Make RDF data more inter-connectable*

Yasunori Yamamoto[1,*,†], Takatomo Fujisawa[2,‡]

[1]Database Center for Life Science, ROIS-DS, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, JAPAN
[2]National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, JAPAN

Abstract

RDF data show their values the most when built in a distributed manner and linked to each other from several aspects with URIs as their keys. However, we have seen several URI mismatches across RDF datasets that should be identical such as the cases of using different prefixes and code systems. In this situation, we need to develop an infrastructure in which these URIs are treated identically by using an URI rewriting dictionary constructed to be tailored to each RDF dataset. Here, we show some examples of these synonymous URIs and propose an architecture to rewrite some URIs when retrieving RDF data from multiple SPARQL endpoints. As a result, users can obtain properties as to a consolidated URI, which otherwise get ones explicitly asserted as triples only.

Keywords

RDF, Web of Data, Data curation

1. Introduction

Several works to represent huge and diverse life science data in the Resource Description Framework (RDF) have emerged since the late 2000s, and the number of newly built RDF data is increasing even now. Currently, 65 SPARQL endpoints are listed at the Umaka-Yummy Data[1] where you can learn the status of each endpoint such as how stable it is, how fast it returns a result, and so on. RDF performs at its maximum potential when each URI denotes one concept and vice versa, since a URI is a global identifier. Multiple RDF datasets built in a distributed manner can be easily joined if this is true. However, there are several URI discrepancies among them. First of all, there are some typos and misprints within a dataset, such as the following:

- `http://www.w3.org/2000/01/rdf-schema#Label`
- `http://www.w3.org/2000/01/rdfschema#label`

These issues can be taken care by *RDF-doctor*[2] which we have developed. Secondly, We have seen several synonymous URI cases including the following examples.

- `http://identifiers.org/taxonomy/9606`
- `http://purl.obolibrary.org/obo/NCBITaxon_9606`

15th International SWAT4HCLS Conference, Feb 26 – 29, 2024, Leiden, The Netherlands
*Corresponding author.
†These authors contributed equally.
yy@dbcls.rois.ac.jp (Y. Yamamoto); tf@nig.ac.jp (T. Fujisawa)
https://researchmap.jp/yayamamo (Y. Yamamoto); https://researchmap.jp/takatomo (T. Fujisawa)
0000-0002-6943-6887 (Y. Yamamoto); 0000-0001-8978-3344 (T. Fujisawa)

[1]https://yummydata.org/
[2]https://pypi.org/project/rdf-doctor/

- `http://rdf.ncbi.nlm.nih.gov/pubchem/taxonomy/TAXID9606`
- `http://www.ncbi.nlm.nih.gov/taxonomy/9606`
- `http://purl.org/obo/owl/NCBITaxon#9606`
- `http://mbgd.genome.ad.jp/rdf/resource/organism/hsa`

All of these URIs denote *Homo sapiens*. We consider this issue to be due to the nature of a distributed way of building RDF datasets. Multiple groups and institutions are involved in building. Therefore, in addition to calling community's attention, we need to construct an infrastructure to minimize these mismatches as much as possible with the help of machines. Here, we propose an infrastructure where synonymous URIs are treated as identical. While there are already related works such as *sameAs*[3], *Identifiers.org*[4], and TogoID[5], there is no attempt to date that aims at providing consolidated results by rewriting URIs in the life science domain.

2. URI consolidation

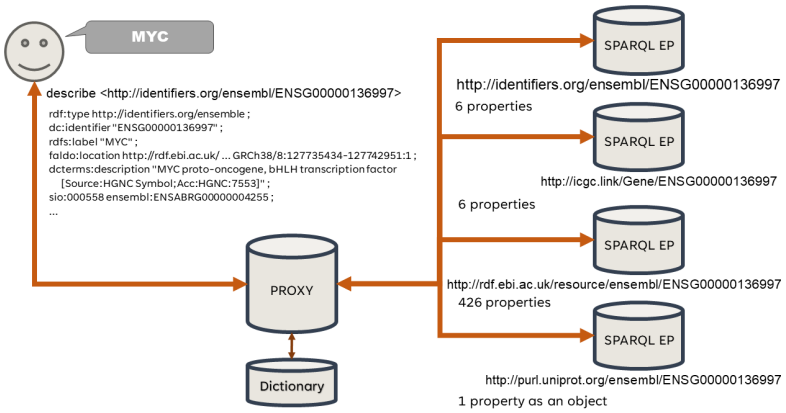

Figure 1: Overall architecture of RDF consolication infrastructure

Figure 1 describes an overall architecture of RDF consolidation infrastructure. Here, we assume that there are multiple datasets and SPARQL endpoints, where a dataset means a RDF graph. We call a consolidating system *proxy*, which looks up a rewriting dictionary to see if a given URI is in it and, if any, rewrites it into its corresponding one for a pair of an endpoint URI and a graph name. Then the proxy issues a query for each endpoint, and shows the consolidated results.

Acknowledgments

This work was supported under the Life Science Database Integration Project, NBDC of Japan Science and Technology Agency.

[3]http://sameas.org/
[4]https://identifiers.org/
[5]https://togoid.dbcls.jp/