UmakaData extension: Toward Realization of a Practical SPARQL Endpoint Discovery Service for Life Sciences

Norio Kobayashi*[1], Yasunori Yamamoto*[2], and Atsuko Yamaguchi[2]

[1] Head Office for Information Systems and Cybersecurity (ISC), RIKEN,
2-1 Hirosawa, Wako, Saitama, 351-0198 Japan
norio.kobayashi@riken.jp
[2] 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. UmakaData shows a list of SPARQL endpoints that provide life science data with reliability scores, called Umaka scores, concerned with properties such as data freshness, accessibility, and performance. UmakaData monitors 72 SPARQL endpoints and scores these endpoints by executing SPARQL queries daily. Recently, in order to realize a class and property catalogue service for each endpoint that helps users write suitable SPARQL queries, an RDF data schema explorer called LOD Surfer crawler accessed SPARQL endpoints that were ranked in the top 50 for Umaka scores. This poster presents our current progress on the Umaka data service and its recent extension.

Keywords: SPARQL endpoint discovery, endpoint federation, RDF data quality check

1 Introduction

Discovering SPARQL endpoints publishing RDF data that is suitable for a user's data analysis is an essential function. In the life sciences, since a wide variety of RDF data is published having classes and properties defined by various ontologies, SPARQL endpoint discovery is a difficult task. In particular, when writing a federated search query, a user may find that classes and properties have different URIs even though the URIs should be the same among SPARQL endpoints. However, checking whether there are differences in classes and properties for each instance is generally quite an expensive task. In order to solve these problems comprehensively, we introduce upper level ontologies by extracting at most several hundred classes from a single ontology having various kinds of classes.

Another problem is practical availability of SPARQL endpoints. In order to address this problem, we have already developed a service called 'UmakaData' [1]

* These two authors contributed equally to this work.

that shows a list of life science SPARQL endpoints and their properties, including availability, performance and data freshness. Our current issue is the selection of the best properties and computational method for a ranking score that reflects users' practical data analysis. This poster reports our trial extension of UmakaData to address the issues described above.

2 UmakaData extension with detail SPARQL endpoint metadata

The UmakaData currently provides endpoint metadata to both RDF data consumers and providers for their mutual understanding. These metadata include their running history, update information, processing speed, support for the four principals of Linked Data, and usage of ontologies that are well known or more common in life science RDF data. In addition, since UmakaData also obtains inter-endpoint relationships, it can provide information on links between RDF data of any pair of SPARQL endpoints. Therefore, UmakaData could provide relationships among classes and properties once it finds a triple which has owl:sameAs as its predicate or any classes whose instance's URI is identical over SPARQL endpoints.

Furthermore, in order to achieve more powerful class-discovering functionality when writting a SPARQL query, we have been working on an extension of Umaka metadata by introducing LOD Surfer[1] metadata that describes the LOD graph structure of a SPARQL endpoint including class-class relationships with statistics including numbers of triples and instances. Since a single instance may relate to different concept classes among different SPARQL endpoints, we introduce upper-level conceptual classes using a part of public ontology that covers wide and deep concepts. For the SPARQL endpoints ranked in the top 50 Umaka scores, the LOD Surfer metadata crawler was executed to extract upper-level concepts. This resulted in the selection of the top 114 Medical Subject Headings and 42 semanticscience integrated ontology concepts associated with 2,724 and 1,133 concepts among 35,248 concepts extracted from the 50 SPARQL endpoints without crawler error.

Our future work will include periodical execution of the LOD Surfer metadata crawler, its tuning to reduce computational complexity, introduction of other upper-level ontologies, and evaluation of the effectiveness of our extended UmakaData metadata using practical applications such as the LOD Surfer.

Acknowledgements

This work has been supported by JSPS KAKENHI grant numbers 17K00434, 17K00424 and 18K19766.

References

1. Yamamoto, Y., Yamaguchi, A., and Splendiani, A.: YummyData: providing high-quality open life science data. Database, Vol. 2018, bay022, 2018.

[1] `http://github.com/LODSurfer/lodsurfer-metadata`