**Tracking data reuse: Following one thousand datasets from public repositories into the published literature**

Heather Piwowar

submitted for SIGUSE research proposal award, Sept 2010

Research questions 1

Motivation and Background 1

Significance 2

Proposed method 3

*Select ten repositories 3*

*Select one thousand datasets 3*

*Identify reuse of datasets in the published literature 4*

*Analysis of reuse environment 4*

Anticipated contributions 6

Limitations 6

Future work 6

Timeline and Budget 7

References 9

**Research questions**

* **How often is data from repositories used in the published literature?    
  What is the distribution of use across datasets and time?**
* **Who reuses data?    
  Are investigators who reuse repository datasets similar to investigators who deposit data?**
* **What is data reused for?    
  How similar are studies that reuse data to studies that deposit data?**

# **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 data repositories becomes a routine part of a researcher’s methods.  Several surveys estimate the opportunities lost due to data withholding (Campbell, 2000; Vogeli et al., 2006).

The current proposal suggests supplementing 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 behaviour will require intensive searches and manual curation.  Although this will leave us far short of a full understanding of the value of data reuse patterns, it will provide valuable evidence, attention, and methods for further investigation.

# **Significance**

There have been few, if any, prior published studies on data reuse patterns from a repository viewpoint.  Such findings will be useful to a variety of stakeholders:

* **Repository planners and developers.** Repositories struggle to achieve ongoing funding (Ball et al., 2004; Beagrie et al., 2010; Goldstein & Ratliff, 2010). Levels of reuse are key in building a business case for funders and investors (Fry et al., 2009).
* **Funders and Journals**.  Funding agencies and journals are increasingly requiring or requesting data sharing (e.g. (National Science Foundation, 2010; Whitlock et al., 2010)).  Because data sharing mandates can be controversial (Campbell, 1999; King, 1995), evidence of value strengthens the ethical case for compelling investigators to bear the personal cost of sharing data. Indeed, several studies have found that researchers are motivated to share their data when they know it will actually be useful to others (Borgman et al., 2007; Hedstrom, 2006).
* **Standards community.**  Knowing how data is reused is important for the development of real-world standards that address current issues.
* **Data reusers.** Understanding patterns of use (and lack of use!) will inform tools, best practices, and reward structures for investigators who reuse data.
* **Data producers.**  Last but not least, analysis into reuse will likely benefit investigators who collect data.  Investigators are increasingly asked to share their data with little evidence to inform appropriate levels of effort; informed decisions by policy makers will decrease wasted effort and point to practical incentives.

In isolation, this study is unlikely to definitively resolve any issues for the stakeholders above.  I hope, however, it will serve as a first step toward useful evidence.  By bringing attention to this area, uncovering hypothesis-generating results, and collecting data that will help identify scalable methods for further investigation, this study will encourage continued investigation into the realized value of data reuse.

# **Proposed method**

I propose to follow one thousand datasets from data repositories into the published literature.  Studying the reuse patterns of 100 datasets from 10 repositories will facilitate analysis across domains, datatypes, and repository structures.  Analysis will focus on the relative levels and timing of data reuse, attributes of investigators who reuse data compared to those who deposit data, and topics studied through data reuse.

## **Select ten repositories**

Ten repositories will be selected for study.  Candidate repositories will be drawn from a recent review (Marcial & Hemminger, 2009), repositories identified in *Nature’s* recommendations (http://www.nature.com/authors/editorial\_policies/availability.html), and other repositories recommended to the investigator.  Candidates will be filtered to include only those that are large, navigable in English, publicly available, and established prior to 2004. The list will be further narrowed to repositories that support querying datasets by deposit date, and those that allow data submitters to annotate their submissions with citations to publications describing data collection.   
  
Based on this short list (estimated to be about 50 candidate repositories), ten repositories will be chosen with a preference for those that offer accession numbers or unique IDs that have an easily queriable prefix (dois, hdls, or an alpha-numeric string), have clear links to the associated literature describing data collection, allow bulk download of their submission metadata or automated data collection from their websites, and those that suggest a citation policy to their users. Attempts will be made to achieve diversity across domain (social science, medicine, biology, astronomy, geosciences, chemistry, physics) and diversity across repository type (institutional, domain, datatype, or journal-based).  
  
Based on preliminary investigations, the ten repositories will likely include these five candidates plus five more, to be determined:

* IQSS Dataverse network
* Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)
* PANGAEA® - Publishing Network for Geoscientific and Environmental Data
* NCBI’s Gene Expression Omnibus
* EBI’s ArrayExpress

## **Select one thousand datasets**

One hundred datasets will be randomly chosen from each of the ten repositories.

For each repository, all datasets submitted in 2005 will be itemized:  datasets will be continually chosen, randomly, until 100 are found with a link to an accompanying data collection publication.  
  
For each dataset, the name and affiliation of the submitter will be gathered, as well as date of publication, journal, authors, affiliations, abstract, and keywords from the accompanying data creation publication.

## **Identify reuse of datasets in the published literature**

Evidence of reuse in the published literature will be collected for each of the one thousand datasets.  For this purpose, the published literature will include preprints and whitepapers.  Collection will include at least the following steps:

1. Manually inspect citations to the data collection paper using Scopus or ISI Web of Science.  For data collection papers with hundreds of citations, a random sample will be selected for manual review.
2. Query for the dataset accession number or unique identifier in the bibliography of published papers using Scopus or ISI Web of Science.
3. Query for the dataset accession number or unique identifier in the full text of published literature.  Google Scholar plus one or more additional portals will be used, chosen based on the repository discipline (for example, PubMed Central and Highwire Press will be searched for biomedical repository IDs).
4. Query for both the repository name and the author’s name in the full text of published literature, using the same portals as above.

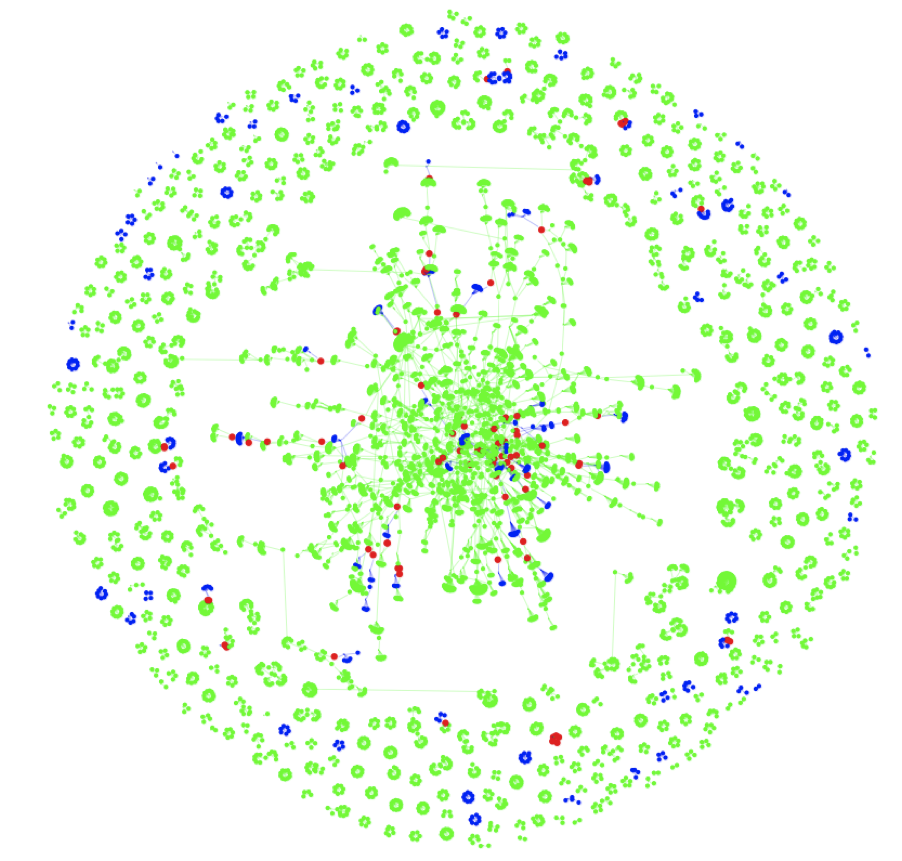
The search strategies that finds each reuse instance will be logged.  In addition, notes will be kept on the number and type of false hits for each search. Date, journal, authors, affiliations, abstract, and keywords will be extracted for all reuse publications.

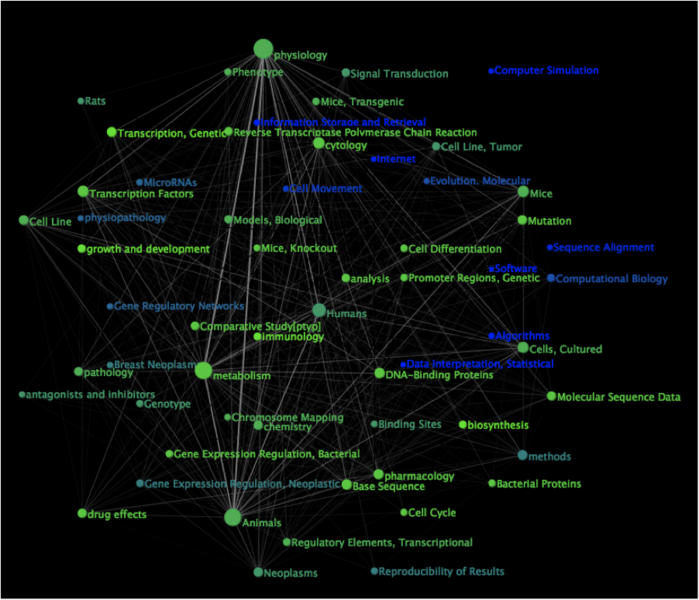
## **Analysis of reuse environment**

The number of reuse publications found for each dataset will be aggregated for each source repository, by dataset and publication year.  This information will be used to plot histograms of reuses per dataset.  Graphs will also be made showing the dates at which reuse publications appear; confidence intervals around median lag times will be compared to understand how the lag differs by repository.  
  
For each repository, the authors of papers that submitted datasets will be compared to authors that reused data.  For datasets where accessing the author list of all data submission papers is simple (NCBI’s Entrez databases, for example), the entire list of data submission authors may be used rather than just those of the sample datasets.  Chi-squared and Fisher’s exact tests will be used to test whether the distribution of countries, institutions, and departments (categorized into disciplines) is different for data creation papers than data reuse papers.

Analysis will be done on the topics of the reuse papers.  Manual review, guided by keyword heuristics, will be done to identify reuse papers that primarily describe the development of a new tool or algorithm.  All papers will be clustered using k-means clustering and/or nearest neighbor algorithms, based on keywords and abstracts features, to see how often the data reuse papers cluster with their data creation papers.

Finally, the data will be used to experiment with visualizations of data reuse patterns. Figure 1 and 2 illustrate possible visualizations.



**Figure 1: Sample co-authorship network, colored to highlight the relationships between authors who have created data (green), reused data (blue), and both created and reused data (red).**

**Figure 2: Sample topic co-occurrence network, colored to highlight the relative frequency of indexing term application to studies that collect original data rather than reuse data : the gradient extends from green (for indexing terms applied only to publications based on primary collection data) to   
blue (for terms applied only to publications that reuse data).**

# **Anticipated contributions**

Anticipated contributions, in the form of research artifacts, include:

* A poster summarizing findings at the ASIS&T 2011 annual meeting
* A journal publication submission after ASIS&T 2011, incorporating feedback from discussions at ASIS&T
* Collected data, statistical scripts, and data collection software made publicly available
* An open notebook. The process of this research will be done as ongoing “open notebook science” (http://en.wikipedia.org/wiki/Open\_Notebook\_Science) with notes, development, and results available openly online as the research is done.

# **Limitations**

The proposed approach, like all approaches, 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 proposed approach is 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.

Finally, it is not clear how generalizable these results will be: the studied repositories will all be large, publication-related, and established at least 6 years ago. They also cover a wide range of subdisciplines with little redundancy, so it will be difficult to distinguish patterns related to repositories from those of subdisciplines.

# **Future work**

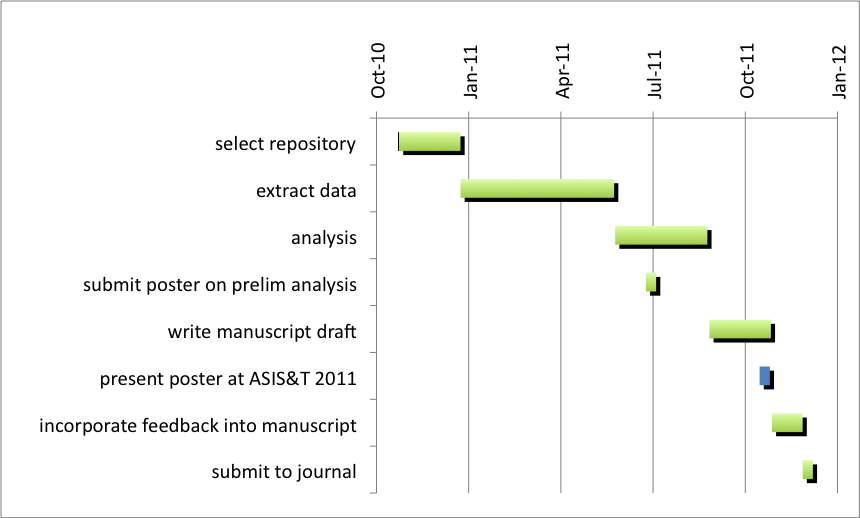
Several projects are likely to follow from this work. First, the results will be used to identify repositories and search methods that are conducive to a larger, ideally automated, collection of reuse instances across time. Large collections of reuse instances could support future efforts to confirm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

The attribution patterns themselves could be catalogued to inform education about best practices, continuing the work of Sieber and Trumbo (1995) and estimate the citation benefit of data sharing, continuing the work of Piwowar, Day, and Fridsma (2007).

It would be informative to compare the datasets that are reused to those that aren’t reused to assess the role of metadata and study topic in reusability, looking at this issue from a different angle than Zimmerman (2007). Finally, it would be informative to correlate download statistics from repositories with reuses in the published literature to understand the strength of this relationship, since downloads statistics are relatively easy to collect.

# **Timeline and Budget**

The proposed work will be completed within a year. Details:



I do not anticipate expenditures related to printing, computer time, fees to subjects, keypunching, statistical consulting, photography, art work, or typing.

This project will require help with data collection. Though my institution I can hire qualified student workers at approximately $4/hour (the remainder of their wage is covered by a student-work program at the institution), for a maximum of 10 hours/week per student. I plan to hire and train two students to do the data collection over the Spring 2011 semester. This is 16 weeks, for a total of 320 hours.

The workload and cost of data collection help is estimated below. The baseline scenario, estimated at $953 and 322 hours, assumes that half of the data collection papers have no citation trail in Scopus/ISI Web of Science and that 500 instances of reuse are found. It is not known how many instances of reuse will be found. If there are 100 instances of reuse, the budget becomes $687; if 1000, the budget becomes $1287.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | minutes per datapoint | datapoints | estimated hours | rate | total |
| full-text queries | 5 | 1000 | 83 | 4 | 333 |
| Scopus/ISI Web of Science queries | 5 | 500 | 42 | 4 | 167 |
| reuse sentence extraction (download full-text PDF into Mendeley, extract reuse attribution sentences and citations) | 10 | 500 | 83 | 4 | 333 |
| training |  |  | 10 | 4 | 40 |
| contingency |  |  | 20 | 4 | 80 |
| **Grand Total** |  |  | **322** |  | **$953** |

If it takes less time than expected for data extraction, the remaining hours will be devoted to cataloging the reuse attribution patterns. If data extraction takes longer than expected, this project for will be nominated for inclusion in a summer student internship program. Another contingency plan is to scale down the number of repositories and scale up the number of datasets selected from quick-to-query repositories, to maintain 1000 total datapoints.

# **References**

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

Anonymous (2009). Data producers deserve citation credit. *Nature genetics, 41*, 1045.

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

Ball, C., Sherlock, G., & Brazma, A. (2004). Funding high-throughput data sharing. *Nat Biotechnol, 22*, 1179-1183.

Beagrie, N., Eakin-Richards, L., & Vision, T. (2010). *Business models and cost estimation: Dryad repository case study*. Paper presented at the iPRES2010.

Borgman, C.L., Wallis, J.C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries, 7*, 17-30.

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

Campbell, P. (1999, April 16). Controversial proposal on public access to research data draws 10,000 comments*. The Chronicle of Higher Education,* p. A42,

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

Fry, J., et al. (2009). Identifying benefits arising from the curation and open sharing of research data produced by uk higher education and research institutes. *JISC report*, 1-89.

Goldstein, S.J., & Ratliff, M. (2010). Dataspace : A funding and operational model for long-term preservation and sharing of research data. *EDUCAUSE Live!*

Hedstrom, M. (2006). Producing archive-ready datasets: Compliance, incentives, and motivation. *IASSIST Conference 2006: Presentations*.

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

King, G. (1995). A revised proposal, proposal. *PS: Political Science and Politics, XXVIII*, 443-499.

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

Marcial, L., & Hemminger, B. (2009). Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science*.

National Science Foundation (2010). Press release 10-077: Scientists seeking nsf funding will soon be required to submit data management plans.

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., Day, R.B., & Fridsma, D.S. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE, 2*.

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.

Whitlock, M.C., et al. (2010). Data archiving. *Am Nat, 175*, 145-146.

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

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries, 7*, 5-16.