Finding RDF data you need by Umaka Suite

Yasunori Yamamoto and Atsuko Yamaguchi

Database Center for Life Science,
Joint Support-Center for Data Science Research, Research Organization of Information and Systems
178-4-4, Wakashiba, Kashiwa, Chiba 277-0871, Japan
{yy,atsuko}@dbcls.rois.ac.jp

Abstract. Umaka suite consists of three tools for RDF data consumers to find the best RDF data of their interests. One is to search for SPARQL endpoints relevant to given keywords. Second is to find an endpoint that provides reliable data. Third is to learn a data structure of an endpoint. These are our solution proposal to issues of hindering further propagation of Linked Open Data in Life Sciences.

Keywords: RDF data discovery, RDF data use

1 Introduction

Semantic Web technology has been adopted in Life Sciences since its early stage, and lots of works have been done to ease the burden of utilizing heterogeneous datasets in an integrated manner. Thanks to these efforts, we can find the designated data easier more than ever by using SPARQL queries over multiple SPARQL endpoints (we call them just endpoints hereafter). Even though learning SPARQL may not be easy, once done it you can search for any datasets through endpoints. The issues are to find right endpoints that provide the designated data. In addition, some endpoints have datasets that are similar to each other. In this situation, we want to access the endpoint that is more reliable. Even if one can find the right endpoint, we want to learn the data structure or schema quickly enough to issue SPARQL queries to retrieve the designated data.

CEUR-WS.org/Vol-2849/paper-29.pdf

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Umaka Suite

The Umaka Suite is our solution to these issues. It consists of three tools: Umaka Search, Umaka-YummyData, and Umaka Viewer. We briefly introduce them.

Umaka Search enables us to search for right endpoints. We issue keywords to it, which returns a list of endpoint URLs relevant to them. This service is currently under development, and we are releasing its alpha version within the next year. A related service is Datao[1], but the source code is not open, hence we cannot tailor it to our purposes.

Umaka-YummyData[1] is a service to find reliable endpoints and facilitate mutual understandings between data providers and data consumers. Umaka-YummyData introduces Umaka score to quantify a reliability of an endpoint. The score is based on several aspects such as update frequency, query processing speed, running history, ontology usage, and so on. While we do not consider that Umaka score is the only index to evaluate an endpoint, it can be a reference to choose one. In addition, we believe that it can be a trigger to begin communication between data providers and data consumers.

Umaka Viewer[2] shows us a graphical representation of data structures of a given RDF dataset. Data structure here means class hierarchies along with a predicate list and statistic data such as the numbers of triples. Umaka Viewer provides an interactive GUI, and we can zoom-in and zoom-out to learn the class hierarchies. In addition, we can learn which predicate links what classes.

3 Future Plans

We intend to release Umaka Search, which covers as many endpoints in life sciences as possible. As obtaining an entire RDF dataset that an endpoint serves is inappropriate for the endpoint, we try to index the RDF dataset that can be bulk downloadable.

References

1. Yasunori Yamamoto, Atsuko Yamaguchi, Andrea Splendiani. "YummyData: providing high-quality openlife science data", Database, 2018

[1] http://search.datao.net/

[2] https://umaka-viewer.dbcls.jp/