Identifying Data Sharing in Biomedical Literature

Heather A. Piwowar and Wendy W. Chapman
University of Pittsburgh, Pittsburgh, PA

Submitted to AMIA Annual Symposium 2008 (American Medical Informatics Association).

This extended abstract will be archived at Nature Precedings, March 2008.

Data from this study will be shared on our website.

For more information on our data sharing research please email or visit!

Abstract

Many policies and projects now encourage investigators to share their raw research data with other scientists. Unfortunately, it is difficult to measure the effectiveness of these initiatives because data can be shared in such a variety of mechanisms and locations. We propose a novel approach to finding shared datasets: using NLP techniques to identify declarations of dataset sharing within the full text of primary research articles. Using regular expression patterns and machine learning algorithms on open access biomedical literature, 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%). We believe our results demonstrate the feasibility of this approach and hope to inspire further study of dataset retrieval techniques and policy evaluation.

Introduction and Motivation

Reusing primary research data has many benefits for the progress of science. For example, new studies advance more quickly and inexpensively when duplicate data collection is reduced, rare conditions can often be explored only through combining several datasets, and new computational methods can be evaluated through re-analysis.

Recognizing the value of data reuse, many initiatives actively encourage investigators to make their raw data available for other researchers. The NIH recently passed a policy requiring data sharing from all genome-wide association studies, supplementing their general policy which requires a data sharing plan for all grants over $500,000. Journals often require data sharing as a condition of publication. Public databases provide a centralized home for many datatypes, while projects such as caBIG™ provide methods for sharing data within a federated architecture. Various organizations are working towards responsible data sharing; Science Commons is designing strategies and tools for increasing data sharing (http://sciencecommons.org), the Microarray and Gene Expression Data Society has generated standards to facilitate data exchange (http://www.mged.org), and an AMIA initiative is working towards a framework for responsible sharing and reuse of healthcare data.

There is a well known adage: you cannot manage what you do not measure. For those with a goal of promoting responsible data sharing, it would be helpful to evaluate the effectiveness of requirements, recommendations, and tools. When data sharing is voluntary, insights could be gained by learning which datasets are shared, on what topics, by whom, and in what locations. When policies make data sharing mandatory, monitoring is useful to understand compliance and unexpected consequences.

Unfortunately, it is difficult to monitor data sharing because data can be shared in so many different ways. Previous assessments of data sharing have included manual curation, investigator self-reporting, and the analysis of citations within database submission entries. These methods are only able to identify instances of data sharing and data withholding in a limited number of cases and contexts.

We propose an alternative approach: using natural language processing (NLP) techniques to identify declarations of dataset sharing within the full text of primary research articles. Although this approach will not identify all shared datasets, we hypothesize that it will identify links between full text and datasets beyond those in current databases and thus add value.

Method

We developed a pilot NLP application to identify references to data sharing in the biomedical literature and compared its predictive performance against a reference standard of bibliographic citations associated with dataset submissions. Below we describe which shared datasets our approach could potentially identify, the reference standard we compiled, regular expression and statistical algorithms we used to identify data sharing, and the evaluation we performed.

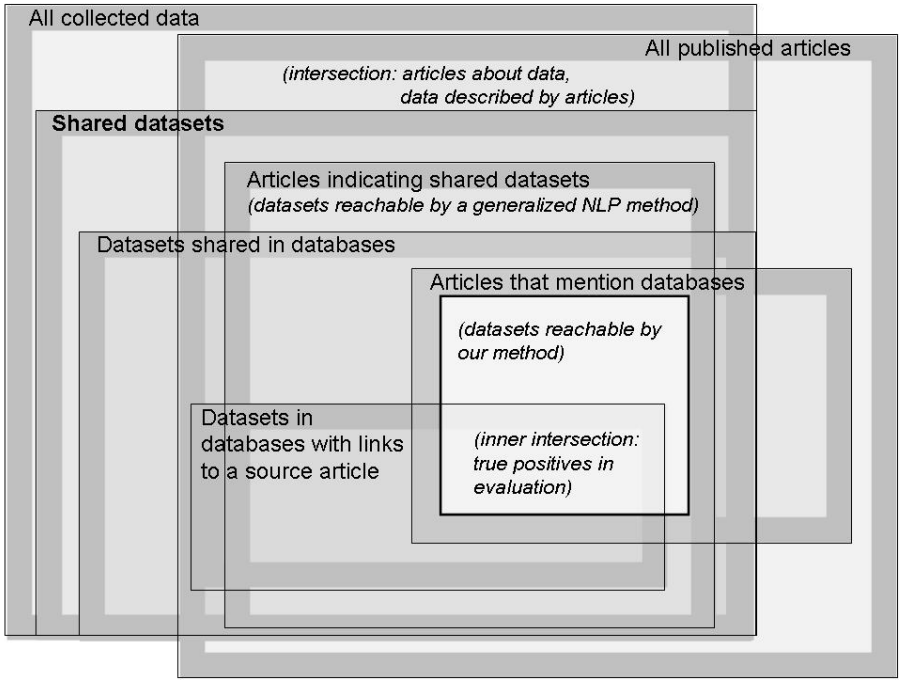

Figure 1. Venn diagram of conceptual relationships between collected data and published articles.

Potential Scope

Ideally we would like to identify all *Shared datasets* (refer to Figure 1 for an illustration of the italicized phrases), preferably linked to the source literature (*Shared datasets* intersected with *All published articles*). Today, this is usually approximated by searching for *Datasets in a database with links to a source article*.

We propose to identify a greater proportion of shared data by analyzing *Articles indicating shared datasets*. In this study we limited our search to the intersection with *Articles that mention databases*, but in theory our approach could be widened to *All published articles*. For this study, we started with all

Articles that mention databases. For each article, we applied several algorithms to predict whether or not the article is an *Article indicating a shared dataset* and compared this prediction against the reference standard described below.

Reference Standard

For each article in our literature cohort we assigned a reference standard classification specifying whether the article indicated that the investigators had deposited their primary data in one of five databases. The reference standard was generated in four stages:

(1) Downloaded literature: We used PubMed Central to download the full text of articles published in journals entitled "BMC*", "PLoS*" or "Nucleic Acids Research." The search resulted in 24,317 articles across 70 journals.

(2) Identified database submission links to literature: We investigated the sharing of three datatypes across five databases: nucleic acid sequences in Genbank, protein structures in the Protein Data Bank (PDB), and gene expression microarray data in Gene Expression Omnibus (GEO), ArrayExpress (AE) and the Stanford Microarray Database (SMD). From each database we extracted the PubMed IDs or bibliographic citations that were associated with dataset submissions. We classified a [articles, database] case as positive for data sharing when the article was associated with a database submission.

(3) Manually filtered articles for additional positive classifications: We anticipated that a portion of the articles without links from databases were nonetheless articles that mention sharing data in the databases. To estimate the prevalence of this occurrence and accurately evaluate our classification algorithms, we manually adjudicated the “sharing status” for 598 cases where articles were not referenced from the database but did match our precise lexical pattern filter, described below. Author HP examined these full-text phrases and reclassified 167 of the negative cases as positive (63 in the test set), as described in the results section.

(4) Selected articles within the scope of our method: Since not all articles that are linked from database submissions have a corresponding mention of the database submission within their full text, an NLP application operating on the articles cannot hope to achieve complete coverage in identifying the dataset submissions. We made the assumption that articles indicating shared datasets would include text about depositing that data in a database and explicitly include the name of the database. Based on that assumption, our subsequent analysis considered only those articles that included one or more occurrences of a database name.

Of the 24,317 articles, 6099 (25%) included at least one of the five database names somewhere within their full-text, including 1238 articles (5%) which mentioned two or more databases for a total of 7463 [article, database] cases. We randomly divided the cases into three subsets: a development set (4435 cases), a training set (2000), and a test set (1028).

NLP Algorithms for Identifying Data Sharing

We implemented two approaches for classifying articles as either containing or not containing text indicating a database submission: a set of regular expression patterns to identify relevant lexical cues and a machine learning approach.

Manually derived regular expression patterns: We manually examined articles in the development set and iteratively developed regular expression patterns to identify phrases that indicated data sharing. The patterns were applied in a 300-character window surrounding each occurrence of a database name within

the full-text articles. Multiple windows within an article, due to a repeated database name, were concatenated. Patterns included the single word “accession”, a regular expression for an accession number (specific for each database), a regular expression for a website URL, a set of lexical patterns for clauses and phrases, and a subset of these lexical patterns chosen to attain higher precision. The full regular expressions can be found at http://www.dbmi.pitt.edu/piwowar .

As an example of our lexical patterns, the regular expression “*accession*$_{.\{0,20\}}$*(for|at)*$_{.\{0,100\}}$*(is|are)”* matches the text *“The Gene Expression Omnibus* ***accession*** *number* ***for*** *the array sequence* ***is*** *GSE546”* from PubMed ID 261870. The GEO database contained a citation to this article within the entry for dataset GSE546. Thus, we considered the case [261870, GEO] a true positive when we evaluated this pattern. Unfortunately, the pattern also matches *“The Genbank* ***accession*** *numbers* ***for*** *the paralogs used in Figure 5* ***are*** *AvrB (P13835)*.” Since this article (PubMed ID: 1839166) did not generate any shared data (it is instead reusing and referencing data someone else had previously shared), [1839166, Genbank] was a false positive for this pattern.

A lexical pattern we chose to include in the precise list is *“(we|was|were|is|are|be|been|have|has) (accessioned|added|archived|assigned|deposited|entered|imported|included|inserted|loaded|lodged| placed|posted|provided|registered|reported_to|stored|submitted|uploaded_to)".*
This pattern matches the true positive sentence, *”Coordinates have* ***been deposited*** *with the Protein Data Bank under the accession code 2AVT.”* False positives also occur, but relatively infrequently.

Machine learning classifiers: We trained machine learning algorithms with three sets of features: binary (match/no match) lexical features from our manually derived patterns, a bag-of-words approach, and finally a combination of both sets of features. Twenty bag-of-word features were chosen using automatic feature selection on the 300-character window surrounding each database name occurrence (unstemmed, including stopwords and bigrams), then tuned by manual removal of 6 features specific to the datatype domains (i.e., “cdna_sequence.,” “of_protein”). We applied a variety of machine learning algorithms (trees, rules, Naïve Bayes, and support vector machines) and found similar performance; we report the results with J48 trees since it had the best performance and trees are transparent, portable, and easy to implement.

Evaluation Method

We calculated recall and precision for classifications assigned by the NLP applications when compared against reference standard classifications. Recall represents the proportion of positive [articles, database] cases that are classified as positive by the application. Precision represents the proportion of [article, database] cases classified as positive by the application that are truly positive. We used the NLTK toolkit version 0.9.1 in Python 2.5.1 for text processing, and Weka via TagHelper Tools for machine learning applications.

Results

The number of articles that mention the given database as a percentage of those known to have shared data within the database (i.e., their PubMed ID is listed within the database) varies from 47% for ArrayExpress to 95% for PDB (Table 1).

Database	Proportion of articles referenced from database that mention the database within the article full text
Genbank	86% (319/369)
PDB	95% (75/79)
GEO	81% (116/143)
ArrayExpress	47% (21/45)
SMD	89% (16/18)

Table 1. The proportion of articles with shared datasets that are within the scope of our algorithm.

Our manual filter for additional positive classifications identified more cases in some databases than others: we reclassified 19% of [article,database] cases from ArrayExpress as positive despite an omitted literature link, compared to 11%, 7%, 2%, and 1% for GEO, Genbank, PDB, and SMD respectively (see Table 2 for raw number of cases). The most common situations included: the database entry listed a citation for another paper by the same authors, the entry listed an erroneous PubMed ID, the entry included a citation without a PubMed ID, or the entry had a blank citation field.

Manually-derived regular expression patterns: The lexical cues that effectively identified articles with shared data varied across databases (Table 2).

	Overall	**Gen-bank**	**PDB**	**GEO**	**AE**	**SMD**
N	**1028**	505	347	104	29	43
Prevalence	**23%**	29%	9%	43%	41%	16%
The word "accession"						
Precision	**.31**	.40	.10	.84	.91	0
Recall	**.88**	.91	.97	.84	.83	0
<accession regular expression pattern>						
Precision	**.47**	.42	.38	.85	.91	1.00

Recall	**.74**	.85	.45	.64	.83	.14
<url regular expression pattern>						
Precision	**.34**	.35	.10	.59	.50	.46
Recall	**.40**	.30	.23	.64	.83	1.00
<lexical regular expression patterns>						
Precision	**.49**	.50	.26	.82	.61	.50
Recall	**.86**	.80	.81	.98	.92	.86
<precise lexical regular expression patterns>						
Precision	**.56**	.58	.31	.83	.85	.63
Recall	**.75**	.68	.81	.89	.92	.71

Table 2. Number of [article, database] cases (N), Prevalence of cases with data sharing, Precision and Recall of manually constructed regular expression patterns evaluated on the test set.

The word “accession,” for example, is included in most articles with shared data (recall 0.88), however it is also included in many articles that do not share data (precision 0.31), particularly in the largest and most established databases, Genbank and PDB.

Machine learning classifiers: Machine learning classification performance using matches to the manual regular expression patterns as features achieved much higher precision than any of the regular expression patterns did, at the expense of recall (Table 3). A classifier trained with the bag-of-words feature set had the highest precision (0.88), exceeding 0.85 for all databases. The classification tree that was learned from a combination of the regular expression cues and the bag-of-word cues had the highest overall recall (0.61) while maintaining high precision.

	Overall	**Gen-bank**	**PDB**	**GEO**	**AE**	**SMD**
ML with Regular expression features						
Precision	**.78**	.74	.68	.96	1.00	1.00
Recall	**.55**	.59	.42	.51	.83	.14
ML with Bag-of-words features						
Precision	**.88**	.85	.94	.95	.88	1.0
Recall	**.52**	.54	.55	.44	.58	.14
ML with Regular expression + Bag-of-words features						
Precision	**.80**	.76	.73	.94	1.00	0
Recall	**.61**	.62	.61	.69	.92	0

Table 3. Precision and Recall of machine learning (ML) applications evaluated on the test set.

The most precise classifier identified 59 cases of data sharing from articles not cited within databases, demonstrating one aspect of the potential value of this approach.

Discussion

Our results suggest it is possible to identify data sharing from the literature using relatively simple techniques. Machine learning methods achieved higher precision, whereas manually derived regular expression patterns showed higher recall. Acceptable precision/recall performance for this problem has not been established and will depend on use and context.

The descriptions of data sharing were surprisingly varied among databases. For example, articles that share data in Genbank almost always mention an accession number, while those that share data in SMD almost never do. This difference is likely due to journal policies which may explicitly require accession numbers from certain databases. These policies probably also contribute to the relatively low precision of accession numbers for identifying data sharing in established databases, since accession numbers are often mentioned in the context of data reuse and reanalysis as well as data sharing. Adding cues to identify data reuse would help improve the precision of the current classifier and also provide an interesting dataset for future study.

In addition to facilitating our primary goal of policy evaluation, broadly identifying shared data has other potential uses. A tool to help database curators populate citation fields would be valued, as demonstrated by the recent PDB data uniformity project. Investigators who wish to reuse data would benefit from retrieval mechanisms that are location-agnostic and allow queries based on MeSH terms, citations, or even article full text. Finally, broad identification of shared datasets would allow the reuse of datasets that are not easily found otherwise, and thus unleash the potential of these underutilized resources.

Our study has several limitations. Our dependence on database systems to provide a gold standard resulted in a database-centric classifier, involving cues such as “accession” and “deposited,” which are not necessarily applicable to sharing data on websites or in supplementary information. The method requires a set of database names, though perhaps a named-entity recognition system could be trained to eliminate this requirement. Our literature cohort included only open access articles; these authors may be more inclined to share data and could possibly discuss their shared datasets differently. The evaluation standard screening was performed by the system developer. Finally, the approach requires access to literature full-text.

To our knowledge, this is the first evaluation of a strategy for finding shared data and the first time NLP has been applied to detecting phrases of data sharing. Future work could apply related methods such as semi-supervised learning to derive lexical cues and automatically expanding a set of cue phrases through bootstrapping.

We are encouraged by the feasibility of identifying data sharing automatically from full text, and hope our approach reduces a barrier in evaluating and refining policies that encourage data sharing.

Data and code availability

Complete patterns, classifiers, datasets, and code will be posted at http://www.dbmi.pitt.edu/piwowar .

Acknowledgments

HP is supported by NLM training grant 5T15-LM007059-19 and WC is funded through NLM grant 1 R01LM009427-01.

References

1. 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.
2. Piwowar HA, Chapman WW. A review of journal policies for sharing research data. International Conference on Electronic Publishing 2008 [To appear]; Toronto, Canada.
3. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: An AMIA white paper. J Am Med Inform Assoc. 2007;14(1):1-9.
4. Noor MA, Zimmerman KJ, Teeter KC. Data sharing: How much doesn't get submitted to Genbank? PLoS biology. 2006;4(7):e228.
5. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS ONE. 2007;2(3):e308.
6. Blumenthal D, Campbell EG, Gokhale M, Yucel R, Clarridge B, Hilgartner S, et al. Data withholding in genetics and the other life sciences: prevalences and predictors. Academic medicine : journal of the Association of American Medical Colleges. 2006;81(2):137-45.
7. Loper E, Bird S. Nltk: The natural language toolkit. 2002. http://arxiv.org/abs/cs/0205028.
8. Witten I, Frank E. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann; 1999.
9. Donmez P, Rose C, Stegmann K, Weinberger A, Fischer F. Supporting CSCL with automatic corpus analysis technology. CSCL '05. International Society of the Learning Sciences; 2005. p. 125-34.
10. Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, et al. The PDB data uniformity project. Nucleic Acids Res. 2001;29(1):214-8.
11. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: First steps in translational bioinformatics. AMIA Annu Symp Proc. 2006:106-10.
12. Medlock B. Exploring hedge identification in biomedical literature. J Biomed Inform. 2008.
13. Abdalla R, Teufel S. A bootstrapping approach to unsupervised detection of cue phrase variants. Association for Computational Linguistics; 2006. p. 921-8.