RIKEN Meta Database: a life-science metadata database based on the Semantic Web

Norio Kobayashi[1,2], Kai Lenz[1], and Hiroshi Masuya[2,1]

[1] Advanced Center for Computing and Communication (ACCC), RIKEN,
2-1 Hirosawa, Wako, Saitama, 351-0198 Japan
{norio.kobayashi, kai.lenz}@riken.jp
[2] BioResource Center (BRC), RIKEN,
3-1-1 Koyadai, Tsukuba, Ibaraki, 305-0074 Japan
hmasuya@brc.riken.jp

Abstract. We introduce a database platform called 'RIKEN Meta Database' that provides information on RIKEN's various life-science databases to help researchers around the world make full use of RIKEN's research results. Since RIKEN procures various and large life-sciences datasets through the genomes and phenomes of various species, including datasets for sequence and image data, such resultant research data and databases are heterogeneous and tend to be published in data-domain specific interfaces and formats. To achieve data re-usability, including facilitating data/database discovery and data description as life-science knowledge for further intelligent data analysis, the database platform is implemented to manage metadata of both data and database using Semantic Web technologies with standardised vocabularies and life-science ontologies. As of September 2015, 153 databases, including external public databases and 15 ontologies, are integrated on the platform, and these data can be access via a SPARQL endpoint and a browser-based graphical user interface that shows metadata in table and graph format.

Keywords: Semantic Web, life-science database platform, database profile

1 Overview

In recent life-science research, theses research fields have become deeper and more specialised, and collaboration among different research fields, such as informatics and material sciences, are actively promoted to comprehensively understand life phenomena. RIKEN is a comprehensive science institute with research centres and laboratories, including physics, chemistry, informatics and life sciences laboratories, that produces various and large data as individual databases. Even in RIKEN, which represents the science community in microcosm, such data and databases are difficult for researchers to find, and intelligent and integrated analyses has not been realised. To address this problem, we have been developing a database platform called the 'RIKEN Meta Database' (http://metadb.riken.jp) to facilitate the management and publication of database metadata according to the Semantic Web. The principal databases published by RIKEN have generated metadata that can be accessed by SPARQL and a browser-based graphical user interface (GUI).

2 System Architecture

The RIKEN Meta Database is designed to be a lightweight database platform. The core components of the platform are 1) a SPARQL endpoint, 2) a GUI web server that provides a GUI for data input and display and 3) a resource description framework (RDF) data converter. The SPARQL endpoint manages RDF data and is a server for a user's SPARQL client and the GUI web server. The GUI web server provides reader-friendly web pages for each database, class and entities in table or graph format by accessing the SPARQL endpoint. The web server also provides an interface that accepts RDF datasets for data uploads. The data converter is a Java application that accepts a Microsoft Excel file, which is converted to RDF format. This converter is used for wet researchers who are willing to publish their data but are not familiar with the RDF format. Currently, components 1) and 2) have been deployed on Linux virtual machines with a 1 GB flash memory drive having 96 GB and 8 GB memory, respectively.

3 Available data

As described, RIKEN is a comprehensive science institute that can perform data editing in close cooperation with biologists and informaticians. In such an environment, careful selection of public ontologies and data classes has been performed for each database. As of September 2015, with 15 public ontologies and 153 databases, including RIKEN's principal databases such as 'FANTOM' (mammalian [1]), 'FOX Hunting' (plant [2]) and 'Metadata of BRC cell resources' (bioresources [3]), which have 704 classes, 1,141 properties, 27 million entities and 102 million triples, are available on the platform. In addition, to assist users in easily finding data and databases, the platform provides the 'RIKEN Database Directory'. This directory provides a collection of database profiles, such as W3C's Health Care and Life Sciences description profile data (http://www.w3.org/TR/hcls-dataset/), including statistics data, and the SPARQL Builder Metadata (http://sparqlbuilder.org/), to find relationships between classes and corresponding triple paths.

References

1. The FANTOM Consortium and the RIKEN PMI and CLST (DGT): A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014)
2. Ichikawa T., Nakazawa M., Kawashima M., Iizumi H., Kuroda H., Kondou Y., Tsuhara Y., Suzuki K., Ishikawa A., Seki M., Fujita M., Motohashi R., Nagata N., Takagi T. Shinozaki K., Matsui M.: The FOX hunting system: an alternative gain-of-function gene hunting technique. Plant J. 45, 974–985 (2006)
3. Nakamura Y.: Bio-resource of human and animal-derived cell materials. Exp. Anim. 59(1), 1–7 (2010)