Search Datasets in Literature: A Case Study of GWAS

Xiao Dong, M.D., M.S., Yaoyun Zhang, Ph.D., Hua Xu, Ph.D.

1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

One of the missions of the NIH BD2K (Big Data to Knowledge) initiative is to make data discoverable and promote the re-use of existing datasets. Our ultimate goal is to develop a scalable approach that can automatically scan millions of scientific publications and identify underlying data sets. Using Genome-Wide Association Studies (GWAS) as a use case, we conducted an initial study to identify GWAS dataset attributes in MEDLINE abstracts, by developing a hybrid approach that combines domain dictionaries and pattern-based rules. The automatic GWAS dataset attribute recognition system achieved an F-measure of 84.85%. We further applied the GWAS attribute recognition system to indexing MEDLINE abstracts and built an online GWAS dataset search engine called "GWAS Dataset Finder". Our evaluation showed that the GWAS Dataset Finder outperformed PubMed significantly in retrieving literature with desired datasets. Our study demonstrates the potential application of text mining methods in building the data discovery index. It can create a better index of literature linked with their underlying data sets, thus improving data discoverability.

Introduction

The global biomedical research community generates large volumes of digital data of different kinds, of different modalities, and in different formats. For researchers to take full advantage of existing research data, they must be findable, accessible, interoperable and reusable.

Current efforts to fulfill this demand mainly focus on indexing and linking existing datasets in multiple repositories. For example, the bioCADDIE project, as part of the NIH BD2K Initiative developed a data index and search engine, based on dataset metadata extracted from large biomedical repositories. [1] BD2K Aztec builds a global biomedical resource discovery index that allows users to simultaneously search a diverse array of tools.[2] The Omics Discovery Index (OomicsDI) provides dataset discovery across a heterogeneous, distributed group of Transcriptomics, Genomics, Proteomics and Metabolomics data resources from eight repositories.[3] ELIXIR archives large amounts of biological data produced by life science experiments in Europe.[4]

One frequent scenario for dataset retrieval is via searching biomedical literature. For example, a user may want to find any published studies containing GWAS (Genome-wide Association Studies) datasets for breast cancer with sample size over 1000. Current literature search engines such as PubMed mainly use keywords as queries, which do not provide the function of search dataset-specific information such as the size of a GWAS dataset. Although search engines in current data index consortiums provide a function to show the linkage between datasets and literature, most of the linkage information is directly imported from data repositories or only linked to open access scientific literature.[1, 2, 4, 5] Therefore, the search scope is very limited to give a comprehensive coverage of literature that generates and in secondary use of the research data. To build a general literature search engine for dataset, a
scalable approach that can automatically scan millions of scientific publications and identify underlying data sets would be highly desirable, due to the exponential growth of biomedical publications.

Our ultimate goal is to develop a general literature search engine covering a broad range of biomedical datasets. This study takes the initial step to validate the feasibility of automatic extraction of dataset mentions from literature and search relevant datasets based on extracted metadata information of datasets. Specifically, we take GWAS datasets as a use case. Targeting the associations between genetic polymorphisms and human phenotypes/traits (e.g., diseases), GWAS has a wide range of biomedical applications (e.g., targeted drug and diagnostics development [6]) and an exponential increase of related publications in recent years. [7] Moreover, the unified case-control study design [8] makes it a typical and general use case for the initial exploration of our study. The datasets used for GWAS usually contain population-level phenotype information, and the experimental results output from GWAS usually contain genotype-phenotype associations (i.e., SNP-traits associations).[9] However, current data repositories such as GWAS Catalog [7] mainly focus on providing search functions for traits and study results (i.e., SNPs, p-value) extracted from published GWAS papers, instead of information about the original cohort datasets used in study. [10] Moreover, although GWAS Catalog incorporates automatic trait and SNP extraction into the curation pipeline, it still relies on manual curation by domain experts, which is time-consuming and labor-intensive. Pilot works of linking literature with genomic-level information, such as the GWAS Integrator[11] and Phenotype–Genotype Integrator (PheGenI) [12], are also built upon the manually curated GWAS Catalog.

To build a literature search engine for datasets used in GWAS research, this study has two aims: 1) to automatically recognize GWAS dataset attributes mentioned in MEDLINE abstracts; and 2) to build a literature search engine with dataset attribute index, so that users could use diverse metadata combinations as queries to find datasets of interest. Our study is the first study trying to automatically index GWAS literature with the attributes of the underlying population-level datasets. Experimental results demonstrate that our automatic attribute recognition method and the search engine based on it are promising.

**Figure 1.** Workflow for GWAS Dataset Finder development.
Methods

The workflow of our study includes the following steps: (1) data collection and corpus annotation; (2) automatic dataset attributes recognition; (3) MEDLINE indexing; and (4) literature search. (Figure 1)

Corpus Annotation

Data sources

A total of 12,883 GWAS abstracts were retrieved by PubMed using the Mesh Term ‘Genome-wide association study’ as the query (by 05/2015). In addition, the 2,079 Medline abstracts (2005/03-2014/07) manually collected in GWAS Catalog were also collected (including 363 abstracts without the Mesh Term of ‘Genome-wide association study’). Finally, we built a repository of 13,196 GWAS abstracts.

Dataset attributes annotation

As mentioned previously, this study focused on the recognition of datasets used for GWAS research, which will be indexed for literature search. Therefore, the annotated attributes (metadata) should be most representative of the datasets. Besides, the study design of GWAS research should also be taken account. After a manual analysis of the GWAS abstracts and metadata in dbGaP ([http://www.ncbi.nlm.nih.gov/gap](http://www.ncbi.nlm.nih.gov/gap)), we identified seven attributes for the datasets: traits (phenotypes) of the cohort, ethnicity of the cohort, source of the dataset, study stage, case size and control size, and for the aim to build the literature search engine, the genotyping platform was also adopted in the study (Table 1).

Table 1. Examples of dataset attributes in GWAS literature

| Attributes       | Example(s)                                                                 |
|------------------|---------------------------------------------------------------------------|
| trait            | Genome-wide SNP analysis of [lung cancer](http://www.ncbi.nlm.nih.gov/gap) risk detects ... |
| ethnicity        | A [Norwegian](http://www.ncbi.nlm.nih.gov/gap) population...; The [Caucasian](http://www.ncbi.nlm.nih.gov/gap) cohort... |
| source           | ... participants from the [Framingham Heart Study (FHS)](http://www.ncbi.nlm.nih.gov/gap) ... |
| stage            | A [replication](http://www.ncbi.nlm.nih.gov/gap) series of ...; The [phase 1](http://www.ncbi.nlm.nih.gov/gap) study contains ... |
| case             | a population of [338](http://www.ncbi.nlm.nih.gov/gap) lung cancer cases ... |
| control          | ... and [335](http://www.ncbi.nlm.nih.gov/gap) controls for the study... |

The attributes of datasets were manually annotated in 300 GWAS abstracts from PubMed. Especially, since most of the traits are mentioned in the title or the first sentence of the abstract, we limited our annotation of trait in these two positions in order to reduce redundancy. If a trait could not be found from these two locations, we then further explored the rest of the text. Figure 2 illustrates an example of our attributes in one GWAS abstract.
Figure 2. Examples of dataset attribute annotations in GWAS abstracts.

**Automatic dataset attribute recognition**

*Dictionary-based attribute recognition:* mentions of trait, ethnicity, platform and source were recognized by fuzzy matching to terms in existing domain ontologies/dictionaries. Specifically, the trait ontology created by GWAS Catalog, disease ontology and the experimental factor ontology developed by EBI was used for trait recognition. [7, 12, 13] Mentions of ethnicities were recognized using the ethnic group of SNOMED-CT. [14] In addition, UMLS and the trait and ethnicity lists collected by UCSD, based on GWAS entries in dbGAP were also employed. [15]

The list of existing genotyping platforms, i.e., Affymetrix, Illumina, Perlegen, Sequenom were used to match the employed platforms. [10] As for sources of datasets, lists of cohort/study names (e.g., British 1958 Birth Cohort, Cardiovascular Health Study) were collected from different resources, such as NCI Cohort Consortium and Asia Cohort Consortium. [16, 17]

*Pattern-based attribute recognition:* mentions of sizes and stages were recognized by keywords and patterns of surrounding contexts. Since multiple numbers in the abstracts could be considered as candidates for size attributes, context information was necessary to accurately recognize case and control sizes. Taking the sentence in Figure 2 as an example, the control size is relatively straightforward to recognize, while the recognition of case size needs to take account its co-occurrence with trait and control size in the same sentence. Furthermore, the candidate numbers were constraint in the study design description section, excluding the experimental results and discussion section using keywords like “Experiment:” wherever possible. Similarly, for stages, some common words such as “further” could also indicate a study stage (replicate stage in this case). Therefore, the context of such keywords should also be considered.

**Literature search engine for dataset**

To build a search engine for dataset retrieval from literature, the recognized dataset attributes were used to build index for GWAS abstracts. Accordingly, our search engine (GWAS Dataset
Finder) accepts a combination of dataset attributes as the queries. Moreover, synonyms provided in related ontologies [12-14] were used to expand the trait and ethnicity, so that to increase the recall of our system. The prototype of GWAS Dataset Finder is integrated into BioCADDIE[1] and can be accessed at https://datamed.org/gwas/gwas_index.php.

Evaluation

**Dataset attribute recognition:** The 300 GWAS abstracts were randomly split into two sets. 60% of the abstracts were used for training the automatic extraction system; 40% were used for test. Precision, recall and F-score were used for dataset attribute recognition. Especially, since trait, platform and ethnicity could have multiple mentions across the whole abstract, and different expressions may be used referring to the same trait and ethnicity (e.g., Italian, Italy), we used an abstract-level evaluation for these three attributes. In another word, it’s not necessary to recognize every mention of these attributes in the abstract.

**Table 3.** Queries used for performance evaluation of GWAS Finder and PubMed.

| Query | Trait                  | Trait + Platform      | Trait + Platform + Case size |
|-------|------------------------|-----------------------|------------------------------|
| 1     | lung cancer            | lung cancer+illumina   | lung cancer+illumina+case size>1000 |
| 2     | type 2 diabetes        | type 2 diabetes+illumina | type 2 diabetes+illumina+case size >1000 |
| 3     | ankylosing spondylitis | ankylosing spondylitis+illumina | Ankylosing spondylitis+illumina+case size>1000 |
| 4     | Alzheimer's disease    | Alzheimer's disease+illumina | Alzheimer's disease+illumina+case size>1000 |
| 5     | multiple sclerosis     | multiple sclerosis+illumina | multiple sclerosis+illumina+case size>1000 |
| 6     | schizophrenia          | schizophrenia+Affymetrix | schizophrenia+Affymetrix+case size >1000 |
| 7     | Crohn's disease        | Crohn's disease+Affymetrix | Crohn's disease+Affymetrix+case size>1000 |
| 8     | smoking                | smoking+Affymetrix     | smoking+Affymetrix+case size>100(no results for >1000) |
| 9     | Prostate cancer        | prostate cancer+Affymetrix | prostate cancer+Affymetrix+case size>100(no results for >1000) |
| 10    | blood pressure         | blood pressure+Affymetrix | blood pressure+Affymetrix+case size>1000 |

**Dataset retrieval:** We evaluated the performance of GWAS Dataset Finder with three types of queries: ‘trait’, ‘trait+ platform’ and ‘trait+ platform+ case size’. For each type, 10 queries were used for evaluation. Average precision, recall and F-score were used as the evaluation criteria. Furthermore, to validate the utility and advantage of GWAS Dataset Finder, a systematic comparison was made between it and PubMed, which is the mainstream literature search
engine. Since PubMed does not have the function to search for a range of number, so we only performed two types of queries, i.e., ‘trait’ and ‘trait+ platform’. For a fair comparison, we use the MeSH term “Genome-wide association study” to limit the search results to GWAS papers in PubMed, which was the same dataset used for GWAS Finder.

Results

Dataset attribute distribution

The distribution of each annotated attribute is illustrated in Figure 3. As can be seen, trait and ethnicity have the highest frequencies, due to multiple times of mentions in the abstracts. On the other hand, the frequencies of platform, dataset source and study stage are among the lowest, which indicate that only using abstracts may not be sufficient for obtaining such types of attributes, full-text articles need to be explored instead.

![Figure 3. Distribution of dataset attributes in GWAS abstracts.](image)

Dataset attribute recognition performance

The performance of dataset attribute recognition was illustrated in Table 4. The performance of trait and platform was relatively high (F-score 0.910 and 1.000), while the ethnicity recognition only yielded a F-score of 0.792. Among all the attributes, source got the lowest F-score of 0.564, probably due to the low coverage of the current cohort/study lists we employed. The overall macro F-score of the whole attribute set is 84.85%.

Dataset retrieval performance

The performance results of GWAS Dataset Finder and PubMed are shown in Table 5. We can see that GWAS Dataset Finder achieved much higher precision in finding intended GWAS literature than PubMed. Despite that the two systems yielded comparative F-score using ‘Trait’ as query, our system outperformed PubMed significantly in each evaluation criterion when using “Trait+platform” as the query.
Table 4. Recognition performance of dataset attributes in GWAS abstracts (%).

|          | Trait | Platform | Ethnicity | Case Size | Control Size | Source | Stage |
|----------|-------|----------|-----------|-----------|--------------|--------|-------|
| Precision | 91.2  | 100.0    | 76.3      | 82.9      | 83.4         | 61.1   | 75.0  |
| Recall   | 90.8  | 100.0    | 82.2      | 80.7      | 85.2         | 52.4   | 88.0  |
| F-score  | 91.0  | 100.0    | 79.2      | 81.8      | 84.3         | 56.4   | 81.0  |

Table 5. Performance comparison between GWAS Dataset Finder and PubMed (%).

| Query type                  | GWAS Dataset Finder | PubMed |
|------------------------------|---------------------|--------|
|                             | P       | R       | F-score | P       | R       | F-score |
| Trait                       | 83.29   | 75.80   | 79.37   | 65.45   | 97.47   | 78.32   |
| Trait+platform              | 83.07   | 89.42   | 86.13   | 66.47   | 86.06   | 75.00   |
| Trait + platform + case size| 92.71   | 75.00   | 82.92   | #       | #       | #       |

Prototype of GWAS Dataset Finder

The prototype version of GWAS Dataset Finder provides the function to search GWAS literature by trait, platform and the minimum case size. These three attributes are chosen based on their performance of automatic extraction and their significance in GWAS. An example of search results of GWAS Dataset Finder is shown below, with the query “breast cancer + illumina + 1000” (Figure 4). The returned results should have “breast cancer” as trait, “illumina” as platform and the case size beyond 1000. GWAS Dataset Finder explicitly located the studies with the trait and platform of our interest, which also have the case size larger than the provided threshold.

Discussion

This pilot study is the first attempt to automatically index GWAS literature with the attributes of the datasets used for research. The automatic GWAS dataset attribute recognition system achieved an overall F-measure of 84.85%. The performance of GWAS Dataset Finder suggests that it is feasible to index the publications with the dataset attributes and thus to establish a first step to link the literatures with the underlying datasets. Our evaluation showed that the GWAS Dataset Finder outperformed PubMed significantly in retrieving literature with desired datasets. More importantly, as a valuable complementary to existing search tools,[7, 10] our study takes the initiative to find the datasets “in used”, other than the datasets “resulted in” from the literature, which facilitates the reproducibility of biomedical studies.

As expected, GWAS Dataset Finder outperformed PubMed significantly in retrieving literature with desired datasets. However, it is found that the more common traits we used for the query, the more likely that both the search engines returned false positive results. For example, with the query ‘smoking + Affymetrix’, both search engines returned the same false positive abstract, which contained a phrase of “with no significant differences in smoking history”. This problem
could be resolved in the attribute recognition stage, by excluding attributes within negation assertions.

Figure 4. An example of the search result of GWAS Dataset Finder.

Despite the evident strengths of our study, it has several limitations. Firstly, there is still space for improvement of the attribute recognition method. The current interface of GWAS Dataset Finder only contains three out of the six attributes as the query fields, which limited our evaluation strategy and further enhancements are needed. Besides, automatic extraction method needs to be explored for full-text articles the next step, so that to obtain more complete, especially for attributes of platforms and dataset sources. Another limitation for this pilot project is that we only took the GWAS as the use case. To achieve our final goal to develop a general literature search tool for datasets in the biomedical domain, more data types need to be explored.
Furthermore, biomedical data are very comprehensive. Even for a specific type like GWAS, there are many different ways to describe attributes of GWAS datasets. In addition to developing automated methods to extract datasets and their attributes from literature, a complementary method could be to develop systems to standardize the submission of biomedical datasets. There are different ways for users to search for biomedical datasets. Searching literature for biomedical datasets, which is the aim of the present study, is an important aspect of data discovery process. More work should be devoted towards this direction.

**Conclusion**

Our study is the first study indexing GWAS literature using the corresponding dataset attributes, which demonstrates the potential application of text mining methods in building the data discovery index. It can create a better index of literature linked with their underlying datasets, thus improving data discoverability.

**Acknowledgements**

Funding: National Institutes of Health (NIH) through the NIH Big Data to Knowledge, Grant 1U24AI117966-01. We would like to thank the bioCADDIE team for working with us to implement GWAS Dataset Finder at DataMed.org.

**References**

1. *bioCADDIE | biomedical and healthCAre Data Discovery and Indexing Ecosystem*. Available from: https://biocaddie.org/.
2. AZTEC. Available from: http://aztec.bio/home/overview.
3. Perez-Riverol, Y., et al., *Ommics Discovery Index-Discovering and Linking Public Omics Datasets*. bioRxiv, 2016: p. 049205.
4. ELIXIR Data for life. Available from: https://www.elixir-europe.org/.
5. Perez-Riverol, Y., et al., *Ommics Discovery Index-Discovering and Linking Public Omics Datasets*. bioRxiv, 2016: p. 049205.
6. Iadonato, S.P. and M.G. Katze, *Genomics: Hepatitis C virus gets personal*. Nature, 2009. 461(7262): p. 357-358.
7. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations*. Nucleic Acids Res, 2014. 42(Database issue): p. D1001-6.
8. Manolio, T.A., *Cohort studies and the genetics of complex disease*. Nat Genet, 2009. 41(1): p. 5-6.
9. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease*. New England Journal of Medicine, 2010. 363(2): p. 166–176.
10. Leslie, R., C.J. O'Donnell, and A.D. Johnson, *GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database*. Bioinformatics, 2014. 30(12): p. i185-94.
11. Yu, W., et al., *GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies*. Eur J Hum Genet, 2011. 19(10): p. 1095-9.
12. Schriml, L.M., et al., *Disease Ontology: a backbone for disease semantic integration*. Nucleic Acids Research, 2012. 40(D1): p. D940-D946.
13. Malone, J., et al. *Developing an application focused experimental factor ontology: embracing the OBO Community*. in Proceedings of ISMB 2008 SIG meeting on Bio-ontologies. 2008.
14. Bodenreider, O., et al., *Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies*. Artificial intelligence in medicine, 2007. 39(3): p. 183–195.
15. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes*. Nature genetics, 2007. 39(10): p. 1181–1186.
16. NCI Cohort Consortium. Available from: http://epi.grants.cancer.gov/Consortia/cohort.html.
17. *The Asia Cohort Consortium.* Available from: https://www.asiacohort.org/about/index.html.