RDF-Based Integration with SPARQL Building System for Life Science Database Archive

Atsuko Yamaguchi[1], Katsuhiko Okubo[2], Norio Kobayashi[3], Kouji Kozaki[4], Sadahiro Kumagai[2], Kai Lenz[3], Tomoe Nobusada[5], Hongyan Wu[1], Yasunori Yamamoto[1], and Hideki Hatanaka[5]

[1] Database Center for Life Science (DBCLS), ROIS,
178-4-4 Wakashiba, Kashiwa, Chiba, 277-0871 Japan
{atsuko, wu, yy}@dbcls.rois.ac.jp
[2] Information & Telecommunication Systems Company, Hitachi Ltd,
6-26-2 Minami Oi, Shinagawa-ku, Tokyo 140-8573 Japan
{katsuhiko.okubo.yh, sadahiro.kumagai.jj}@hitachi.com
[3] Advanced Center for Computing and Communication (ACCC), RIKEN,
2-1 Hirosawa, Wako, Saitama, 351-0198 Japan
{norio.kobayashi, kai.lenz}@riken.jp
[4] The Institute of Scientific and Industrial Research (ISIR), Osaka University,
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047 Japan
kozaki@ei.sanken.osaka-u.ac.jp
[5] National Bioscience Database Center, JST,
5-3, Science Plaza 7F, Yonbancho, Chiyoda-ku, Tokyo 102-8666 Japan
{nobusada, hideki}@biosciencedbc.jp

Abstract. The Life Science Database Archive (LSDB Archive, https://dbarchive.biosciencedbc.jp/) is a service to collect, preserve and provide databases generated by life-science researchers in Japan. As of September 2015, the LSDB Archive includes 103 databases and all the databases can be downloadable with appropriate licenses and metadata. Although a simple keyword search tool is available for the databases, more flexible retrieval system to obtain relevant data from heterogeneous databases is required. Therefore, we first converted the databases into RDF datasets, uploaded in a triple store. Then, we developed a prototype of retrieval system using SPARQL Builder. Because SPARQL Builder assists users in writing queries, the prototype enables users without knowledge of RDF to access the datasets.

Keywords: SPARQL, RDF, database archive, life-science databases

1 Introduction

In life sciences, many kinds of data have been generated as results of experimental research. Although they are often organized and provided as databases, we found that many databases in Japan were not maintained anymore after fundings of projects end as of 2006 [1]. Even in case that databases are maintained, because they were unclear with respect to the terms of use, or were not

downloadable, they may not be fully used. To address these issues, the Life Science Database Archive (LSDB Archive, https://dbarchive.biosciencedbc.jp), a service to maintain, store and provide downloadable databases with appropriate licences and unified descriptions, started from 2009. As of September 2015, the LSDB Archive includes 103 heterogeneous databases generated by life-science researchers during national projects in Japan.

Although a simple keyword search tool is available for the databases, more flexible retrieval system to obtain relevant data from heterogeneous databases is required. To do so, we decided to use semantic web technologies to integrate the databases. We converted the databases into RDF datasets, and uploaded in a triple store with a SPARQL endpoint as a trial. However, a SPARQL query construct is intractable to users who are unfamiliar with semantic web technologies although a SPARQL endpoint is very flexible retrieval system. As a system for assisting users to write SPARQL query, we employed SPARQL Builder [2] that is a semiautomatic SPARQL query generation system. Using this system, we developed a prototype of a search interface for LSDB Archive that enable users to extract semantically related data.

2 Method and Result

To generate initial RDF datasets from databases in LSDB Archive, we used TogoDB [3], which accepts tabular formatted data and generates RDF datasets. By indicating a class for each column in TogoDB, rdf:type for objects are automatically attached. To type subjects, we added an additional data including rdfs:domain for each property corresponding to a column. Because all the classes used in the RDF datasets are introduced from external ontologies, we extracted necessary part of those ontologies. Then we generated SPARQL Builder metadata for the RDF datasets together with extracted ontologies.

We then developed a prototype of a search interface on RDFized LSDB Archive using SPARQL Builder. Using this interface, by selecting two classes and a relationship between the two classes, users can search for desired data using SPARQL query from combined heterogenious RDF datasets in LSDB archive.

Acknowledgments. This work was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

1. Ministry of Education, Culture, Sports, Science, and Technology (MEXT) Integrated Database Project: http://lifesciencedb.mext.go.jp/en/
2. Yamaguchi A., Kozaki K., Lenz K., Wu H, Kobayashi N.: An Intelligent SPARQL Query Builder for Exploration of Various Life-science Databases, CEUR Workshop Proceedings 1279, The 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014), Riva del Garda, Italy.
3. TogoDB: http://togodb.org/