

Linking database submissions to primary citations with PubMed Central

Heather A. Piwowar and Wendy W. Chapman

Department of Biomedical Informatics, University of Pittsburgh

Background: *Dataset submissions are growing exponentially. Links between dataset submissions and primary literature that describe the data collection are useful for many reasons: rich documentation, proper attribution, improved information retrieval, and enhanced text/data integration for analysis. Unfortunately, many database submissions do not include primary citation links, as database submissions are often made prior to publication. We suggest that automated tools can be developed to help identify links between dataset submissions and the primary literature. These tools require full text to differentiate cases of data sharing from data reuse and other contexts. In this study, we explore the possibility that deep analysis of full text may not be necessary, thereby enabling the querying of all reports in PubMed Central.*

Methods: *We trained machine learning tree and rule-based classifiers on full-text open-access article unigram vectors, with the existence of a primary citation link from NCBI's Gene Expression Omnibus (GEO) database submission records as the binary output class. We manually combined and simplified the classifier trees and rules to create a query compatible with the interface for PubMed Central.*

Results: *The query identified 40% of non-OA articles with dataset submission links from GEO (recall), and 65% of the returned articles without dataset submission links were manually judged to include statements of dataset deposit despite having no link from the database (applicable precision).*

Conclusion: *We hope this work inspires future enhancements, and highlights the opportunities for simple full-text queries in PubMed Central given the mandated influx of NIH-funded research reports.*

Introduction

The expected deluge of full-text biomedical research articles into PubMed Central (PMC), as mandated by recent NIH policy¹, creates many opportunities for improving research tools and processes. Most biomedical text mining and natural language processing (NLP) has been limited to titles and abstracts: these are available in abundance in PubMed. Analysis of machine-readable full text would permit a much

deeper and wider scope of study, but assembling a corpus has been hindered by the complex, disparate, and decentralized access processes and licenses of publisher websites. While PMC does not permit automated downloading of non-Open Access full text (as per publisher licenses), full text can be queried from the PMC interface. Integrating the ability to query the full text of all future NIH-funded research reports in combination with MeSH terms and other NCBI Entrez database links offers exciting possibilities.

In this report, we explore the potential of one such application: linking articles that describe data collection to their database submission entries. Databases that store research datasets often include citation links to the articles that describe the initial generation and use of the datasets. As we discuss below, these links are valuable, often missing, and time-consuming to manually derive. We previously developed several NLP systems to identify declarations of database submission within research articles², however these systems required access to complete full text for feature extraction. To take advantage of the PMC resource, here we develop a system restricted to rules that can be expressed within the PMC query interface.

We apply our system to gene expression microarray studies deposited in NCBI's Gene Expression Omnibus (GEO) database³. Gene expression data are expensive to collect, often but not always shared, and valuable for reuse. The GEO database is the largest repository for gene expression datasets, is well integrated with PMC query results, and contains links from submitted datasets to primary citation reports.

Methods

Our goal was to develop a PMC query for retrieving articles that mention depositing a dataset into GEO. We developed the query using a selection of Open Access (OA) articles, and evaluated it on non-OA articles.

We used a gold standard based on our previous work.² Positive cases came from two sources: all OA articles that were linked from the GEO DataSet primary submission field, plus articles without a primary citation link from the GEO

database that were nonetheless judged to have deposited data into GEO. Manual judgment for the selected OA articles was based on reviewing the full-text reports. Negative cases were considered those articles that were not linked from GEO Datasets and were manually classified as lacking any indication within their full text that they had deposited a dataset into GEO.

We used NCBI's Entrez E-Utilities, PubMed Central, Python, TagHelper Tools⁴, and Weka⁵ to remove rare words (<40 occurrences) and stopwords, create unigram bag-of-word vectors, automatically select features, and build tree (J48) and rule (PART) machine learning classifiers for a variety of parameter values. We manually derived a PMC-compatible query based on the most robust feature selection and classifier results.

Recall was calculated by determining what percentage of the non-OA (since OA was used in training) PMC articles with links to Gene Expression Datasets were found by the query. We evaluated applicable precision by manually reviewing the non-OA query hits for articles that are not currently linked to GEO datasets and determining whether they indeed included statements of dataset submission to GEO.

Finally, we compared the current count of NIH-funded, GEO-linked articles in PubMed to those currently within PMC to project the possible impact our query might have once all NIH-funded datasets are deposited in PMC.

Results

The training set was composed of open-access articles, including 550 positive examples (articles that had links from the GEO primary citation fields or were manually determined to have shared data in GEO) and 165 negatives (articles without links from GEO). We combined the rules and tree branches that occurred most frequently across the trained machine learning classifiers to compose the following PubMed Central query:

```
(geo OR omnibus)
AND microarray
AND "gene expression"
AND accession
NOT (databases
OR user OR users
OR (public AND accessed)
OR (downloaded AND published))
```

This query retrieved 772 articles, of which 455 were not open access. The results included 385 of the 966 PubMed Central non-OA articles with links to the GEO Datasets ("**pmc gds**"[filter] NOT "**open access**"[filter]), for a recall of 40%.

Next, we limited the query to non-OA articles without a PMC link to the GEO Datasets. We manually determined that 44 of the 68 results included a statement of dataset submission to GEO within their full-text report. This indicates an overall query precision of 94% (385+44/455) for retrieving articles that have deposited datasets into GEO and an applicable precision of 65% (44/68) for retrieving articles that don't have PMC links but should. Our error analysis of the 24 false positives found that 13 of the articles referenced GEO datasets in the context of dataset reuse rather than submission (including 2 reusing their own work), 4 referenced GEO in the context of platform descriptions rather than datasets, and 5 didn't reference the GEO database at all but rather mentioned the word "geo" for another purpose, usually the beta-geo gene.

The PubMed database contains 4291 articles with links from GEO DataSets (in PubMed: "**pubmed gds**"[filter]). Thus, the addition of an estimated 115 (177*65%) novel true positive links would increase the current number of dataset-submission-to-primary-citation links by about 2.6%.

We also estimated how the query impact might increase once new NIH-funded articles are deposited in PMC. PMC contains 202 articles published in 2007, funded by the NIH, and linked from GEO DataSets. In comparison, the PubMed database contains 596 such articles—almost three times as many. Our query returned 39 hits for NIH-funded articles published in 2007 that were *not* linked from GEO DataSets. If all NIH-funded articles were in PMC, and if similar patterns exist for microarray papers that share their data but do not currently have links from

the GEO database, we estimate our query could return roughly 117 (39*3) new articles per year identifying data sharing not included in primary citations, of which 76 (117*65%) might be true positives. This would increase the annual count of primary citation links by about 5.5% (1310 NIH and non-NIH PubMed articles with GEO Database links in 2007 + 76 projected additions).

The trivial query, "**gene expression omnibus**" **AND (submitted OR deposited)** resulted in a 34% recall and 90% overall precision.

Discussion

Database submissions often include a link to the research article that describes the original data collection conditions and interpretations. Our results suggest a simple query on full-text can automatically identify database submission primary citation links with a precision of 94% and recall of 40%. A trivial full-text query identified articles with 90% precision and 34% recall. Precision for the subset of articles without existing links from the GEO database was 65%. The methods we describe can be used to develop queries for identifying primary citations across a wide variety of datatypes and databases.

The approach outlined in this study is much more practical than a complex regular-expression classifier running on article full text. Processing full-text articles requires not only access licenses and reuse permissions (or a limitation to open access content) but also the maintenance of a text repository and classification system. Querying full text through PubMed Central, in contrast, is publicly available, requires no infrastructure beyond an internet connection, and covers all OA and non-OA articles within PMC.

We imagine this query could be used in two ways. It could be used by dataset-seekers, by appending it onto PubMed or PMC queries to find articles with shared datasets. Alternatively, it could be used by biocurators as a tool for identifying primary citations that may be missing from their database submission fields. This latter use has broad implications, which we discuss further below.

Links between shared datasets and primary citations have many purposes. First, the citation

serves as rich documentation for the dataset, whether as free text or as meta-data mark-up as illustrated by the BioLit PDB Clone (<http://biolit.ucsd.edu/pdb/>). Second, the citation provides a crucial mechanism for attributing recognition to the originators of the dataset upon reuse.⁶ Third, the citation provides a link for enhanced information retrieval or text/data integration pathways.^{7,8}

Unfortunately, links to primary citations are often missing from database submission entries because datasets are usually submitted before publication details are known.⁹ Evidence suggests that a significant number of links from database submissions to the primary literature may be missing. For example, the PDB data uniformity project of 2000 found that 33% of submission entries lacked a citation. Half of these were recovered automatically using the list of submitter names, 40% through manual searches of PubMed and the Thomson ISI databases, and 10% (3% of total) were presumed to represent work that was never published.¹⁰ More recently, another large-scale PDB remediation project looked at improving the quality of many fields, including primary citations. As of May 2005, 8508 (27%) of the 31663 database submissions required remediation due to inconsistent or missing PubMed IDs and citation information. A report near the end of the remediation process¹¹ estimated that manual searches found PubMed IDs for 1226 entries, citation information without PubMed IDs for 387 entries, and about 700 (2% of 31663) were presumed unpublished. These examples suggest that a sizeable number of entries may be missing citation fields, and that most of them are recoverable. Unfortunately, these efforts are time-intensive and thus difficult to incorporate into the workflow of busy biocurators.¹² NLP is already being used to aid database curation in a variety of tasks¹³, and we believe it can also help biocurators identify missing links to primary citations.

Procedurally, our query results could be manually confirmed and then used to update database records. GEO asks for omitted citations (<http://www.ncbi.nlm.nih.gov/geo/info/ucitations.html>); we have sent them our findings and they have updated their database to include the missing links identified in this study. Other databases, however, consider the submission record the property of the submitter¹⁴ and are

thus unlikely to add citations without permission. Perhaps in these situations an automated system could be developed to email submitters requesting they add or permit the addition of the citation.

The performance of our query could undoubtedly be improved through systematic refinement.¹⁵ Future work could involve deriving additional cues through bootstrapping and semi-supervised learning, including stemmed words with wildcards, and refining the query based on error analysis (for example, excluding hits on beta-geo). Additional improvements could be achieved if the PMC query capabilities were enhanced. For this application, it would be particularly useful to remove negation and modal verbs from the stop word list (<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43>) so they could be included within query n-grams.

Linking shared datasets to primary citations increases their value: the datasets become easier to find, easier to understand, easier to responsibly acknowledge, and easier to integrate with other information. Dataset deposits are growing exponentially. As NIH-funded research makes its way to PMC, the opportunity for creating links between datasets and full-text articles increases enormously. We hope this study provides a useful preliminary tool and inspires further research in this area.

Our manual annotation results are available at <http://www.dbmi.pitt.edu/piowar>.

Funding

National Library of Medicine (5T15-LM007059-19 to HAP, 1R01-LM009427-01 to WWC)

References

1. NOT-OD-08-033 Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research.
2. Piowar, H.A. & Chapman, W.W. Identifying Data Sharing in Biomedical Literature. Available from *Nature Precedings* <<http://hdl.handle.net/10101/npre.2008.1721.1>> (2008).
3. Barrett, T., et al. NCBI GEO: mining tens of millions of expression profiles--

4. Rose, C.P., et al. Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning, *International Journal of Computer Supported Collaborative Learning* (In Press).
5. Witten, I.H. & Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco (2005).
6. Compete, collaborate, compel. *Nat Genet* **39**(2007).
7. Butte, A.J. & Chen, R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA Annu Symp Proc*, 106-110 (2006).
8. Muller, H.M., Kenny, E.E. & Sternberg, P.W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* **2**(2004).
9. Piowar, H.A. & Chapman, W.W. A review of journal policies for sharing research data. Available from *Nature Precedings* <<http://hdl.handle.net/10101/npre.2008.1700.1>>, (2008).
10. Bhat, T.N., et al. The PDB data uniformity project. *Nucleic Acids Res* **29**, 214-218 (2001).
11. PDBj News Letter. in *Volume 7, March 2006* <http://www.pdbj.org/NewsLetter/newsletter_vol7_e.pdf> (2006).
12. Burkhardt, K., Schneider, B. & Ory, J. A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS Computational Biology* **2**(2006).
13. Karamanis, N., et al. Natural Language Processing in aid of FlyBase curators. *BMC Bioinformatics* **9**(2008).
14. Pennisi, E. DNA DATA: Proposal to 'Wikify' GenBank Meets Stiff Resistance. *Science* **319**, 1598-1599 (2008).
15. Zhang, L., Ajiferuke, I. & Sampson, M. Optimizing search strategies to identify randomized controlled trials in MEDLINE. *BMC Medical Research Methodology* **6**, 23 (2006).