Open knowledge graph on clinical trials

Sandeep Kunkunuru[*] [1,2]

[1]IIPH Hyderabad
[2]VaidhyaMegha Private Limited

[*] Correspondence: Sandeep Kunkunuru <hi18sk@iiphh.org>

Abstract

Using any clinical trial id from across the globe find the associated diseases, interventions, research articles and genes. Also discover relationships between various medical topics through co-occurrences in articles. Query the graph using SparQL from cli or GraphQL using any API client tool ex: Postman or curl

Introduction

A knowledge graph could be defined as below per [3]

```
...
Herein we adopt an inclusive definition, where we view a knowledge graph as a graph
of data intended to accumulate and convey knowledge of the real world, whose nodes
represent entities of interest and whose edges represent relations between these
entities. The graph of data (aka data graph) conforms to a graph-based data model,
which may be a directed edge-labelled graph, a propertygraph, etc
...
By knowledge, we refer to something that is known. Such knowledge may be accumulated
from external sources, or extracted from the knowledge graph itself.
```

The objective of this project was to build a knowledge graph with clinical trial ids at the heart and below feature list in addition :

Feature list

- Using GraphQL API knowledge graph can be queried using any API client tool ex: curl or Postman.
- Graph includes trials from across the globe. Data is sourced from WHO's ICTRP and clinicaltrials.gov
- Links from trial to MeSH vocabulary are added for conditions and interventions employed in the trial.

- Links from trial to PubMed articles are added. PubMed's experts curate this metadata information for each article.
- Added MRCOC to the graph for the selected articles linked to clinical trials.
- Added PheGenI links i.e. links from phenotype to genotype as links between MeSH DUI and GeneID.
- Added SparQL query execution feature. Adding CLI mode. Adding a count SparQL query for demo.
- 5 co-existing bi-partite graphs together comprise this knowledge graph. Bi-partite graphs are between
  - trial–> condition
  - trial–> intervention
  - trial –> articles
  - article –> MeSH DUIs
  - gene id –> MeSH DUIs

Sources

- WHO's ICTRP
  - Registries covered in ICTRP include :
- AACT Clinicaltrials.gov
- NLM MeSH
- NLM MRCOC
- NLM PubMed
- NLM PheGenI

Methods

This section lists of high level steps that were executed to build each of the 5 bi-partite graphs.

Linking clinical trials to conditions

- AACT offers clinicaltrials.gov's clinical trial registration data as a downloadable database snapshot.
- This snapshot includes all the trial registration data along with MeSH literals for conditions.
- Database snapshot is in the form a PostgreSQL database dump.
- Snapshot was restored into a PostgreSQL database and the table 'browse_conditions' was queried to retrieve MeSH literals for conditions.
- MeSH literals were then queried within MeSH RDF to retrieve MeSH DUIs.
- Trial id and MeSH DUIs were used to create an edge in the knowledge graph using Apache Jena [1]. [2].

Linking clinical trials to interventions

- The same AACT snapshot as above includes all the trial registration data along with MeSH literals for interventions too.
- Snapshot was restored into a PostgreSQL database and the table 'browse_interventionss' was queried to retrieve MeSH literals for conditions.
- MeSH literals were then queried within MeSH RDF to retrieve MeSH DUIs.
- Trial id and MeSH DUIs were used to create an edge in the knowledge graph using Apache Jena [1]. [2].

Collecting trials from across the globe

- While several approaches exist to keep the design simple, ICTRP's full export dataset along with its weekly incremental files were used.
- These files were imported into a PostgreSQL database as a table.
- After de-duplicating trial ids across AACT's database a global unique list of trial ids, across the globe, is formulated.
- This final list of trials is exported as an intermediate artifact in a CSV file format.

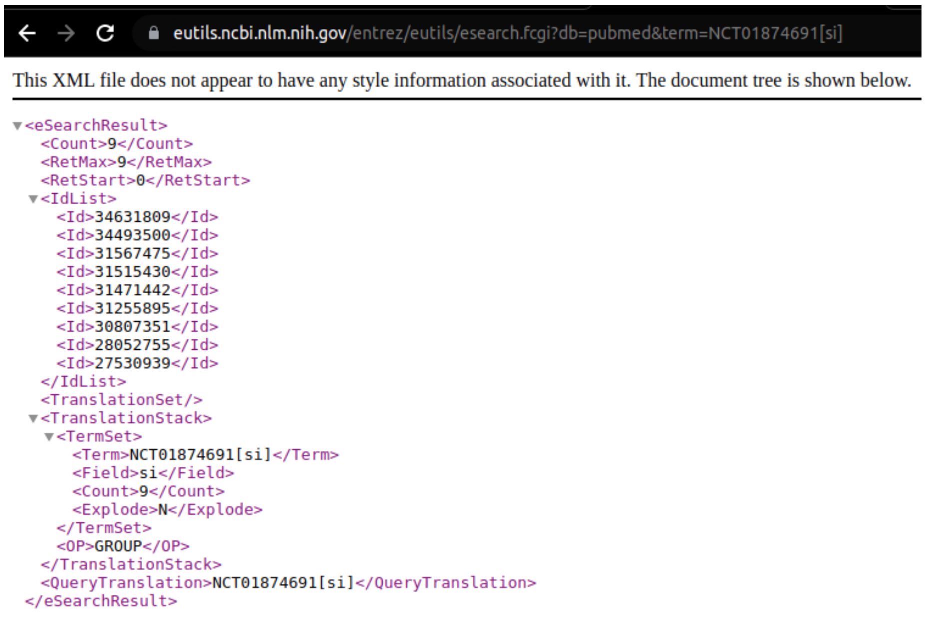
Linking clinical trials to research articles

- The final list of trials exported above as CSV file is read one-line at a time.
- The NLM (The world's largest medical library, the U.S. National Library of Medicine is part of the National Institutes of Health) extracts trail ids from an article and places them into the article's metadata in secondary id field.
- To retrieve journal articles related to a clinical trial id ex: NCT00000419, use PubMed's API called e-Utils with clinical trial id as shown below:

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=NCT01874691[si]`

- In the above URL "[si]" refers to Secondary ID which can be used to search within article's metadata.
- All the journal articles related to the trial id are collected from the above API invocation.
- Output contains PMIDs (pubmed records) of respective clinical trials.

- Using Spring WebClient, JAXB, Jackson and Lambok response XML is automatically parsed and PMID list is constructed in-memory.
- The PMID list is then written into RDF along with trial id using Apache Jena.
- The PMIDs are also persisted into database along with trial id, irrespective of whether any linked articles are found or not.


eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=NCT01874691[si]

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<eSearchResult>
   <Count>9</Count>
   <RetMax>9</RetMax>
   <RetStart>0</RetStart>
  ▼<IdList>
     <Id>34631809</Id>
     <Id>34493500</Id>
     <Id>31567475</Id>
     <Id>31515430</Id>
     <Id>31471442</Id>
     <Id>31255895</Id>
     <Id>30807351</Id>
     <Id>28052755</Id>
     <Id>27530939</Id>
   </IdList>
   <TranslationSet/>
  ▼<TranslationStack>
    ▼<TermSet>
       <Term>NCT01874691[si]</Term>
       <Field>si</Field>
       <Count>9</Count>
       <Explode>N</Explode>
     </TermSet>
     <OP>GROUP</OP>
   </TranslationStack>
   <QueryTranslation>NCT01874691[si]</QueryTranslation>
 </eSearchResult>
```

Figure 1: Trials-Articles

Linking articles to MeSH DUIs

- NLM provides a MeSH term co-occurrence file which provides co-occurrences of MeSH terms (DUIs) at the article level in a pipe separated file format.
- This file is currently at 183 GB size with approximately 1.7 Billion rows.
- The necessary 3 columns need for the purposes of linking articles to MeSH DUI were selected from the file using linux commands. Columns include article id, MeSH DUI 1 and MeSH DUI 2
- The resultant filtered list is sorted on article id column, to allow for efficient search operations.
- This filtered and sorted MRCOC file currently is of 40 GB size with same row count i.e. 1.7 Billion rows
- List of trials along with articles, persisted above in database is used to a build a sorted list of article ids.
- This list of article ids is read one id at a time, i.e. it is read in a streaming fashion.
- Filtered and sorted MRCOC file is also read in a streaming fashion in a file co-parsing patterns.
- Matches are found between articles and MRCOC file records while linearly parsing both.
- All matches are saved into knowledge graph as edges using Apache Jena [1]. [2].

Linking genes to MeSH DUIs

- NLM provides PheGenI, a search tool and database for linking Phenotype MeSH literals to Gene IDs.
- Phenotype/trait along with Gene Id 1 and Gene ID 2 are selected from the PheGeni file.
- If trait is already present in the knowledge graph i.e. if the trait is found to be linked to any trial, then below edges are added using Apache Jena [1]. [2].
  - trait –> Gene Id 1
  - trait –> Gene Id 2

Results

Querying knowledge graph using SparQL

```
java -jar -Xms4096M -Xmx8144M target/vaidhyamegha-knowledge-graphs-v0.9-jar-with-dependencie
    -m cli -q src/main/sparql/1_count_of_records.rq
...
Results:
--------
5523173^^http://www.w3.org/2001/XMLSchema#integer
```

Querying knowledge graph using GraphQL (via Hyper-GraphQL)

Start server

```
  java -cp "target/vaidhyamegha-knowledge-graphs-v0.9-jar-with-dependencies.jar:lib/*" \
    com.vaidhyamegha.data_cloud.kg.App -m server
```

From Postman

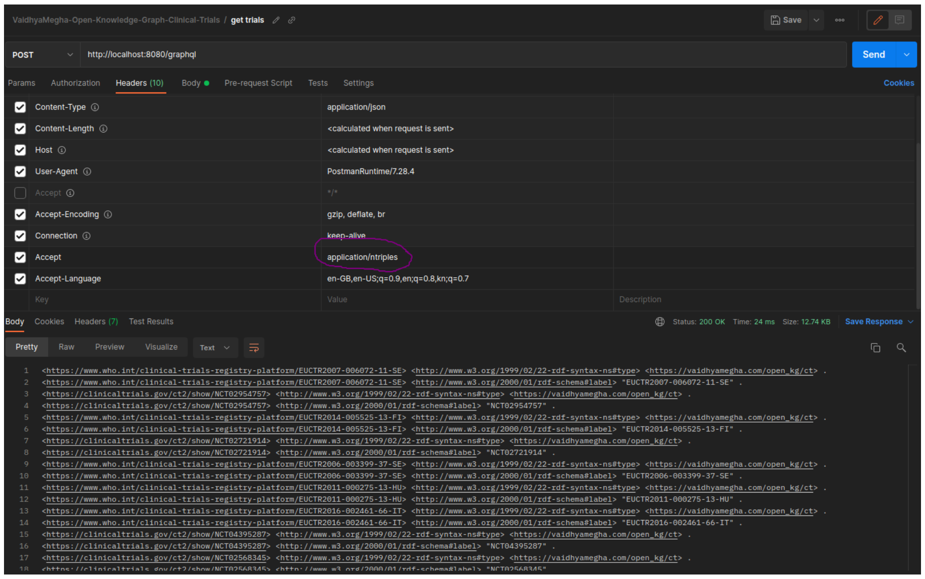
- With ntriples response

- With json response

Start client

In a separate terminal execute GraphQL query using curl (alternatively use Postman)

```
$ curl --location --request POST 'http://localhost:8080/graphql' \
    --header 'Accept: application/ntriples' \
    --header 'Accept-Language: en-GB,en-US;q=0.9,en;q=0.8,kn;q=0.7' \
    --header 'Content-Type: application/json' \
    --data-raw \
    '{"query": "{\n trial_GET(limit: 30, offset: 1) {\n label\n }\n \n}","variables":{}}'
```



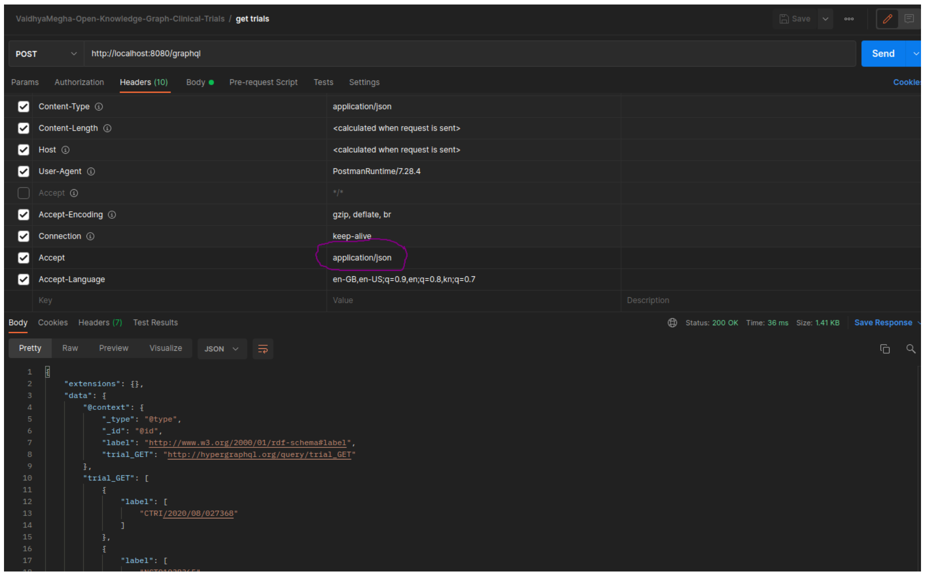
http://localhost:8080/graphql

✓ Content-Type	application/json	
✓ Content-Length	<calculated when request is sent>	
✓ Host	<calculated when request is sent>	
✓ User-Agent	PostmanRuntime/7.28.4	
Accept	*/*	
✓ Accept-Encoding	gzip, deflate, br	
✓ Connection	keep-alive	
✓ Accept	application/ntriples	
✓ Accept-Language	en-GB,en-US;q=0.9,en;q=0.8,kn;q=0.7	
Key	Value	Description

Status: 200 OK Time: 24 ms Size: 12.74 KB

```
<https://www.who.int/clinical-trials-registry-platform/EUCTR2007-006072-11-SE> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2007-006072-11-SE> <http://www.w3.org/2000/01/rdf-schema#label> "EUCTR2007-006072-11-SE" .
<https://clinicaltrials.gov/ct2/show/NCT02954757> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT02954757> <http://www.w3.org/2000/01/rdf-schema#label> "NCT02954757" .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2014-005525-13-FI> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2014-005525-13-FI> <http://www.w3.org/2000/01/rdf-schema#label> "EUCTR2014-005525-13-FI" .
<https://clinicaltrials.gov/ct2/show/NCT02721914> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT02721914> <http://www.w3.org/2000/01/rdf-schema#label> "NCT02721914" .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2006-003399-37-SE> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2006-003399-37-SE> <http://www.w3.org/2000/01/rdf-schema#label> "EUCTR2006-003399-37-SE" .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2011-000275-13-HU> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2011-000275-13-HU> <http://www.w3.org/2000/01/rdf-schema#label> "EUCTR2011-000275-13-HU" .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2016-002461-66-IT> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2016-002461-66-IT> <http://www.w3.org/2000/01/rdf-schema#label> "EUCTR2016-002461-66-IT" .
<https://clinicaltrials.gov/ct2/show/NCT04395287> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT04395287> <http://www.w3.org/2000/01/rdf-schema#label> "NCT04395287" .
<https://clinicaltrials.gov/ct2/show/NCT02568345> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT02568345> <http://www.w3.org/2000/01/rdf-schema#label> "NCT02568345"
```

Figure 2: GraphQL - NTriples response



http://localhost:8080/graphql

✓ Content-Type	application/json	
✓ Content-Length	<calculated when request is sent>	
✓ Host	<calculated when request is sent>	
✓ User-Agent	PostmanRuntime/7.28.4	
Accept	*/*	
✓ Accept-Encoding	gzip, deflate, br	
✓ Connection	keep-alive	
✓ Accept	application/json	
✓ Accept-Language	en-GB,en-US;q=0.9,en;q=0.8,kn;q=0.7	
Key	Value	Description

Status: 200 OK Time: 36 ms Size: 1.41 KB

```
{
    "extensions": {},
    "data": {
        "@context": {
            "_type": "@type",
            "_id": "@id",
            "label": "http://www.w3.org/2000/01/rdf-schema#label",
            "trial_GET": "http://hypergraphql.org/query/trial_GET"
        },
        "trial_GET": [
            {
                "label": [
                    "CTRI/2020/08/027368"
                ]
            },
            {
                "label": [
```

Figure 3: GraphQL - JSON response

```
<https://www.who.int/clinical-trials-registry-platform/EUCTR2007-006072-11-SE>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2007-006072-11-SE>
    <http://www.w3.org/2000/01/rdf-schema#label>
    "EUCTR2007-006072-11-SE"^^<http://www.w3.org/2001/XMLSchema#string> .
<https://clinicaltrials.gov/ct2/show/NCT02954757>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT02954757>
    <http://www.w3.org/2000/01/rdf-schema#label>
    "NCT02954757"^^<http://www.w3.org/2001/XMLSchema#string> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2014-005525-13-FI>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <https://vaidhyamegha.com/open_kg/ct> .
<https://www.who.int/clinical-trials-registry-platform/EUCTR2014-005525-13-FI>
    <http://www.w3.org/2000/01/rdf-schema#label>
    "EUCTR2014-005525-13-FI"^^<http://www.w3.org/2001/XMLSchema#string> .
<https://clinicaltrials.gov/ct2/show/NCT02721914>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <https://vaidhyamegha.com/open_kg/ct> .
<https://clinicaltrials.gov/ct2/show/NCT02721914>
    <http://www.w3.org/2000/01/rdf-schema#label>
    "NCT02721914"^^<http://www.w3.org/2001/XMLSchema#string> .
...
<http://hypergraphql.org/query> <http://hypergraphql.org/query/trial_GET>
    <https://www.who.int/clinical-trials-registry-platform/EUCTR2016-002461-66-IT> .
<http://hypergraphql.org/query> <http://hypergraphql.org/query/trial_GET>
    <https://www.who.int/clinical-trials-registry-platform/CTRI/2020/08/027368> .
<http://hypergraphql.org/query> <http://hypergraphql.org/query/trial_GET>
    <https://www.who.int/clinical-trials-registry-platform/EUCTR2013-001294-24-DE> .
```

Discussions

Acknowledgements

Declarations

Appendix

Tables

Web resources

Knowledge graph

- Introduction to knowledge graphs

MeSH

- Tree view
- Record view along with 'Mesh Tree Structures'

PheGenI

- What is PheGenI
- PheGenI: The Phenotype-Genotype Integrator demo
- Downstream analysis of PheGenI results demo

Apache Jena

- Download
- Read RDF files
- Write RDF files
- Querying a model
- SparQL
- index RDF files
- query RDF files using indexes
- count triples query
- Execute SparQL query programmatically

GraphQL

- Bridges between GraphQL and RDF

Java

- Configure Java heap size and if needed stack size.

- Invoke Entrez API using Spring webclient
  - Handling XML response
- JAXB impl in Java 11
- JAXB marshalling objects in lists
- Java multiple resources with autocloseable try-with-resources
- escape pipe character in grep
- Other older relevant links on Java and Entrez
  - download-pubmed-abstracts-in-java
  - Github repo : pubmed_ws_client
- Read file into a string

Linux commands

- Check if string exists in file in bash
- Prevent grep from exiting when match is not found

PostgreSQL

- PostgreSQL array columns
  - JDBC insert into array columns
- PostgreSQL date column with default value
- PostgreSQL upsert statement
- PostgreSQL - pg_restore - restore only one selected schema
- PostgreSQL - Array functions
- Execute query on PostgreSQL using psql non-interactively
- Save psql inline query output to a file
- In PostgreSQL formulate a query to get all items in an array column

Postman

- Tutorial for using GraphQL with Postman

References

[1] Carroll, J.J. et al. 2004. Jena: Implementing the semantic web recommendations. *Proceedings of the 13th international world wide web conference on alternate track papers & posters* (New York, NY, USA, 2004), 74–83.

[2] Grobe, M. 2009. RDF, jena, sparql and the 'semantic web'. *Proceedings of the 37th annual acm siguccs fall conference: Communication and collaboration* (New York, NY, USA, 2009), 131–138.

[3] Hogan, A. et al. 2021. Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge.* 12, 2 (2021), 1–257.