

Foundational Studies for Measuring the Impact, Prevalence, and Patterns of Publicly Shared Biomedical Research Data

Heather Piwowar

Department of Biomedical Informatics

University of Pittsburgh

A. Description of the Research	2
A.1 Background	2
A.2 Significance	3
A.3 Methodology	5
Aim 1 – Does sharing have benefit for those who share?	5
Aim 2 – Can sharing and withholding be systematically measured?	7
Aim 3 – How often is data shared? What predicts sharing? How can we model sharing behavior?	9
A.4 Future Plans	10
A.5 Bibliography	11
B. Schedule of Completion	12
C. Budget	12
D. Other Support	13
E. Advisor Endorsement	13
F. Curriculum Vitae	14 - 15

Foundational Studies for Measuring the Impact, Prevalence, and Patterns of Publicly Shared Biomedical Research Data

A. Description of the Research

A.1 Background

Many initiatives encourage research data sharing in hopes of increasing research efficiency and quality, but the effectiveness of these early initiatives is not well understood. Reusing research data has many benefits for the scientific community: new research hypotheses can be tested more quickly and inexpensively when duplicate data collection is reduced. Shared data can be aggregated to study otherwise-intractable issues, and a more diverse set of scientists can become involved when analysis is opened beyond those who collected the original data. Publicly available data helps to identify errors, discourages fraud, and is useful for training new researchers.

Funders, publishers and academic organizations -- eager to realize such benefits -- have developed tools, resources and policies to encourage and require data-producing investigators to make their datasets publicly available. Despite these investments of time and money, **we do not have a firm grasp on the prevalence or patterns of data sharing and reuse, the effectiveness of initiatives, or the costs, benefits, and impact of repurposing biomedical research data.**

Previous assessment methods for assessing data sharing prevalence have included manual curation (Gleditsch & Strand, 2003; McCain, 1995; McCullough, McGeary, & Harrison, 2008; Noor, Zimmerman, & Teeter, 2006; Ochsner et al., 2008; Piwowar, Day, & Fridsma, 2007) and investigator self-reporting (Blumenthal et al., 2006; Campbell et al., 2002). Models of knowledge sharing have emerged from the information science and management of information systems communities, usually derived from case studies or survey instruments (Bock, Zmud, & Lee, 2005; Cabrera, Collins, & Salgado, 2006; Constant, Kiesler, & Sproull, 1994; Harder, 2008; Hedstrom, 2006; Hsu et al., 2007; Kolekofski, 2003; Lee, Dourish, & Mark, 2006; Liang, Liu, & Wu, 2008; Samieh & Wahba, 2007; Seonghee & Boryung, 2008; Siemsen, Roth, & Balasubramanian, 2008; Wasko & Faraj, 2005). These approaches provide insight into motivation, but are subject to an intention-action gap (Kuo & Young, 2008) and are labor-intensive to repeat in multiple subdisciplines and over time to monitor changes in behavior.

The proposed research will build on and supplement previous work through an analysis of observed variables, thereby providing an alternative perspective for understanding and monitoring data sharing behavior.

A.2 Significance

This proposal describes a new, exploratory, and innovative research project that could significantly impact the adoption of data sharing in biomedical research. You can not manage what you do not measure: understanding the rewards, prevalence, and patterns of data sharing and withholding will facilitate effective refinement of data sharing initiatives to better address real-world needs.

The proposed evaluation of current data sharing behavior will be useful in three ways. First, an estimate of the prevalence with which data is shared, either voluntarily or under mandate, will provide a valuable baseline for assessing future adoption and continued intervention. Second, analyses of current behavior will likely identify subfields (perhaps research areas with a particular disease or organism focus, or those in well funded research groups) with relatively high prevalence of data sharing; digging into these can illuminate valuable best practices. Third, the same analyses will likely reveal subareas in which researchers rarely share their research datasets. Future research could focus on these challenging areas, to understand their unique obstacles for data sharing and refine future initiatives accordingly.

Progress will also be of significant value to a broad cross-section of disciplines, including:

- **Funders, policy makers and thought leaders.** Although some results of this analysis 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 policy makers and thought leaders.
- **Information science and digital library community.** Data use behavior and resource use metrics are active research topics in information science and digital library research; the proposed research applies metrics to real-world policy questions. My proposed bibliometric emphasis supplements the more common survey and case-study approaches to studying data sharing behavior. It also provides an example of research that leverages publicly-available bibliometric tools.
- **Database, software, and data standard developers.** The usage patterns of those who share data provide critical requirement specification feedback for developing and refining databases, software, and standards to support data sharing and reuse. Learning who does not currently share data can provide insight into failings of current tools and opportunities for improvements.

- **Biomedical informatics community.** Informatics involves evaluation of the generation, use, and value of information resources; this research addresses this topic from a novel perspective. The biomedical informatics field will also benefit by exposure to methods it does not commonly apply. For example, my plans to apply natural language processing techniques to the biomedical literature through full-text portals has far-reaching applicability. Finally, the general biomedical informatics community will benefit if and when this research leads to initiatives that increase the rate of data sharing.
- **Open Science community.** Grassroots movements to increase openness and transparency in science will benefit from rigorous, quantitative assessments of current data sharing behavior.
- **Primary Investigators.** Last but not least, I expect that this research will help inspire investigators to share their data and help inform the creation of tools that help them. As data sharing is 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.

Expected contributions, taking the form of papers and associated datasets, include:

1. an assessment of the observed and measured rewards, prevalence, and patterns of gene expression microarray dataset sharing
2. a publicly available dataset associating microarray study publications with data sharing status
3. reusable open-source software for collecting and analyzing the bibliometric data in this study
4. a generalizable approach for developing practical, real-world natural language tools for information retrieval and extraction within a wide selection of biomedical literature
5. preliminary models of data sharing behavior using observed variables

Although I plan to limit this study to one datatype to allow an in-depth analysis, I believe the approach and results to be largely generalizable across domains.

As further support for the significance of this work, our preliminary publications have been enthusiastically welcomed by peer reviewers; reviews have frequently declared the research to be “relevant and timely.”

A.3 Methodology

Aim 1 – Does sharing have benefit for those who share?

Goal: Measure the association between an article's publication citation rate and whether its authors made their gene expression datasets publicly available. The results of Aim 1 provide motivation for Aim 2 and preliminary work for Aim 3.

Importance: While the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Demonstrating a boost in citation rate would be a potentially important motivator for publication authors. To my knowledge, this is the first study to investigate a relationship between citation rate and biomedical data availability. This work also serves as preliminary work for measuring sharing prevalence and patterns.

Dataset and Methods: We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data.

Status: I have completed and published a study addressing Aim 1:

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. [doi:10.1371/journal.pone.0000308](https://doi.org/10.1371/journal.pone.0000308)

The complete paper will be included in the final dissertation.(Piwowar, Day, & Fridsma, 2007)

Results: Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. We found the 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), as illustrated in Figure 1.

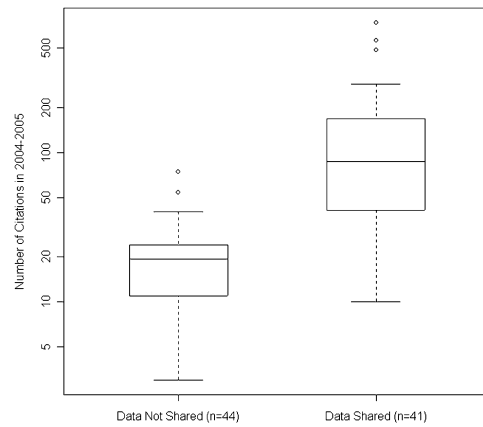


Figure 1: Distribution of 2004-2005 citation counts of 85 trials by data availability.

Whether a trial's dataset was made publicly available was significantly associated with the log of its 2004–2005 citation rate (69% increase in citation count; 95% confidence interval: 18 to 143%, $p = 0.006$), independent of journal impact factor, date of publication, and US authorship. Detailed results of this multivariate linear regression are given in Table 1. This result held even for lower-profile publications and thus is relevant to authors of all trials.

Table 1: Multivariate regression on citation count for 85 publications.

	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

We calculated a multivariate linear regression over the citation counts, including covariates for journal impact factor, date of publication, US authorship, and data availability. The coefficients and p-values for each of the covariates are shown here, representing the contribution of each covariate to the citation count, independent of other covariates.

doi:10.1371/journal.pone.0000308.t002

Conclusion: Research consumes considerable resources from the public trust. As data sharing gets easier and benefits are demonstrated for the individual investigator, hopefully authors will become more apt to share their study data and thus maximize its usefulness to society.

Aim 2 – Can sharing and withholding be systematically measured?

Goal: Develop and evaluate methods for automatically identifying biomedical research data sharing and withholding. This will involve two sub-aims, discussed separately below. The results of Aim 2 will be used to generate a dataset for use in Aim 3.

Aim 2a – Identify studies that create data

Background: Manual classification is not scalable, and indexing terms do not have sufficient recall and precision to identify studies that generate gene expression microarray data. Natural language processing (NLP) can be used to automatically extract this methodological information from article full. Acquiring a local repository of article full-text is complex. Luckily, the full text of many biomedical studies can be queried through full-text portals, including PubMed Central, Highwire Press, and Scirus. I estimate that when their results are pooled, these portals query the full text of 85% of gene expression articles published more than 12 months ago, relative to full-text available for download through the University of Pittsburgh library subscriptions.

Method: I propose to use statistical and lexical NLP approaches to design a query for use in full-text portals, retrieving articles that have generated gene expression microarray datasets. I anticipate that “biology wet lab” words such as *isolate*, *hybridize*, and *probe* will be relevant, alone and in short phrases.

Evaluation: I will use the manually-curated dataset of Ochsner et al (2008) as a reference standard to evaluate recall and precision of my full-text query. Ochsner identified 400 true positives (studies that generated microarray data) from an inclusion set of 800 (query on titles, abstracts, and index terms on 20 journal publications in 2007). Since there are no established performance requirements for this task, I will consider performance adequate if the precision of my full-text query is above 70% against this reference standard. Recall performance must be high enough that use of the query in Aim 3 will retrieve enough datapoints to power the subsequent regression analysis. Opinions differ on how many datapoints are needed to adequately power a regression analysis, but a conservative estimate suggests that 1250 articles (40 datapoints x 30 regression coefficients) should be sufficient (Nunnally & Bernstein, 1978).

Risks and Contingency Plans: The largest technical risk in the research plan is that it may be unexpectedly difficult to automatically identify dataset production 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.

Aim 2b – Identify studies that share their data

Background: Identifying shared datasets is a difficult problem, in general. Within the domain of microarray analysis, the task is easier. The Gene Expression Omnibus (Barrett et al., 2007) and ArrayExpress (Parkinson et al., 2007) databases have emerged as the dominant centralized repositories for sharing gene expression microarray data (Piwowar & Chapman, 2008a). Importantly, they contain links from submitted datasets to the PubMed identifiers of primary citation reports. Our preliminary work suggests that database citation links have high recall for retrieving articles with data shared in centralized databases. Database citation links have the added benefits of almost-perfect precision, a wide scope without the need for access to full text, and no bias introduced through community norms in lexical statements of data sharing within full text.

Method: I propose to identify shared data using citation links from GEO through the PubMed filter “pubmed_gds[filter]”, and from ArrayExpress through a search for PubMed IDs in their query interface.

Evaluation: I will calculate the recall of database citation links to estimate the proportion of shared datasets it retrieves. Using 300 random articles from the Ochsner et al. (Ochsner et al., 2008) review as a reference standard, I will calculate:

$$\text{recall} = \frac{\text{the number of articles identified by Ochsner as having shared data that also are linked from GEO}}{\text{the total number of articles found by Ochsner as having shared data}}$$

I will also compare the characteristics of the retrieved studies to ensure they are not a skewed subset.

Status: Data collection and statistics completed. Drafting manuscript, to be submitted to BMC Bioinformatics.

Risks and Contingency Plans: If the query evaluation suggests insufficient recall (< 70%), I will develop and apply NLP filters similar to our previous work (Piwowar & Chapman, 2008c) to sacrifice some precision for recall.

Limitations and Assumptions: This approach includes sharing to the predominant centralized database and excludes sharing to databases for which there is no citation link within the submission entry. Although this will lead to underestimating the prevalence of data sharing, I don't expect it to bias our estimates of data sharing patterns.

Aim 3 – How often is data shared? What predicts sharing? How can we model sharing behavior?

Goal: Measure current data sharing and withholding behavior, and associate these sharing decisions with features that may predict or influence an investigator's choice.

Dataset: The dataset will be comprised of all articles reachable by full text query within PubMed Central, Highwire Press, and Scirus using the query developed in Aim 2a. Articles with datasets found by the queries of Aim 2b will be considered to have shared data; the rest of the articles will be considered to have withheld data.

Proposed features: I selected a set of features to collect and analyze, chosen based on degree of directness it serves as a proxy, completeness for which they are available, and ease of collection within the scope of this project. I hypothesize the following variables will be associated with an increased prevalence of data sharing:

Author characteristics

(for both the first and last authors, separately**)

- career citations in PMC
- number of prior gene expression publications
- number of prior publications
- years since first publication
- published in open access journals before?
- published papers with shared microarray data before
- previously reused gene expression datasets from GEO
- is an NIH PI
- gender

*** I will attempt to disambiguate author names through use of the Author-ity system (Torvik & Smalheiser)*

Study characteristics

- organism under study
- disease under study

Environmental characteristics

- year of publication
- number of co-authors
- corresponding author country
- number of funding sources
- journal prestige
- is an open-access journal
- research-orientation of university
- number of datasets submitted by this institution
- university vs. other type of institution
- relative amount of tech transfer from this institution

Policy characteristics

- strength of journal data sharing policy
- strength of funder data sharing policy

Method: I will calculate prevalence of dataset sharing within our full sample, then adjust this raw estimate based on the relevant precision and recall values estimated in Aim 2, to account for over- and under- estimates in retrieval numbers due to query imprecision. Next, I will compute the univariate odds for each of the features to assess the degree to which they are associated with sharing datasets that have been produced, using the nonparametric Wilcoxon Mann-Whitney and Fisher's exact tests, as appropriate. For the core analysis, I will use multivariate logistic regression to compute the independent association of each variable to the probability of dataset sharing, and report the coefficients and 95% confidence intervals. As the last component of this project, using exploratory factor analysis I intend to derive a model of data sharing behavior based on observed variables.

Pilot Studies: Our pilot studies, using a subset of articles and covariates, suggest the feasibility of this approach (Piwowar & Chapman, 2008a, 2008b; Piwowar & Chapman, 2009).

Limitations and Assumptions: Because associations do not imply causation, the proposed research 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.

The study has additional limitations. Although restricting the investigation to only microarray allows an in-depth analysis of specific facets of data sharing, future work should apply the methodology and lessons learned to other datatypes to quantify generalizability. The study is limited by the accuracy with which I can identify dataset creation and data sharing. The study is limited to published articles with queryable full text, and thus will omit some older articles or those published in more obscure journals. I am not considering datasets as shared if they are available upon request or published online in another venue than a major database, and may thereby discount an important and effective sharing mechanism. This approach assumes that my list of observed variables includes proxies for many of the actual decision-making influences. My attempts to unambiguously identify authors will contain errors, and thus my estimations of previous publishing, data sharing behavior, and grant information will contain errors. This analysis assumes that the first and last authors are the main decision makers about whether or not to share datasets. Finally, several variables are NIH-centric, which erodes our ability to understand the influences of institutional factors or funding levels in the rest of the world (about 54% of microarray papers have non-USA contact addresses).

A.4 Future Plans

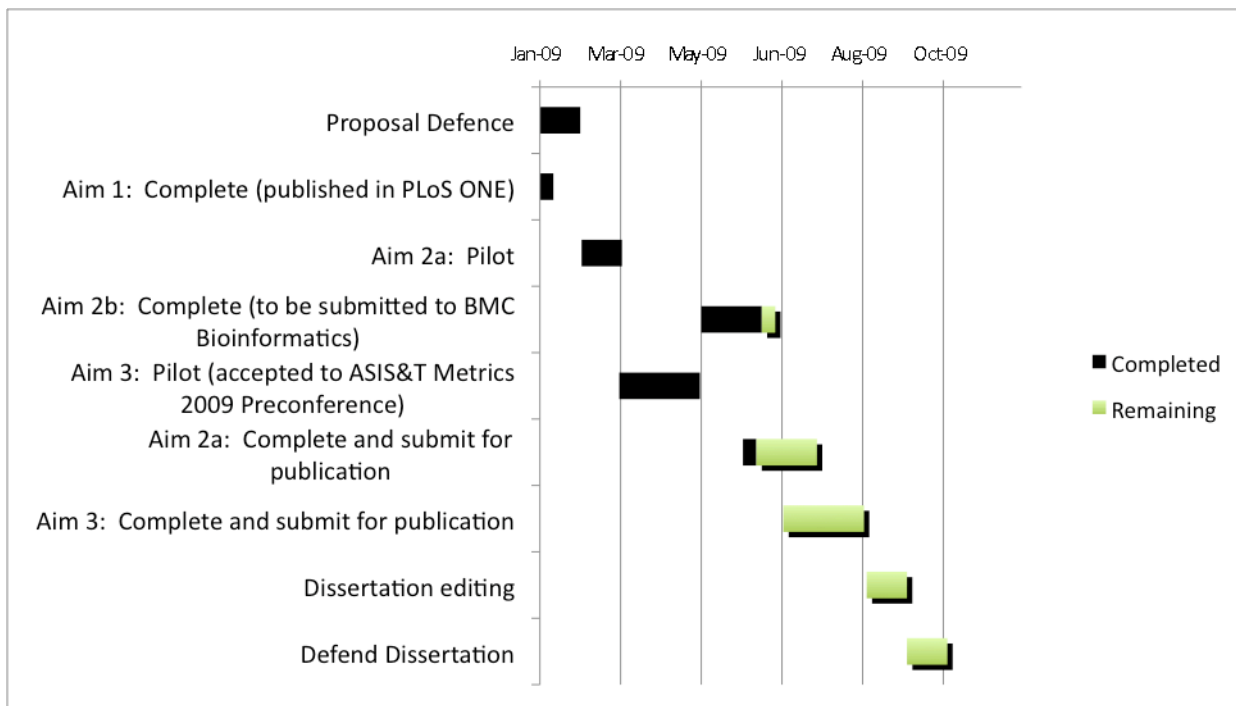
This work can be extended in many ways, beyond the scope of this thesis. I plan to leverage the foundations and tools established in this thesis project to study wide-scale patterns and prevalence of data *reuse*, since many of the benefits of data sharing can only be realized if shared datasets are indeed used by other investigators. A focus on wide-scale trends will compliment case-studies of data reuse, such as that of Zimmerman (2003).

I also hope to extend full-text information retrieval work: I plan to generalize the method for refining full-text queries using open access publications, and I have begun quantifying the strengths and weaknesses of using PubMed Central as a source of citation information.

A.5 Bibliography

- Barrett, T., et al. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, 35(Database issue).
- Blumenthal, D., et al. (2006). Data withholding in genetics and the other life sciences: prevalences and predictors. *Acad Med*, 81(2), 137-145.
- Bock, G., Zmud, R. W., & Lee, J. (2005). Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate. *MIS Quarterly*, 29(1), 87-112.
- Cabrera, A., Collins, W. C., & Salgado, J. F. (2006). Determinants of individual engagement in knowledge sharing. *The International Journal of Human Resource Management*, 17(2), 245-264.
- Campbell, E. G., et al. (2002). Data withholding in academic genetics: evidence from a national survey. *JAMA*, 287(4), 473-480.
- Constant, D., Kiesler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research*, 5(4), 400-421.
- Gleditsch, N. P., & Strand, H. (2003). Posting your data: will you be scooped or will you be famous? *International Studies Perspectives*, 4(1), 89-97.
- Harder, M. (2008). How Do Rewards and Management Styles Influence the Motivation to Share Knowledge? *Center for Strategic Management and Globalization, SMG Working Paper(6)*,
- Hedstrom, M. (2006). Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation. *IASSIST Conference 2006: Presentations*.
- Hsu, M., et al. (2007). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human-Computer Studies*, 65(2), 153-169.
- Kolekofski, K. (2003). Beliefs and attitudes affecting intentions to share information in an organizational setting. *Information & Management*, 40(6), 521-532.
- Kuo, F., & Young, M. (2008). A study of the intention-action gap in knowledge sharing practices. *Journal of the American Society for Information Science and Technology*, 59(8), 1224-1237.
- Lee, C. P., Dourish, P., & Mark, G. (2006). The human infrastructure of cyberinfrastructure. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*.
- Liang, T.-P., Liu, C., & Wu, C.-H. (2008, May). Can Social Exchange Theory Explain Individual Knowledge Sharing Behavior? A Meta Analysis, from <http://www.whiceb.com/download/whiceb2008/semimar/Ting-Peng%20Liang.pdf>
- McCain, K. (1995). Mandating Sharing: Journal Policies in the Natural Sciences. *Science Communication*, 16(4), 403-431.
- McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008). Do Economics Journal Archives Promote Replicable Research? *Canadian Journal of Economics*, 41(4), 1406-1420.
- Noor, M. A., Zimmerman, K. J., & Teeter, K. C. (2006). Data Sharing: How Much Doesn't Get Submitted to GenBank? *PLoS Biol*, 4(7).
- Nunnally, J. C., & Bernstein, I. H. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Ochsner, S. A., et al. (2008). Much room for improvement in deposition rates of expression microarray datasets. *Nature Methods*, 5(12), 991.
- Parkinson, H., et al. (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue).
- Piwowar, H. A., & Chapman, W. W. (2008a). A review of journal policies for sharing research data, *ELPUB*.
- Piwowar, H. A., & Chapman, W. W. (2008b). *Identifying Data Sharing in Biomedical Literature*. Paper presented at the AMIA annual symposium.
- Piwowar, H. A., & Chapman, W. W. (2008c). Linking database submissions to primary citations with PubMed Central. *BioLINK Workshop at ISMB*.
- Piwowar, H. A., & Chapman, W. W. (2009). *Public sharing of research datasets: a pilot study of associations*. Paper presented at the Metrics PreConference at ASIS&T 2009 [to appear].
- Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3).
- Samieh, H. M., & Wahba, K. (2007). Knowledge Sharing Behavior From Game Theory And Socio-Psychology Perspectives. *Hawaii International Conference on System Sciences*.
- Seonghee, K., & Boryung, J. (2008). An analysis of faculty perceptions: Attitudes toward knowledge sharing and collaboration in an academic institution. *Library* 30(4), 282-290.
- Siemens, E., Roth, A., & Balasubramanian, S. (2008). How motivation, opportunity, and ability drive knowledge sharing: The constraining-factor model. *Journal of Operations Management*, 26(3), 426-445.
- Torvik, V., & Smalheiser, N. (2009). Author Name Disambiguation in MEDLINE. *Transactions on Knowledge Discovery from Data*, [to appear].
- Wasko, M. M., & Faraj, S. (2005). Why Should I Share? *MIS Quarterly*, 29, *Special Issue on IT and Knowledge Management*(March).
- Zimmerman A. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. PhD Thesis, Information and Library Science, University of Michigan.

B. Schedule of Completion



C. Budget

I foresee expenditures related to dissemination of my research results. I anticipate having three journal-quality publications, in addition to the already-published Aim 1. I aim to publish the results of Aim 2a and 2b in domain journals, and Aim 3 in an information science journal.

Because this research depends on study publication in “gold” open access journals, I feel compelled to publish the results of this study in biomedical gold open access journals wherever appropriate. The relevant biomedical journals have author-fees, as outlined below.

In addition, I plan to present the results of this work at the JASIS&T conference, to solicit feedback from the information science community. I plan to do this by presenting the pilot work for Aim 2b at the ASIS&T 2009 Metrics preconference (accepted), and hopefully the Aim 3 model at ASIS&T 2010. My estimated travel expenses are outlined below.

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

Anticipated Expenditures	Estimated Cost
Professional Travel to present results	
ASIS&T 2009 (will present Aim 3 pilot at Metrics preconference)	\$1,200.00
	Registration, airfare, and shared hotel.
ASIS&T 2010 (publication goal for Aim 3 factor model)	\$175.00
	Registration only. (I live in the conference city)
Open Access Publishing in domain journals	
PLoS Computational Biology (publication goal for Aim 2a)	\$2,200.00
BMC Research Methods (publication goal for Aim 2b)	\$1,560.00
Total Estimated Expenditures	\$5,135.00

D. Financial Support

I received tuition support and a training stipend through National Institute of Health, National Library of Medicine training grant number 5T15-LM007059-19. This fellowship began on July 1, 2005 and ends on June 30, 2009. I have been awarded several travel grants to present preliminary work at various conferences.

I have been granted a Graduate Student Researcher stipend through the Department of Biomedical Informatics at the University of Pittsburgh from July 1, 2009 until October 31, 2009. This stipend does not include funding for professional travel or publishing fees.

I do not have other employment, scholarships, or assistantships.

E. Advisor Endorsement

My dissertation advisor, Dr Wendy Chapman, Assistant Professor of Biomedical Informatics at the University of Pittsburgh, endorses this proposal as indicated in the accompanying letter.

Other members of my dissertation committee include:

- Dr Ellen Detlefsen, Associate Professor, School of Information Sciences, University of Pittsburgh
- Dr Madhavi Ganapathiraju, Assistant Professor of Biomedical Informatics, University of Pittsburgh
- Dr Brian Butler, Associate Professor of Business Administration, University of Pittsburgh
- Dr Gunter Eysenbach, Associate Professor Dept of Health Policy, Management, and Evaluation at the University of Toronto

Heather Piwowar

hpiwowar@gmail.com

Research Passions	<i>Improving biomedical research progress by leveraging the full value of data resources.</i> <i>Current focus: evaluating data sharing and reuse policies and behaviors</i> <i>Related interests: natural language processing, machine learning, data mining, bibliometrics, simulation, genomics, personalized cancer therapy, open science...</i>
Education	University of Pittsburgh Pittsburgh, PA <u>Doctoral candidate</u> anticipating graduation in 2010 <u>Masters of Science</u> in 2006 Department of Biomedical Informatics, concentration in Bioinformatics, GPA 3.9/4.0 Massachusetts Institute of Technology Cambridge, MA <u>Bachelors of Science</u> in 1995, <u>Masters of Engineering</u> in 1996 Majored in Electrical Engineering and Computer Science, GPA 4.9/5.0 Concentration in Digital Signal Processing (DSP) Humanities and mathematics each comprised 20% of curriculum
Research History & Accomplishments	<u>Piwowar</u> , Becich, Bilofsky, Crowley (2008) Towards a data sharing culture: Incentives for Leadership from Academic Health Centers PLoS Med 5(9): e183. doi:10.1371/journal.pmed.0050183
Journal articles	<u>Piwowar</u> , Day, Fridsma (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate PLoS ONE 2: 3. e308 doi:10.1371/journal.pone.0000308
Conference papers	<u>Piwowar</u> , Chapman (2009) Public sharing of research datasets: a pilot study of associations. To appear in: Metrics Pre-conference workshop at ASIS&T 2009 <u>Piwowar</u> , Chapman (2008) Linking database submissions to primary citations with PubMed Central In: BioLINK 2008. <u>Piwowar</u> , Chapman (2008) Identifying Data Sharing in Biomedical Literature In: AMIA 2008 Annual Symposium. <u>Piwowar</u> (2008) Prevalence and Patterns of Biomedical Research Data Reuse In: JCDL 2008 Doctoral Consortium. <u>Piwowar</u> , Chapman (2008) A review of journal policies for sharing research data In: ELPUB 2008. Harkema, <u>Piwowar</u> , Amizadeh, Dowling, Ferraro, Haug, Chapman. (2008) A Baseline System for the i2b2 Obesity Challenge, In: i2b2 2008 Workshop.
Posters	<u>Piwowar</u> , Chapman (2008) Envisioning a Biomedical Data Reuse Registry AMIA <u>Piwowar</u> , Chapman (2008) Prevalence and Patterns of Microarray Data Sharing PSB 2008 <u>Piwowar</u> , Fridsma (2007) Examining the uses of shared data ISMB 2007
Invited Presentations	(2009) Measuring the Adoption of Open Science. PSB Open Science Workshop, Hawaii (2009) Intro to Test Driven Development. DBMI Computer Users Group, U of Pittsburgh (2008) Why study Data Sharing? (+ why share your data) DBMI Colloquium, University of Pittsburgh (2007) Sharing Detailed Research Data is Associated with Increased Citation Rate . NLM Trainee Conference, Stanford University
Recent Awards & Grants	(2009) Doctoral proposal accepted (2008) Best Trainee Poster Award, Department of Biomedical Informatics, University of Pittsburgh (2008) DBMI ELPUB travel grant (2008) ASIS&T SIGUSE doctoral student travel award (2007) Doctoral Comprehensive Exam passed with High Honors (2007) NSF ISMB travel grant (2007) Best Trainee Paper Award, Department of Biomedical Informatics, University of Pittsburgh (2005-2009) NLM Biomedical Informatics training fellowship
Teaching Experience	(2008) DBMI Intro to Research: Discussion on Open Science (2006, 2007, 2008) DBMI Intro to Biomedical Informatics course lecture: Informatics in Industry (2007) Series of informal Brown Bag Skill Seminars on various topics (RSS, Entrez tools, Excel tricks)

Service (2006-2008) DBMI Student Representative to Admissions and Training Program Core Committees

Online Presence Self-archived research papers and data: <http://www.dbmi.pitt.edu/piwowar>
Research blog: <http://researchremix.wordpress.com>

Industrial Experience & Accomplishments
2001 – 2005

Precision Therapeutics, Inc. **Pittsburgh, PA**
35 (now >60) people, provides personalized cancer therapy information to oncologists
Consultant, 2004-2005

- Contributed to a new, strategically critical bioinformatics algorithm
- Director of Clinical Informatics, 2003-2004
- Co-led the design and implementation of all clinical trials analysis
- Senior Systems Developer, 2001-2003
- Designed, implemented, and supported HIPAA-compliant intranet web applications for corporate data viewing and editing, using Java/J2EE and Oracle
 - Contributed ideas, analyses, and tools for use throughout the organization

1998 – 2001

Vocollect, Inc. **Pittsburgh, PA**
30 (now >350) people, develops portable voice interface computers
Senior Software Developer, 1998–2001

- Ported an embedded system to Windows CE and continued to develop and maintain C and assembly code, as a member of a small team
 - Improved the speech recognition accuracy of our product, as a member of a small team
 - Co-led a skunk works project which envisioned and prototyped our real-time enterprise management product as a web application
 - Contributed to department learning lunches and company vision statement
- Manager of Customer Service, 1999
- Managed a 12-person Customer Service team for 6 months

1996 – 1998

Ascend Communications, Inc. (now Lucent) **Alameda, CA**
200 people, develops internet remote access concentrators
Software Engineer, 1996–1998

- Responsible for implementation of Japanese mobile phone data service software, working directly with NTT
- Worked to develop, as a member of a small team, a competitor to Rockwell's 56K embedded concentrator modem chips

1995 (6 months)
1994 (3 months)
1993 (3 months)

Schlumberger Austin Research Center **Austin, TX**
300 people, refined tools used for oil exploration and extraction
Student Intern, 1 year

- Developed an adjustable mathematical model of the Schlumberger Digital Telemetry system (Master's thesis)

Computational Experience

Programming languages, statistical packages, and tools:

- Active: Python, R, Weka, SQLite, Subversion
- Previous: Java, Lisp, C, Analog Devices DSP assembly code, StrongARM assembly code, Matlab, Perl, PHP, Visual BASIC, SQL, CVS, ...

Operating systems: Mostly Windows. A little Unix, OS/X, VMS, embedded realtime OSs.

Professional Development & Memberships

American Society for Information Science and Technology (SIGUSE, SIGMETRICS, SIG/STI),
2008 – present

International Society for Computational Biology, 2004 – present

American Medical Informatics Association, 2004 – present

Attended multiple seminars on grant writing, public speaking, publishing, and teaching.

Actively learning and adopting test-driven development practices.

Personal Interests

Spending time with my husband and young daughter, traveling, long-distance cycling, walking, reading, brainstorming solutions, learning new things.
Canadian Citizen, USA Permanent Resident.