LOD Surfer Metadata: Essential LOD catalogue data for traversing life-science LOD amongst multiple SPARQL endpoints

Norio Kobayashi[1] and Yasunori Yamamoto[2]

[1] Head Office for Inrormation 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@dbcls.rois.ac.jp

Abstract. Researchers in the life sciences are advanced developers and publishers of Linked Open Data (LOD). Because the datasets generated by various apparatuses and experimental methods are heterogeneous and diverse, LOD descriptions involve not only widely and commonly used but also uses original data classes and ontology terms. To resolve this problem, we have revised the currently proposed LOD Surfer Metadata, which describe detailed LOD graph structures such as class–class relationships and statistics, by introducing upper-level conceptual classes based on Medical Subject Headings (MeSH) terms. With these descriptions, users can easily discover and integrate multiple LOD over classes of different ontologies. Furthermore, the availability and frequency of data updates can be checked by a public SPARQL endpoint monitoring service. Accordingly, we have been developing a practical LOD-discovery service for integrative data analysis such as SPARQL federate searching. This poster presents our developmental progress on the LOD Surfer Metadata and LOD discovery service.

Keywords: LOD Surfer Metadata, dataset catalogue, LOD discovery service, class–class relationship, life-science data integration

1 Introduction

Researchers in the life sciences have used Linked Open Data (LOD) to publish their research results and to integrate multiple datasets. As the life-science data described as LOD are highly heterogeneous, the number of ontologies is increasing and similar concepts exist in different ontologies. This situation inhibits the discovery of target data and their integrated analysis over different LOD datasets, and (together with other database utilisation problems) necessitates an LOD data catalogue and an LOD discovery service. To this end, we

have been developing LOD Surfer Metadata as a catalogue of LOD datasets published through SPARQL endpoints. Originally, this metadata was defined for a SPARQL query building service called LOD Surfer. This poster presents the details of the LOD Surfer Metadata and their implementation in a LOD discovery service.

2 LOD Surfer Metadata

The LOD Surfer tool discovers linked connections amongst life-science LOD published at different SPARQL endpoints by traversing a path of class–class relationships. LOD Surfer Metadata briefly describe the corresponding LOD graph structure of a SPARQL endpoint, and extract the class–class relationships along with their statistics (such as numbers of triples and instances) for analysing the comprehensiveness of a dataset. However, when implementing LOD Surfer, several problems with LOD Surfer Metadata emerge: (1) SPARQL queries for the statistical data of huge LOD require huge computational cost, and the statistics do not necessarily improve the LOD Surfer performance, (2) a single instance can relate to different concept classes amongst different SPARQL endpoints, and (3) one cannot track the availability and data update rates of SPARQL endpoints. To resolve (1), we introduced description levels depending on the crawling cost of each SPARQL endpoint, where level 1 is equivalent to a VoID (`https://www.w3.org/TR/void/`) description, level 2 describes simple class–class relationships, and level 3 describes detailed class–class relationships with statistics. To resolve (2), we have been introducing upper-level conceptual classes using part of the Medical Subject Heading (MeSH) terms, which broadly categorise the particular ontology terms in the LOD datasets. Finally, problem (3) will be resolved by developing an LOD discovery service, as described in the next section.

3 LOD discovery service

The revised LOD Surfer Metadata include essential data for discovering LOD datasets and their associated SPARQL endpoints, enabling users to search class and class–class relationships using broadly conceptual keywords taken from MeSH terms. In contrast, the availability, performance and freshness of the SPARQL endpoints collected and curated by SPARQL endpoint monitoring services, such as YummyData (`https://yammydata.org`), are available for practical use. We have just begun developing a novel LOD discovery service that integrates LOD Surfer Metadata with YummyData, providing the advantages of both systems. Preliminary, we have selected 35 popular SPARQL endpoints with YummyData scores above 50, and have begun crawling these SPARQL endpoints periodically.

Acknowledgements

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