A method to track dataset reuse in biomedicine: filtered GEO accession numbers in PubMed Central

Heather A Piwowar
National Evolutionary Synthesis Center
2024 W. Main Street, Suite A200
Durham, NC 27705
hpiwowar@nescent.org

ABSTRACT
Reusing research data has important potential benefits: generative science and efficient resource use. Tracking the reuse of research datasets would allow us to understand whether the potential benefits are indeed realized, enable recognition of investigators who produce, annotate, and share useful data, and inform data sharing and reuse initiatives, tools, and policies.

Unfortunately, the lack of clear attribution practices for data make automated tracking of data reuse difficult.

I present a method for tracking research data reuse that takes advantage of the community norms around gene expression microarray data sharing and the rich NCBI Entrez resources. Specifically, I search the full-text of papers stored in PubMed Central for accession numbers of datasets archived in NCBI's Gene Expression Omnibus (GEO) repository. Studies known to have created microarray data are excluded through automated filters and guided manual curation. Reuse studies authored by investigators who were also authors or submitters of the original dataset can be flagged. MeSH terms attached to the data creation and data reuse studies provide additional information for analysis. Finally, I extrapolate the findings to all of PubMed.

Automated portions of this method have been implemented in python and are openly available. Although imperfect, this dataset is a valuable initial resource for research into patterns of data reuse.

Keywords
data sharing, data reuse, method, bioinformatics, bibliometrics.

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

INTRODUCTION
Tracking data reuse would facilitate recognition of authors who produce, annotate, and share useful datasets. Identifying areas with frequent reuse could highlight best practices for research agendas, tools, standards, and repositories. Such analysis could also identify areas that have yet to receive major benefits from shared data initiatives.

Despite this need, little research has been done on patterns and prevalence of data reuse, outside case study descriptions (Zimmerman, 2003).

Unfortunately, there are no well-established attribution standards for datasets. Furthermore, datasets do not have unambiguous identifiers, data attribution is often within full text and thus difficult to query across journals and disciplines, and it is challenging to disambiguate the mention of a dataset in the context of reuse from the mention of a dataset deposit.

Studying the reuse of one datatype, gene expression microarray data, allows us to mitigate several of these issues through convenient community norms and rich NCBI resources. Most shared gene expression microarray data is deposited into a single central repository: the NCBI's Gene Expression Omnibus (GEO). It is common practice to refer to datasets by their GEO accession numbers, and GEO accession numbers have a unique format which is easily queriable. In addition, most creations and reuses of gene expression microarray data in the published literature are indexed by PubMed and increasingly (as per NIH mandate) available for full-text query within PubMed Central (PMC). NCBI's eUtils web service facilitates automated queries, filtering based on links between datasets and articles, and extraction of standard indexing metadata.

I describe a preliminary implementation of this protocol for data collection, an early validation, and limitations of this approach.

METHOD
The method for collecting a list of dataset reuses involves several querying and filtering steps:

1. Query GEO for a list of accession numbers for all datasets deposited within specified date range

2. Query the full-text of publications within PubMed Central for each of these accession numbers

   I search for several variants of each accession: space or no space after the prefix (GDS or GSE), and the full number or the number minus the "*20000*" offset.

3. Exclude studies that mention GEO accession numbers in the context of data deposition

   The Entrez filter "NOT pmc_gds" excludes many but not all papers with deposits in GEO. To catch data deposition studies overlooked by the filter, I query all studies for evidence of data-creation and data-sharing language in their full text (Piwowar & Chapman, 2008; Piwowar 2010) and use these heuristics to guide manual review.

4. Identify author overlap between reuse and data creation

   I look for author last name and institutional overlap, manually inspecting cases that are unclear.

5. Extrapolate to all of PubMed

   I extrapolate estimates of reuse in PubMed Central to all of PubMed, using the current yearly proportion of articles with the MeSH term "gene expression profiling" in PMC relative to all of PubMed, given in Table 1.

Table 1: Percent of PubMed microarray articles in PMC

year	01	02	03	04	05	06	07	08	09
% in PMC	18	15	16	17	18	20	23	31	26

IMPLEMENTATION

This method has been implemented in Python. The data, source code, and notes for this project are openly available in the spirit of Open Notebook Science at http://openwetware.org/wiki/DataONE.

A preliminary validation compared the results of this method to the GEO third-party usage page at http://www.ncbi.nlm.nih.gov/projects/geo/info/ucitations.html. The proposed method found 256 of the 618 reuse articles listed by GEO staff (41%). The method also found 802 articles not on the GEO list. A quick inspection suggests that most of these are novel uses rather than errors, but a formal evaluation is needed.

It is worth noting that the GEO third-party usage list has not been updated for the past seven months, no doubt due to manpower constraints. An accurate, automated system would surely help maintain this and similar resources.

DISCUSSION

I note several limitations of this method. First, the approach captures a only subset of all dataset reuses. Data reuse is sometimes attributed without using accession numbers, through citations to the data producing paper or simply reporting the criteria of a repository query. This method also overlooks studies that both create and reuse data: for efficiency I eliminate studies all that create data, recognizing that I may be unfortunately excluding some studies that both create and reuse data. Importantly, the method doesn't capture reuse outside the peer-reviewed literature, such as datasets used in education and training.

Extrapolations based on this data may be biased. Papers in PubMed Central may not be representative of all biomedical literature, depending on the relative degree of open access adoption and NIH funding of various communities.

Finally, of course the data does not identify reuses after the timeframe of the data collection. GEO has existed for less than ten years: its data may be useful and used for a long time yet.

The method could be enhanced in several ways. I collect reuses through both GDS and GSE accession numbers, but these are actually different levels of granularity for the same data components. Modeling their relationships will facilitate a more robust interpretation. Also, determining author identify through last name is insufficient: Author-ity clusters would allow more accurate determination of author identity (Torvik & Smalheiser, 2009).

Hopefully improved standards will make tracking data reuse more straightforward in the future. In the mean time, I believe this dataset, although imperfect, will permit preliminary investigations into data reuse behaviour.

ACKNOWLEDGMENTS

This research was conducted under the auspices of DataONE, funded by a Cooperative Agreement through the NSF DataNET program (OCI-0830944).

REFERENCES

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

Piwowar HA and Chapman WW. (2008) Linking database submissions to primary citations with PubMed Central. *BioLINK Workshop at ISMB.*

Torvik VI and Smalheiser NR. (2009) Author Name Disambiguation in MEDLINE. *ACM Trans Knowl Discov Data.* Jul 1; 3(3).

Zimmerman A. (2003) Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. (2003) *PhD Dissertation, University of Michigan.*