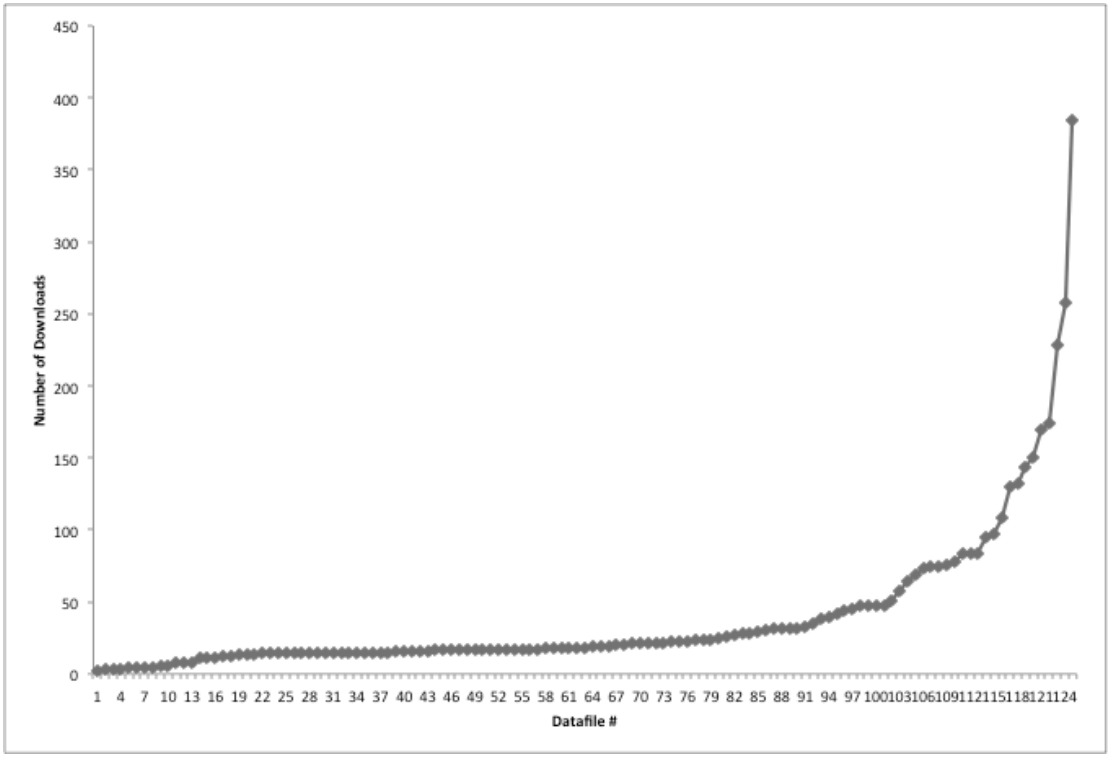

Figure 9. Datafile # (#1-125 for 125 datafiles in that year) vs. the number of downloads for each datafile.

4.2 Data Grouped by Time Period

In addition to being analyzed when grouped by individual year, the data were also analyzed when they were grouped by a range of years. The 1930s-1950s are grouped because there are only three years with downloaded production date datafiles from the 1930s, two from the 1940s, and nine from the 1950s. Moreover, none of these years had a high number of datafiles (Figure 10). The rest of the years are grouped by decade because starting in the 1960s, the number of datafiles and downloads increases significantly. There is a spike in the 1980s, sensible considering the high numbers of datafiles produced in both 1988 and 1989. A graph of the total number of downloads for each year shows a similar trend (Figure 11), although it does seem like datafiles from the 2010s are downloaded proportionally more frequently than those in other decades.

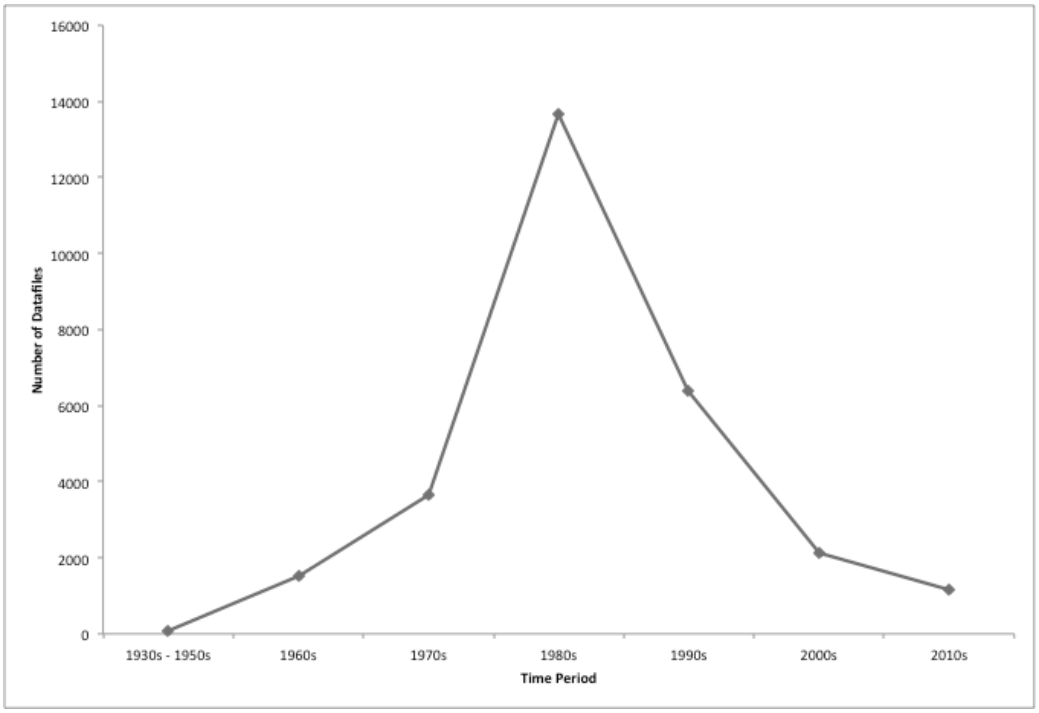

Figure 10. Time period vs. number of datafiles produced in each time period.

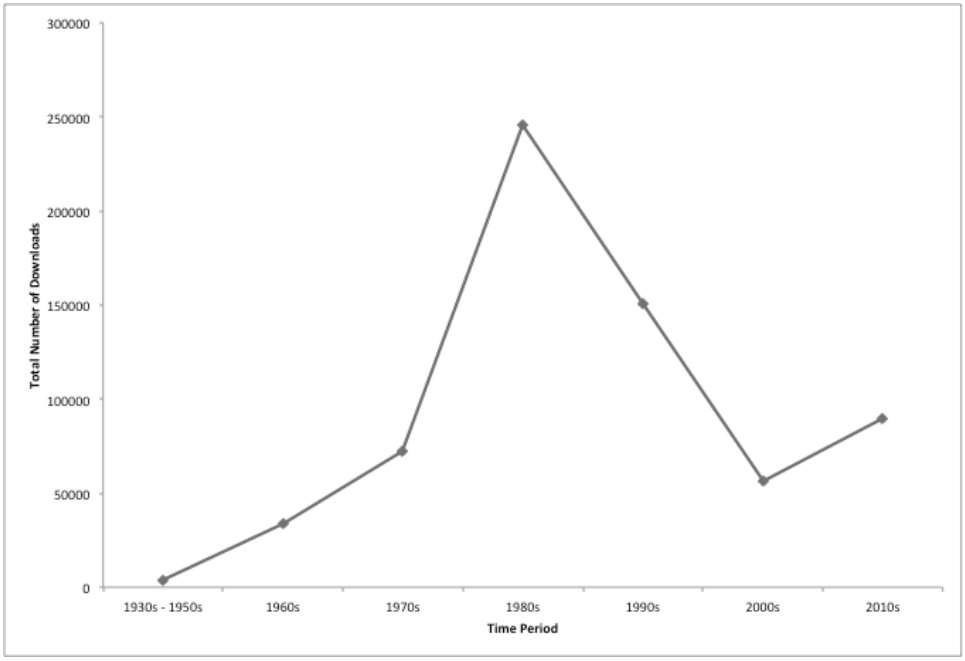

Figure 11. Time period vs. number of downloads for datafiles produced in each time period.

The average number of downloads was also determined for each time period (Figure 12). As with the data from the individual years, the average number of downloads evens out fairly nicely. The standard deviations, however, are greater than the mean for three out of seven of the time periods, including the 2010s, rendering the above observation insignificant. T-tests were also run on these data, and in this case, they were all significant. These data are also not included because the data are so skewed that, while the p-values indicate significance, the difference between the means is, in all likelihood, not actually significant.

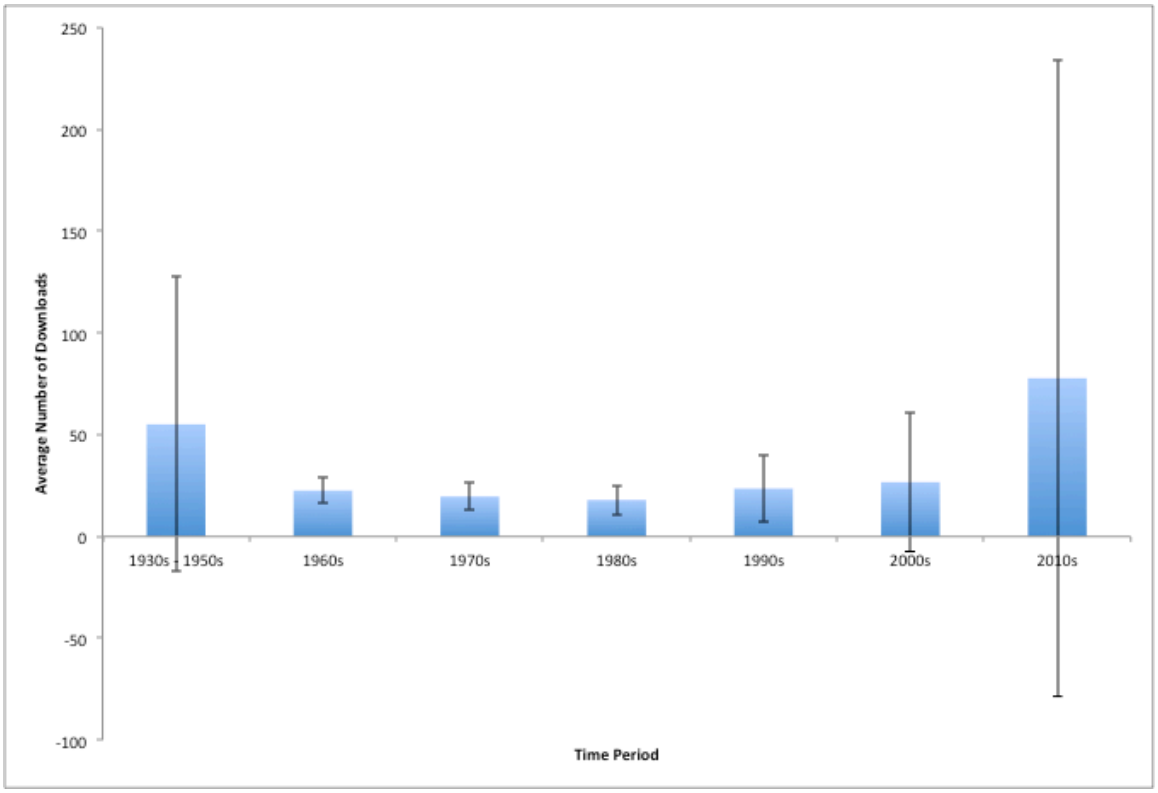

Figure 12. Time period vs. average number of downloads for datafiles produced in each time period. Error bars are the standard deviation of the mean.

5. Discussion

The original goals of this study were to determine usage patterns for the datasets, rather than the datafiles, in the Odum Institute Dataverse. Since there are not additional metadata for each datafile in a dataset, I originally planned to analyze downloads on the dataset level. The decision to analyze the data on the datafile rather than dataset level was made after I realized that a dataset with 20 datafiles probably has a larger number of downloads than a dataset with two datafiles (as one would need to download all datafiles to accurately use the data), which could lead to false conclusions. On the other hand, there are problems with the chosen method. Analyzing the data on the datafile level leads to a larger sample size for each year, which could also lead to false conclusions. Unfortunately, there is no convenient way to determine how many datasets are in each datafile. Otherwise, attempts would have been made to normalize the download numbers to account for discrepancies in the number of datafiles in each dataset.

As mentioned in the methods section, the non-anonymous downloads show that the same person downloaded some datasets multiple times. As there is no way to account for this for the anonymous datasets, the instances of repeat downloads were left alone. This could also have clouded the results by leading to falsely large numbers of downloads for some of the datafiles.

The datafiles were analyzed based on Production Date (datasetfieldtype_id = '41' – see Appendix D) for a few reasons. This bears more weight than Deposit Date

(datasetfieldtype_id = '57') or Distributed Date (datasetfieldtype_id = '55'), since production date is inherent to the datafile, while deposit date and distributed date are external pieces of metadata. Another option was Date of Collection, but when the query in Appendix D was run for Date of Collection, switching out datasetfieldtype_id = '41' for datasetfieldtype_id = '61', no values appeared. This is likely because Date of Collection is an optional piece of metadata, and the depositors probably didn't feel like it was necessary to include it. There is the problem that the required metadata fields consist of title, author name, email address, text/description of study, and subject, none of which are date fields. Thus, a number of the datafiles lacked a production date, and there is no easy way to identify those or gather an alternate date from them. It's possible that the datafiles that lack a production date could alter the significance findings above, but there is no way of knowing this with the data that are available.

There were also no clear patterns in datafile usage. The number of datafiles in any given year varies considerably, from two (1930, 1945, 1948, 1950, 1952, 1954, 1957, and 1959) to 6,744 (1988). The number of downloads in any given year also varies considerably, from 25 (1952) to 112,901 (1988). The large number of datafiles in 1988 (the year with the most datafiles and downloads) is due to a number of Computer Assisted Panel Studies (CAPS) datasets that were produced in that year. CAPS datasets all contain a large number of datafiles, and, as described above, one should download all datafiles for accurate usage of a dataset.

There is a remarkably and surprisingly consistent average number of downloads for the individual years. If there had been an obvious pattern (i.e. datasets produced in 2010 were downloaded at much more frequent rates than those produced in any year in

the 1960s), and all of the data were normally distributed, Odum could make an effort to collect datasets that were produced in 2010. As is, this analysis doesn't lend itself well to informing collection development strategy. One wonders if there is a better method for analyzing the data, or if the data are so skewed that any test of significance is not useful.

As this is the first analysis of usage patterns for any repository, I needed to come up with the methods from scratch, and date metadata were relatively easy to organize. Keywords were also considered as a topic for analysis, but this would have involved a significantly larger amount of organization than the analysis of usage patterns by date, since Excel can't organize keywords by category as it can with dates. For this analysis, it would have been important to keep keywords in multiple categories (i.e. if respondents were polled about their opinions about current events, which included opinions about the Chicago 7 and abortion, this datafile would be placed in both the defense and health categories for keywords). The researcher conducting this analysis would have had to be mindful of their inherent bias and the fact that their keyword grouping might differ from another researcher's. It might thus be useful to ask multiple people to initially group the keywords and see if the results are the same. Results from this analysis, if significant, also would have had collection development implications: if datasets about genetically modified organisms are downloaded more than datasets with vaccine trial data, efforts could be made to collect datasets with information about GMOs.

One factor that comes with using a database where researchers have entered their own metadata is that the metadata can be very messy. It is important that users can enter their own metadata, but having a controlled vocabulary could make these metadata more consistent (Moine et al., 2014). As Odum has three tiers of support, and the self-service

option (free) has users create their own metadata, it is reasonable to suspect that many users go that route rather than paying $3000 for the guided service, the next level up. A few times during the process of cleaning the datasetfieldvalue table, entire studies were, inexplicably, included in the metadata in the database. These took up thousands of lines, and the entire studies were deleted such that the rest of the data could be put into the database so they wouldn't occupy multiple entries (see Methods section for a description of why this is a problem). One of the metadata fields included the term 'CFFACE Training Yadda Yadda Data Blah,' and one wonders who was responsible for the metadata for this particular project.

6. Conclusion

As this study was the first of its kind for any repository, its main goal was to develop a methodology for this type of analysis. Dates are not the most interesting unit for analysis, but they did lend themselves to easy sorting, and were thus selected for analysis. In the future, it would be interesting to analyze usage patterns based on dataset keyword/subject, or dataverse source (Odum, Harvard, DataCarolina, etc.). One would also have to be mindful about skewness and small sample sizes for another type of this sort of analysis. Another path for analysis would be to figure out which datafiles are not being used; this would involve a bit more coding, as there is no table in DV4 for the datafiles that have not been downloaded like the guestbookresponse table for downloads. Both of these avenues of research could inform future collection development practices for Odum, if they are seeking to add datasets to their collection that would get the most use.

In addition, Odum does not currently track time of download in its guestbookresponse table, but analyzing datafile use based on timestamps would also be a potential path for future research. One wonders if, after the Supreme Court legalized gay marriage in 2014, if there was an upsurge in downloads about public opinion about same-sex marriage. Alternatively, another time to examine would have been right after the Affordable Care Act was passed for increases in downloads of datasets about health care and health insurance. It would be interesting to see if a large number of people were interested in datafiles that had to do with health care.

This study also demonstrates the dangers of letting users enter their own metadata. While this is vital to a repository's livelihood (it would be next-to-impossible to hire enough staff members to enter all the metadata for every deposited study), it would be prudent to establish best practices for researchers who wish to enter their own metadata. Entering their own metadata also gives researchers a sense of ownership over the data that they are sharing, but data sharing does take time, and errors in metadata input are likely to happen. One way to solve this might be to have staff members look over user-entered metadata after the users enter to it ensure compliance with the best metadata standards.

7. Appendix A

Create Table Statements Used to Create the Relational Database

```
/*1. Create dataset table - datasetversion and guestbookresponse refer to dataset via FK*/

Create table dataset (
id varchar(10) primary key
);

/*2. Create datasetversion table - datasetfield refers to datasetversion via FK*/

Create table datasetversion(
id varchar(10) primary key,
dataset_id varchar(10),
Constraint datasetFK1 foreign key (dataset_id) references dataset (id)
);

/*3. Create datasetfield table - datasetfieldvalue refers to datasetfield via FK*/

Create table datasetfield(
id varchar(10) primary key,
datasetversion_id varchar(10),
Constraint datasetversionFK foreign key (datasetversion_id) references datasetversion
(id)
);

/*4. Create guestbookresponse table - guestbookresponse refers to dataset via FK*/

Create table guestbookresponse(
id varchar(10) primary key,
datafile_id varchar(10),
dataset_id varchar(10),
Constraint datasetFK2 foreign key (dataset_id) references dataset (id)
);

/*5. Create datasetfieldvalue table - datasetfieldvalue refers to datasetfield via FK*/

Create table datasetfieldvalue(
id varchar(10) primary key,
fieldvalue varchar(10000),
```

```
datasetfield_id varchar(10),
Constraint datasetfieldFK foreign key (datasetfield_id) references datasetfield (id)
)


/*NOTE: Forgot a column in datasetversion:*/
ALTER TABLE datasetversion
ADD COLUMN versionstate varchar(20);

/*NOTE 2: Forgot to add column and values in it in datasetfield */

ALTER TABLE datasetfield
ADD COLUMN datasetfieldtype_id varchar(4);

/*Format for each Update statement */

Update datasetfield Set datasetfieldtype_id='NEW THING' Where id='PK';

Find ^
Replace Update datasetfield Set datasetfieldtype_id='

Find \t
Replace ' Where id='

Find $
Replace ';
```

8. Appendix B

Regular Expression Statements Used to Populate the Relational Database

Beginning of Lines
Find ^
Replace Insert into TABLENAME values ('

Tabs
Find \t
Replace ', '

End of Lines
Find $
Replace ');

9. Appendix C

Queries Used to Generate the Summary Statistics

```
Select distinct(datafile_id), count(*)
From guestbookresponse
Group by datafile_id;

Select distinct(dataset_id), count(*)
From guestbookresponse
Group by dataset_id;
```

10. Appendix D

Query Used to Produce the Data for Analysis

```
Select distinct(datafile_id), count(*), fieldvalue
From datasetfieldvalue join datasetfield on datasetfieldvalue.datasetfield_id =
datasetfield.id join datasetversion on datasetfield.datasetversion_id = datasetversion.id
join dataset on datasetversion.versionstate = dataset.id join guestbookresponse on
dataset.id = guestbookresponse.dataset_id
Where datasetversion.dataset_id = 'RELEASED' and datasetfieldtype_id = '41'
Group by datafile_id, fieldvalue;


/* NOTE: The versionstate column and the dataset_id column in the datasetversion table
had their column names switched accidentally (the versionstate column was one of the
ones that was added after the database was created).*/
```

11. Literature Cited

Borgman, C. L. (2010). *Research Data: Who Will Share What, with Whom, When, and Why?* (SSRN Scholarly Paper No. ID 1714427). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1714427

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

Borgman, C. L. (2015). If Data Sharing is the Answer, What is the Question? *ERCIM News*, (100). Retrieved from http://ercim-news.ercim.eu/en100/special/if-data-sharing-is-the-answer-what-is-the-question

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3-4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Ceci, S. J. (1988). Scientists' Attitudes toward Data Sharing. *Science, Technology, & Human Values*, *13*(1/2), 45–52.

Cliggett, L. (2013). Qualitative Data Archiving in the Digital Age: Strategies for Data Preservation and Sharing. *ResearchGate*, *18*(How-to-Article), 1–11.

Crabtree, J. (2013). Building on the Work of Colleagues: A Moment of Reflection. *IASSIST Quarterly*, *37*(1-4), 57–61.

Crabtree, J., & Donakowski, D. (2007). Building Relationships Project Update 2007. *Journal of Digital Information*, *8*(2). Retrieved from https://journals-tdl-org.libproxy.lib.unc.edu/jodi/index.php/jodi/article/view/191/172

Crosas, M. (2011). The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine*, *17*(1/2). https://doi.org/10.1045/january2011-crosas

DMP Preview: DMPTool. (n.d.). Retrieved September 29, 2016, from https://dmptool.org/plans/20539/preview

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

Frank, R. D., Kriesberg, A., Yakel, E., & Faniel, I. M. (2015). Looting Hoards of Gold and Poaching Spotted Owls: Data Confidentiality Among Archaeologists & Zoologists. *Proceedings of the Association for Information Science & Technology*, *52*(1), 1–10.

Freese, J. (2007). Replication Standards for Quantitative Social Science Why Not Sociology? *Sociological Methods & Research*, *36*(2), 153–172. https://doi.org/10.1177/0049124107306659

Gómez, N.-D., Méndez, E., & Hernández-Pérez, T. (2016). Social Sciences and Humanities Research Data and Metadata: A Perspective from Thematic Data

Repositories. *Aslib Journal of Information Management*, *68*(4), 545–555. https://doi.org/10.3145/epi.2016.jul.04

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). 10 Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, *10*(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Green, A., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services*, *23*(1), 35–53. https://doi.org/10.1108/10650750710720757

Guss, S. A. (2009). Assessing the Effects of Institutional Review Boards on Social Science data Archiving in Digital Repositories: Assessing the Effects of Institutional Review Boards on Social Science Data Archiving in Digital Repositories.

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications. *Aslib Journal of Information Management*, *68*(4), 478–494. https://doi.org/10.1108/AJIM-01-2016-0008

Holdren, J. P. (2013, February 22). Memorandum for the heads of executive departments and agencies. Office of Science and Technology Policy, Executive Office of the President. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management*, *35*(4), 408–418. https://doi.org/10.1016/j.ijinfomgt.2015.04.007

Kim, Y., & Stanton, J. M. (2012). Institutional and Individual Influences on Scientists' Data Sharing Practices. *Journal of Computer Science Education*, *3*(1), 47–56.

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424

King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301

Klump, J. (2011). Criteria for the Trustworthiness of Data Centres. *D-Lib Magazine*, *17*(1/2). https://doi.org/10.1045/january2011-klump

Kowalczyk, S. T. ., skowalczyk@dom. ed. (2014). Where Does All the Data Go: Quantifying the Final Disposition of Research Data. *Proceedings of the Association for Information Science & Technology*, *51*(1), 432–441.

Moine, M.-P., Valcke, S., Lawrence, B. N., Pascoe, C., Ford, R. W., Alias, A., … Guilyardi, E. (2014). Development and exploitation of a controlled vocabulary in support of climate modelling. *Geosci. Model Dev.*, *7*(2), 479–493. https://doi.org/10.5194/gmd-7-479-2014

National Academy of Sciences. (2009). *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/12615

National Research Council. (1985). *Sharing Research Data*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/2033

National Research Council. (1999). *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/9692

Nelson, B. (2009). Data sharing: Empty archives. *Nature News*, *461*(7261), 160–163. https://doi.org/10.1038/461160a

Nicholson, S. W., & Bennett, T. B. (2011). Data Sharing: Academic Libraries and the Scholarly Enterprise. *Portal: Libraries and the Academy*, *11*(1), 505–516.

Nielsen, M. (2011). *Open science now!* Retrieved from http://www.ted.com/talks/michael_nielsen_open_science_now?language=en

Odum Institute. (2016). Odum Institute Dataverse Network. Retrieved August 29, 2016, from http://arc.irss.unc.edu/dvn/

Odum Institute: Data Management Services. (n.d.). Retrieved October 17, 2016, from http://www.odum.unc.edu/odum/contentSubpage.jsp?nodeid=11

Peer, L., Green, A., & Stephenson, E. (2014). Committing to Data Quality Review. *International Journal of Digital Curation*, *9*(1), 263–291. https://doi.org/10.2218/ijdc.v9i1.317

Pienta, A. M., Alter, G., & Lyle, J. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. *The Organisation, Economics and Policy of Scientific Research Workshop*. Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/78307/pienta_alter_lyle_100331.pdf?sequence=1

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Re3Data. (2016, August 28). Search | re3data.org. Retrieved August 28, 2016, from http://service.re3data.org/search?query=&subjects%5B%5D=12%20Social%20and%20Behavioural%20Sciences

Recker, A., Müller, S., Trixa, J., & Schumann, N. (2015). Paving the Way For Data-Centric, Open Science: An Example From the Social Sciences. *Journal of Librarianship & Scholarly Communication*, *3*(2), 1–17. https://doi.org/10.7710/2162-3309.1227

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30, Supplement 1*, S19–S31. https://doi.org/10.1016/j.giq.2012.06.011

Sturges, P., Bamkin, M., Anders, J. H. S., Hubbard, B., Hussain, A., & Heeley, M. (2015). Research data sharing: Developing a stakeholder-driven model for journal policies. *Journal of the Association for Information Science and Technology*, *66*(12), 2445–2455. https://doi.org/10.1002/asi.23336

Tannenbaum, E., & Taylor, M. (1991). Developing social science data archives. *International Social Science Journal*, *43*, 225–234.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M.

(2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide: e0134826. *PLoS One*, *10*(8). https://doi.org/http://dx.doi.org.libproxy.lib.unc.edu/10.1371/journal.pone.0134826

Vision, T. J. (2010). Open Data and the Social Contract of Scientific Publishing. *BioScience*, *60*(5), 330–331. https://doi.org/10.1525/bio.2010.60.5.2

Xia, J., & Liu, Y. (2013). Usage Patterns of Open Genomic Data. *College & Research Libraries*, *74*(2), 195–206.

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*, *8*(1), 143–156. https://doi.org/10.2218/ijdc.v8i1.251

Yoon, A. (2014). End users' trust in data repositories: definition and influences on trust development. *Archival Science*, *14*(1), 17–34. https://doi.org/10.1007/s10502-013-9207-8

Yoon, A., & Tibbo, H. (2011). Examination of Data Deposit Practices in Repositories with the OAIS Model. *IASSIST Quarterly*, *35*(4), 6–13.

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists* (Ph.D.). University of Michigan, United States -- Michigan. Retrieved from http://search.proquest.com.libproxy.lib.unc.edu/docview/287907131/abstract/CA79ABA76EDF4372PQ/1