FAIR Service Descriptions: enriching life science SPARQL endpoints

Jerven Bolleman[1,*], Alan Bridge[1] and Nicole Redaschi[1]

[1]SIB Swiss Institute of Bioinformatics

Abstract

SPARQL service descriptions allow for rich information schemas describing the data inside SPARQL endpoints. Rewriting information schema (re)-discovery queries to queries using an existing one can give major performance benefits. Rich service descriptions have many use cases beyond query rewriting.

Keywords

SPARQL, RDF, Information schema, Query rewriting

A significant challenge for users of SPARQL endpoints is discovering the shape and quantity of the data exposed inside them. The W3C standards for SPARQL allow for a Service Description (SD), enumerating the capabilities and capacities of SPARQL endpoint. The Swiss-Prot group provides extensive service descriptions for it's SPARQL endpoints: (https://hamap.expasy.org/sparql, https://beta.swisslipids.org/sparql,https://sparql.rhea-db.org/sparql and https://sparql.uniprot.org/sparql).

A SD contain metadata about a SPARQL endpoint, such as when it was updated and which ontologies it uses. Such a SD can be seen as an information schema for a SPARQL endpoint. Using the Service Description [1], VoID [2] and VoID-Ext [3] vocabularies. We store these in in-dependant named graphs, which we always name as address of the SPARQL endpoint + ./well-known/void. e.g. https://sparql.rhea-db.org/.well-known/void.

FAIR SDs have many use cases, such as:

- Query optimization and dataset visualizations. The tool SPEX which generates entity relationship diagrams uses these in part if they are available.
- Generating ShACL files describing the shape of the data in a SPARQL endpoint.
- Generate APIs in languages such as R or Python to access the data in the SPARQL endpoint. To be demonstrated in the CHIST-ERA: Open Research Data - TRIPLE project.
- License and last updated information for FAIR data monitors.

As an example: a common SPARQL query people are thought to use is to discover how many distinct classes there are in a SPARQL endpoint shown in listing:1. For large datasets like UniProt this is a non-trivial. Imagine running it as a classical unix pipeline like listing:2. Then be surprised that this takes a few days to run if you have enough disk space and memory that

SWAT4HCLS 2024: Bridging Life Sciences and Technology, February 26-29, Leiden, The Netherlands

*Corresponding author.

jerven.bolleman@sib.swiss (J. Bolleman); alan.bridge@sib.swiss (A. Bridge); nicole.redaschi@sib.swiss (N. Redaschi)

0000-0002-7449-1266 (J. Bolleman); 0000-0003-2148-9135 (A. Bridge); 0000-0001-8890-2268 (N. Redaschi)

is. This is because there are more than 140 billion distinct triples in UniProt. Of course having such a SD is not enough as the people who are used to using such queries won't change to use a different query on an "information schema" by default. This means we need to rewrite the query (listing:1) to a query in the form of (listing:3). Query rewriting needs to take into account variations in prefix, white-space and variable naming. We solve this by using a SPARQL parser from the RDF4j project use the abstract SPARQL algebra for the query matching and rewrite. The original query with is redirected to a new location with a new query (http 301).

Listing 1: "Count distinct classes used in a SPARQL endpoint."

```
SELECT (COUNT(DISTINCT ?class) AS ?classes)
WHERE { ?subject a ?class . }
```

Listing 2: "Simple pipeline to count the unique classes in an ntriples file."

```
sort -u all_triples_in_uniprot.nt | grep rdf:type | sort -u |
    wc -l
```

Listing 3: "Rewritten SPARQL query to retrieve the count of the distinct classes in the endpoint."

```
SELECT (COUNT(DISTINCT ?classesRaw) AS ?classes)
FROM <http://sparql.uniprot.org/.well-known/void>
WHERE { [] <http://rdfs.org/ns/void#class> ?classesRaw . }
```

Acknowledgments

The Swiss-Prot group is part of the SIB Swiss Institute of Bioinformatics and of the UniProt Consortium. Swiss-Prot group activities are supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI and UniProt is supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Institute of Mental Health (NIMH), and National Cancer Institute (NCI) of the National Institutes of Health (NIH) under grant U24HG007822.

References

[1] Sparql 1.1 service description, 2013. URL: https://www.w3.org/TR/sparql11-service-description/.

[2] M. H. J. Z. Keith Alexander, Richard Cyganiak, Describing linked datasets with the void vocabulary, 2011. URL: https://www.w3.org/TR/void/.

[3] E. Mäkelä, Aether – generating and viewing extended void statistical descriptions of rdf datasets, in: V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, A. Tordai (Eds.), The Semantic Web: ESWC 2014 Satellite Events, Springer International Publishing, Cham, 2014, pp. 429–433.