Discovery and Reuse of Open Datasets: An Exploratory Study

Authors: Sara Mannheimer, Leila Belle Sterman, Susan Borda

Mannheimer, Sara, Leila B. Sterman, and Susan Borda. 2016. "Discovery and Reuse of Open Datasets: An Exploratory Study." Journal of eScience Librarianship 5(1): e1091. http://doi.org/10.7191/jeslib.2016.1091

Journal of eScience Librarianship

Volume 5 | Issue 1 Article 5

2016-07-19

Discovery and Reuse of Open Datasets: An Exploratory Study

Sara Mannheimer
Montana State University-Bozeman

Leila Belle Sterman
Montana State Univeristy-Bozeman

Susan Borda
Montana State University-Bozeman

Recommended Citation

Mannheimer, Sara, Leila B. Sterman, and Susan Borda. 2016. "Discovery and Reuse of Open Datasets: An Exploratory Study." *Journal of eScience Librarianship* 5(1): e1091. http://dx.doi.org/10.7191/jeslib.2016.1091

ISSN 2161-3974

JeSLIB 2016; 5(1): e1091
doi:10.7191/jeslib.2016.1091

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Discovery and Reuse of Open Datasets: An Exploratory Study

Sara Mannheimer, Leila Belle Sterman and Susan Borda

Montana State University, Bozeman, MT, USA

Abstract

Objective: This article analyzes twenty cited or downloaded datasets and the repositories that house them, in order to produce insights that can be used by academic libraries to encourage discovery and reuse of research data in institutional repositories.

Methods: Using Thomson Reuters' Data Citation Index and repository download statistics, we identified twenty cited/downloaded datasets. We documented the characteristics of the cited/downloaded datasets and their corresponding repositories in a self-designed rubric. The rubric includes six major categories: basic information; funding agency and journal information; linking and sharing; factors to encourage reuse; repository characteristics; and data description.

Results: Our small-scale study suggests that cited/downloaded datasets generally comply with basic recommendations for facilitating reuse: data are documented well; formatted for use with a variety of software; and shared in established, open access repositories. Three significant factors also appear to contribute to dataset discovery: publishing in discipline-specific repositories; indexing in more than one location on the web; and using persistent identifiers. The cited/downloaded datasets in our analysis came from a few specific disciplines, and tended to be funded by agencies with data publication mandates.

Conclusions: The results of this exploratory research provide insights that can inform academic librarians as they work to encourage discovery and reuse of institutional datasets. Our analysis also suggests areas in which academic librarians can target open data advocacy in their communities in order to begin to build open data success stories that will fuel future advocacy efforts.

Correspondence: Sara Mannheimer: sara.mannheimer@montana.edu

Keywords: open data, data discovery, data reuse, institutional data repositories

Introduction and Background

A fundamental role of libraries is that of information access provider, and at its core, data discovery is simply a form of information access. 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. Building from this foundation of information access, libraries are well-equipped to suggest data description practices and repository features that will encourage discovery and reuse (Witt, Carlson, Brandt, and Cragin 2009; Wallis, Mayernik, Borgman, and Pepe 2010; Faniel, Minor, and Palmer 2014). In this article, we analyze twenty cited or downloaded datasets and the repositories that house them, in order to produce insights that can be used by academic libraries to encourage discovery and reuse of research data in institutional repositories.

Todd Vision describes data as “a classic example of a public good, in that shared data do not diminish in value” (2010, 330). This sentiment is a guiding tenet of the Open Data movement, which aims to make research data freely and publicly available. The movement has been strengthened in the United States by two recent policy developments: first, major funding agencies have begun to require data management plans (NIH 2003; NSF 2011), and second, several prominent journals (Dryad 2011; PLOS 2014) and the Office of Science and Technology Policy (Holdren 2013) have issued policies requiring that supporting data be published alongside associated articles. These policies encourage the practice of open data publishing for two key reasons. First, open data supports reproduction and validation of research (Santer, Wigley, and Taylor 2011; Lutter, Barrow, Borgert, Conrad, Edwards, and Felsot 2012). Second, open data encourages the repurposing of research data in order to promote new discoveries and advance science (Kelder 2005; Faniel and Jacobsen 2010).

The ascendency of the Open Data movement has resulted in a growing number of repositories providing access to research data in addition to publications. The Registry of Research Data Repositories[1] currently includes over eight hundred repositories run by institutions in the United States and over fifteen hundred repositories worldwide (Registry of Research Data Repositories 2016). Discipline-specific repositories like National Center for Biotechnology Information[2] and Worldwide Protein Data Bank[3] facilitate disciplinary data sharing, while general-purpose repositories like Dryad[4] and Figshare[5] attempt to fill the gaps by housing data from a range of disciplines. These general data are often “long tail data” — described by Wallis, Rolando, and Borgman as tending to be “small in volume, local in character, intended for use only by [the research team], and less likely to be structured in ways that allow data to be transferred easily between teams or individuals” (2013). Institutional repositories in academic libraries — initially built to provide open access to publications (Crow 2002; Lynch 2003) — are a natural fit for institutionally-produced research data, and especially long tail data that may not fit the scope of a discipline-specific repository.

Data sharing culture varies between scientific disciplines. Strong cultures of data sharing exist in geophysics, molecular biology, and ecology (Nelson 2009). Social scientists and medical researchers — who often produce human subject data or other sensitive data that requires

[1] http://re3data.org
[2] http://www.ncbi.nlm.nih.gov
[3] http://www.wwpdb.org
[4] http://datadryad.org
[5] https://figshare.com

more effort to prepare for publication—are less likely to share their research data online (Tenopir et al. 2011). Although the practice of open data sharing is on the rise, the literature has yet to clearly demonstrate whether published datasets are being discovered and reused. As Wallis, Rolando, and Borgman (2013) inquire, “if we share data, will anyone use them?” Publishing data openly is only the first step toward successful data sharing. To realize the goals of the Open Data movement, published datasets must be discoverable and reusable.

Promoting discovery of datasets is a complex process. Researchers have traditionally found data by reading published literature, talking with professional peers, or searching trusted data repositories (Zimmerman 2007). Recent work has also explored semantic web applications to promote web-scale discovery of open access repository resources through implementation of the Research Description Framework (RDF) and schema.org metadata (Latif, Borst, and Tochtermann 2014). However, RDF and schema.org metadata implementation have only been preliminarily explored in the specific context of data repositories (Rosati and Mayernik 2013). Beyond discoverability, once a researcher finds an applicable dataset, the data must also be reusable. White et al. (2013) suggest three strategies to encourage data reuse: (1) document data well; (2) format data for use with a variety of software; (3) share data in established repositories with open licenses. In this article, we aim to identify common characteristics of cited/downloaded datasets and their repositories. We propose that these common characteristics can be used to provide insights for academic librarians looking to increase discovery and reuse of datasets published in institutional data repositories.

Methods

Measuring reuse of datasets is a difficult endeavor. Researchers have used several different methods to attempt to track reuse. Piwowar, Carlson, and Vision (2012) searched Google Scholar for the accession number, DOI, and journal name for 100 datasets, in order to find studies that mention dataset reuse in the text of the article. Chao (2012) identified the affiliated publications for datasets in order to take advantage of more traditional bibliometric measurement methods. In a smaller-scale study, Belter (2014) used a combination of Web of Science, full-text search capabilities provided by journal publishers' websites, and Google Scholar in order to measure reuse of oceanographic data sets.

In order to identify datasets that have been discovered and reused, our research team opted to use data citation counts and data download counts. We consider citation count to be a more accurate measure of reuse than download count because citations are proof of use, whereas downloads simply hint at use. As Konkiel writes, “we cannot be sure if downloads mean that the dataset has been used in any way, just as we cannot be sure that downloads of journal articles guarantee a paper has been read” (Konkiel 2013). Consequently, citations are our preferred metric of measurement. However, we also consider download counts to be a useful metric for measuring reuse, partly due to the results of a 2015 survey conducted by Kratz and Strasser. The authors write:

> “We asked what metrics researchers would most respect when evaluating a dataset’s impact. Respondents considered number of citations to be the most useful metric; 49% (n = 119) found citation count highly or extremely useful. Unexpectedly, a substantial 32% (n = 77) felt the same way about number of downloads” (Kratz and Strasser 2015).

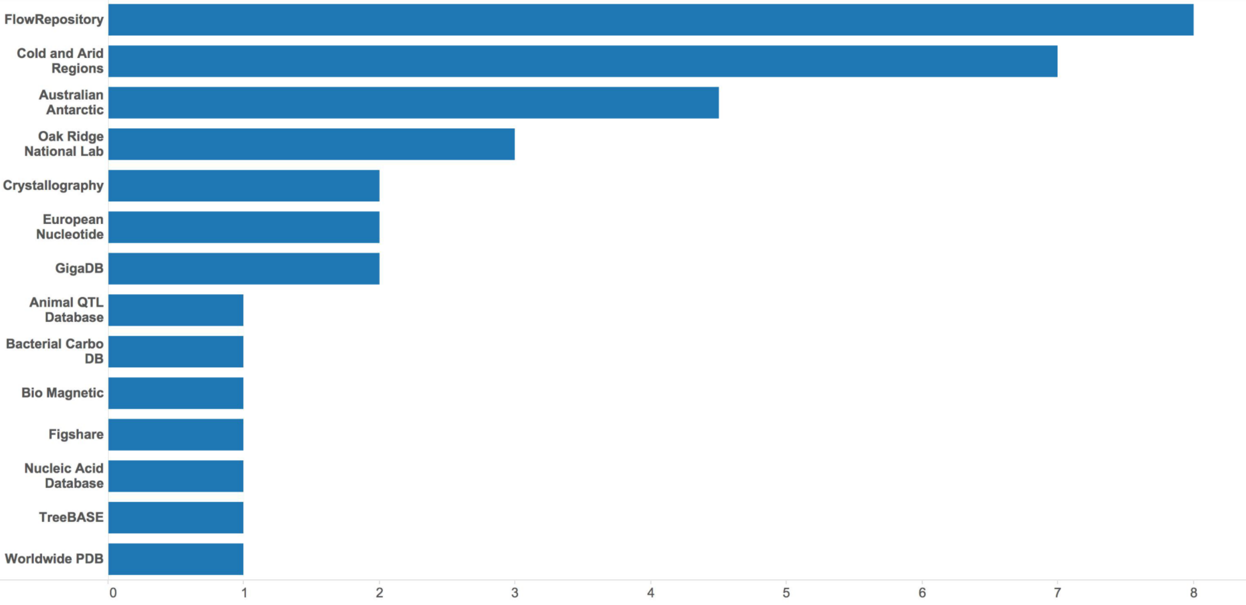
While download counts are a less concrete measure of reuse than citations, this survey result indicates that researchers believe that download metrics can reflect reuse.

Tracking downloads in addition to citations in this study was also necessary from a practical standpoint. Including datasets published in institutional repositories was important to gain a broader understanding of dataset characteristics that may influence discovery and reuse. Since few citation statistics were available for institutional data repositories, we were compelled to include downloads as a measure of dataset reuse for these repositories.

We used Thomson Reuters' Data Citation Index (DCI) (Thomson Reuters 2016a) to identify cited datasets. The DCI is a subscription-based database on the Web of Science that indexes data repositories and reports the number of articles that cite individual datasets. In order to index data repositories, the DCI requires that the repositories be "demonstrably active, whether by continued maintenance and curation of the data sets held, or by addition of new materials, evidenced by data deposition statistics" (Force and Robinson 2014). When choosing data repositories to index, the DCI also looks for robust metadata, evidence of repository persistence, funding statements, peer review, and links between datasets and the research literature. The DCI continually monitors the repositories it indexes for availability, quality, and relevance to the DCI. The DCI does not track citations itself, but rather aggregates this information as collected by data repositories (Thomson Reuters 2016b).

In order to select datasets for our analysis, we assumed that more data has been published and reused in recent years, due to data archiving mandates from academic journals and funding agencies. In order to provide a sufficient amount of time for datasets to be discovered, used, and cited, we limited our results to data published in 2013 (three years prior to this study). It is important to note that since our search was limited to datasets published in 2013 and indexed in the DCI, each dataset chosen for analysis in this paper may not be the highest-cited dataset in its repository — it is merely the highest-cited dataset that was published in

Figure 1: Median Citations by Repository



2013 and shows citations in the DCI. Our initial search returned 763,057 cited datasets. Since the DCI limits the amount of data a user can extract to blocks of five hundred cited datasets, we downloaded the top-cited one thousand datasets. From these one thousand datasets, we chose the fourteen repositories with the highest median citations per dataset in the DCI (see Figure 1). We then conducted our exploratory analysis using the top-cited dataset from each of these fourteen repositories.

Among the repositories with the highest median citations in the DCI in 2013, the number of citations drops quickly from a median citation of eight in the Australian Antarctic Data Center to a median of one citation per dataset in the Animal QTL Database, as illustrated by Figure 1.

Since there were no institutional data repositories with citations reported in the DCI for 2013, we produced a convenience sample of six Digital Library Federation member institutions:

1. Cornell University eCommons[6]
2. Johns Hopkins Data Archive Dataverse Network[7]
3. Oregon State University ScholarsArchive@OSU[8]
4. Purdue University Research Repository (PURR)[9]
5. Data Repository for the University of Minnesota (DRUM)[10]
6. University of New Mexico LoboVault[11]

If the institutional data repository indicated a most-downloaded or most-cited dataset, we used that dataset for our analysis. If no repository-wide download statistics were available, we selected a highly-downloaded dataset.

The datasets in our final twenty results reflect either citations in the DCI or a high number of downloads in an institutional repository. We documented the characteristics of the cited/ downloaded datasets and their corresponding repositories by reviewing publicly-available information on repository websites and inputting our observations into a self-designed rubric. The rubric addresses the characteristics of cited/downloaded datasets and their repositories by grouping them into six major categories: basic information; funding agency and journal information; linking and sharing; factors to encourage reuse; repository characteristics; and data description (see Appendix A for blank rubric; the completed rubric is available from Montana State University ScholarWorks http://doi.org/10.15788/m2059z). The rubric allowed us to identify common characteristics of cited/downloaded datasets.

[6] https://ecommons.cornell.edu
[7] https://archive.data.jhu.edu/dvn
[8] https://ir.library.oregonstate.edu/xmlui
[9] https://purr.purdue.edu
[10] https://www.lib.umn.edu/datamanagement/drum
[11] https://repository.unm.edu

Results

From our sample of twenty cited/downloaded datasets and their corresponding repositories, we identified the following characteristics from which we can gain insight into factors that may encourage discovery and reuse.

Our analysis reveals that the cited/downloaded datasets in our sample generally comply with the basic recommendations for facilitating reuse, outlined by White et al. (2013):

- The data are documented well. 95% (19/20) of the datasets analyzed have a readme file or extensive metadata.
- The data are formatted for use with a variety of software. 80% (16/20) are available in non-proprietary file formats.
- The data are shared in established repositories with open licenses. All twenty datasets are published in established repositories, and all twenty datasets are openly accessible. However, only four out of the twenty datasets have explicit licenses, all of which are Creative Commons Licenses — three are licensed CC BY, and one is placed in the public domain using CC 0.

Beyond these best practices for reuse, we identify additional factors that appear to influence dataset discovery (see Table 1).

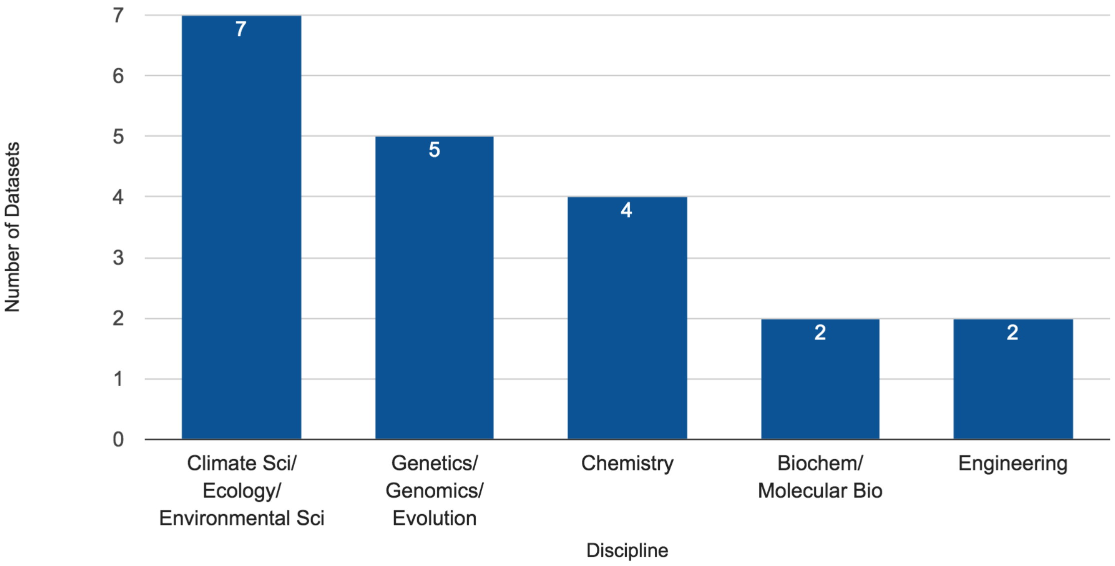
- 60% (12/20) of the datasets analyzed are indexed in more than one location on the web. For example, the most-cited dataset in our results is available from the Australian Antarctic Data Centre; additionally, metadata and a link to the dataset are available from the Global Change Master Directory.
- A persistent identifier also appeared to influence discovery and reuse; all (20/20) of these cited/downloaded datasets have a persistent identifier, eight of which are Digital Object Identifiers (DOIs).
- Data mandates also appear to contribute to citations and downloads; of the fifteen datasets that disclose an external funding source, nine (60%) are funded by agencies that require data archiving.

Table 1: Characteristics of Cited/Downloaded Datasets

Criterion	Value	Percent
Readme/Extensive Metadata	19	95%
Non-Proprietary File types	16	80%
External Indexing	12	60%
Persistent Identifier	20	100%
Funder requires data archiving	9 (out of 15 that disclosed funding)	60%

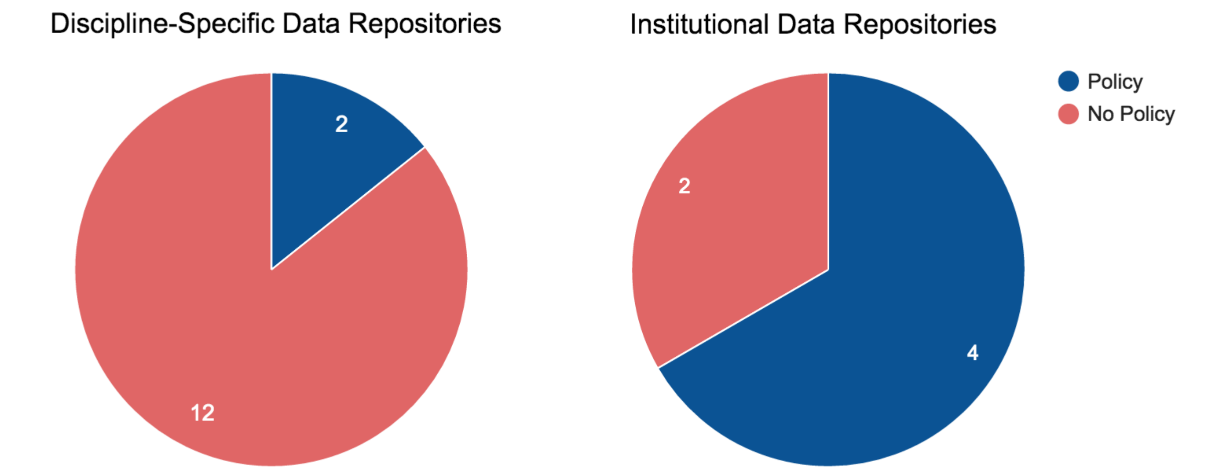
The cited/downloaded datasets in our results can be grouped into five broad disciplines: Climate Science, Ecology, & Environmental Science; Genetics, Genomics, & Evolution; Chemistry; Biochemistry & Molecular Biology; and Engineering (see Figure 2). This finding reinforces existing research showing that some disciplinary cultures support data sharing and reuse more than others (Nelson 2009; Tenopir et al. 2011). This culture of reuse extends to the creation of discipline-specific repositories in these disciplines. If data repositories are established elements of the disciplinary research ecosystem, researchers are more likely to discover and reuse data from those repositories, regardless of metadata, file type, or other factors.

Figure 2: Academic Disciplines



This finding suggests that datasets are most easily discoverable in discipline-specific repositories. It seems to follow that libraries should recommend that researchers deposit in major disciplinary data repositories. Unfortunately, our research showed that, of the fourteen disciplinary repositories in our study, only two (~14%) had preservation policies. This stands in contrast to the institutional repositories in our study, four out of six (~67%) of which had preservation policies (see Figure 3). Revisiting the datasets analyzed in our rubric five months after initial data collection, two out of the fourteen datasets in discipline-specific repositories (~14%) were unavailable online. On January 24, 2016, the Treebase Repository produced a 502 error, and the Animal QTL Database reported that the persistent identifier for the cited dataset in our analysis could not be found. While this trend suggests a conflict between discovery and preservation, our small sample size of 20 repositories limited the scope of our results; a larger study would allow for more conclusive results. Still, a key library mission is to ensure long-term preservation of information. Since the value of research datasets will persevere, preservation is an important consideration. Librarians should carefully evaluate discipline-specific repositories through the lens of preservation before making recommendations to researchers.

Figure 3: Data Repository Preservation Policies



Upon closer examination, two seemingly significant characteristics — whether a repository provides a suggested data citation, and whether a dataset underlies an associated research publication — are less clearly related to discovery and reuse. While 55% (11/20) of cited/ downloaded datasets are published in repositories that offer a suggested citation, 20% (4/20) of the repositories we analyzed suggested citing the associated publication, rather than providing a suggested citation specific to the dataset itself. We produced a similarly unclear result when analyzing whether cited/downloaded datasets are associated with a specific research publication. While 75% (16/20) of the cited/downloaded datasets in this study are associated with a publication, only 37.5% (6/16) of those publications cite or link to the associated dataset. Past research has suggested that researchers find data by reading published articles (Zimmerman 2007); this practice does not appear to be reflected in our research. However, while the absence of data references in the associated publications analyzed here is an interesting result, our sample size is too small to seriously call into question whether an associated publication directly leads to dataset discovery.

Discussion

The results of this exploratory study generated insights that may help academic libraries encourage discovery and reuse of institutional datasets.

From our analysis, it appears that the following factors may facilitate dataset reuse:

- Robust data description
- Non-proprietary file types
- Publication in open access repositories

Many libraries already provide guidance in these areas. Our research suggests that extending and expanding these services would be beneficial.

It appears that the following factors may facilitate dataset discovery:

- Publication in prominent, discipline-specific repositories (after evaluating for sustainability and preservation activities)
- Cross-indexing between institutional data repositories, discipline-specific repositories, and discipline-specific metadata catalogs
- Persistent identifiers, especially DOIs

This research suggests that datasets published in discipline-specific repositories may be more discoverable. However, discovery is only one part of the open data equation; librarians should carefully evaluate repositories' preservation activities before making recommendations. Once datasets are deposited and published in trustworthy discipline-specific repositories, institutional data repositories can provide metadata records for these datasets in order to further encourage discovery. Correspondingly, libraries can request that datasets published in the institutional data repository be indexed in appropriate discipline-specific data repositories or catalogs. DOIs and other persistent identifiers also appear to facilitate discovery and reuse.

Finally, our analysis suggests that cited/downloaded datasets are:

- Funded by agencies that require data publication
- Produced by researchers in a few specific disciplines

These findings suggest that academic librarians may be able to target open data advocacy in their communities, directly soliciting datasets from certain disciplines and from grant awardees whose funders require data publication. By providing these targeted services, libraries are well-positioned to encourage discovery and reuse, and to begin to build open data success stories that will fuel future advocacy efforts.

Limitations and Future Directions

This exploratory research produced promising insights into the factors that influence data discovery and reuse. We note a number of limitations to our study. Our exploratory approach included a small sample of datasets (n = 20), including a convenience sample of institutional data repositories. Also, the Data Citation Index is an imperfect tool to measure data citations. We note three major limitations related to our use of the DCI to conduct this research. First, the DCI relies on direct reporting from repositories. This limits our discipline-specific repository results to data citations that have been reported to the DCI. Second, the DCI did not report citations for data in institutional data repositories for 2013. In order to include institutional data repositories in our sample, we had to extend our reuse metrics to download statistics — an even less clear-cut measure of reuse. Third, for several of the datasets with a single citation reflected in the DCI, the dataset and the citing article are created by the same author. The fact that these datasets are indexed in the DCI suggests better discoverability; however, while the single citation reflected in the DCI for these datasets does indicate use, it may not indicate reuse. Our decision to search the DCI only for datasets published in 2013 likely affected the number of citations per dataset. Future research should investigate all datasets in the DCI, in order to better identify datasets with high numbers of citations.

Lastly, our study was limited by an unfortunate reality: in the data sharing community, there is an absence of standard data citation practices or other data reuse metrics (Parsons, Duerr, and Minster 2010). While the DCI evaluates indexed repositories for overall quality (Thomson Reuters 2016b), it does not evaluate the accuracy of each repository's data citation tracking methods. In order to gauge the efficacy of the strategies suggested by this research — and in order to conduct future, more conclusive research — we must first be able to reliably measure data reuse.

A 2011 editorial by Michael Whitlock suggests that scientists who reuse data could go so far as to offer co-authorship to original data creators. "At the very least," he writes, "whenever data are reused, researchers must cite not only the paper or papers in which they were originally described, but also the data package itself" (Whitlock 2011, 63). Proper attribution is also an important incentive for data sharing. Tenopir et al.'s 2011 study found that one of the most important conditions that scientists had for sharing data was that they receive proper citation credit from those who use the data. The Digital Curation Center recommends that data be cited in the manner of traditional publications — as entries in an article's reference list (Ball and Duke 2015). However, this recommendation has yet to receive full uptake from the scholarly community — datasets rarely receive traditional citations (Sieber and Trumbo 1995; Mooney and Newton 2012; Robinson-Garcia, Jiménez-Contreras, and Torres-Salinas 2015). In 2010, a preliminary study found that only 33% of repositories, 6% of journals, and .02% of funders suggested a best practice for data citation (Enriquez et al. 2010). Even if datasets are cited, these citations are often difficult to track, due to the lack of established practices for documenting data reuse (Mayernik 2013).

Progress is being made to promote data citation practice. DOIs provide a single, persistent identifier that can be used for citation and access of datasets (Simons 2012). The DataCite project further facilitates discovery and machine-readability by providing dataset DOIs that include an underlying dataset-specific XML metadata scheme (Starr and Gastl 2011). CODATA-ICSTI released reports on the landscape of data citation in 2012 (National Research Council 2012) and 2013 (CODATA-ICSTI Task Group 2013), and a cooperative working group released the Force11 Joint Declaration of Data Citation Principles in 2014, stating that "data should be considered legitimate, citable products of research" (Data Citation Synthesis Group 2014). The Making Data Count project (Making Data Count 2016) further examines and encourages data citation practices and data reuse metrics. As these initiatives help establish standards for data citation, data reuse will become easier to track. With better data citation tracking, more robust conclusions can be reached regarding how to support the discovery and reuse of datasets.

Conclusion

The Open Data movement aims to encourage data availability for the purpose of discovery and reuse. Through analysis of cited/downloaded datasets and their corresponding repositories, the exploratory research described in this paper reveals two complementary insights:

- The common characteristics of cited/downloaded datasets and their corresponding repositories can provide direction for librarians looking to facilitate discovery and reuse of datasets published in institutional data repositories.

- Data sharing and reuse appear to happen relatively rarely, and even when researchers do reuse data, citation practices are inconsistent.

Academic libraries can facilitate dataset reuse by encouraging researchers to document data thoroughly, to use non-proprietary file types, and to publish data in open access repositories. Dataset discovery can also be facilitated by encouraging researchers to publish data in trustworthy, discipline-specific repositories, by cross-indexing institutional repositories with discipline-specific repositories and discipline-specific metadata catalogs, and by using persistent identifiers. Moreover, academic librarians — with their data management expertise and their knowledge of scholarly communication — are well-positioned to help develop standardized data citation practices, to advocate for open data, and to educate researchers about data sharing best practices. Drawing upon the insights gained from this research, academic librarians can begin to tailor services and direct future research, ultimately improving the landscape surrounding dataset discovery and reuse.

Supplemental Content

Appendix A
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2016.1091 under "Additional Files".

Disclosure

The authors report no conflict of interest.

References

Ball, Alex, and Monica Duke. 2015. "How to Cite Datasets and Link to Publications." *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Last modified July 30. http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Belter, Christopher W. 2014. "Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets." *PLOS One* 9(3): e92590. http://doi.org/10.1371/journal.pone.0092590

CODATA-ICSTI Task Group on Data Citation Standards and Practices. 2013. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data." *Data Science Journal* 12, pp.CIDCR1–CIDCR7. http://doi.org/10.2481/dsj.OSOM13-043

Chao, Tiffany C. 2011. "Disciplinary Reach: Investigating the Impact of Dataset Reuse in the Earth Sciences." *Proceedings of the American Society for Information Science and Technology* 48(1): 1-8. http://doi.org/10.1002/meet.2011.14504801125

Crow, Raym. 2002. "The Case for Institutional Repositories: A SPARC Position Paper." *ARL Bimonthly Report* 223: 1-15. http://www.arl.org/storage/documents/publications/arl-br-223.pdf

Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*, Edited by Maryann Martone. San Diego CA: FORCE11, 2014. https://www.force11.org/datacitation

Enriquez, Valerie, Sarah Walker Judson, Nicholas M. Walker, Suzie Allard, Robert B. Cook, Heather A. Piwowar, Robert J. Sandusky, Todd J. Vision, and Bruce E. Wilson. 2010. "Data Citation in the Wild." *Nature Proceedings*. http://doi.org/10.1038/npre.2010.5452

Faniel, Ixchel M., and Trond E. Jacobsen. 2010. “Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues’ Data.” *Computer Supported Cooperative Work (CSCW)* 19(3): 355-375. http://doi.org/10.1007/s10606-010-9117-8

Faniel, Ixchel M., David Minor, and Carole L. Palmer. 2014. “Putting Research Data into Context: Scholarly, Professional, and Educational Approaches to Curating Data for Reuse.” *Proceedings of the American Society for Information Science and Technology* 51(1): 1-4. http://doi.org/10.1002/meet.2014.14505101016

Force, Megan M., and Nigel J. Robinson. 2014. “Encouraging Data Citation and Discovery with the Data Citation Index.” *Journal of Computer-Aided Molecular Design* 28(10): 1043-1048. http://doi.org/10.1007/s10822-014-9768-5

Holdren, John P. 2013. “Increasing Access to the Results of Federally Funded Scientific Research.” *White House Office of Science and Technology Policy.* Released February 22. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Kelder, Jo-Anne. 2005. “Using Someone Else's Data: Problems, Pragmatics and Provisions.” *Forum: Qualitative Social Research* 6(1). http://www.qualitative-research.net/index.php/fqs/article/view/501

Konkiel, Stacy. 2013. “Tracking Citations and Altmetrics for Research Data: Challenges and Opportunities.” *Bulletin of the American Society for Information Science and Technology* 39(6): 27-32. http://doi.org/10.1002/bult.2013.1720390610

Kratz, John E., and Carly Strasser. 2015. “Researcher Perspectives on Publication and Peer Review of Data.” *PLOS One* 10(2): e0117619. http://doi.org/10.1371/journal.pone.0117619

Latif, Atif, Timo Borst, and Klaus Tochtermann. 2014. “Exposing Data from an Open Access Repository for Economics as Linked Data.” *D-Lib Magazine* 20(9): 2. http://doi.org/10.1045/september2014-latif

Lutter, Randall, Craig Barrow, Christopher J. Borgert, James W. Conrad, Debra Edwards, and Allan Felsot. 2012. “Data Disclosure for Chemical Evaluations.” *Environmental Health Perspectives* 121(2): 145-148. http://doi.org/10.1289/ehp.1204942

Lynch, Clifford A. 2003. “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age.” *portal: Libraries and the Academy* 3(2): 327-336. http://doi.org/10.1353/pla.2003.0039

Making Data Count: The Data Level Metrics Project. 2016. “Data Metrics Survey Results Published.” *Making Data Count Blog*. Accessed June 23. http://mdc.lagotto.io/blog

Mannheimer, Sara, Susan Borda, and Leila B. Sterman. 2016. “Cited/Downloaded Dataset and Repository Characteristics [dataset].” *Montana State University ScholarWorks*. http://doi.org/10.15788/m2059z

Mayernik, Matthew S. 2013.“Bridging Data Lifecycles: Tracking Data Use via Data Citations Workshop Report.” *National Center for Atmospheric Research Technical Note NCAR/TN-494+ PROC*. http://doi.org/10.5065/D6PZ56TX

Mooney, Hailey, and Mark P. Newton. 2012. “The Anatomy of a Data Citation: Discovery, Reuse, and Credit.” *Journal of Librarianship and Scholarly Communication* 1(1), p.eP1035. http://doi.org/10.7710/2162-3309.1035

National Institutes of Health. 2003. “Data Sharing Policy and Implementation Guidance”. Last modified March 5. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Research Council. *For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*, Edited by Paul E. Uhlir. Washington, DC: National Academies Press, 2012. http://doi.org/10.17226/13564

National Science Foundation. 2011. “NSF Data Management Plan Requirements.” http://www.nsf.gov/eng/general/dmp.jsp

Nelson, Bryn. 2009. “Data Sharing: Empty Archives.” *Nature* 461: 160-163. http://doi.org/10.1038/461160a

Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. 2010. “Data Citation and Peer Review.” *Eos Transactions American Geophysical Union* 91(34): 297-298. http://doi.org/10.1029/2010EO340001

Piwowar, Heather A., Jonathan D. Carlson, and Todd J. Vision. 2012.“Beginning to Track 1000 Datasets from Public Repositories into the Published Literature.” Proceedings of the American Society for Information Science and Technology 48(1): 1-4. http://doi.org/10.1002/meet.2011.14504801337

PLOS. 2014. “Data Availability.” http://journals.plos.org/plosone/s/data-availability

Registry of Research Data Repositories. 2016. “re3data.org Reaches a Milestone & Begins Offering Badges.” Accessed June 23. http://www.re3data.org/2016/04/re3data-org-reaches-a-milestone-begins-offering-badges/

Robinson-García, Nicolas, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 2015. “Analyzing Data Citation Practices Using the Data Citation Index.” *Journal of the Association for Information Science and Technology*. http://doi.org/10.1002/asi.23529

Rosati, Antonia, and Matthew Mayernik. 2013. “Facilitating Data Discovery by Connecting Related Resources.” *Semantic Web Journal*. (Open peer-review decision: Reject and resubmit.) http://www.semantic-web-journal.net/content/facilitating-data-discovery-connecting-related-resources

Santer, B. D., T. M. L. Wigley, and K. E. Taylor. 2011. “The Reproducibility of Observational Estimates of Surface and Atmospheric Temperature Change.” *Science* 334(6060): 1232-1233. http://doi.org/10.1126/science.1216273

Sieber, Joan E., and Bruce E. Trumbo. 1995. “(Not) Giving Credit Where Credit Is Due: Citation of Data Sets.” *Science and Engineering Ethics* 1(1): 11-20. http://doi.org/10.1007/BF02628694

Simons, Natasha. 2012. “Implementing DOIs for Research Data.” *D-Lib Magazine* 18(5/6). http://doi.org/10.1045/may2012-simons

Starr, Joan, and Angela Gastl. 2011. “isCitedBy: A Metadata Scheme for DataCite.” *D-Lib Magazine* 17(1/2). http://doi.org/10.1045/january2011-starr

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. “Data Sharing by Scientists: Practices and Perceptions.” *PLOS One* 6(6): e21101. http://dx.doi.org/10.1371/journal.pone.0021101

Thomson Reuters. 2016a. “About the Data Citation Index.” Accessed June 21. http://wokinfo.com//products_tools/multidisciplinary/dci/

Thomson Reuters. 2016b. “Data Citation Index: The Repository Selection Process.” Accessed June 21. http://wokinfo.com//products_tools/multidisciplinary/dci/selection_essay/

Vision, Todd J. 2010. “Open Data and the Social Contract of Scientific Publishing.” *BioScience* 60(5): 330-331. http://doi.org/10.1525/bio.2010.60.5.2

Wallis, Jillian C., Matthew S. Mayernik, Christine L. Borgman, and Alberto Pepe. 2010. “Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality.” *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, Gold Coast, QLD, Australia June 21-25: 333-340. http://doi.org/10.1145/1816123.1816173

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 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. http://doi.org/10.1371/journal.pone.0067332

White, Ethan P., Elita Baldridge, Zachary T. Brym, Kenneth J. Locey, Daniel J. McGlinn, and Sarah R. Supp. 2013. “Nine Simple Ways to Make It Easier to (Re)Use Your Data.” *Ideas in Ecology and Evolution* 6(2): 1-10. http://doi.org/10.4033/iee.2013.6b.6.f

Whitlock, Michael C. 2011. “Data Archiving in Ecology and Evolution: Best Practices.” *Trends in Ecology & Evolution* 26(2): 61-65. http://doi.org/10.1016/j.tree.2010.11.006

Witt, Michael, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin. 2009. “Constructing Data Curation Profiles.” *International Journal of Digital Curation* 4(3): 93-103. http://doi.org/10.2218/ijdc.v4i3.117

Zimmerman, Ann. 2007. “Not By Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse.” *International Journal on Digital Libraries* 7(1): 5-16. http://doi.org/10.1007/s00799-007-0015-8