Beginning to track 1000 datasets from public repositories into the published literature

Heather A. Piwowar
DataONE and Dryad Digital Repository
National Evolutionary Synthesis Center
hpiwowar@nescent.org

Jonathan D. Carlson
School of Library and Information Studies
University of Wisconsin—Madison
jcarlson4@wisc.edu

Todd J. Vision
Department of Biology
University of North Carolina-Chapel Hill
tjv@bio.unc.edu

ABSTRACT
Data sharing provides many potential benefits, although the amount of actual data reused is unknown. Here we track the reuse of data from three data repositories (NCBI's Gene Expression Omnibus, PANGAEA, and TreeBASE) by searching for dataset accession number or unique identifier in Google Scholar and using ISI Web of Science to find articles that cited the data collection article. We found that data reuse and data attribution patterns vary across repositories. Data reuse appears to correlate with the number of citations to the data collection article. This preliminary investigation has demonstrated the feasibility of this method for tracking data reuse.

Keywords
data reuse, data sharing, data archiving, bibliometrics, scholarly communication, human information behavior.

MOTIVATION AND BACKGROUND
The potential benefits of data sharing are impressive: less money spent on duplicate data collection, reduced fraud, diverse contributions, better tuned methods, training, and tools, and more efficient and effective research progress. Many datasets have now been publicly archived. Have the potential benefits been realized? Are the data sets reused? Have they saved money? Enabled new science? Enabled diverse contributions? Is data sharing worth the effort? We don't know. There are certainly some superstar success stories that need no analysis: Data in Genbank and the Protein Data Bank are heavily reused and have resulted in fundamental scientific advances not otherwise possible. These repositories are so successful, though, that they are discounted as special cases. What do reuse patterns look like for datasets in other repositories?

Zimmerman (2003) has done seminal work in data reuse, investigating how ecologists locate and strive to understand data for secondary analysis. Sandusky (2007) has studied the use of figures and other data components within full text articles within research and teaching. Hine (2006) looked at citation mentions of repositories and assessed the degree to which data repositories become a routine part of a researcher's methods. Several surveys estimate the opportunities lost due to data withholding (Campbell, 2000; Vogeli et al., 2006; Piwowar 2011).

Our current study supplements this prior work by tracking individual datasets from repositories into the published literature and analyzing the environments of reuse. Tracking data reuse is difficult due to inconsistency in attribution practice (Sieber & Trumbo, 1995) and ambiguity between attributions describing data submission and data reuse (Piwowar, 2010). Efforts are underway to improve the citation of datasets through unique identifiers and standard citation practices (Altman & King, 2007; 2009; Cook, 2008; Pollard & Wilkinson, 2010; Vision, 2010), but these improvements are not yet in common practice. As a result, examining current behavior requires intensive searches and manual curation. Although this will leave us far short of a full understanding of the value of data reuse patterns, it provides valuable evidence, attention, and methods for further investigation.

METHOD
Here we report results for tracking datasets from the first three (out of planned 10) repositories: NCBI's Gene Expression Omnibus, PANGAEA, and TreeBASE.

Identifying datasets for tracking
From each repository we randomly chose 100 datasets, selecting from all datasets submitted in the year 2005 that were associated with a published data collection study. We

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

chose to study datasets deposited in 2005 because many repositories were firmly established at that time and we felt that five years would be sufficient for a range of data reuse studies to be conducted, published, and indexed.

Identifying reuse candidates

Two approaches were used to identify possible reuse in the published literature (including preprints and whitepapers) over the period 2005-2010.

First, authors sometimes attribute dataset reuse by mentioning the identifier of the reused datasets in the full text of their studies. We used Google Scholar to find studies that attribute reuse this way. For each dataset in our sample, we queried Google Scholar using the dataset accession number, DOI, or other unique identifier with an "AND" and the repository name. The relevant hits were recorded and imported to a Mendeley group.

Second, authors often attribute data reuse by citing the paper that describes the original collection of the dataset (the "data collection article"). We used ISI Web of Science to identify studies that used this method of data reuse attribution. For each dataset, we located the data collection article within ISI Web of Science and exported the list of all articles that cite this data collection article. This list of all citations was processed to subselect 150 random citations, stratified by the total number of times the data collection article had been cited. The subselection of the ISI WoS results was saved as a BibTeX file then uploaded to the Mendeley group.

Confirming reuse instances

Manual review was performed for each instance of potential data reuse. We located the article full text, read the relevant sections of the papers, and manually determined if the data from the associated dataset had been reused within the study. Tags were applied to the Mendeley citation to indicate data reuse, no data reuse, or data reuse ambiguous as well as a confidence level of high, medium, or low. We also applied a tag indicating location of the attribution, and the search strategy used to find the instance of reuse.

Annotation and analysis

Notes were kept on the number and type of false hits for each search. Date, journal, authors, affiliations, abstract, and keywords were collected for all reuse publications.

When an instance of data reuse was found by more than one method we counted it only as an "attribution in text" for the purposes of the analysis. Low confidence reuses are not included in this analysis.

We extrapolated findings from our subsample of citations to data collection papers by weighting all instances of reuse identified through citations by the ratio of (total number of citations to data collection papers / number of citations to data collection papers included in manual annotation subsample).

RESULTS

We estimate that 100 random datasets deposited into GEO in 2005 were reused approximately 550 times in the following five years, in aggregate. Similarly, we estimate that 100 random datasets deposited to each of Pangaea and TreeBase in 2005 have been used 588 and 32 times, respectively (see Table 1).

	GEO	Pangaea	TreeBase
Dataset ID in text (observed)	114	7	4
Data collection article cited (extrapolated)	436	581	28
Total (estimate)	**550**	**588**	**32**

Table 1: Number of times 100 randomly-chosen datasets from each of three repositories have been reused in the published literature
(datasets submitted in 2005, lit search covered 2005-2010)

Of the 100 datasets per repository that we tracked, we directly observed reuse of least 35 datasets from GEO, 15 from Pangaea, and 4 from TreeBase. These numbers represent lower bounds on the true number of datasets reused because they have not been extrapolated beyond our citation subsample.

As seen in Table 1 and Figure 1, attribution patterns vary across repositories. Reuse of data from Pangaea was almost always accompanied by a citation in the reference list to the data collection article, whereas at least 10% of attributions for reuse of GEO and TreeBase data are made through mentions of the dataset identifiers in the paper full text. GEO data was sometimes attributed in footnotes and tables.

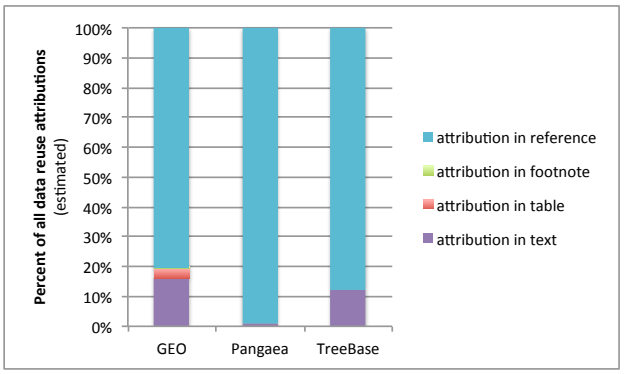

Figure 1: Location of data attribution in published studies that reused datasets

For two of the three repositories, we observed a relationship between the number of citations a data collection paper received and the number of times we observed reuse of its associated dataset. Articles that reused data appear within a

year of data submission and continued to accumulate through 2010 (Figure 2).

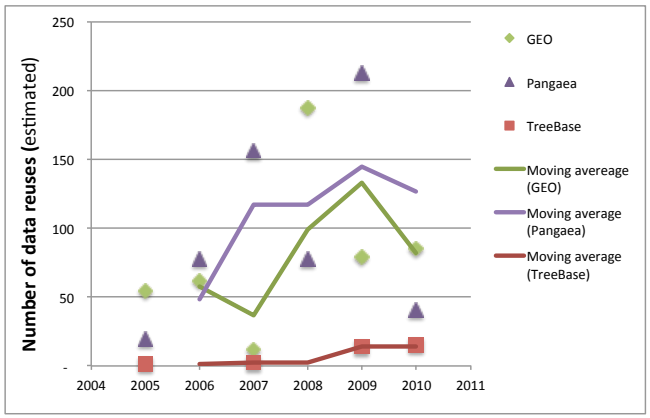

Figure 2: Aggregate data reuse by publication date of the article that reused the data

LIMITATIONS

Our approach for studying data reuse has limitations. Its focus on reuse in the published research literature overlooks other valuable reuse in education, policy, unpublished validation, and private study. Furthermore, there are benefits to data sharing and archiving even if the data are never reused: for example, sharing detailed datasets likely discourages fraud.

The methods were particularly conducive to locating reuse in literature openly available on the web, available in full-text databases, and published by authors or in journals that choose robust data citation practices. This may introduce bias relative to all reuses.

Our data pool is incomplete and may be missing several examples of data reuse. For example, our reliance on ISI Web of Science for citations to data collection articles failed to identify reuses in preprints, theses, dissertations and journals outside its index.

Results as presented here do not reflect the uncertainty of our extrapolation estimates.

FUTURE WORK

This work will extend to track one thousand datasets in total: 100 datasets from each of 10 repositories. Further analysis will look at patterns across time, journal, authors, and topic.

The results will also be used to identify repositories and search methods that are conducive to a larger, ideally automated, collection of reuse instances across time. Large collections of reuse instances could support future efforts to confirm the rarity of analysis duplication (Bachrach & King, 2004), misinterpretation (Liotta et al., 2005), and scooping.

DATA AND CODE AVAILABILITY

This project represents an experiment in open science. Interested readers are invited to reuse data, view code, share ideas, and follow this project's future iterations at

https://notebooks.dataone.org/tracking1000datasets/

ACKNOWLEDGMENTS

This research was conducted under the auspices of DataONE, funded by a Cooperative Agreement through the NSF DataNET program (OCI-0830944). Additional support for data collection has been provided through several sources, including an ASIS&T SIG USE Elfreda A Chatman Research Proposal Award, a Discovery grant to Michael Whitlock from the Natural Sciences and Engineering Research Council of Canada, and the DataONE Summer 2011 internship program (funded by DataONE and INTEROP: Creation of an International Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences, NSF grant #0753138).

REFERENCES

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13.

Bachrach, C., & King, R. (2004). Data sharing and duplication: Is there a problem? Arch Pediatr Adolesc Med, 158, 931-932.

Campbell, E. (2000). Data withholding in academic medicine: Characteristics of faculty denied access to research results and biomaterials. Research Policy, 29, 303-312.

Cook, R. (2008). Editorial: Citations to published data sets. FLUXNET newsletter, 4, 1-2.

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. Social Studies of Science, 36, 269-298.

Liotta, L., et al. (2005). Importance of communication between producers and consumers of publicly available experimental data. J Natl Cancer Inst, 97, 310-314.

Piwowar, H.A. (2010). Foundational studies for measuring the impact, prevalence, and patterns of publicly sharing biomedical research data. Universtiy of Pittsburgh PhD Dissertation.

Piwowar, H.A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE 6(7): e18657.

Pollard, T.J., & Wilkinson, J.M. (2010). Making datasets visible and accessible: Datacite's first summer meeting. Ariadne, July.

Sandusky, R.J., Tenopir, C., & Casado, M.M. (2007). Uses of figures and tables from scholarly journal articles in teaching and research. Proceedings of the American Society for Information Science and Technology, 44, 1-13.

Sieber, J.E., & Trumbo, B.E. (1995). (not) giving credit where credit is due: Citation of data sets. Science and Engineering Ethics, 1, 11-20.

Vision, T.J. (2010). Open data and the social contract of scientific publishing. BioScience, 60, 330-331.

Vogeli, C., et al. (2006). Data withholding and the next generation of scientists: Results of a national survey. Acad Med, 81, 128-136.

Zimmerman, A. (2003). Data sharing and secondary use of scientific data: Experiences of ecologists. University of Michigan PhD Dissertation.