Searching for Datasets:

The Quest for Relevant Data

Bridget Disney

University of Missouri

Abstract

In the past few decades, research has become more data-driven as funding agencies and journals have begun to require data management plans to be submitted with proposals. Datasets are available to researchers through repositories and web sites but finding pertinent and relevant data can still prove to be difficult. This case study looks at five researchers from different disciplines to see how they approach the search for datasets, recording their insights and noting their best practices through the lens of archival intelligence theory. Past papers of the researchers will be assessed and think aloud strategies for new searches will be examined. The significance of this study is to encourage data reuse and promote innovation by exploring best practices for finding datasets.

Keywords: dataset discovery, information seeking behavior, dataset metadata

Searching for Datasets: The Quest for Relevant Data

Jack Esselink, self described Big Data and Analytics Evangelist at IBM, makes a bold statement when he declares "Today big data and analytics is everywhere for everyone" (2017, p.1). The phrase of the day seems to be "data driven research" as we use data to make decisions in our ever increasing complex and recorded world. Recent developments such as ubiquitous data collection, advances in database storage, computer processing power, predictive analytics as well as breakthroughs in cognitive and natural language processing have converged to provide an environment conducive to data exploration, which is now included in almost all academic disciplines. Additionally, and not surprisingly, major funding agencies (e.g. National Science Foundation, National Institute of Health) and prominent journals (e.g. Dryad Digital Repository, Public Library of Science) have begin to require data management plans to be submitted with proposals (Mannheimer, Sterman, & Borda, 2016), making data inclusion imperative.

Given the plethora of data generated from various sources today, a current and popular trend is to make the generated data available for reuse. The original data can be stored in a formatted dataset that can be shared with colleagues and other interested parties to use for their own research. Pasquetto, Randles, and Borgman (2017) go to great lengths to explain the nuances between use and reuse of data, but generally make the distinction that the data is "used" by the primary consumer for whom the data was intended, and "reused" by everyone else. However, this definition breaks down when datasets are generated by large institutions for general applications such as widespread weather data from National Oceanic and Atmospheric Administra-

tion (NOAA) or DNA sequences collected by the Human Genome Project (HGP), thus not having a primary user. This paper involves two instances of data discovery, one being data created by one researcher and being "reused" by another party, and the other being a general dataset produced via a data collection project meant to be used by the public.

Ronald Dekker (2006), while discussing the importance of having datasets, states that datasets can be the "primary intellectual output of research" ("2.3 Publishing data-sets," para. 1) , especially if the data is original and cannot be reproduced. The dataset becomes the supporting evidence for the analysis and thus contributes to any significant findings as an integral part of the research. Dekker makes a case for dataset reuse to verify findings, promote innovation, bolster longitudinal studies, and even generate further inquiry and collaboration. He and Nahar (2016) also note that dataset reuse contributes to retaining the integrity of data through continual preservation, most notably when public repositories are used. However, they also conclude that most data used for investigation originates from within an affiliated research group and is not obtained by exploring outside of one's own territory. Wallis, Rolando, and Borgman (2013) also mention the importance of dataset reuse which can allow researchers to access baseline data for comparisons, reanalyze and verify previous evidence for verification, and ask new questions about complicated data. Achieving these results can be especially beneficial in the case of "long tail" datasets which represent the more obscure data gathered by small, yet significant projects. Heidorn (2008) holds that this esoteric type of data "is a breeding ground for new ideas and never before attempted science" (p. 282).

This study explores the many ways in which datasets can found by researchers. Chen et al., (2018) accentuate that “making datasets findable is key to promoting the reuse of existing datasets” (p. 301). Even with the all the benefits of data reuse, locating existing dataset collections can be challenging, resulting in a process that can be haphazard and inconsistent. Though it is expected and common for datasets to be in various formats and locations, there is also lack of standardization in the metadata and citations used to describe the datasets. This could make data discovery difficult and may be why Mannheimer et al. (2016) came to the conclusion that “data sharing and reuse appear to happen relatively rarely” (p. 11). This is true even though surveys show that 75% of researchers are willing to share their data (Tenopir et al., 2011). Chen et al. (2018) discloses that “major initiatives have been established to build repositories and knowledge bases for specific types of data and domains” (p. 301). Perhaps this is in response to sentiments expressed in a paper written only five years earlier by Wallis et al. (2013) that remarks that repositories are rarely the first choice for finding datasets, falling after direct requests for data and data posted on web sites. Could the more recent formation of established repositories have provided stable locations for dataset discovery?

Many approaches have been evaluated to locate data, including the adoption of data portals (an online list of datasets that link to the actual data), perusing research papers, using search engines that rely on dataset metadata, employing specialized search engines that extract dataset information from research papers, making use of Linked Open Data (data that is linked to other data in a particular format), and networking within the academic community. In addition, several

models have been proposed for searching within specific domains. However, Noy, Burgess, and Brickley (2019) have concluded that "the tools for querying datasets are not as well developed as Web search engines, and require more human work (e.g., scrolling through more search results returned from short queries) than Web search" ("Related Work," para. 4).

Some research has shown that the use of discipline specific repositories, cross-indexing between repositories, detailed data descriptions, and the use of persistent identifiers such as DOIs (Digital Object Identifier) can improve dataset discovery (Mannheimer et al., 2016). However, support for interdisciplinary problem solving has been found lacking. With the recent and surging accessibility to these resources plus an emphasis on interdisciplinary research, approaches for finding pertinent data could be improved. This case study will observe and analyze the dataset discovery process of academic faculty and students as they pursue data suitable for their research. The intention is to uncover advantageous practices for finding useful data and to ultimately improve detection of relevant datasets.

A secondary consideration will be to explore the particular ways that librarians can assist in dataset discovery. Wilkinson, Pollard, and Farquhar (2010) perceived a need within academic research that could be filled by libraries when they endeavored to "better understand researcher behavior and requirements and provide a first look at resource discovery of datasets alongside other materials" (p. 101). Librarians can bring specialized skills to dataset discovery, revealing previously unknown sources and adding perspectives that enrich the research process. By ob-

serving and interacting with researchers, librarians have the capacity to increase their expertise, resulting in greater proficiency that benefits the university's data management system.

Literature Review

There have been quite a few approaches for dataset discovery. This literature review focuses on the studies for finding datasets, although certain instances of reuse may be included as reuse could imply that a dataset can be located easily and is indeed valuable enough to be discovered. In addition, applicable papers about dataset citation have been acknowledged.

Metadata and Citation

The first matter to be addressed is how generated data is made available and presented to the world. After the production of a dataset, information must be created that describes the data and makes it uniquely identifiable. One way this can be done is with metadata (data about the data), that is generated manually or automatically. Heidorn (2008) cites lack of metadata standards as a barrier of data reuse, suggesting metadata working groups and development of metadata tools to facilitate its creation. He and Nahar (2016) suggest that "descriptive metadata should contain more information about and contexts in which data are generated and their usability" (p. 17), noting that detailed information about the datasets varied greatly when written by authors from differing research backgrounds.

Complicating the issue is that distinct types of datasets may require unique and dissimilar components of metadata. Olfat, Kalantari, Rajabifard, Sentot, and Williamson (2012) describe some of the singular challenges experienced when documenting geospatial datasets. This is em-

phasized when they indicate that “the current data discovery services are not user-friendly and sufficiently efficient to serve the end users to easily find the most appropriate datasets and services to meet their needs in a spatially enabling platform” (para. 1). Another point made by Olfat et al. (2012) is that geospatial metadata is usually compiled after the actual dataset and may not be updated when changed. As such, the user may not be privy to current information that could be significant to their research.

Chen et al. (2018) evaluated a biomedical dataset discovery system (DataMed) under the belief that “metadata from diverse datasets can be mapped to a unified representation model, thus enabling more efficient search across domain-specific repositories and making data more discoverable by users” (p. 301). These authors explore the feasibility and potential of housing datasets in one easily accessible location, using standardized descriptions. They did conclude that this is indeed possible, achieving a 90% success rate when transforming the metadata to its new format. Their study addresses an increasing need to share data across disciplines as multi-domain research is becoming more common and relevant.

Mooney and Newton (2012) advocate accurate data citation to promote data discovery and reuse, commenting that a reluctance to share data among researchers may be due to fear of improper attribution for their work. They analyzed citations according to a newly created rubric called the Data Citation Adequacy Index (DCAI), and concluded that “even across the randomly selected cross-disciplinary sample, citation of digital research data is a rarefied activity” (p. 13).

Mannheimer et al. (2016) stressed the importance of proper attribution and citation for dataset discovery by this statement, “With better data citation tracking, more robust conclusions can be reached regarding how to support the discovery and reuse of datasets” (p. 10). The Data Documentation Initiative (Hoyle et. al, 2016) established by the Inter-university Consortium for Political and Social Research (ICPSR) in 1995 has a goal “to document research datasets and processes thoroughly so that data are independently understandable” (p. 30). It works to establish open structured metadata and citation standards for data. The contributions of this organization can serve to promote dataset discovery by providing systematic methods of finding data through standardized metadata and citations.

Others studied how specific information included in the metadata could provide more successful searching. One example of this is Saeeda and Kremen (2017) who proposed that the addition of temporal information could give clues to the usefulness of a dataset. For instance, if it was known that a particular dataset was obtained from a time period during World War II, certain conclusions could be made about the contents of the data.

Křemen and Nečaský (2019), from the Czechoslovakian Republic, looked at government data exclusively, expanding its metadata to include semantic vocabulary and then comparing it to other similar metadata schemas. Semantic vocabulary goes beyond typical metadata that uses name-value pairs, instead using attribute tags to define specific portions of the data. An example might be defining a political party as “<party>Democrat</party>”, or date as “<date>April 21, 2019</date>”, etc. Each attribute belongs to one of five schemas (basic, public sector, legal,

agenda, dataset vocabularies) which are layered in the architecture of the metadata. The aim of this study was to assess the feasibility using a semantic vocabulary in the narrow scope of government data. They determined that in terms of data discovery, there are advantages and disadvantages to this approach which will need more investigation. They also established that the workload for this task was not oppressive and could be realistically accomplished.

Library Focus

There have also been investigations of cataloging and indexing of datasets. Mannheimer et al. (2016) approached the problem of dataset discovery from an academic library perspective, examining the characteristics of a set of 20 cited or downloaded datasets from Thomson Reuters' Data Citation Index (DCI) or institutional repositories to gain insights into data discoverability and use. Discoverability and reuse was determined by citation and download counts. Each dataset was grouped by six characteristics (basic information, funding agency and journal information, linking and sharing, factors to encourage reuse, repository characteristics). After analysis, they observed that datasets are most easily discovered when they are well documented, formatted in a multitude of ways using different software (non-propriety file types), and exist in an open licensed trusted repository. In addition, they found that dataset discovery is positively influenced by the dataset being indexed in more than one web location, using a persistent identifier (e.g. Digital Object Identifiers, DOI), and being part of a project that requires data archiving due to external funding. They also mentioned that datasets appear to be most easily discovered in discipline specific repositories.

Read et al. (2015) reported on a project at the NYU School of Medicine to create a catalog for internal and external datasets. The metadata for the catalog was set up by librarians in consultation with the population health researchers. The metadata fields were designed to record information for both internal and external datasets but there were cases where different types of metadata were necessary, such as external datasets linking to who was using the data, and internal datasets with links to who collected the data. They did not rule out adding additional metadata fields for atypical datasets even though this would add elements that have little relevance to some of the entries. The reason for this was to provide inclusive information which would include "long tail" data and promote serendipitous encountering of information. While this study proved to be successful in establishing a local community of data sharing, they acknowledged the challenges of maintaining this catalog as it expanded to accommodate more data information in the future.

Swanson and Rinehart (2016) examined case studies where librarians must think outside the familiar "subject headings" paradigm to search for contextual data that meets the needs of their patrons. Furthermore, they explore the shift that must occur as academic libraries become an embedded participant in research. While this study focuses on resolution of various data problems within the library, it underscores possible deficiencies in the traditional approach to locating information about datasets.

Datasets for Teaching

There are times when finding a specialized type of dataset is necessary, to be used as a learning tool to demonstrate or teach particular techniques. Hoti, Francis, and Lancaster (2010) describe some of the challenges of locating teaching datasets, determining their desirable characteristics and possible sources. In this scenario, real datasets are usually preferred over concocted ones as they provide a more authentic experience and can spark meaningful discussions for practical problems. The authors list four beneficial characteristics for teaching datasets: open access, knowledge of time and location, thorough documentation of variables and context, and available publications to present how the data was used. There are also technical issues to consider such as the format of the dataset and associated teaching materials available. The authors suggest sources for teaching datasets but ultimately conclude that a dataset discovery resource is needed. They envision a collaborative effort using the metadata format proposed by the Data Documentation Initiative (DDI).

Search Engines

Certain authors suggested their own version of search engines that could assist the discovery process. Singhal and Srivastava (2017) tackle the problem of the researcher knowing what he/she wants without having the actual name of the dataset, calling this searching for the "application context." They articulate this by stating that the "main challenge of searching datasets with their application context is the lack of relevant information in the text description associated with the datasets on their source pages" (p. 82). To address this dilemma, they proposed two new search engine designs, one that queries based on the user's profile, and another

that develops a database containing dataset names along with their application context. The outcome showed an improvement over baseline when implementing the two elements of the proposed search engine. They also compared their new system to a traditional search engine (Bing) which demonstrated enhanced results.

Lu, Bangalore, Cormode, Hadjieleftheriou, and Srivastava (2012) built a dataset discovery tool that extracts datasets from articles and provides dataset resolution. They undertook the obstacles of identifying and obtaining dataset information from PDF files, offering researchers user friendly web interface to search for datasets.

Even Google has entered into the dataset search arena with Google Dataset Search. Noy, et al. (2019) discuss this specialized search engine which bills itself as an open ecosystem for datasets, meaning that variable metadata is created by dataset owners using the same schema (http://schema.org) format. They describe a particular challenge of creating a such a search engine: “Given a set of Web pages that publish dataset metadata, unknown in advance, build a search engine over this metadata to enable users to find datasets on those pages” (“Defining the Problem,” para. 4). Users can publish their own structured data (metadata) on a web page that describes and provides access to their dataset. That information will be picked up by the Google web crawler and be included in the Google Dataset search engine. Although this might not seem as trustworthy as a trusted repository, which provides reliable, long-term access to managed digital resources (Research Libraries Group, 2002), the popularity of Google might encourage dataset searching and provide some insight into what data is available.

Sharing Data

In her ethnological study, Jillian Wallis (2014) observes information behavior at two scientific conference meetings. For the first one, she attended a discussion between the presenter, Dr. Wonsuck Kim, a geophysicist, and his audience on "what would need to happen in order for modelers to be able to use the experimental data" (p. 102). This data, which emulated natural geophysical events, was produced by Kim and his colleagues. Kim wanted to share the data with other modelers and also capture additional data they had developed. It was available on the Internet but proved difficult to find. Wallis witnessed how Kim addressed this obstacle by encouraging feedback and collaboration from the meeting attendees.

At a second scientific workshop, Wallis (2014) observed a discussion and interaction of dataset producers and users of climate data. In particular, they ran into a problem with "the lack of metadata that would allow them to evaluate the fitness of the model for their application" (p. 104). This is because in global climate models, geographic features are reduced and details for a thorough analysis are not available. Wallis found that in both observations (the geophysical and climate events), the success of the data interchange depended on three things: metadata standards, education of dataset users, and gatekeepers who could assist both parties.

Yi Shen (2016) surveyed faculty at Virginia Tech to learn about their data needs and usage, noting that "significant differences between the colleges in openness of data suggest different cultures of data sharing activities and community practices" (p. 164). Understandably, it would be difficult to find data that was deemed to be sensitive and not intended for public use as

is the case in some disciplines. He noted that an additional hindrance to making data available is the time and effort it takes to publish it. Though there seems to be enthusiasm for sharing data, Shen (2016) confirmed that there is a gap between the perceived value of reusing data and the actual data being shared (p. 170).

Tenopir et al. (2011) conducted an extensive survey of over 1000 scientists, revealing that 75% of researchers would be interested in sharing their data, but believe that only a third of it could be easily accessed. That third is a lower amount that all other types of information (journals, technical reports, etc.) The results obtained from this survey reflect those of Shen (2016) regarding the gap between wanting to share data and actually making it available. Reasons for this include a lack of time and funding, and not having available server space for storage. This implies that there would be more data to discover if there was more of it shared. Interestingly, it was reported that about 78% of datasets do not use a standard metadata description, and there is scant awareness of metadata tools that could facilitate this.

Cragen, Palmer, Carlson, and Witt (2010) interviewed 20 scientists about their data practices for the Data Curation Profile Project. Again, with a focus from the viewpoint of sharing data, the results explained why sometimes data is not available for sharing. These include having too many requests and not having enough time to fulfill them, and having data misused in a way that was harmful to the current research. They also uncovered difficulties of ownership for cross disciplinary data and researchers not being knowledgeable in how to create metadata, reiterating the results of Tenopir et al (2011).

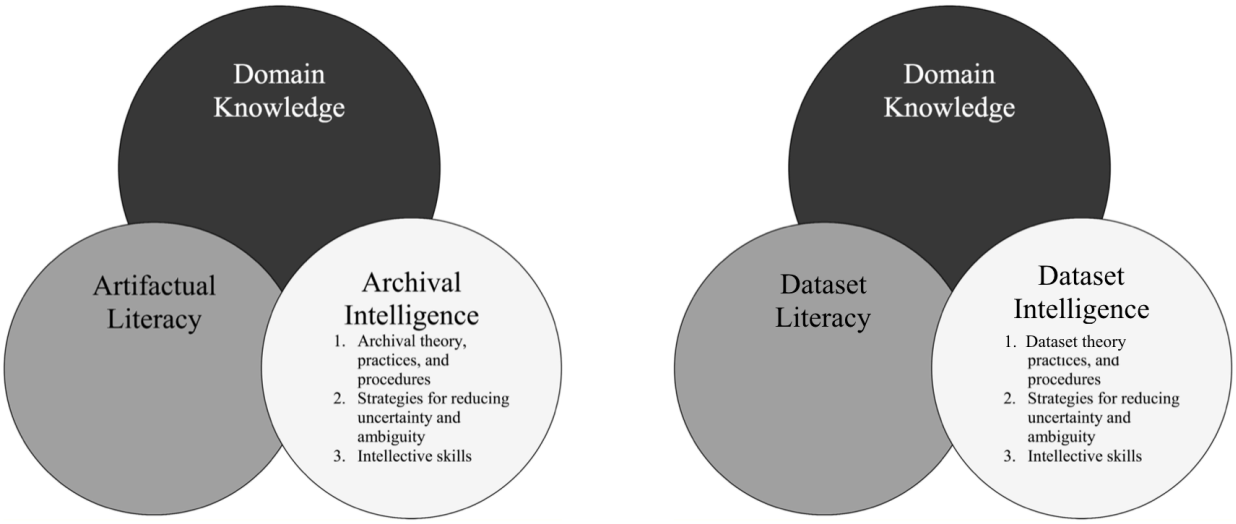
While there have been several articles that approach this topic from many different angles, existing literature is not sufficient to explain the process or mechanisms used by individual researchers as they search for relevant datasets. This case study will attempt to observe researchers in a semi-guided process as they explore, search, and make decisions that apply to their search in order to uncover situations that may have been previously overlooked.

Theoretical Model

This study examines dataset discovery through the lens of archival intelligence. Archival intelligence is "a researcher's knowledge of archival principles, practices, and institutions, such as the reasons underlying archival rules and procedures, how to develop search strategies to explore research questions, and an understanding of the relationship between primary sources and their surrogates" (Yakel and Torres, 2003, p. 52). Searching for and retrieving archived objects in an archival system requires a skill set that includes domain expertise, understanding how the metadata relates to the object, and being knowledgeable about the research environment. In the context of dataset discovery, archival (or dataset) intelligence is very similar. Although the terms are not interchangeable, digital archives and datasets are comparable, both having storage units considered to be "objects." Hoyle et al. (2016) defines a dataset as "a discrete collection of measurements collected via observation, experiments, or analysis, using specified methodologies and instruments, and structured in a manner documented by formal schemas" (p. 32). This compilation of data assets mirrors physical artifacts which may contain many parts assembled in a group. Mannheimer et al. (2016) also noticed the similarities between archiving and data curation stat-

ing, “The expertise developed in libraries is therefore applicable to data discoverability, with traditional cataloging and archiving skills closely paralleling the skills required to curate and preserve data (p. 2).

Image 1. Original vs. Modified Artificial Intelligence Theory



As shown in Image 1, dataset intelligence could have a theory all to itself with its own strategies and intellectual skills to facilitate understanding. In their paper on archival intelligence and user expertise, Yakel and Torres (2003) ask this question: Are there certain characteristics that can be identified to distinguish novice and expert users of archives and primary sources? One could ask this same question about users of datasets. This could also be an area where experienced librarians could contribute.

The theory of information horizons (Fisher, Erdelez, & McKechnie, 2015, p. 191) emphasizes the notion that information behavior is a process, constricted or enabled by its contextual landscape. Although this theory is not directly considered for this paper, a recommended tool from this section of the text, Theories of Information Behavior (Fisher et al., 2015) will be employed. A hand drawn information horizon map, defined as "all information resources, including people, he or she typically access when seeking information in the context that is the focus of the research study" (p. 195), will be used to assist in understanding the resources available to the participant and their significance.

This map, along with its explanation by the participant, will provide detail about the relationships among the resources, revealing concentrated information avenues that may be available to him/her. Alternatively, information boundaries may be exposed due to economics or policies of the data searcher's institution. Awareness of these situations may stimulate a conversation about the participant's satisfaction with a particular method of searching.

Problem Statement and Research Questions

Studies have shown that researchers are interested in availing themselves of existing datasets but have yet to embrace this practice because of the difficulty in locating them. This case study will investigate the experiences of a specific group of faculty and students as they search for datasets to meet their needs. Additionally, an exploratory question of how librarians can assist patrons searching for datasets will be contemplated for use in further studies.

RQ1: How can researchers optimize their search for relevant datasets?

RQ2: What are the opportunities for librarians to assist with dataset searches?

Methodology and Methods

Research Field

Recent published and unpublished manuscripts by faculty and students at the University of Missouri that show the application of at least one dataset will be selected. The authors of those manuscripts will be contacted to see if they are interested in participating in this study. If so, they will be invited to spend an hour at the Allen Institute for Research on Learning, Information & Technology on the University of Missouri campus. Follow up interviews will be conducted if necessary. There will be approximately five participants, from different disciplines if possible.

Data Collection Method

Data will be collected from four sources. For each participant, at least one article referencing datasets that they have written will be read. The invitee will also be interviewed using demographic and open-ended questions (see Appendix A) about the search techniques they have used in the past. During this interview, they will be asked to sketch an information horizon map (Fisher et al., 2005, p.195) to outline their search process. After the interview, they will be asked to assume that a previously specified dataset was not available and to demonstrate on a computer or on paper the strategies that they might employ to find a similar one. This think aloud portion of the exchange will be video recorded, with hand written field notes and reflective memos documented by the observer.

Data Analysis and Results

Data collected for each case will be assembled in separate participant folders. Each folder will contain the person's article(s), interview, video, and a copy of their hand drawn horizon map. Transcriptions of the interview and think aloud video, as well as the observer's notes and memos will be added to the folder. A table representation for the horizon map will also be included indicating the participant's preferences for information resources. This map will be analyzed to determine preferences and strong resource paths of the participant. Each case will be reviewed with the author adding additional notes/memos, resulting in a writeup describing the content and the results of the horizon map analysis. The report for each case will be reviewed by the informant to verify accuracy. Themes and patterns will be generated using a within-case analysis (Creswell and Poth, 2018, p. 100) and triangulated between the sources. The themes generated for each case will then be added to a spreadsheet with notes to compare all the cases using a cross-case analysis (Creswell and Poth, 2018, p. 100). The theme generation will continue in an iterative manner, adding additional concepts as necessary.

Finally, an overall analysis of the themes will be performed to determine optimal practices for finding datasets. Exceptions between disciplines and experience level will be weighed, and meaningful quotes will be extracted. At this time, the inductive design will be evaluated to see if any of the specific strategies can be applied broadly to dataset searching and if there are ways that libraries could assist in this search.

Limitations

The main limitation will be that the participants will have differing levels of knowledge and expertise. Certain disciplines encountered in this study may promote less or a greater use of datasets, causing a disparity in the findings. There is a chance that certain methods of searching might be more popular at this location, the University of Missouri, and not include other useful techniques. This study will not include all possible areas of research and therefore may not be a comprehensive representation of dataset searching.

Timeline

Activity	Length	Date
Proposal	16 weeks	May 15, 2019
Search for University of Missouri articles using dataset	4 weeks	June 15, 2019
Contact authors and invite them to participate	4 weeks	July 15, 2019
Conduct interviews	4 weeks	August 15, 2019
Transcribe data	4 weeks	September 15, 2019
Data analysis	4 weeks	October 15, 2019
Write up	4 weeks	November 15, 2019

Conclusion/Expected Results/Implications

This project will demonstrate commonly used techniques used for dataset searching. By interviewing and observing researchers in an open format, it will not only provide insight into the processes followed and issues encountered, but will also allow new innovative ideas to be un-

covered. Further studies could include a wider variety of disciplines and methods, or conversely, drill down to focus on one or two of the approaches that seem to be the most promising.

As a follow up to this study, the results could be compared with data services offered by academic libraries that could support researchers searching for datasets.

References

Chen, X., Gururaj, A. E., Ozyurt, B., Liu, R., Soysal, E., Cohen, T., … Xu, H. (2018). DataMed – an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association*, *25*(3), 300–308. Retrieved from https://doi.org/10.1093/jamia/ocx121

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023-4038.

Creswell, J. W., & Poth, C. N. (2017). Qualitative inquiry and research design: Choosing among five approaches. Sage publications.

Dekker, R. (2006). The importance of having datasets. Retrieved from https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1760&context=iatul

Esselink, J. (2017, August 8). *Today big data and analytics is everywhere for everyone*. Retrieved from https://www.ibm.com/blogs/think/nl-en/2017/03/08/today-big-data-analytics-everywhere-everyone/

Fisher, K. E., Erdelez, S., & McKechnie, L. (Eds.). (2005). *Theories of information behavior*. Information Today, Inc..

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository. *Aslib Journal of Information Management*, *68*(4), 478-494.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library trends*, *57*(2), 280-299.Hoti, V., Francis, B., & Lancaster, G. (2010, July). Resource discovery for teaching datasets. Proceedings of the Eighth International Conference on Teaching Statistics: *Data and Context in Statistics Education: Towards an Evidence-based Society.*

Hoti, V., Francis, B., & Lancaster, G. (2010, July). Resource discovery for teaching datasets. Proceedings of the Eighth International Conference on Teaching Statistics: *Data and Context in Statistics Education: Towards an Evidence-based Society.*

Hoyle, L., Vardigan, M., Greenfield, J., Hume, S., Ionescu, S., Iverson, J., ... & Witt, M. (2016). DDI and enhanced data citation. *IASSIST Quarterly*, *39*(3), 30-30.

Křemen, P., & Nečaský, M. (2019). Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary. *Journal of Web Semantics*, *55*, 1–20. https://doi.org/10.1016/j.websem.2018.12.009

Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M., & Srivastava, D. (2012). A dataset search engine for the research document corpus. In *2012 IEEE 28th International Conference on Data Engineering* (pp. 1237–1240). Arlington, VA, USA: IEEE. https://doi.org/10.1109/ICDE.2012.80

Mannheimer, S., Sterman, L. B., & Borda, S. (2016). Discovery and reuse of open datasets: an exploratory study. *Journal of eScience Librarianship*, *5*(1).

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit.

Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem.

Olfat, H., Kalantari, M., Rajabifard, A., Senot, H., & Williamson, I. P. (2012). Spatial metadata automation: A key to spatially enabling platform.

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt4xf018wx/qt4xf018wx.pdf?t=p7cttq

Read, K., Athens, J., Lamb, I., Nicholson, J., Chin, S., Xu, J., ... & Surkis, A. (2015). Promoting data reuse and collaboration at an academic medical center. *IJDC*, *10*(1), 260-267.

Research Libraries Group. (2002). Trusted digital repositories: Attributes and responsibilities. Retrieved from https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

Saeeda, L., & Kremen, P. (2017). Temporal knowledge extraction for dataset discovery. In *PROFILES@ ISWC*.

Shen, Y. (2016). Research data sharing and reuse practices of academic faculty researchers: A Study of the Virginia Tech Data Landscape. *International Journal of Digital Curation*, *10*(2), 157-175.

Singhal, A., & Srivastava, J. (2017). Research dataset discovery from research publications using web context. *Web Intelligence*, *15*(2), 81–99. https://doi.org/10.3233/WEB-170354

Swanson, J., & Rinehart, A. K. (2016). Data in context: Using case studies to generate a common understanding of data in academic libraries. *The Journal of Academic Librarianship*, *42*(1), 97-101.

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.

Wallis, J. (2014). Data Producers Courting Data Reusers: Two Cases from Modeling Communities. *IJDC*, *9*(1), 98-109.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS one*, *8*(7), e67332.

Wilkinson, J. M., Pollard, T., & Farquhar, A. (2010). British library dataset programme: supporting research in the library of the 21st century. *Liber Quarterly*, *20*(1).

Yakel, E., & Torres, D. (2003). AI: archival intelligence and user expertise. *The American Archivist*, *66*(1), 51-78.

Appendix A

Interview Questions

1. What is your role at the University of Missouri?
2. How long have you been in that position?
3. What kind of research have you done?
4. Can you tell me about this article? (Show chosen article written by participant.)
5. What kind of data did you use for your article?
6. How did you come to use the data that you did?
7. Can you draw an information horizon map? (May need to explain this.)
8. What kind of assistance did you have to help you obtain this dataset?
9. Were you satisfied with the data that you obtained?
10. What would you do differently next time?