

# Using SPARQL to unify queries over data, ontologies, and machine learning models in the PhenomeBrowser knowledgebase

Ali Syed<sup>1</sup>[0000-0002-5103-9058], Şenay Kafkas<sup>1</sup>[0000-0001-7509-5786], Maxat Kulmanov<sup>1</sup>[0000-0003-1710-1820], and Robert Hoehndorf<sup>1</sup>[0000-0001-8149-5890]

<sup>1</sup>Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division, King Abdullah University of Science and Technology, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia  
{ali.syed, senay.kafkas, maxat.kulmanov, robert.hoehndorf}@kaust.edu.sa

**Abstract.** We have developed the PhenomeBrowser knowledge base to integrate phenotype associations from a variety of sources into a single knowledge base. We use SPARQL as a unifying query language to access RDF data, perform Description Logic queries over ontologies, and compute the semantic similarity between entities in the knowledge base.

**Keywords:** Phenotype · Semantic Similarity · SPARQL

## 1 Introduction

The similarity between phenotypes associated with entities studied in the life sciences can be used to reveal interactions between biomedical entities at a molecular level [12]. Entities that are similar phenotypically are often related to each other on a molecular level as well [15], and this principle can be used to suggest or discover novel relations between these entities. There are several databases that have been developed for integrating phenotypes and exploring relations between them such as Online Mendelian Inheritance in Men (OMIM) [5] and ClinVar [9], as well as integrated databases such as Monarch [10]. Key challenges in integrating and exploring phenotype data is the use of integrated phenotype vocabularies or ontologies that can systematically relate phenotype classes between different contexts such as the entity studied or the species in which phenotypes are observed (human, model organism, or non-model organism) [4]; the computation of semantic (phenotypic) similarity or relatedness [11]; and the ability to query phenotype-related information using a uniform and (ideally) standardized query language.

We developed the PhenomeBrowser knowledgebase as a semantic framework that combines an RDF-based knowledge base of phenotype associations collected from community resources and from in-house curation with the ability to perform Description Logic queries over phenotype (and other) ontologies as well as to perform some basic operations on a type of machine learning model.

The framework used to develop PhenomeBrowser is based on using SPARQL as query language for any structured data and Apache Lucene indices and queries (implemented in the form of ElasticSearch) for natural language information.

PhenomeBrowser currently contains over four million phenotype associations for genes, diseases, drugs, pathogens and chemical entities (metabolites). We incorporate the Vec2SPARQL method [8] over the Bio2Vec knowledge graph embedding repository (<https://bio2vec.cbrc.kaust.edu.sa>) to perform queries incorporating semantic similarity and machine learning, and we rely on the AberOWL services [7] to perform Description Logic Queries within SPARQL queries. The interface for PhenomeBrowser is based on these SPARQL queries and we provide access to PhenomeBrowser through SPARQL. The Phenomebrowser web portal further implements queries for specific tasks such as finding gene–disease associations, host–pathogen or drug–target interactions, all based on phenotypic similarity. The PhenomeBrowser software and underlying components are available as Free Software ([phenomebrowser.net](http://phenomebrowser.net)) and can serve as an initial model on how to combine graph databases, Description Logic queries, and machine learning within a single framework unified through SPARQL as query language.

## 2 Design and Implementation

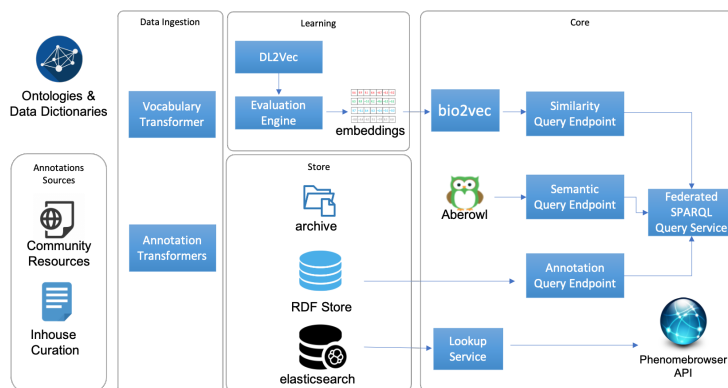


Fig. 1: Core Components of the PhenomeBrowser framework.

Core components of the PhenomeBrowser architecture are shown in Figure 1. Phenotype annotations data from community resources and in-house curation is transformed into RDF format. The transformed data is subsequently stored in an RDF store. We implemented a data intake workflow using snakemake [1] to achieve reproducible and robust automation. As data model for phenotype associations, we rely on community standards for phenotypes developed by the OBO Foundry initiative [14] and the Monarch project [10]. We use the Dublin

Core vocabularies [2] to encode provenance information, and the OBO Relation Ontology [13] to encode relations. We use the integrated phenotype ontology PhenomeNet [6] to integrate phenotypes across different contexts. The data intake workflows also generate text indices for entity as well as classes and relations from the PhenomeNet ontology. Text indices are Apache Lucene indices implemented in Elasticsearch, and we make search of text indices available through a REST API that is complementary to the SPARQL endpoint for querying structured data.

Data that is ingested from public sources is passed to the ontology-based machine learning method DL2Vec [3] to generate embeddings for entities (and ontology classes) that can be used to compute similarity. The embeddings are added to the Bio2Vec repository which stores the embeddings and makes them available for similarity-based queries through a REST API and a special SPARQL endpoint implementing the Vec2SPARQL extensions [8].

When querying data, we use the AberOWL [7] SPARQL endpoint to execute queries that incorporate deductive inference over Semantic Web ontologies. AberOWL is an ontology repository that provides reasoning over ontologies as a service. The queries further federate to the Vec2SPARQL endpoints provided by Bio2Vec, and therefore combine querying RDF phenotype data, phenotype ontologies (through AberOWL), and semantic similarity (through Bio2Vec).

### 3 Querying using SPARQL

One application of computing phenotype similarity is ranking candidate genes for a disease [15]. Using PhenomeBrowser’s integrated SPARQL endpoint, we can perform this operation through SPARQL and therefore suggest gene–disease associations. Figure 2 shows a query for finding genes that are phenotypically similar to *ventricular septal defect* (HP:0011623). In the first section of the query, the content of the *FILTER* block performs a Description Logic Query to retrieve all classes that are equivalent to or subclasses of the *ventricular septal defect* phenotype from the HPO; the query is performed using the AberOWL ontology repository and reasoning service which expands the query and replaces it with the URIs of the classes resulting from the query. Subclasses of *ventricular septal defect* in the HPO include *Tetralogy of Fallot* as well as several more specific forms of Tetralogy of Fallot, and also includes *ventricular septal defect* itself (as the query is reflexive).

The second section of the query contains a federated query to the Bio2Vec SPARQL endpoint and uses the `mostSimilar` function; the `mostSimilar` function is implemented by the Vec2SPARQL method and executes the phenotypic similarity search for the diseases selected in the first section of the query on the Bio2Vec SPARQL endpoint. The *mostSimilar* function is a custom SPARQL function that takes as arguments the dataset identifier in Bio2Vec, the identifier for the entity within the dataset, the number of entities to retrieve (in our case, we retrieve the three most similar entities to our query), and the type (using `rdf:type`) of the entity (in our case, the entity type is `gene`). In the third sec-

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pb: <http://phenomebrowser.net/>
PREFIX b2v: <http://bio2vec.net/function#>
PREFIX b2vd: <http://bio2vec.net/dataset#>>

SELECT ?simGene ?simGeneLabel ?genePhenotype ?genePhenotypeLabel
WHERE {
  {
    SELECT ?disease
    FROM <http://phenomebrowser.net>
    WHERE {
      ?association rdf:type rdf:Statement .
      ?association rdf:object ?phenotype .
      FILTER ( ?phenotype in (
        OWL subeq <http://phenomebrowser.net/sparql> <HP> {
          'ventricular septal defect'
        }
      ) ) .
      ?association rdf:subject ?disease .
      ?disease rdf:type pb:Disease .
    } LIMIT 20
  } .
  SERVICE <https://bio2vec.cbrc.kaust.edu.sa/ds/query> {
    (?simGene ?val ?x ?y) b2v:mostSimilar(b2vd:dataset_4 ?disease 3 pb:Gene) .
  }
  GRAPH <http://phenomebrowser.net> {
    ?simGene rdfs:label ?simGeneLabel .
    ?geneAssociation rdf:subject ?simGene .
    ?geneAssociation rdf:object ?genePhenotype .
    ?genePhenotype rdfs:label ?genePhenotypeLabel .
  }
} ORDER BY asc(?simGeneLabel)

```

Fig. 2: SPARQL query finding genes that are phenotypically similar to *ventricular septal defect*.

tion of the query, we add labels to genes found in the second section of the query and their associated phenotypes.

## 4 Conclusion

We developed the PhenomeBrowser knowledgebase as a semantic framework that integrates querying over graph databases, ontologies, and knowledge graph embeddings, using SPARQL as a unifying and standardized query language. PhenomeBrowser is accessible at <http://phenomebrowser.net>.

## Acknowledgements

We acknowledge use of the resources of the KAUST Supercomputing Core Laboratories.

## Funding

This work was supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3790-01-01, URF/1/4355-01-01, FCC/1/1976-08-01, and FCC/1/1976-08-08.

## References

1. Snakemake.
2. Dcmi metadata terms, Jan 2020.
3. Jun Chen, Azza Althagafi, and Robert Hoehndorf. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics*, 37(6):853–860, 10 2020.
4. Georgios V Gkoutos, Paul N Schofield, and Robert Hoehndorf. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, 19(5):1008–1021, April 2017.
5. Ada Hamosh, Alan F. Scott, Joanna Amberger, David Valle, and Victor A. McKusick. Online mendelian inheritance in man (omim). *Human Mutation*, 15(1):57–61, 2000.
6. Robert Hoehndorf et al. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119, 2011.
7. Robert Hoehndorf, Luke Slater, Paul N Schofield, and Georgios V Gkoutos. AberOWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16:26, 2015.
8. Maxat Kulmanov, Senay Kafkas, Andreas Karwath, Alexander Malic, Georgios V Gkoutos, Michel Dumontier, and Robert Hoehndorf. Vec2SPARQL: integrating SPARQL queries and knowledge graph embeddings. *bioRxiv*, 2018.

9. Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 2013.
10. Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, Erin Foster, J. P. Gourdine, Julius O. B. Jacobsen, Dan Keith, Bryan Laraway, Suzanna E. Lewis, Jeremy NguyenXuan, Kent Shefchek, Nicole Vasilevsky, Zhou Yuan, Nicole Washington, Harry Hochheiser, Tudor Groza, Damian Smedley, Peter N. Robinson, and Melissa A. Haendel. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, 45(D1):D712–D722, Jan 2017.
11. Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):e1000443, July 2009.
12. Slavé Petrovski and David B. Goldstein. Phenomics and the interpretation of personal genomes. *Science Translational Medicine*, 6(254):254fs35–254fs35, 2014.
13. B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol*, 6(5):R46, 2005.
14. Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe R. Serra, Alan Ruttenberg, Susanna A. Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.
15. Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, 7(11):1–20, 11 2009.