A SPARQL Endpoint Profiler for Efficient Question Answering Systems

Yasunori Yamamoto
Database Center for Life Science (DBCLS)
Various SPARQL Endpoints
You can get any data necessarily and sufficiently from them...?
A SPARQL editor you might often see...

Almost nothing here...

Use the force
We’d like to make this happen.
Question Answering

LODQA (Linked Open Data Question Answering) is an open source project aiming at developing a system to generate SPARQL queries from natural language.

- A tutorial on natural language interfaces for SPARQL endpoint for SWAT4LS in Berlin (Dec. 9, 2014)
- NLIWoD 2014 workshop was collocated with ISWC2014 in Trentino (Oct. 19, 2014).
- OKBQA 2014 hackathon was held in Jeju (Feb. 17-21, 2014).
- LODQA launched as an open source project (Dec. 20, 2013).

System demo

Configure

Choose a configuration

- QALD-Biomed

It targets DrugBank, SIDER, and DisGeNet as included in QALD.

Query

Enter a natural language query.

Sample queries

- what genes are associated with alzheimer disease?
- what drugs are used for diseases associated with the gene ALD?
“What genes are associated with Alzheimer disease?”

```
SELECT ?t1
WHERE {
  ?t1 [:isa] [genes] .
  ?t2 [:isa] [alzheimer disease] .
}
```

Mapping to existing vocabularies

- [genes] = a [gene] class
- [alzheimer disease] = instance of a [disease] class

Out of this study’s focus

PubDictionary / identifiers.org
Example: BIO2RDF

[gene] Class = bio2rdf:omim_vocabulary:Gene
[disease] Class = bio2rdf:umls_vocabulary:Resource
[disease] Class = bio2rdf:omim_vocabulary:Phenotype

A QA system wants to know if there is any path between them.
Finding EPs that can compose a path between a [gene] class and [alzheimer disease].

gene/omim/umls?

Naive method

Slow and inefficient
Our approach

Previously gathers metadata of each endpoint

QA System

Metadata

SPARQL Endpoint

SPARQL Endpoint

SPARQL Endpoint

SPARQL Endpoint
Effective and efficient
Ideal way

Each endpoint provides its metadata

QA System

Metadata

SPARQL Endpoint

Metadata

SPARQL Endpoint

Metadata

SPARQL Endpoint

Metadata

SPARQL Endpoint

Metadata
VoID SD

Yes, they’re fine!

Vocabulary of Interlinked Datasets
SPARQL 1.1 Service Description
But, we cannot get...

Data about relationships between classes

Gene: Cap9, APCS, PRNP

Phenotype: EEC SYNDROME I, TOXEMIA OF PREGNANCY, DIABETES
SPARQL Builder Metadata

Extensions to VoID / SD
Obtain all the relationships
Predicate first vs. Class first
<table>
<thead>
<tr>
<th>Criterion</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>1,976.29</td>
<td>1</td>
<td>50,303</td>
<td>5.0</td>
<td>509,884</td>
</tr>
<tr>
<td>Properties</td>
<td>33.48</td>
<td>1</td>
<td>885</td>
<td>21.0</td>
<td>9,608</td>
</tr>
</tbody>
</table>

http://stats.lod2.eu/stats
Obtain metadata using SPARQL

A locally declared class

any resource that appears as an object of rdf:type in the dataset.

:sycamore rdf:type :Tree .

A locally undeclared class (LUC) instance

any resource that does not appear as an subject of rdf:type in the dataset.
Obtain metadata using SPARQL

```
SELECT ?p (COUNT(?p) AS ?rc)
WHERE {
  ?f a :Class_A .
  ?t a :Class_B .
}
GROUP BY ?p
```
Improvement

$N^2$ times ($N$: number of the classes)
Even a dataset of many classes has only a few relationships.
→ inefficient

SELECT ?p ?c (COUNT(?p) AS ?pc)
WHERE {
  ?f a :Class_A ;
  ?p ?t ;
  !rdf:type ?t .
  ?t a ?c .
}
GROUP BY ?p ?c
→ $N$ times
Gathering and providing metadata

Triple Data Profiler Repository

- About Triple Data Profiler (Japanese)
- SBM vocabulary information
- Source Code

Enter SPARQL Query

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sd: <http://www.w3.org/ns/sparql-service-description#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sbm: <http://www.sparqlbuilder.org/2014/05/rdf-metadata-schema#>

```sparql
SELECT * {{ ?siendpoint ?j;
  ?sd:defaultDataset/edm:NamedGraph/edm:name ?g}}
```

Example Queries

- Query 1
  
  Find any classes that have the string "gene".

Results per page: 50

Output: HTML

Generated by the LODStar linked data browser from the Functional Genomics Production Team (FGPT)
Searching example

[gene] Class = bio2rdf:omim_vocabulary:Gene
[disease] Class = bio2rdf:umls_vocabulary:Resource

OR
From [disease]

From [gene]

<table>
<thead>
<tr>
<th>Count</th>
<th>Predicate</th>
<th>Object class</th>
</tr>
</thead>
<tbody>
<tr>
<td>608220</td>
<td>bio2rdf-omim:x-gi</td>
<td>bio2rdf-gi:Resource</td>
</tr>
<tr>
<td>98639</td>
<td>bio2rdf-omim:article</td>
<td>bio2rdf-pubmed:Resource</td>
</tr>
<tr>
<td>95443</td>
<td>bio2rdf-omim:refers-to</td>
<td>bio2rdf-omim:Resource</td>
</tr>
<tr>
<td>81438</td>
<td>bio2rdf-omim:vocabulary:refers-to</td>
<td>bio2rdf-omim:Gene</td>
</tr>
<tr>
<td>28032</td>
<td>bio2rdf-omim:vocabulary:gene-symbol</td>
<td>bio2rdf-hgnc:Resource</td>
</tr>
</tbody>
</table>
Things to do

To improve inter-operability
Integrating vocabularies with other related parties.

To improve efficiency
Calling for data providers to provide metadata.

To realize a QA system using metadata
Discussing with LODQA developers.
Thank you!