

Proposed Foundations for Evaluating Data Sharing and Reuse in the Biomedical Literature

Heather A. Piwovar

Department of Biomedical Informatics, University of Pittsburgh
200 Meyran Avenue
Pittsburgh, PA 15260
1.412.647.7113

hpiwovar@gmail.com

ABSTRACT

Science progresses by building upon previous research. Progress can be most rapid, efficient, and focused when raw datasets from previous studies are available for reuse. To facilitate this practice, funders and journals have begun to request and require that investigators share their primary datasets with other researchers. Unfortunately, it is difficult to evaluate the effectiveness of these policies. This study aims to develop foundations for evaluating data sharing and reuse decisions in the biomedical literature by developing tools to answer the following research questions, within the context of biomedical gene expression datasets: What is the prevalence of biomedical research data sharing? Biomedical research data reuse? What features are most associated with an investigator's decision to share or reuse a biomedical research dataset? Does sharing or reusing data contribute to the impact of a research article, independently of other factors? What do the results suggest for developing efficient, effective policies, tools, and initiatives for promoting data sharing and reuse? I suggest a novel approach to identifying publications that share and reuse datasets, through the application of natural language processing techniques to the full text of primary research articles. Using these classifications and extracted covariates, univariate and multivariate analysis will assess which features are most important to data sharing and reuse prevalence, and also estimate the contribution that sharing data and reusing data make to a publication's research impact. I hope the results will inform the development of effective policies and tools to facilitate this important aspect of scientific research and information exchange.

Categories and Subject Descriptors

H.1.1 [Systems and Information Theory]: Value of information;
H.3.5 [Online Information Services]: Data sharing; J.3 [Life and Medical Sciences]: Biology and genetics, Health

General Terms

Measurement, Human Factors

Keywords

data sharing, data reuse, evaluation, policy, bioinformatics, bibliometrics

1. INTRODUCTION

Sharing information facilitates science. Reusing previously-collected data in new studies allows these valuable resources to contribute far beyond their original analysis.[1] In addition to being used to confirm original results, raw data can be used to explore related or new hypotheses, particularly when combined with other publicly available data sets. Real data is indispensable when investigating and developing study methods, analysis techniques, and software implementations. The larger scientific community also benefits: sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and patient population resources by avoiding duplicate data collection.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals require the submission of detailed biomedical data to publicly available databases as a condition of publication.[2,3] Since 2003, the NIH has required a data sharing plan for all large funding grants and has more recently introduced stronger requirements for genome-wide association studies[4,5]; other funders have similar policies. Several government whitepapers[1,6] and high-profile editorials[7-12] call for responsible data sharing and reuse, large-scale collaborative science is providing the opportunity to share datasets within and outside of the original research projects[13,14], and tools, standards, and databases are developed and maintained to facilitate data sharing and reuse.

Despite these investments of time and money, we do not yet understand the prevalence and patterns of data sharing and reuse, the effectiveness of initiatives, or the costs, benefits, and impact of repurposing biomedical research data.

The goal of this study is to build foundational tools, datasets, and analyses for identifying and evaluating data sharing and reuse decisions within the biomedical literature.

2. FINDINGS AND CHALLENGES IN RESEARCH ON SHARING AND REUSE

This section highlights a few major findings and challenges in research on biomedical data sharing and reuse.

2.1 Understanding Attitudes and Behavior

The largest body of knowledge about motivations and predictors for data sharing and withholding comes from Campbell and co-authors[15-17]. They surveyed researchers, asking whether they have ever requested data and been denied, or themselves denied other researchers from access to data. Results indicated that participation in relationships with industry, mentors' discouragement of data sharing, negative past experience with data sharing, and male gender were associated with data withholding.[15] In another survey, among geneticists who said they intentionally withheld data related to their published work, 80% said it was too much effort to share the data, 64% said they withheld data to protect the ability of a junior team member to publish, and 53% withheld data to protect their own publishing opportunities.[16]

Occasionally, the administrators of centralized data servers publish feedback surveys of their users. As an example, Ventura[18] reports a survey of researchers who submitted and reviewed microarray studies in the *Physiological Genomics* journal after its mandatory data submission policy had been in place for two years. Almost all (92%) of authors said that they believed depositing microarray data was of value to the scientific community and about half (55%) were aware of other researchers reusing data from the database.

2.2 Identifying Instances of Data Sharing and Reuse

While surveys have provided insight into sharing and reuse behavior, other issues are best examined by studying the demonstrated behavior of scientists. Unfortunately, observed measurement of data behavior is difficult because of the complexity in identifying all episodes of data sharing and reuse. Although indications of sharing and reuse usually exist within a published research report, the descriptions are in unstructured free text and thus complex to extract.

Data sharing can sometimes be inferred from the "primary citation" field of database submission entries, however these references often missing when data is submitted prior to study publication. Populating the submission citation fields retrospectively requires intensive manual effort, as demonstrated by the recent Protein Data Bank remediation project[19], and thus is not usually performed. No effective way exists to automatically retrieve and index data housed on personal or lab websites or journal supplementary information.

Identifying instances of data reuse is even more difficult. There are few collections or queries that identify studies which reuse data, with the exception of meta-analyses.

Reuse often (but not always) includes a citation reference to the study that produced the data. Mercer et al.[20] is one of several researcher teams who have derived a classification schema for citation contexts. Several groups have methods of automatically classifying citation contexts using natural language processing (NLP) techniques; Teufel et. al[21] uses cue phrases to classify

citations into several groups, including a broad "adapts or modifies tools/methods/data" category.

2.3 Estimating the Costs and Benefits of Data Sharing and Reuse

Estimating the costs and benefits of data sharing and reuse would be challenging even with a comprehensive dataset of occurrences. A complete evaluation would require comparing projects that shared or reused with other similar projects that did not, across a wide variety of variables including person-hours-till-completion, total project cost, received citations and their impact, the number and impact of future publications, promotion, success in future grant proposals, and general recognition and respect in the field.

Pienta[22] is currently investigating these questions with respect to social science research data and publications. Zimmerman[23] has studied the ways in which ecologists find and validate datasets to overcome the personal costs and risks of data reuse.

Examining variables for their benefits on research impact is a common theme within the field of bibliometrics. Research impact is usually approximated by citation metrics, despite their recognized limitations.[24]

2.4 Evaluating the Impact of Data Sharing Policies

Studying the impact of data sharing policies is difficult because policies are often confounded with other variables. If, for example, impact factor is positively correlated with a strong journal data sharing policy as well as a large research impact, it is difficult to distinguish the direction of causation. Evaluating data sharing policies would ideally involve a randomized controlled trial, but unfortunately this is impractical.

Despite many funder and journal policies requesting and requiring data sharing, the impact of these policies have only been measured in small and disparate studies. McCain manually categorized the journal "Instruction to Author" statements in 1995.[2] A more recent manual review of gene sequence papers found that, despite requirements, up to 15% of articles did not submit their datasets to Genbank.[25]

2.5 Related Fields

Evaluation of data sharing and reuse behavior is related to a number of other active research fields: code reusability in software engineering, motivation in open source projects and corporate knowledge sharing, tools for collaboration, evaluating research output, the sociological study of altruism, information retrieval, usage metrics, data standards, the semantic web, open access, and open notebook science.

3. RESEARCH QUESTIONS

Within the scope of this project, I plan to address the following questions:

- What is the prevalence of biomedical research data sharing? Biomedical research data reuse?
- What features are most associated with an investigator's decision to share or reuse a biomedical research dataset?
- Does sharing or reusing data contribute to the impact of a research article, independently of other factors?

- What do the results suggest for developing efficient, effective policies, tools, and initiatives for promoting data sharing and reuse?

I will consider these questions within the context of gene expression microarray data. Microarray data provides a useful environment for investigation: despite being valuable for reuse and costly to collect, is not yet universally shared.

4. PROPOSED METHODOLOGY

I propose to address the research questions by (1) collecting a cohort of articles about gene expression microarray data, (2) developing a system to automatically identify instances of dataset sharing and reuse within the cohort, and (3) analyzing the instances of dataset sharing and reuse for univariate and multivariate predictors. These steps are explained in further detail below.

4.1 Data Collection

The cohort will consist of English-language non-review research articles indexed in PubMed under the MeSH term “gene expression profiling,” published between 2000 and 2007 (21000 articles). Using a combination of automated and manual steps I will obtain the full text of all articles that are available electronically in machine-readable format with a University of Pittsburgh HSL account. The final article count will depend on the availability of machine-readable articles and permission to download articles in bulk from publisher websites.

For each article, I will record many potentially relevant covariates, including number of authors, sources of funding, MeSH terms related to organism and disease of study, journal impact factor, journal subdiscipline, journal data sharing policy (or lack thereof), and whether the article was originally published by the journal as open-access.

Finally, I will record the citation history of the article from the ISI Web of Science. I will attempt to remove self-citations and reuse citations by investigators who previously co-authored a paper with the original research team. This will hopefully eliminate reuse due to restricted data sharing “behind the scenes” with current and former colleagues and students.

4.2 Data Classification

4.2.1 Criteria for classification

For the purposes of this study, I will consider data “shared” if it is publicly available on the internet. I will use a variety of mechanisms to classify each article as Dataset-Producing or Dataset-NonProducing, Dataset-Sharing or Dataset-NonSharing, and Dataset-Reusing or Dataset-NonReusing.

An article will be considered Dataset-Producing if the full text indicates the execution of a wet-lab gene expression microarray experiment.

All Dataset-Producing articles will be assessed for Dataset-Sharing status. I will consider an article to have shared its dataset if: (a) its PubMed ID or citation is included in a dataset submission record within the Gene Expression Omnibus (GEO)[26], ArrayExpress[27], or SMD[28] databases, or (b) it contains lexical phrases indicating data submission to a database or website.

All cohort articles will be considered as potentially Dataset-Reusing. I will consider an article to have reused a previously

shared microarray dataset if: (a) the article’s PubMed ID or citation is included on a list of data reuse studies (such as the partial list of reused GEO-datasets maintained on GEO’s website), (b) the article has MeSH terms suggesting it is a meta-analysis, or (c) the article’s full text contains lexical phrases and/or citations indicating data reuse.

4.2.2 Automatic classification system

I will manually classify a random subset of cohort articles according to the above criteria, and use this gold standard to develop and validate a natural language processing (NLP) systems to do the classifications automatically. The NLP approach will be similar to the preliminary work on Dataset-Sharing classification described in Section 5.4. I also plan on experimenting with additional NLP techniques such as semi-supervised training[29], bootstrapping cue phrases[30], and boosting classifiers as necessary.

I expect the Dataset-Producing classification problem to be relatively straightforward because standard and relatively-specific terms are used to describe the method for a gene expression experiment (e.g., RNA extraction, hybridization, imaging). I expect Dataset-Reuse classification to be fairly challenging, because there are so many diverse ways to acknowledge data reuse provenance within free text.

4.3 Analysis to address Research Questions

4.3.1 Prevalence of sharing and reuse

I will compute the prevalence of sharing by dividing the number of articles identified as Dataset-Sharing by the number identified as Dataset-Producing. The prevalence of reuse is simply the number of articles identified as Dataset-Reusing divided by the total number of articles in the cohort.

4.3.2 Features associated with sharing and reuse

For each of the three cohort classifications (Dataset-Producing, Dataset-Sharing, and Dataset-Reusing), I will compute the univariate odds ratio for each of the covariates described in Section 4.1. I will also compute a multivariate logistic regression using these covariates for each of the three classifications.

4.3.3 Effect of sharing and reuse on article impact

I will use regression to assess the association between data decisions and research impact (approximated by citation count), independently of other covariates known to impact citation count. I will compute a multivariate linear regression over the logarithmic-transform of each article’s yearly citation count, including as independent variables the collected covariates, the three binary covariates representing the Dataset-Producing classification, Dataset-Sharing classification, and Dataset-Reusing classification, and interaction terms.

4.3.4 Implications

I will consider the analysis results in light of the current data sharing policy environment to highlight potential implications.

5. PRELIMINARY RESULTS

This project proposal involves integrating and extending the preliminary work described below.

5.1 Data Sharing Impact in a Pilot Cohort

We conducted a preliminary investigation into the citation impact of data sharing by a small, homogeneous cohort of studies.[31] Using linear regression, we found that studies with publicly shared microarray data were associated with a 69% increase in citation count compared to studies without shared datasets, independently of journal impact factor, date of publication, and author country of origin (see Table 1).

Table 1. Multivariate regression on citation count for 85 clinical cancer microarray publications. Reproduced from [31].

| | Percent increase in citation count (95% confidence interval) | p-value |
|---|--|---------|
| Publish in a journal with twice the impact factor | 84% (59 to 109%) | <0.001 |
| Increase the publication date by a month | -3% (-5 to -2%) | <0.001 |
| Include a US author | 38% (1 to 89%) | 0.049 |
| Make data publicly available | 69% (18 to 143%) | 0.006 |

The project extends this preliminary analysis by considering a larger and more heterogeneous cohort, additional covariates, an analysis to predict sharing prevalence, and the additional endpoint of dataset reuse.

5.2 Impact of Journal Policy

We conducted a pilot study to understand the current state of data sharing policies within journals, the features of journals that are associated with the strength of their data sharing policies, and whether the strength of data sharing policies impact the observed prevalence of data sharing.[3] We measured data sharing prevalence as the proportion of papers with submission links from NCBI's Gene Expression Omnibus (GEO) database. We conducted univariate and linear multivariate regressions to understand the relationship between the strength of data sharing policy and journal impact factor, journal subdiscipline, journal publisher (academic societies vs. commercial), and publishing model (open vs. closed access). Of the 70 journal policies, 53 made some mention of sharing publication-related data within their Instruction to Author statements. Of the 40 policies with a data sharing policy applicable to microarrays, we classified 17 as weak and 23 as strong. Policy strength was positively associated with measured data sharing submission into the GEO database: the journals with no data sharing policy, a weak policy, and a strong policy had median data sharing prevalence of 8%, 20%, and 25% respectively (see Figure 1).

This preliminary work suggests that journal policy is an important factor, and (when a policy exists) is extractable from a journal's Information to Author statement.

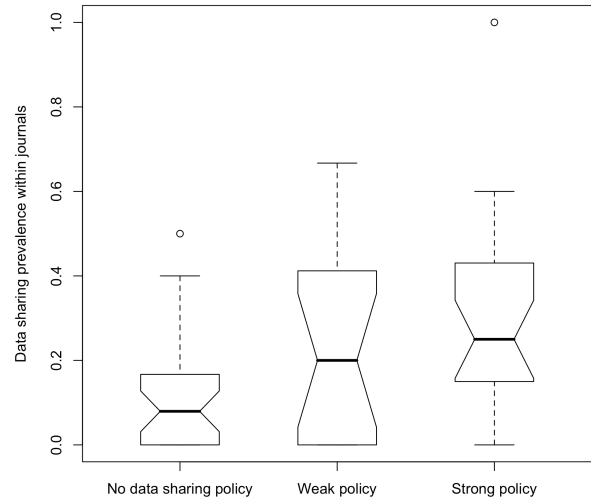


Figure 1. A boxplot of the relative data-sharing prevalence for each journal, grouped by the strength of the journal's data-sharing policy. For each group, the heavy line indicates the median and the box encompasses the interquartile. NOTE: Prevalence analysis has not been restricted to data-producing articles, and so must be considered relatively and not to an absolute of 100%. Reproduced from [3].

5.3 Prevalence across Research Topic Areas

We performed a preliminary investigation of rough keyword-based Dataset-Producing and Dataset-Reusing classifiers, trained and tested on a manually-labeled set of documents (PLoS articles prior to January 2007 containing the word "microarray," n=200).[32] We compared the Medical Subject Heading (MeSH) terms of the articles identified as Dataset-Reusing to those identified as Dataset-Producing to estimate the odds that a specific MeSH term would be used given all studies with original microarray data, compared to the odds of the same term describing studies with re-used data. Publications with reused data did involve a relatively high proportion of studies involving fungi (odds ratio (OR)=2.4), and a relatively low proportion involving rats, bacteria, viruses, plants, or genetically-altered or inbred animals (OR<0.5) compared to publications with original data.

We also assessed the prevalence and patterns of Dataset-Sharing, using only links from within the GEO or ArrayExpress database[33]. Of the 2503 articles, 440 (18%) articles had links from either GEO or ArrayExpress. Interestingly, studies with free full text at PubMed were twice (OR=2.1) as likely to be linked as a data source within GEO or ArrayExpress than those without free full text, as seen in Figure 2. Studies with human data were less likely to have a link (OR=0.8) than studies with only non-human data. The proportion of articles with a link within these two databases has increased over time: the odds of a data-source link for studies was 2.5 times greater for studies published in 2006 than 2002. As might be expected, studies with the fewest funding sources had the fewest data-sharing links:

only 28 (6%) of the 433 studies with no funding source were listed within GEO or ArrayExpress. In contrast, studies funded by the NIH, the US government, or a non-US government source had data-sharing links in 282 of 1556 cases (18%), while studies funded by two or more of these mechanisms were listed in the databases in 130 out of 514 cases (25%).

These studies demonstrate that funding source, organism, and open-access are important covariates in data behavior.

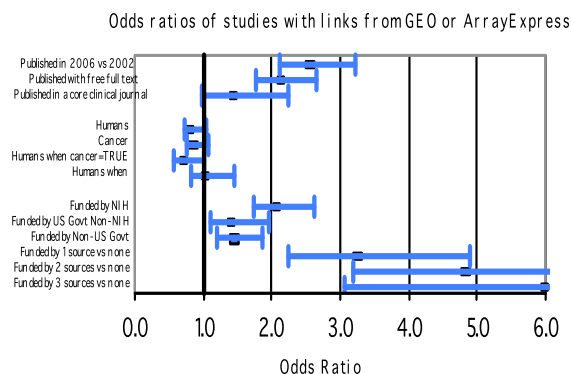


Figure 2. Preliminary Indications of Data-Sharing Patterns with 95% confidence intervals. Reproduced from [33].

5.4 Automatic Identification of Data Sharing

A pilot NLP system has been developed and validated for one of the three proposed cohort classifications, Dataset-Sharing versus Dataset-NonSharing.[34] Using regular expression patterns and machine learning algorithms on open access biomedical literature published in 2006, our system was able to identify 61% of articles with shared datasets with 80% precision. A simpler version of our classifier achieved higher recall (86%), though lower precision (49%).

These results demonstrate the feasibility of using an NLP approach to automatically identify instances of data sharing from biomedical full text research articles.

6. LIMITATIONS AND RISKS

We note an important limitation of this proposal: associations do not imply causation. The research here will not be sufficient to conclude, for example, that a policy change associated with increased data sharing will in fact *cause* increased sharing. It would be possible that both factors stem from a common cause.

This study has several other limitations. Although restricting the study to one datatype allows an in-depth analysis of many specific facets of data sharing and reuse, future work should apply the methodology and lessons learned to other datatypes to quantify generalizability. My sample will omit some articles: I might not find all of the datasets that have been shared in niche databases or on personal or lab websites and not all papers will be available in machine-readable full text, particularly for early years. I am not considering datasets as shared if they are available upon request and may thereby discount an important and effective sharing mechanism. Although broadly used, citations are a rough and imperfect measurement of research impact, in part because they may include negative critiques of an article or its associated data. Citations do not consider reuse in

the context of education and training, and thus undervalue the impact of data sharing reused for this purpose.

The largest technical risk in the research plan is that it may be unexpectedly difficult to automatically identify reuse with acceptable precision and recall. In this case, I plan to supplement the automated classification with manual curation, possibly resulting in a smaller cohort of articles for analysis.

7. ANTICIPATED CONTRIBUTIONS

I anticipate several important contributions arising from this novel research application.

First, I would, of course, make my dataset publicly available (limited only by licensing restrictions). This would provide a foundational data sharing and reuse dataset for further study by other researchers. I could imagine future work extending and refining my analysis, using the data to investigate novel questions such as whether the data-sharing community has members in common with the data-reuse community – interesting, and also relevant to developing incentives and policies. I envision the reuse data forming the backbone of a Data Reuse Registry, providing a prototype system for ongoing prospective data reuse attribution and cataloguing.[35]

Although some of the analysis results may be intuitive (a stronger journal data sharing policy results in more data sharing, or shared data permits reuse and thus supports a higher citation rate), these relationships have not yet been demonstrated. Concrete, supporting – or contradictory! – evidence will be of value to a wide spectrum of decision-makers.

I hope this research will identify sub-communities with frequent practices of sharing and reuse. Examining these situations can highlight best practices to be used when developing research agendas, tools, standards, repositories, and communities in areas that have yet to receive major benefits from shared data.

Finally, but most importantly, I believe this research will inspire further work in this area. There is a common adage: “You can not manage what you do not measure.” Research consumes considerable resources from the public trust. As data sharing and reuse are evaluated and policies and incentives improved, hopefully investigators will become more apt to share and reuse study data and thus maximize its usefulness to society.

8. ACKNOWLEDGMENTS

I thank my advisor Dr Wendy Chapman for her support and discussion of these ideas, the 2008 Joint Conference on Digital Libraries Doctoral Consortium reviewers and participants for their insightful feedback, Virginia Tech for travel funding, and the NLM for training support through grant 5T15-LM007059-19.

9. REFERENCES

- [1] Fienberg, S. E., Martin, M. E. and Straf, M. L. Sharing research data. National Academy Press, Washington, D.C., 1985.
- [2] McCain, K. Mandating Sharing: Journal Policies in the Natural Sciences. *Science Communication*. 1995;16(4):403-431.
- [3] Piwowar, H. A. and Chapman, W. W. A review of journal policies for sharing research data. *ELPUB* 2008.

- [4] NIH Data Sharing Policy and Implementation Guidance. 2003.
- [5] NIH. NOT-OD-08-013: Implementation Guidance and Instructions for Applicants: Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS). 2007.
- [6] Cech, T. Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. National Academies Press, Washington, D.C., 2003.
- [7] Got data? *Nat Neurosci.* 2007;10(8):931-931.
- [8] Compete, collaborate, compel. *Nat Genet.* 2007;39(8).
- [9] Democratizing proteomics data. *Nat Biotech.* 2007; 25(3):262-262.
- [10] Time for leadership. *Nat Biotech.* 2007; 25(8):821-821.
- [11] How to encourage the right behaviour. *Nature.* 2002;416(6876)1.
- [12] Altman, R. B., et al. Genetic nondiscrimination legislation: a critical prerequisite for pharmacogenomics data sharing. *Pharmacogenomics.* 2007;8(5):519.
- [13] Kakazu, K. K., Cheung, L. W. and Lynne, W. The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. *Hawaii Med J.* 2004;63(9 Sep):273-275.
- [14] GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet.* 2007;39(9):1045-1051.
- [15] Blumenthal, D., Campbell, E. G., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S. and Holtzman, N. A. Data withholding in genetics and the other life sciences: prevalences and predictors. *Acad Med.* 2006; 81(2):137-145.
- [16] Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A. and Blumenthal, D. Data withholding in academic genetics: evidence from a national survey. *JAMA.* 2002; 287(4):473-480.
- [17] Vogeli, C., Yucel, R., Bendavid, E., Jones, L. M., Anderson, M. S., Louis, K. S. and Campbell, E. G. Data withholding and the next generation of scientists: results of a national survey. *Acad Med.* 2006; 81(2):128-136.
- [18] Ventura, B. Mandatory submission of microarray data to public repositories: how is it working? *Physiol Genomics.* 2005;20(2) Jan 20:153-156.
- [19] PDBj. Report on the Remediation of Primary Citations of PDB data. PDBj News Letter, Volume 7, March 2006. Available at <http://www.pdbj.org/NewsLetter/newsletter_vol7_e.pdf>
- [20] Mercer, R. and Di Marco, C. The Importance of Fine-Grained Cue Phrases in Scientific Citations. Canadian Conference on AI 2003.
- [21] Teufel, S., Siddharthan, A. and Tidhar, D. Automatic classification of citation function. *ACL* 2006.
- [22] Pienta, A. 1R01LM009765-01 Barriers and Opportunities for Sharing Research Data. NIH Grant. 2007.
- [23] Zimmerman, A. Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries.* 2007;7(1-2):5-16.
- [24] Seglen, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ.* 1997;314(7079):498-502.
- [25] Noor, M. A., Zimmerman, K. J. and Teeter, K. C. Data Sharing: How Much Doesn't Get Submitted to GenBank? *PLoS Biol.* 2006; 4(7).
- [26] Edgar, R., Domrachev, M. and Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30(1):207-210.
- [27] Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. and Brazma, A. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2005;33, Database issue Jan 1:D553-555.
- [28] Sherlock, G., Boussard, H., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. and Cherry, J. M. The Stanford Microarray Database. *Nucleic Acids Res.* 2001; 29(1):152-155.
- [29] Medlock, B. Exploring hedge identification in biomedical literature. *J Biomed Inform.* 2008; Aug 41(4):636-54.
- [30] Abdalla, R. and Teufel, S. A bootstrapping approach to unsupervised detection of cue phrase variants. *ACL* 2006.
- [31] Piwowar, H. A., Day, R. S. and Fridsma, D. B. Sharing detailed research data is associated with increased citation rate. *PLoS ONE.* 2007; 2(3):e308.
- [32] Piwowar, H. A. and Fridsma, D. B. Examining the uses of shared data. Poster at ISMB 2007. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2007.425.3>>
- [33] Piwowar, H. A. and Chapman, W. W. Prevalence and Patterns of Microarray Data Sharing. Poster at PSB 2008. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2008.1701.1>>
- [34] Piwowar, H. A. and Chapman, W. W. Identifying Data Sharing in Biomedical Literature. *AMIA* 2008.
- [35] Piwowar, H. A. and Chapman, W. W. Envisioning a Biomedical Data Reuse Registry. Poster at *AMIA* 2008.