Enhancement and Reusage of Biomedical Knowledge Graph Subsets

Jose Emilio Labra Gayo[1], **Carolina González-Cavazos**[2], **Seyed Amir Hosseini Beghaeiraveri**[2], **Ammar Ammar**[4], **Andra Waagmeester**[5], **Sabah Ul-Hasan**[2], **Egon Willighagen**[4], **and Nils Hoffmann**[6]

1 WESO research group, University of Oviedo, Spain **2** Gene Wiki Project, Scripps Research Institute, USA **3** Heriot-Watt University, UK **4** Maastricht University, The Netherlands **5** Micelio/Gene Wiki, Belgium **6** Forschungszentrum Juelich, Germany

BioHackathon series:
BioHackathon Europe 2022
Paris, France, 2022
Enhancement and Reusage of Biomedical Knowledge Graph Subsets

Submitted: 14 Nov 2022

Abstract

Knowledge Graphs (KGs) such as Wikidata act as a hub of information from multiple domains and disciplines, and is crowdsourced by multiple stakeholders. The vast amount of available information makes it difficult for researchers to manage the entire KG, which is also continually being edited. It is necessary to develop tools that extract subsets for domains of interest. These subsets will help researchers to reduce costs and time, making data of interest more accessible. In the last two BioHackathons (BH20, BH21), we have created prototypes to extract subsets easily applicable to Wikidata, as well as to define a map of the different approaches used to tackle this problem. Building on those outcomes, we aim to enhance subsetting in both definitions using Entity schemas based on Shape Expressions (ShEx) and extraction algorithms, with a special focus on the biomedical domain. Our first aim is to develop complex subsetting patterns based on qualifiers and references for enhancing credibility of datasets. Our second aim is to establish a faster subsetting extraction platform applying new algorithms based on Apache Spark and new tools like a document-oriented DBMS platform.

Introduction

Knowledge graphs (KGs) like Wikidata are successfully employed to represent and link data. Although Wikidata is openly-accessible and editable by anyone, the amount of information captured is continually increasing, making difficult to handle by researches. Therefore, there is a current need to develop extraction subset tools that could make Wikidata knowledge more accesible for domain experts. In previous BioHackathons we have reviewed the functionality and requirements of multiple tools that can be used to create subsets of knowledge graphs (ShEx+slurp, WDumper, WDSub, SparkWDSub, Wikibase dump filter, KGTK). During this BioHackathon we aim to continue exploring the requirements of these tools and establish use-case examples on the biomedical doamin.

This paper describes the activities that have been done during our participation at Biohackathon-Europe 2022, project 11. The main goal of the project was to create subsets of Wikidata and to analyse those subsets. The project is a continnuation from work that started in past hackathons held at Biohackathon 2021 Europe and SWAT4HCLS. - Biohackathon Europe 2021: Handling Knowledge graphs subsets - Virtual Biohackathon 2020 - SWAT4HCLS 2019

Activities done during Biohackathon-Europe 2022

1 Setup Wdsub library

The wdsub library was taken as the tool that can generate the subsets. It is available in docker as a command line tool so in order to run it is only required to have docker installed.

The library has 2 main inputs: a Wikidata dump in JSON format, and a Shape Expression that describes the contents of the target subset.

Wikidata dumps can be downloaded from this link. That page contains a link to the latest dump as well as a link to previous dumps since 2014 from Internet Archive

Shape Expressions are based on the ShEx language and describe the RDF serialization from Wikidata. Those Shape Expressions are compatible with the Entity schemas namespace from Wikidata. The wdsub tool can also use as input WShEx schemas. WShEx is a ShEx variant that is adapted to the Wikibase data model. More information about WShEx can be found at: https://www.weso.es/WShEx/.

The shape expressions used during the biohackathon are available at this github repo.

2 Creating "hello world" subsets

To evaluate the setup Andra created two shape expressions gene.shex and disease.shex. The disease shape was written to generate a smaller subset, since the gene shape generated a large subset that was to big to be loaded in a (free-tier) public sparql endpoint on either data.world or TriplyDB, both have limitation in their free tier. However, the resulted disease subset from Wikidata which is 36 Mb (gzipped) was also to big to fit in the free-tier packages. Future work should include finding (affordable) hosting for subsets.

These two Shape Expressions described both the gene and disease items in Wikidata, by describing the types (ie. using P31 (instance of)) and the relavant identifiers. Not all items on Genes nor Diseases use instance-of (P31) statements. Quite some of those statements are wrongly annotated as such and need further curation. We considered describing these incorrect (negative) patterns (e.g. an item that is annotated as both a gene and a protein), but decided that this is best done post subsetting. The aim of a subset from our perspective is to describe a subset-profile which would extract a subset with high recall. Curation often needs complex queries which time-out on the Wikidata Query Serivce. The curation can subsequently happen as either an additional Shape Expression validation pipeline, or through a designated SPARQL endpoint where the subset included.

Preliminary results show that WDSub allows describing subset descriptions that allow multiple shapes. This is particularly interesting in Wikidata, where the same concept can be described using different sets of properties or shapes. Here we described genes and diseases using both their types and identifiers, but another example is how in Wikidata ship wrecks are described. Wikidata items on shipwrecks are either annotated using P31 (instance of) Q852190 (shipwreck), or by annotating ships with a statement using P793 (significant event) Q906512 (shipwrecking).

Identifying all variants of shapes describing a concept within Wikidata is a challenging task, however these shapes on genes and diseases show that we are now able to extract subsets from wikidata using multiple shapes that define a subset.

caveat

With the current implementation of wdsub traversing a graph to identifiy a subset is not yet implemented. A use-case where this is needed it the subset of taxonomic clade of lepidoptera, which consists of all butterflies and moths. Extracting subsets on taxonomic clades requires a recursive shape expression. Currently, WDSub does not support recursive shapes.

This requires a different implementation. WDSub now feeds on the daily dumps of Wikidata. These dumps are concatenated JSON representations of all wikidata items. Traversing wikidata to extract subsets such as the one for all butterflies and moths would require graph traversal which, as said, is not yet implemented in WDSub

3 Creating GeneWiki subset

The biohackathon assisted in an evaluation of existing nodes and edges on drug-target interactions categories within Wikidata. We built machine readable schemas of drug-target interactions in Wikidata for future data reuse (link).

4 Creating LIPID MAPS Wikidata subset

During the BioHackathon, one of the use cases formulated to experiment with subsetting was about lipid chemical compounds. LIPID MAPS is a project curating knowledge about lipids and WikiPathways captures knowledge about biological processes. Complementing the information from this curation project with knowledge from Wikidata defines a unique new knowledge graph to support lipidomics research, e.g. in EpiLipidNET.

5 Uploading subsets to local SPARQL endpoint

The gene and disease subsets are turtle (RDF) files, which can be stored in any SPARQL endpoint. We have install GraphDB from Ontotext locally to load the generated gene and disease subsets.

6 Analysing the subsets using SPARQL queries

Once the SPARQL endpoint was publicly available, we were able to create several SPARQL queries to analyse if the results were the expected ones. An example SPARQL query is the following, which obtains all medications (Q12140) and counts the links to diseases (Q12136) for which they are the *medical condition treated* (P2175):

```
prefix rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
prefix wd:       <http://www.wikidata.org/entity/>
prefix wdt:      <http://www.wikidata.org/prop/direct/>


SELECT ?x ?xLabel ?countY WHERE {
  ?x wdt:P31 wd:Q12140  .
 { select ?x (count(?y) as ?countY) where {
     ?x wdt:P2175 ?y .
     ?y wdt:P31 wd:Q12136 .
 }}

 Optional { ?x rdfs:label ?xLabel }
 filter(
    lang(?xLabel)='en'
  )
}
```

The query can be run online and some results:

```
"x"                                      ,"xLabel"             ,"countY"
"http://www.wikidata.org/entity/Q411240" ,"mifepristone"       ,"3"
"http://www.wikidata.org/entity/Q415304" ,"desogestrel"        ,"2"
"http://www.wikidata.org/entity/Q416331" ,"pentoxifylline"     ,"6"
"http://www.wikidata.org/entity/Q417222" ,"methylprednisolone" ,"3"
"http://www.wikidata.org/entity/Q419652" ,"danazol"            ,"2"
"http://www.wikidata.org/entity/Q1752915","trifluoperazine"    ,"2"
"http://www.wikidata.org/entity/Q181354" ,"biotin"             ,"1"
"http://www.wikidata.org/entity/Q23767"  ,"calcium carbonate"  ,"5"
"http://www.wikidata.org/entity/Q407548" ,"magnesium hydroxide","6"
"http://www.wikidata.org/entity/Q419991" ,"Magaldrate"         ,"5"
```

```
"http://www.wikidata.org/entity/Q2601832","ticarcillin"        ,"2"
"http://www.wikidata.org/entity/Q1060922","oxybutynin"         ,"2"
"http://www.wikidata.org/entity/Q221361" ,"clozapine"          ,"3"
"http://www.wikidata.org/entity/Q411461" ,"pilocarpine"        ,"2"
"http://www.wikidata.org/entity/Q412323" ,"rituximab"          ,"1"
"http://www.wikidata.org/entity/Q417813" ,"oxymetazoline"      ,"2"
. . .
```

Conclusions

- WDSub proved to be able to extract subsets from daily wikidata dumps in an acceptable timeframe (9-12 hours)
- By defining subset profiles that consists of multiple shapes, it is possible to extract high recall subsets from Wikidata. Within Wikidata, with its more then 10k properties, concepts are described using many property sets and graph patterns. By collecting all the possible variations as shapes in a single Shape Expressions file, it is possible to extract high recall subsets from wikidata.
- Building on the work done in previous hackathon (Both biohackathon and SWAT4HCLS), we now have several subset mechanisms to generate subsets from Wikidata.

Future work

- Wikidata subsetting as a service:
  - We could create a curated list of ShEx/subsets that are regularly generated for specific purposes.
  - Service like WDumper that takes ShEx as input.
- Evaluation of data quality:
  - Investigate the quality of the data (references and qualifiers)
- Integrate extracted subset with other sources:
  - Use extracted subsets as starting points as a seed to create new knowledge graphs, integrating their information with other sources. Example: 1- Map with data from other ontologies, KGs, etc. 2- Use BioSchemas.
- Drug repurpusing:
  - GeneWiki subset representing drug-target interactions in Wikidata could be used as input to find new uses of drugs.

Challenges to consider

- Update online subsets periodically:
  - Need to update ShEx.
  - Download dumps, and update the subset.
  - We could probably save the ShEx in a github repo and take the latest version from there.

References