RIKEN MetaDatabase: a database platform as a microcosm of linked open data cloud in the life sciences

Norio Kobayashi[1,2,3], 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
[3] RIKEN CLST-JEOL Collaboration Center, RIKEN,
6-7-3 Minatojima-minamimachi, Chuo-ku, Kobe 650-0047, Japan

Abstract. Life sciences research produces numerous heterogeneous datasets at a rapid rate, and researchers in this field face serious difficulties in handling and publishing such datasets. These problems occur not only in the life sciences but also within a research institute such as RIKEN, the largest Japanese comprehensive science institute. To address this issue, we developed RIKEN MetaDatabase, which is a database platform built on our private cloud infrastructure and based on the Resource Description Framework (RDF). The platform was released in April 2015, and 110 databases including mammal, plant, bioresource and image databases with 21 ontologies have been published through this platform as of July 2016. This presentation focuses on the user interface and biological application examples of the RIKEN MetaDatabase.

Keywords: Semantic Web, database cloud platform, database integration, life sciences

1 Introduction

Recent developments in the life sciences produce a large number of heterogeneous datasets that are specialised for various life sciences subfields. This makes it difficult for researchers to discover, use and publish such datasets. To address these difficulties, the following major tasks are required: 1) the realisation of rich and useful data integration in a sustainable way and 2) the realisation of easy, flexible and low-cost operation to enable researchers to perform data integration tasks. These challenges also occur in the data management of RIKEN, Japan's largest comprehensive science institute, which generates large-scale life sciences datasets in various fields and is confronted with issues regarding the promotion of collaborative research promotion amongst different fields. This situation can be thought of as a microcosm of the linked open data cloud in the life sciences.

To solve these issues, we developed RIKEN MetaDatabase, which is a database platform based on the Resource Description Framework (RDF) that provides metadata management at low cost, systematic data integration and global publication on the Web. This presentation provides overview of the platform, focusing on the user interface and biological application examples.

2 Data model

Life sciences datasets, including the RIKEN databases, are often represented in tabular form or hosted by a relational database system. To realise a simple and user-friendly database platform, we restricted the RDF data handled by RIKEN MetaDatabase to tabular-type database data and tree-type ontology data.

A tabular-type database data describes the RDF data wherein all RDF resources are associated with an RDF class. A table is generated for each class of subject instances of RDF triplets and described in a spreadsheet, as shown in Fig. 1. In this spreadsheet, the second column comprises a list of instances of class 'Background strain', the third column includes a list of literal values of rdf:langString and the fourth column includes a list of 'Taxon' classes as instances of owl:Class. A spreadsheet can be converted into RDF using our application.

A tree-type ontology data comprises Web Ontology Language (OWL) ontology data, which represents the concepts and data classes of the database data using a conceptual hierarchy.

	1	2	3	4
1	English Attribution	Background strain	name	taxon
2	日本語属性	背景系統	名称	生物種
3	Property URI		rdfs:label	obo:RO_0002162
4	Data type	animal:0000004	rdf:langString	owl:Class
5		animal:0000004_7	"AIZ [Mus musculus molossinus]"@en	NCBITaxon:57486
6		animal:0000004_10	"AKT [Mus musculus musculus]"@en	NCBITaxon:39442
7		animal:0000004_12	"AST [Mus musculus musculus (wagneri)]"@en	NCBITaxon:39442
8		animal:0000004_23	"BFM/2 [Mus musculus domesticus]"@en	NCBITaxon:10092
9		animal:0000004_51	"Car [Mus caroli]"@en	NCBITaxon:10089

Fig. 1. Example spreadsheet describing the RDF data of class 'Background strain' (`http://metadb.riken.jp/db/rikenbrc_mouse/animal_0000004`) in tabular form.

3 Implementation

RIKEN MetaDatabase was implemented as a database platform comprising two components: (a) a web server providing both a graphical user interface (GUI) and an application programming interface (API) as a SPARQL endpoint and (b) an RDF triplet store, where each component is lightweight and performed on a virtual machine in a private cloud known as RIKEN Cloud Service (`http://cloudinfo.riken.jp`).

The distinctive features of the GUI are its table view and card view, which provide an intuitive RDF data view for a list of instances associated with a class and an instance, respectively, as shown in Fig. 2.

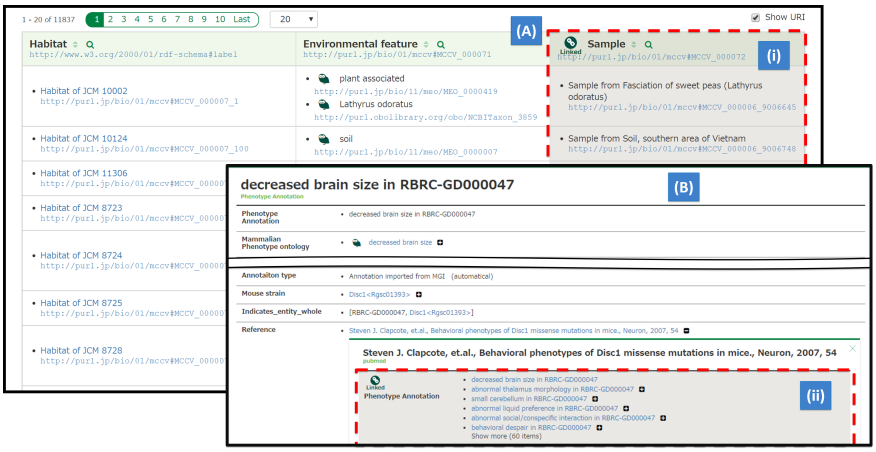

Fig. 2. Snapshots of (A) the tabular view of class 'Habitat' of the 'Japan Collection of Microorganisms (JCM) resource' database (`http://metadb.riken.jp/metadb/db/rikenbrc_jcm_microbe`) and (B) the card view of instance 'decreased brain size in RBRC-GD000047' of class 'Phenotype Annotation' of the 'Metadata of BRC mouse resources and phenotypes' database (`http://metadb.riken.jp/metadb/db/rikenbrc_mouse`). (i) and (ii) are reverse linked instances of the concerned instances, i.e., instances of class 'Habitat' and instance 'reference paper', respectively.

4 Integrated datasets and biological application examples

As of July 2016, 21 public ontologies, including Gene Ontology (GO) and Semantic-science Integrated Ontology (SIO) were selected and published as mirrors. These ontologies refer to 110 databases, of which 59 were RIKEN's original databases. The remaining 51 databases are external databases that are converted from original non-RDF databases and linked from RIKEN's databases. In total, RIKEN MetaDatabase comprises 148 million triplets, 797 classes, 2.94 million instances and 1,352 properties. The original databases pertain to various research fields, e.g., FANTOM (mammals [1]), FOX Hunting (plants [2]), Heavy-atom Database System (proteins [3]) and Metadata of the BioResource Center (BRC) resources (bioresources [4–6]).

More examples of datasets integrated over RIKEN and public datasets are summarised as follows. In the ENU-induced mutations in RIKEN Mutant Mouse Library, next-generation sequencing (NGS) metadata are described in a common RDF scheme, which was developed by a collaboration between DNA Data Bank of Japan (DDBJ) and RIKEN. The Japan Collection of Microorganisms (JCM)

resources (http://metadb.riken.jp/metadb/db/rikenbrc_jcm_microbe) are described using a common RDF scheme for strains of microorganisms known as Microbial Culture Collection Vocabulary (MCCV), which is used in MicrobeDB.jp (http://microbedb.jp/). J-phenome (http://jphenome.info), a portal of animal phenotypes, employs a unified scheme using common phenotype ontologies such as Mammalian Phenotype Ontology (MP) and Phenotypic Quality Ontology (PATO). Furthermore, the International Mouse Phenotyping Consortium (IMPC) RDF data (http://metadb.riken.jp/metadb/db/IMPC_RDF) are provided, which enable the data integration of these phenotype data with other datasets. For instance, to retrieve the phenotype that can be expressed when a specific biological pathway is inactivated, a connection of the IMPC and Reactome datasets (http://www.reactome.org) is used.

5 Conclusions

We discussed an overview of the RIKEN MetaDatabase platform, which allows biologists to generate RDF datasets using a spreadsheet and publish them using various intuitive views, such as table and card views, and as a SPARQL endpoint. We have successfully realised integrated datasets, such as for microorganisms and mouse phenotypes, both within RIKEN and external communities. The database platform is lightweight, and multiple system instances can be launched as virtual machines in a cloud environment in order to develop individual databases for each research project. In the future research, we will focus on the implementation of a practical federation amongst such system instances.

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. Sugahara, M., Asada, Y., Shimada, H., Taka, H., Kunishima, N.: HATODAS II: heavy-atom database system with potentiality scoring. J. Appl. Crystallogr. 42: 540-544 (2009)
4. Yoshiki, A., Ike, F., Mekada, K., Kitaura, Y., Nakata, H., Hiraiwa, N., Mochida, K., Ijuin, M., Kadota, M., Murakami, A., Ogura, A., Abe, K., Moriwaki, K., Obata, Y.: The mouse resources at the RIKEN BioResource center. Exp. Anim. 58(2): 85–96 (2009)
5. Nakamura, Y.: Bio-resource of human and animal-derived cell materials. Exp. Anim. 59(1): 1–7 (2010)
6. Yokoyama, K.K., Murata, T., Pan, J., Nakade, K., Kishikawa, S., Ugai, H., Kimura, M., Kujime, Y., Hirose, M., Masuzaki, S., Yamasaki, T., Kurihara, C., Okubo, M., Nakano, Y., Kusa, Y., Yoshikawa, A., Inabe, K., Ueno, K., Obata, Y.: Genetic materials at the gene engineering division, RIKEN BioResource Center. Exp. Anim. 59(2): 115–124 (2010)