Dug: A Semantic Search Engine Leveraging Peer-Reviewed Literature to Span Biomedical Data Repositories

Alexander M. Waldrop[1]*, John B. Cheadle[2], Kira Bradford[3], Nathan Braswell[1], Matt Watson[3], Andrew Crerar[1], Chris M. Ball[2], Yaphet Kebede[3], Carl Schreep[3], PJ Linebaugh[3], Hannah Hiles[3], Rebecca Boyles[2], Chris Bizon[3], Ashok Krishnamurthy[3,4], Steve Cox[3]*

[1]Center for Genomics, Bioinformatics, and Translational Research, RTI International, Research Triangle Park, NC, USA 27709-2194

[2]Research Computing Division, RTI International, Research Triangle Park, NC, USA 27709-2194

[3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA 27599-7568

[4]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA 27599-3175

*Corresponding authors: awaldrop@rti.org and scox@renci.org

Abstract

Motivation: As the number of public data resources continues to proliferate, identifying relevant datasets across heterogenous repositories is becoming critical to answering scientific questions. To help researchers navigate this data landscape, we developed Dug: a semantic search tool for biomedical datasets that utilizes evidence-based relationships from curated knowledge graphs to find relevant datasets and explain *why* those results are returned.

Results: Developed through the National Heart, Lung, and Blood Institute's (NHLBI) BioData Catalyst ecosystem, Dug can index more than 15,911 study variables from public datasets in just over 39 minutes. On a manually curated search dataset, Dug's mean recall (total relevant results/total results) of 0.79 outperformed default Elasticsearch's mean recall of 0.76. When using synonyms or related concepts as search queries, Dug's (0.28) far outperforms Elasticsearch (0.1) in terms of mean recall.

Availability and Implementation: Dug is freely available at https://github.com/helxplatform/dug. An example Dug deployment is also available for use at https://helx.renci.org/ui.

Contact: awaldrop@rti.org or scox@renci.org

Introduction

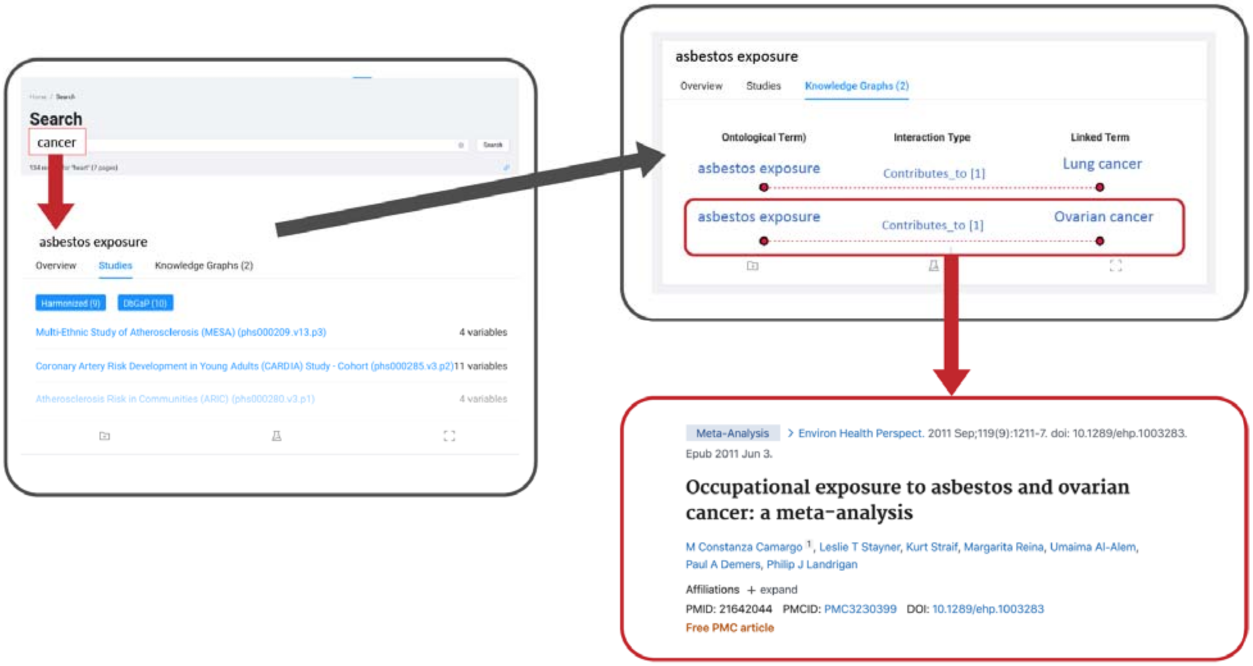
The ability to interrogate large-scale data resources is becoming a central focus of many research efforts. In past decades, the U.S. National Institutes of Health (NIH) and other public funding agencies have supported data generation at unprecedented scales through projects such as Trans-Omics for Precision Medicine (TOPMed) (University of Washington Department of Biostatistics, 2020), All of Us (The “All of Us” Research Program, 2019), and Helping to End Addiction Long-Term (HEAL) (Collins *et al.*, 2018). As a result of these continued data generation efforts, the ability to integrate data within and across often disjoint and complex public data repositories is quickly replacing data scarcity as a primary bottleneck to research progress. While successful data integration efforts have resulted in novel diagnostics, therapies, and prevention strategies, researchers often lack even the most basic tools for navigating this complex data landscape (Powell, 2021).

In particular, there is a growing need for comprehensive search tools that can identify datasets relevant to a researcher's particular scientific question. Despite recent NIH emphasis on making research data more Findable, Accessible, Interoperable, and Re-Usable (“FAIR” data principles) (Wilkinson *et al.*, 2016), the diversity of public data repositories has proven to be a formidable barrier to developing intelligent search strategies. To illustrate this heterogeneity, consider that the NIH alone currently refers data submission to more than 95 domain-specific repositories (NIH Data Sharing Resources, 2020). In many cases, even the more established repositories like the NIH Database of Genotypes and Phenotypes (dbGaP) often only require studies to submit free-text descriptions of experimental variables. One does not have to think of more complex examples than “gender” vs. “sex” or “heart attack” vs. “myocardial infarction” to understand the challenges of identifying relevant datasets among the massive, growing corpus of non-standardized biomedical datasets.

In the absence of widespread adoption of metadata standards, emerging techniques for Natural Language Processing (NLP) are increasingly enabling semantic search over biomedical datasets. Here, we define semantic search as search that considers the intent and context of the query as opposed to a purely lexical approach (Tran *et al.*, 2007). A number of existing tools successfully employ methods for named entity recognition to annotate free text with synonyms or similar ontology terms (Bell *et al.*, 2019; Chen *et al.*, 2018; Canakoglu *et al.*, 2019; Huang *et al.*, 2016; Laulederkind *et al.*, 2012; Pang *et al.*, 2015; Soto *et al.*, 2019). As an example, these tools might annotate a variable called “myocardial infarction” with synonyms like “heart attack” and “cardiac episode” so that any of these search queries would return the underlying dataset.

Despite the utility of these tools, there remains a need for a truly context-aware search tool that recognizes higher-order and potentially more interesting connections between datasets. Consider a researcher seeking datasets related to cancer across a set of repositories with which she is unfamiliar. Obviously, she would expect her results to include datasets that explicitly contain words like ‘cancer,’ ‘carcinoma,’ or ‘sarcoma.’ But what if she was specifically interested in lung cancer and it could show her datasets that measured smoking behavior? Or asbestos exposure? And what if it could even explain exactly *why* it was returning these more speculative results? By expanding our conception of what constitutes a *relevant* result, we can show researchers biological connections that, while not explicitly searched for, might be useful for hypothesis generation or scientific support.

Toward that end, we present Dug (https://github.com/helxplatform/dug): a semantic search tool for biomedical datasets that leverages ontological knowledge graphs to intelligently suggest relevant connections derived from peer-reviewed research. Given a search term, Dug returns lexical matches, semantically equivalent terms, and biologically relevant terms based on connections in curated knowledge graphs. As shown in Fig. 1, Dug is also the first biomedical search engine that can explain *why* it returns what it returns. Here, we discuss Dug's motivations, architecture, functionality, and evaluation, as well as demonstrate its successful deployment in the NHLBI's BioData Catalyst Ecosystem (National Heart Lung and Blood Institute *et al.*, 2020).



asbestos exposure

Multi-Ethnic Study of Atherosclerosis (MESA) (phs000209.v13.p3)	4 variables
Coronary Artery Risk Development in Young Adults (CARDIA) Study - Cohort (phs000285.v3.p2)	11 variables
Atherosclerosis Risk in Communities (ARIC) (phs000280.v3.p1)	4 variables

asbestos exposure

Ontological Term)	Interaction Type	Linked Term
asbestos exposure	Contributes_to [1]	Lung cancer
asbestos exposure	Contributes_to [1]	Ovarian cancer

Meta-Analysis > Environ Health Perspect. 2011 Sep;119(9):1211-7. doi: 10.1289/ehp.1003283. Epub 2011 Jun 3.

Occupational exposure to asbestos and ovarian cancer: a meta-analysis

M Constanza Camargo [1], Leslie T Stayner, Kurt Straif, Margarita Reina, Umaima Al-Alem, Paul A Demers, Philip J Landrigan

Affiliations + expand

PMID: 21642044 PMCID: PMC3230399 DOI: 10.1289/ehp.1003283

Free PMC article

Fig 1: The Dug web portal leverages knowledge graph connections with supporting links to PubMed literature to explain why certain results are relevant to a user's query.

Implementation/Methods

Returning to our researcher above, Dug is designed to find relevant datasets for a query like “lung cancer,” and to allow users to discover datasets they wouldn’t have found using lexical or even synonym-based search engines. In this section, we discuss the computational architecture underpinning this functionality.

Dug consists of two primary components: A **Dug API service** that orchestrates metadata ingestion, indexing, and search, and the **Dug search web portal** for displaying results to end users (Figure 2).

Briefly, the ingestion/indexing pipeline is designed to:

1. Parse heterogenous study metadata formats into a common Dug metadata format. For our researcher, this might be reading in clinical trial datasets from ClinicalTrials.gov or clinical data from COPDGene in dbGaP.

2. Annotate free-text descriptions of study variables with a set of ontological identifiers. For example, Dug might annotate a study variable from one of the datasets above called “cigarette usage” with an ontology term for “smoking behavior”

3. Expand resulting annotations with relevant terms returned from knowledge-graph queries (e.g., “What chemical entities are risk factors for lung cancer?”). In the above example,

we might discover a knowledge graph connection between "smoking behavior" and an ontology term for "lung cancer"

4. Index each study variable, its associated ontological concepts, and the set of knowledge graph answers to an Elasticsearch (Kuć and Rogozinski, 2016) endpoint. Putting it all together, our researcher will now be able to both find relevant datasets to her initial query, as well as explore and discover related datasets.

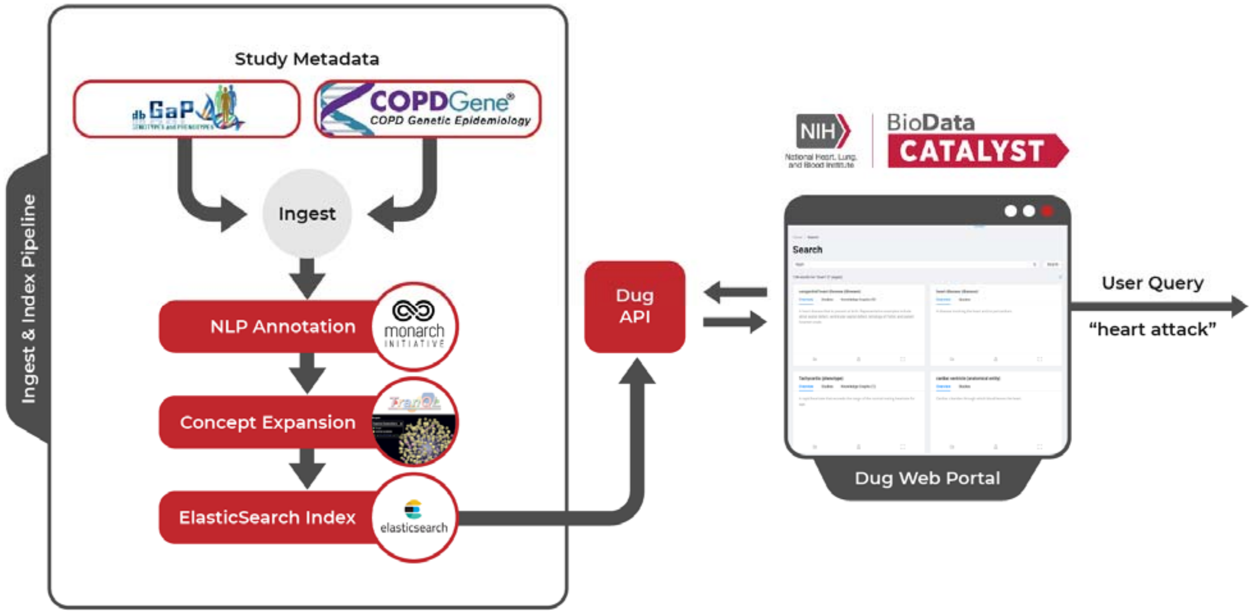


Fig. 2: Dug makes study metadata searchable by parsing heterogenous metadata formats into a common format (ingest), annotating metadata using NLP tools to extract ontology identifiers from prose text (NLP annotation), searching for relevant connections in federated knowledge graphs using Translator Query Language (TranQL) (concept expansion), and finally indexing all this information into an Elasticsearch index. Dug's web portal utilizes a flexible API to query and display search results back to end users.

Finally, the Dug search portal is a web-based application that sits on top of the Dug API and displays the results of user queries, renders auditable knowledge graphs for certain results, and organizes search results by underlying data type (e.g., “genotype” vs “DICOM image”). Below, we discuss each of these components in greater detail.

Ingestion and Indexing Pipeline

Data ingestion

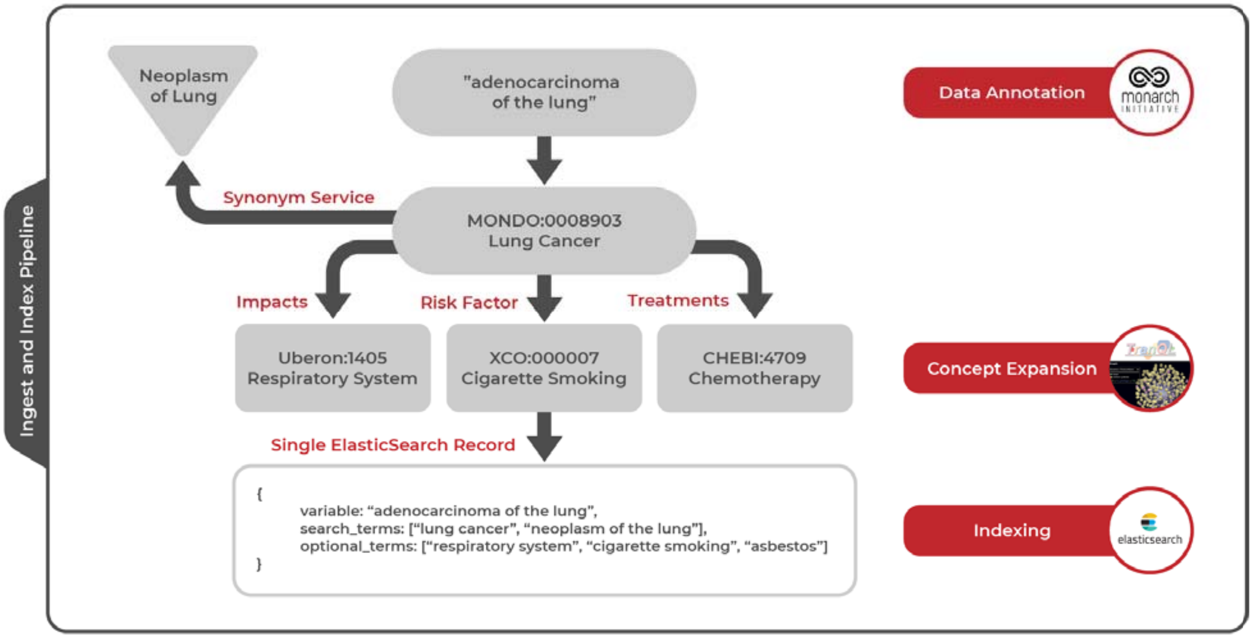
Researchers need insight into studies prior to applying for data use. To accommodate this, Dug ingests and indexes study *metadata* (e.g., text descriptions of study variables, descriptions of clinical images) as opposed to the actual study data, which may be controlled access (e.g., genotypes, clinical phenotypes). To accommodate the diversity of metadata formats available across public data repositories, our ingestion pipeline abstracts out retrieval modes (e.g., local file, network file, FTP, API) and data parsing formats (e.g., dbGaP data dictionary, XML, JSON). Similar to the Data Tags Suite (DATS) metadata schema (Sansone *et al.*, 2017), Dug parses diverse metadata formats into a common *DugElement* metadata model, which defines a standard set of metadata required for indexing (e.g., variable name, description, study/collection name, study description). Dug can be flexibly adapted to ingest nearly any metadata format by extending its plug-in interface to implement a single function parsing input data into *DugElement* objects.

Data Annotation

As shown in Fig 3, the purpose of Dug's data annotation module is to extract a set of biomedical ontology identifiers from ingested metadata elements using tools for named entity recognition. For most instances of Dug, we prefer the functionality of the /nlp/annotate endpoint exposed by the Monarch Initiative's (Mungall *et al.*, 2016) Biolink API (https://github.com/biolink/biolink-api). For context, the underlying Biolink model (https://biolink.github.io/biolink-model) provides a high-level data model for representing biomedical knowledge (Reese *et al.*, 2021) and can be used to integrate across domain-specific ontologies like the Chemical Entities of Biological Interest ontology (ChEBI) (Hastings *et al.*, 2013) or the Human Phenotype Ontology (HPO) (Köhler *et al.*, 2021). Monarch's particular API service accepts prose text as input and returns a set of ontological identifiers with additional information in JSON format.

As with the ingest module, to accommodate the growing number of NLP services and tools for biomedical named entity recognition, Dug also abstracts out the Annotation module. To extend Dug's annotation interface, developers need only create a child class specifying the new API endpoint and define an additional function converting successful API responses into an internal data structure called a *DugIdentifier*, which defines a minimal set of ontological information needed for downstream processing (e.g., id, name, Biolink type).

By converting free text to standardized ontology identifiers, we can leverage the growing number of semantic web services supporting this nomenclature in order to gather additional information about each identifier. Dug utilizes a normalization service to transform identifiers to the preferred equivalents (https://github.com/TranslatorSRI/NodeNormalization), and an ontology metadata service for fetching identifier names, descriptions, synonyms, and Biolink types (https://onto.renci.org/apidocs/).


Ingest and Index Pipeline
Neoplasm of Lung
"adenocarcinoma of the lung"
Synonym Service
MONDO:0008903 Lung Cancer
Impacts
Risk Factor
Treatments
Uberon:1405 Respiratory System
XCO:000007 Cigarette Smoking
CHEBI:4709 Chemotherapy
Single ElasticSearch Record
Data Annotation
monarch INITIATIVE
Concept Expansion
Indexing
elasticsearch

```
{
        variable: "adenocarcinoma of the lung",
        search_terms: ["lung cancer", "neoplasm of the lung"],
        optional_terms: ["respiratory system", "cigarette smoking", "asbestos"]
}
```

Fig. 3: Detailed example of ingest and index pipeline. After ingesting a variable called "adenocarcinoma of the lung" from study metadata, Dug leverages NLP methods for named entity recognition to annotate the variable with an ontology identifier for "Lung Cancer" from the MONDO disease ontology. The resulting identifier is then used to gather synonyms for lung cancer such as "Neoplasm of lung" from an external API service. During concept expansion, Dug leverages TranQL to query knowledge graphs for other ontological concepts related to lung cancer through certain predicates; above we are looking for risk factors, treatments, and anatomical entities impacted by lung cancer. During indexing, all terms discovered through annotation and concept expansion are combined with the original metadata into a single Elasticsearch record so that queries against any of these terms will yield the initial variable measuring "adenocarcinoma of the lung."

Concept Expansion

Dug's ability to retrieve contextualized search results and explain these connections to end users is undergirded by a process we call **concept expansion**. The goal of concept expansion is to further annotate ontological identifiers by identifying relevant connections within ontological knowledge graphs. For example (Fig. 3), we might augment a metadata variable annotated in the previous step as "lung cancer" with ontological identifiers for "asbestos exposure" or "cigarette smoking" based on peer-reviewed evidence supporting those linkages.

Briefly, the core data structure for concept expansion is a knowledge graph, in which biomedical data are organized into a network structure, with nodes representing entity types (e.g., disease, gene, chemical exposure) and edges providing predicates that describe the relationship between entities; Biolink predicates include terms such as 'causes,' 'is associated with,' and 'is expressed in.'

In order to contextualize metadata within a knowledge graph, we leverage data integration approaches developed through the NCATS Biomedical Data Translator (Biomedical Data Translator Consortium, 2019). Chief among these are ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways) (Bizon *et al.*, 2019), a biomedical knowledge graph-based, open-question-answering system that allows users to ask questions such as, "*What chemical entities are associated with lung cancer?*" and TranQL (Translator Query Language; https://tranql.renci.org) (Cox *et al.*, 2020), a query, visualization, and API environment for iterative querying of Translator knowledge graphs.

Dug leverages TranQL to gather an expanded set of ontological concepts related to those extracted via NLP annotation in the previous step. Dug allows platform administrators to define the set of TranQL query templates used to retrieve related ontology identifiers. Below is an example of a query template used to retrieve diseases impacting a specific body part:

```
FIND Disease -> Anatomical_Entity WHERE ANATOMICAL_ENTITY ==
{query_ontology_id}
```

During concept expansion, Dug then uses these templates to substitute actual ontological identifiers extracted from the previous NLP annotation step in order to retrieve a set of relevant terms for a specific variable.

The “answers” returned by TranQL queries are then used to both increase the search relevance of related concepts and provide a basis for including explanations for the links that led to the result. Critically, this includes the ability to point users to peer reviewed literature and curated ontological knowledge for further research. As shown in Fig. 4, TranQL “answers” are actually knowledge graphs themselves: a set of nodes (ontological identifiers), edges (predicates), and metadata about both edges and nodes including names, descriptions, and synonyms of each ontological identifier, as well as any PubMed literature links supporting each edge.

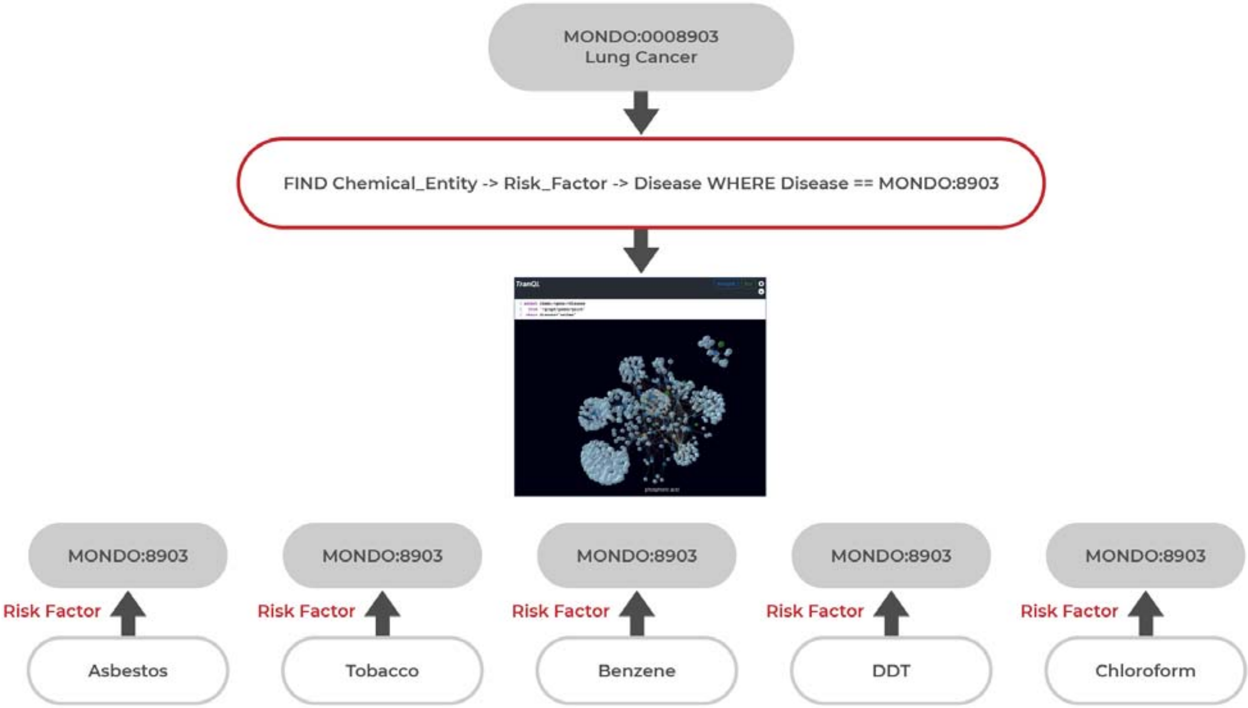


Figure 4: An example TranQL query for chemical risk factors of lung cancer. Each TranQL answer returned is a knowledge subgraph linking one chemical element node back to the query node.

Data Indexing

Once a metadata record has been annotated and undergone concept expansion, the entire resulting data structure must be indexed. To maximize the speed and flexibility of Dug’s search functionality, Dug’s back-end search architecture is implemented as a set of linked Elasticsearch indices. Though Dug’s ingestion and indexing pipelines are relatively time-intensive, Dug’s actual search is remarkably fast and efficient because it searches over pre-computed indices.

As shown in Fig. 3, Dug's semantic capabilities result from combining ingested study metadata with terms harvested through the annotation and concept expansion steps into a single

Elasticsearch record. For each metadata variable ingested, Dug's indexer adds a 'search_terms' field to the original record containing the names and synonyms of every identifier added via the data annotation step. The indexer then adds an 'optional_terms' field by traversing the names of knowledge graph nodes added during concept expansion. For example, a metadata variable originally called "adenocarcinoma of the lung" may now also be labeled with "asbestos exposure" and would be returned by queries against either term.

To allow Dug to quickly organize search results by higher-level concepts, we partition 1) ingested metadata records, 2) core ontological concepts, and 3) expanded knowledge graph answers respectively into three separate Elasticsearch indices. Any ontological identifier extracted during data annotation is added to the Concept index. By indexing study variables with id pointers to these concepts, Dug eliminates the need to calculate these groups dynamically, and eliminates redundant text stored across study variables mapping to the same ontological concept. Each indexed knowledge graph answer contains a JSON representation of the answer sub-graph returned by TranQL, as well as a pointer back to the ontological concept that was used in the original query.

Search Engine

Search Functionality

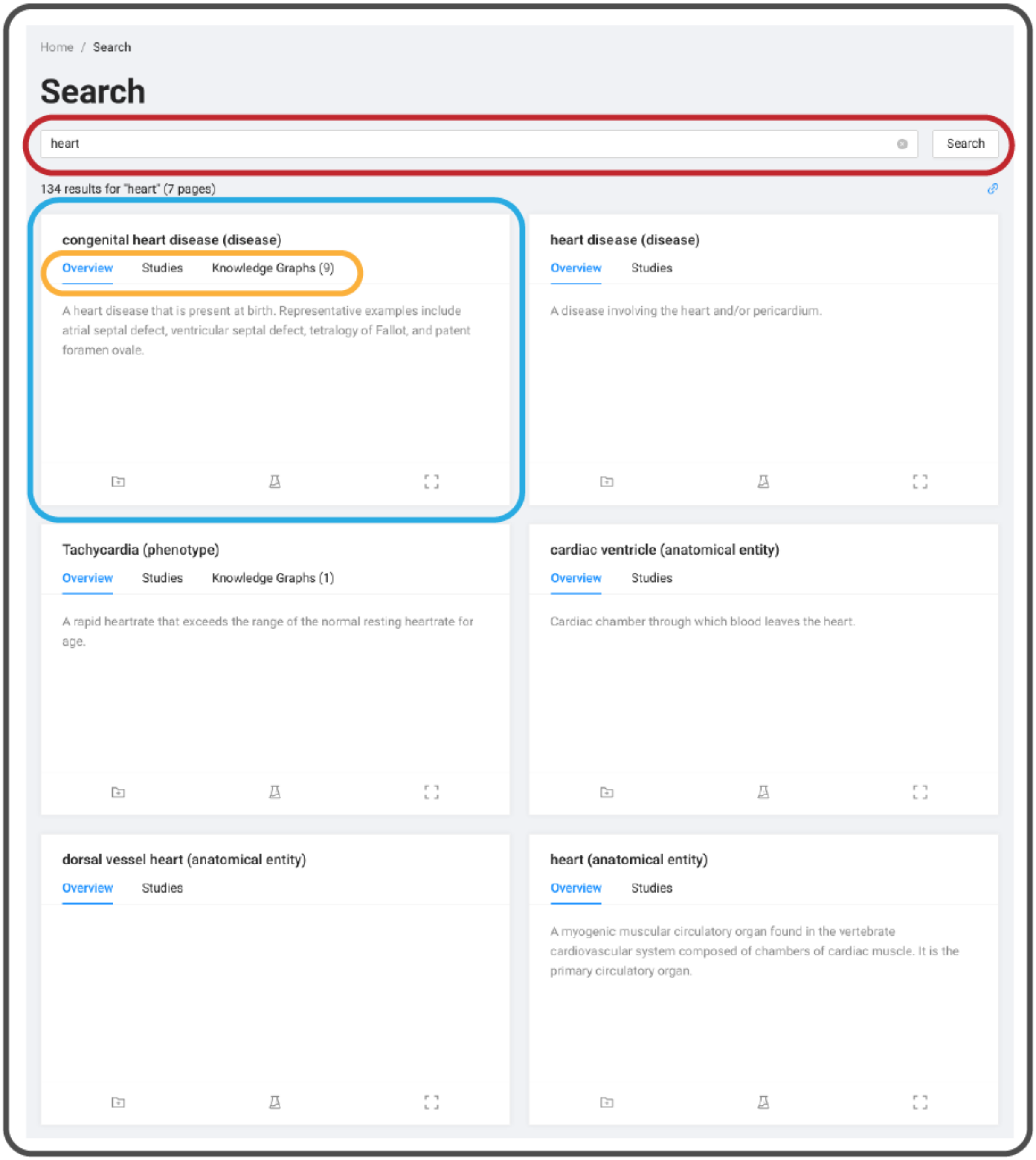
Dug's search API exposes three search endpoints for querying each of its underlying Elasticsearch indices.

- **/search_var** - search for study variables matching a user's query
- **/search_concepts** - search for ontological concepts matching a user's query
- **/search_kg** - search for knowledge graph answers matching user's query and an ontological concept id

Dug's search API uses the default Elasticsearch algorithm (cosine similarity based on vector space model using TF-IDF weighting) (Elasticsearch, 2021) to rank and retrieve indexed metadata records. The search field weighting scheme prioritizes exact matches from the originally ingested text first, followed by synonyms added through annotation, and lastly related terms added through concept expansion.

Search User Interface

Dug's user interface (UI) is a stand-alone React JS-based web application designed to provide an intuitive interface for navigating large collections of data. Dug's minimalist UI design is intended to reduce the burden on users to think through search criteria, and instead, empower them to discover search terms they are interested in exploring. As shown in Fig. 5, Dug provides a simple, Google-like search box that prioritizes exact phrase matching (AND logic) over partial matching (OR logic). Other features inspired by popular search engines include auto-generated tabbing of search results based on data type. For example, if a user's query returns 50 proteomics datasets, 10 genomics datasets, and 3 imaging datasets, Dug automatically creates tabs for each data type.



Home / Search

Search

heart

Search

134 results for "heart" (7 pages)

congenital heart disease (disease)

Overview Studies Knowledge Graphs (9)

A heart disease that is present at birth. Representative examples include atrial septal defect, ventricular septal defect, tetralogy of Fallot, and patent foramen ovale.

heart disease (disease)

Overview Studies

A disease involving the heart and/or pericardium.

Tachycardia (phenotype)

Overview Studies Knowledge Graphs (1)

A rapid heartrate that exceeds the range of the normal resting heartrate for age.

cardiac ventricle (anatomical entity)

Overview Studies

Cardiac chamber through which blood leaves the heart.

dorsal vessel heart (anatomical entity)

Overview Studies

heart (anatomical entity)

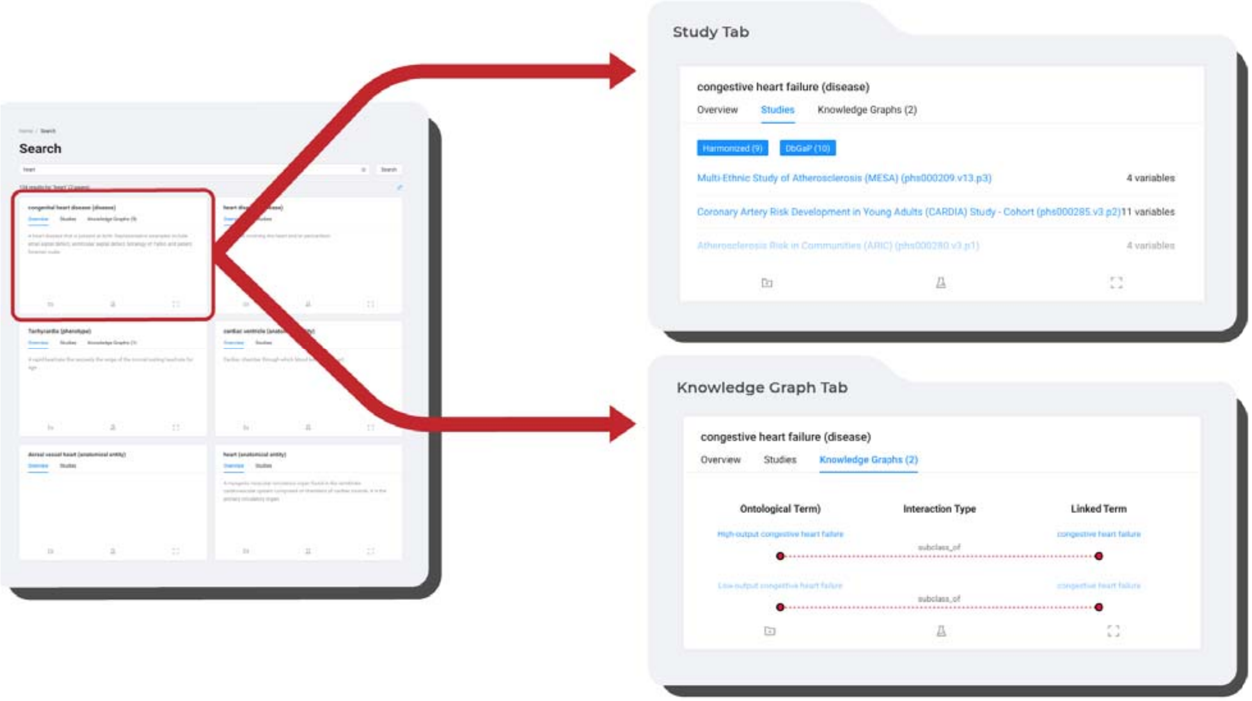
Overview Studies

A myogenic muscular circulatory organ found in the vertebrate cardiovascular system composed of chambers of cardiac muscle. It is the primary circulatory organ.

Fig. 5: Dug's UI aggregates search results for user queries (red) into higher-order ontological concepts (blue) based on NLP annotation. Links to knowledge graphs (orange) explain biological relationships between the query and each concept. Concepts may not contain

knowledge graph links if they are synonymous with the search query (e.g., heart attack vs. myocardial infarction) or if TranQL did not return any answers during concept expansion.

Dug's UI can organize search results in two distinct ways: by variable and by concept. When organized by variable, Dug returns a list of study variables organized by relevance. Each result contains general information about the variable returned: parent study, external links, variable name, and any descriptions parsed from the original metadata file. When organized by concept as in Fig. 5, Dug aggregates results into higher-level ontological concepts. Users can then click on browser disclosures to see datasets of interest from each concept group. In this way, Dug can also be used as a preliminary harmonization step to create *de novo* groups of similar variables based on NLP annotations (Fig. 5). By providing both approaches, Dug's concept-based search gives users a more exploratory look at the data landscape, while its variable-based search allows users to drill down on specific variables of interest.


Search
heart
Search
congenital heart disease (disease)
Overview
Studies
Knowledge Graphs (9)
Tachycardia (phenotype)
cardiac ventricle (anatomical entity)
dorsal vessel heart (anatomical entity)
heart (anatomical entity)
Study Tab
Knowledge Graph Tab

congestive heart failure (disease)

Overview · Studies · Knowledge Graphs (2)

Harmonized (9) · DbGaP (10)

Multi-Ethnic Study of Atherosclerosis (MESA) (phs000209.v13.p3)	4 variables
Coronary Artery Risk Development in Young Adults (CARDIA) Study - Cohort (phs000285.v3.p2)	11 variables
Atherosclerosis Risk in Communities (ARIC) (phs000280.v3.p1)	4 variables

congestive heart failure (disease)

Overview · Studies · Knowledge Graphs (2)

Ontological Term)	Interaction Type	Linked Term
High-output congestive heart failure	subclass_of	congestive heart failure
Low-output congestive heart failure	subclass_of	congestive heart failure

Fig. 6: Dug results show datasets with variables relevant to a user's query, and knowledge graph disclosures for understanding why Dug considered those results relevant.

The defining feature of Dug's UI is its ability to explain *why* it returns certain results (Fig. 6). When Dug returns a result based on text added during concept expansion, the UI fetches and renders the corresponding knowledge graph answer from the Knowledge Graph index to explain the connection. For knowledge graph answers containing PubMed links, Dug also renders links to peer-reviewed literature supporting those specific relationships.

Implementation

The Dug search app is implemented in Python3.7 (https://python.org) and is available via GitHub (https://github.com/helxplatform/dug). Dug packages its core code and 3rd party services into a single Dug container service deployable locally via docker-compose (https://docs.docker.com/compose/) or on Kubernetes (https://kubernetes.io/) clusters via a Dug helm chart. In addition to Elasticsearch, Dug utilizes Redis (https://redis.io/) to cache API requests and minimize redundant calls to external services. Dug's core services are externally configurable from a single file that allows users to specify annotation modules, external API endpoints, ontology normalization services, and query templates that will be used during concept expansion. Dug's ingest architecture leverages the Pluggy framework (https://pluggy.readthedocs.io/en/latest/) to define new metadata parser modules. Dug also provides a makefile script to install the service locally. Lastly, all major commands (e.g., crawl, index, search) can be invoked either from the command line or via API calls to the service directly.

Deployment on BioData Catalyst

To demonstrate Dug's utility in a production environment, we deployed Dug on the NHLBI's BioData Catalyst ecosystem and indexed the TOPMed freeze 5b and 8 studies (excluding parent studies) available on the ecosystem. Public data dictionaries downloaded directly from dbGaP were ingested and indexed for each of the 76 datasets included in these freezes.

Discussed in detail below, 15,911 of these variables were also manually harmonized into 65 higher-order groups called "phenotype concepts" by data curation experts at the TOPMed Data Coordinating Center (DCC) (Stilp *et al.*, 2021). To facilitate browsing by these expertly curated groups within Dug, we annotated ingested variables with phenotype concepts as though they were external ontology identifiers so the underlying variables could be aggregated by these phenotype concepts in the web portal.

.

Evaluation

We evaluated Dug's ability to return relevant study variables using the TOPMed phenotype concepts dataset as our framework for evaluation. The 65 TOPMed harmonized phenotypes developed by Stilp *et al.* (2021) were created to enable interoperability between TOPMed datasets by manually combining semantically similar dbGaP variables under a single term. For instance, the dbGaP variable names of "HOSPITALIZED FOR HEART ATTACK," "NONFATAL MI," and "cardiac episode: ECG" would all relate to the TOPMed phenotype concept, "myocardial infarction."

For each of the 65 TOPMed phenotype concepts, we queried concept titles (e.g., "myocardial infarction") against Dug's indexed collection of dbGaP variables to see how well Dug could recapitulate the manually curated set of variables. To quantify Dug's performance, we used the standard information retrieval metrics of recall, precision, and F-score (F1), which is the weighted harmonic mean of precision and recall. We chose to report each of these scores at the 10th result (P@10, R@10, F1@10), at the 50th result (P@50, R@50, F1@50) and at the nth result (P, R, F1) for every search. While (P, R, F1)@10 is a fairly standard metric, we chose (P, R, F1)@50 as an estimate to the maximum number of results a person would reasonably scroll through, assuming 25 results per page (Jansen and Spink, 2005).

Within this framework, each variable that belongs to a given TOPMed phenotype concept represents a condition positive (P) that we would reasonably expect Dug to return. Stop words (e.g., "the," "in," "of") were removed, and any known abbreviations were expanded and added to the query. For instance, “Resting arm systolic BP” would become “Resting arm systolic BP blood pressure.” True positive (TP) results were defined as variables returned by the TOPMed phenotype concept to which they belonged. Conversely, false positives (FP) were defined as variables returned by any TOPMed phenotype concept to which they did not belong. From these values, the recall, precision, and F-score are calculated at results returned (10, 50, or n) for each TOPMed phenotype concept query.

$$Recall = \frac{TP}{P}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

To contextualize Dug’s retrieval metrics against a more traditional search strategy, we evaluated Dug’s semantic search capabilities against the default, lexical Elasticsearch scoring algorithm that considered only fields available in the original metadata file.

To evaluate Dug’s semantic retrieval capabilities, a secondary evaluation was performed as above, except instead of the expanded phenotype concept names, reasonable synonyms from the Unified Medical Language System (UMLS) (Bodenreider, 2004) were substituted. We specifically chose only synonyms containing no words in common with the original concept query; if no reasonable synonym could be chosen for a given TOPMed phenotype, it was removed from the evaluation. These synonyms were then used as the query for each TOPMed phenotype concept and information retrieval metrics calculated separately as before.

Results

The initial Dug deployment on the NHLBI Biodata Catalyst ecosystem successfully indexed 15,991 study variables from 76 genomics datasets in just over 39 minutes. This deployment provides comprehensive search to BioData Catalyst program members and research fellows over genomics datasets comprising TOPMed freezes 5b and 8. In total, Dug augmented these study variables with 573 ontological concepts and 11,752 knowledge graph answers.

Evaluation

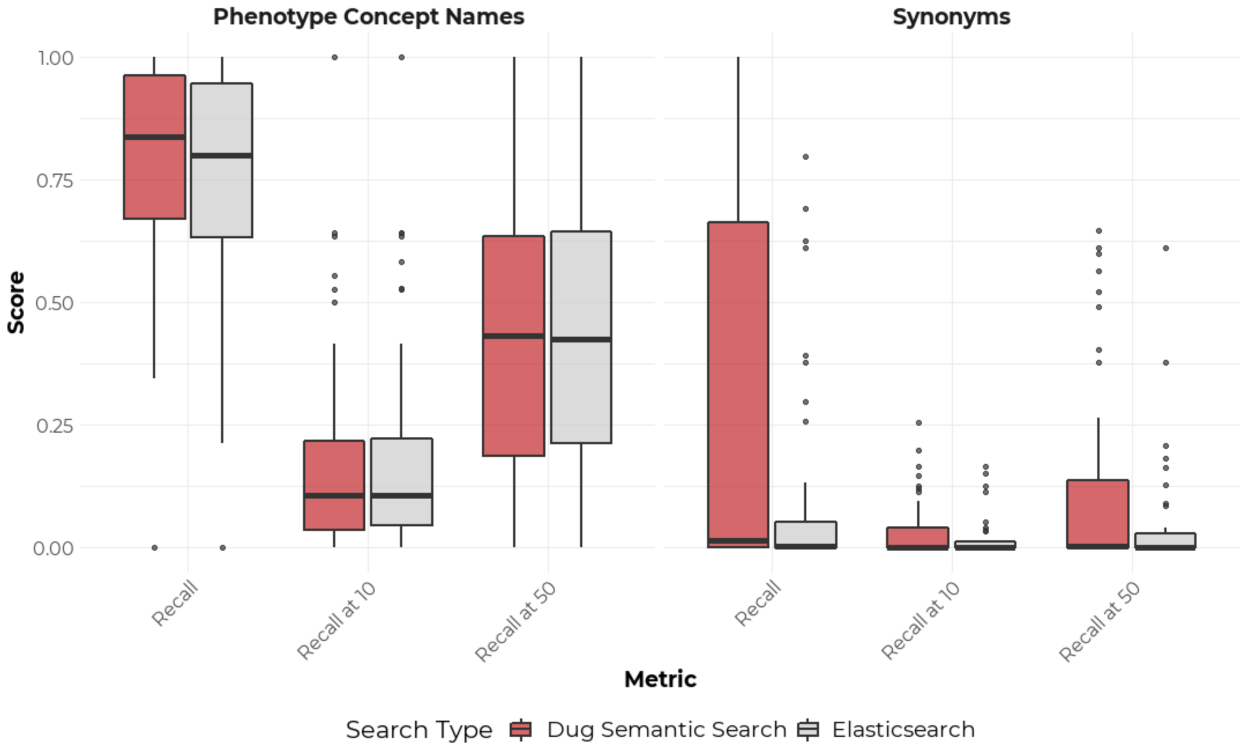


Fig 7: Dug vs default Elasticsearch evaluation metrics. Left: Recall, R@10, R@50 when using TOPMed phenotype concept titles to search for underlying variables. Right: Recall, R@10, R@50 when using UMLS synonyms of TOPMed phenotype concept titles to search for underlying variables. When using phenotype concept names as queries, Dug outperforms lexical Elasticsearch in overall recall, and is similar in R@10 and R@50. When using synonyms of phenotype concept names as queries, Dug outperforms lexical Elasticsearch in every recall metric by a wider margin.

Shown in Fig. 7, Dug search shows superior recall metrics compared to default Elasticsearch when querying both TOPMed phenotype concept names and their synonyms against dbGaP variables indexed on BioData Catalyst. Dug's recall (0.28 mean recall) particularly outperforms default Elasticsearch (0.11 mean recall) when TOPMed phenotype synonyms were used as queries. There was little difference between recall @ 10 and recall @ 50 metrics between search modalities on phenotype concept names. When searching on synonyms, the differences between search types were more apparent. While median recall of synonyms was low for both Dug semantic search and Elasticsearch, the upper quartile in all recall metrics was much higher for Dug semantic search than Elasticsearch. Upon closer inspection, Dug's semantic variable annotation provided synonyms that matched with the synonym queries for a large handful of the phenotype concepts, but missed the mark with others, resembling a binary distribution (Supplemental fig. 1). Overall, Dug matches or exceeds default Elasticsearch on the TOPMed phenotype concept reference dataset, owing to its semantic annotation and concept expansion modules.

Dug also generally returned more results than Elasticsearch for a given search. On one hand, this resulted in a slightly lower precision and F-score for Dug relative to Elasticsearch, however this finding is not unexpected given that Dug was designed explicitly to make exploratory connections. Overall median precision for Dug was 0.06, whereas for Elasticsearch it was 0.28. Overall median F1 was 0.10 for Dug and 0.42 for Elasticsearch. Dug and Elasticsearch metrics @10 and @50 were similar: P@10 and P@50 were 1.0 and 0.74 respectively for Dug; for Elasticsearch they measured 1.0, and 0.82. F1@10 and F1@50 were 0.20 and 0.41 for Dug, 0.20 and 0.50 for Elasticsearch. In general, Dug trades off an increase in overall recall for a decrease in overall precision, with all metrics @10 and @50 relatively similar between the two search modalities.

Discussion

The exponential increase in publicly available datasets over the past decade has created a need for comprehensive search tools that can identify datasets relevant to a researcher's particular scientific question. To help researchers better navigate this new data landscape, we created Dug: a semantic search tool for biomedical datasets that leverages ontological knowledge graphs to intelligently suggest relevant connections derived from peer-reviewed research.

The results of our evaluation demonstrate Dug’s ability to find the correct datasets regardless of how a user lexically expresses a query. Though Dug’s recall increase was modest when using the original TOPMed phenotype concepts as test queries, many of the variables included in these phenotype concepts were lexical matches that wouldn’t benefit from semantic annotation. In hindsight, this could be because the 65 TOPMed Harmonized Phenotypes were created by subject matter experts parsing through dbGaP variables via keyword searches looking for lexical matches. Fortunately, this made our synonym query dataset an excellent test of Dug's semantic recall due to the non-overlapping requirement between synonyms and the original queries. Indeed, for many of our test queries, Dug returns almost the exact same set of results regardless of the lexical expression of the underlying concept. Moreover, because Dug abstracts out the NLP tool used for metadata annotation, Dug will be able to improve over time as new tools for ontological annotation are developed.

Though we present here only the specific deployment of Dug on the BioData Catalyst platform, Dug's modular design and its stand-alone companion web portal can flexibly fit myriad use cases. For example, a centrally hosted version of Dug could index multiple data repositories to service a much larger user base. On the other hand, smaller data coordinating centers like the NIDDK central repository (Rasooly *et al.*, 2015; Cuticchia *et al.*, 2006) or large data ecosystem initiatives like the NIH's HEAL Data Ecosystem (U.S. Department of Health and Human Services) could also use Dug to search across non-standardized data from diverse consortium members via a single portal without the need for significant manual curation.

Current and future work is centered around addressing known limitations and responding to user feedback from the BioData Catalyst consortium. A principal concern at the moment is parallelizing the indexing process to be able to index multiple datasets simultaneously and increase the throughput for larger datasets. Additionally, we are evaluating various strategies for further improving ranking search results by relevance based on input from current users. At present, we believe Dug provides a powerful, flexible tool for searching intuitively across complex data resources that are increasingly common in the biomedical data landscape.

Acknowledgements

The authors would like to thank Ben Heavner, Adrienne Stilp, and Ingrid Borecki for their invaluable feedback during development. We also want to thank members of the BioData Catalyst Fellows program for feedback and testing.

Funding

This work was supported by the National Heart, Lung, and Blood Institute (NHLBI) [grant number 1-OT3-HL142479-01]; the National Center for Advancing Translational Sciences (NCATS) [grant number 1-OT3-TR002020-01]; and the Helping to End Addiction Long-Term (HEAL) Office [grant number 1-OT2-OD031940-01].

Conflict of Interest: none declared.

References

Bell,E. *et al.* (2019) Finding useful data across multiple biomedical data repositories using DataMed. *Nat. Genet.*, **49**, 816–819.

Biomedical Data Translator Consortium (2019) The Biomedical Data Translator Program: Conception, Culture, and Community. *Clin. Transl. Sci.*, **12**, 91–94.

Bizon,C. *et al.* (2019) ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J. Chem. Inf. Model.*, **59**, 4968–4973.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Canakoglu,A. *et al.* (2019) GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database (Oxford)*, **2019**.

Chen,X. *et al.* (2018) DataMed – an open source discovery index for finding biomedical datasets. *J. Am. Med. Informatics Assoc.*, **25**, 300–308.

Collins,F.S. *et al.* (2018) Helping to End Addiction Over the Long-term: The Research Plan for the NIH HEAL Initiative. *JAMA*, **320**, 129–130.

Cox,S. *et al.* (2020) Visualization Environment for Federated Knowledge Graphs: Development of an Interactive Biomedical Query Language and Web Application Interface. *JMIR Med. informatics*, **8**, e17964–e17964.

Cuticchia,A.J. *et al.* (2006) NIDDK data repository: a central collection of clinical trial data. *BMC Med. Inform. Decis. Mak.*, **6**, 19.

Elasticsearch (2021). Theory Behind Relevance Scoring. Retrieved July 1, 2021 from https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html

Hastings,J. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.

Huang,J. *et al.* (2016) OmniSearch: a semantic search system based on the Ontology for MIcroRNA Target (OMIT) for microRNA-target gene interaction data. *J Biomed Semant.*, **7**, 25.

Jansen,B.J. and Spink,A. (2005) Analysis of document viewing patterns of web search engine users. *Web Min. Appl. Tech.*, 339–354.

Köhler,S. *et al.* (2021) The Human Phenotype Ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.

Kuć,R. and Rogozinski,M. (2016) ElasticSearch server Packt Publishing Ltd.

Laulederkind,S.J. *et al.* (2012) Ontology searching and browsing at the Rat Genome Database. *Database (Oxford)*, **2012**, bas016.

Mungall,C.J. *et al.* (2016) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.

National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services (2020). The NHLBI BioData Catalyst. Zenodo. https://doi.org/10.5281/zenodo.3822858

National Institutes of Health, U.S. Department of Health and Human Services (2020). NIH Data Sharing Resources. Retrieved July 1, 2021 from https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html.

Pang,C. *et al.* (2015) BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. *J Am Med Inf. Assoc*, **22**, 65–75.

Powell,K. (2021) The broken promise that undermines human genome research. *Nat. News*.

Rasooly,R.S. *et al.* (2015) The National Institute of Diabetes and Digestive and Kidney Diseases Central Repositories: A Valuable Resource for Nephrology Research. *Clin. J. Am. Soc. Nephrol.*, **10**, 710 LP – 715.

Reese,J.T. *et al.* (2021) KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns (N Y)*, **2**, 100155.

Sansone,S. *et al.* (2017) OPEN DATS , the data tag suite to enable discoverability of datasets. *Sci Data*, 1–8.

Soto,A.J. *et al.* (2019) Thalia: Semantic search engine for biomedical abstracts. *Bioinformatics*, **35**, 1799–1801.

Stilp,A.M. *et al.* (2021) A System for Phenotype Harmonization in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. *Am. J. Epidemiol.*

The “All of Us” Research Program (2019) *N. Engl. J. Med.*, **381**, 668–676.

Tran,T. *et al.* (2007) Ontology-based interpretation of keywords for semantic search. In, *The semantic web*. Springer, pp. 523–536.

U.S. Department of Health and Human Services. What is the HEAL Data Ecosystem? Retrieved July 1, 2021 from https://heal.nih.gov/about/heal-data-ecosystem.

University of Washington Department of Biostatistics (2020) NHLBI Trans-Omics for Precision Medicine WGS-About TOPMed. *About TOPMed*.

Wilkinson,M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.