RIKEN MetaDatabase: a database publication platform for RIKENs life-science researchers that promotes research collaborations over different research area

Kai Lenz[1], Hiroshi Masuya[2,1], and Norio Kobayashi[1,2,3]⋆

[1] Advanced Center for Computing and Communication (ACCC), RIKEN,
2-1 Hirosawa, Wako, Saitama, 351-0198 Japan
{kai.lenz, norio.kobayashi}@riken.jp
[2] RIKEN BioResource Center (BRC),
3-1-1, Koyadai,Tsukuba, Ibaraki, 305-0074 Japan
hmasuya@brc.riken.jp
[3] RIKEN CLST-JEOL Collaboration Center,
6-7-3 Minatojima-minamimachi, Chuo-ku, Kobe 650-0047, Japan

Abstract. To promote data dissemination and integration of life sciences datasets produced by RIKEN, we developed an infrastructure database called "RIKEN MetaDatabase" that enables data publication and integration using the Resource Description Framework. In addition, we implemented a simple workflow for data management and a graphical user interface representing data links across laboratories. Consequently, inter-laboratory collaboration and coordination have been accelerated. Combined with global standardization activities, we expect that this infrastructure will contribute to worldwide data integration.

Keywords: life-sciences database, database integration, Resource Description Framework, Semantic Web

1 Introduction

RIKEN is a comprehensive science research institute that covers physics, chemistry, and life sciences. Conventionally, RIKEN produces various individual life sciences databases. The integrated analysis of the internal and external research data is essential.

Semantic Web technologies based on the Resource Description Framework (RDF) are powerful tools to realize distributed global data integration. Recently, RDF-based public databases, such as Bio2RDF[1], DisGeNet[2], EBI RDF Platform[3], have been published. In this context, we conclude that developing an infrastructure system that encourages RIKEN researchers to participate in RDF-based global data integration is beneficial. However, generating RDF data requires technical knowledge of Semantic Web technologies. To overcome this, we

⋆ To whom correspondence should be addressed: norio.kobayashi@riken.jp have developed a database platform, which we refer to as RIKEN MetaDatabase, that realizes database integration among different life science fields. In this study, we provide an overview of its design and implementation.

2 Requirements

The RIKEN MetaDatabase assumes two user types: data viewers and data publishers. The main requirement of the system is to ensure simple and useful operations by both user types.

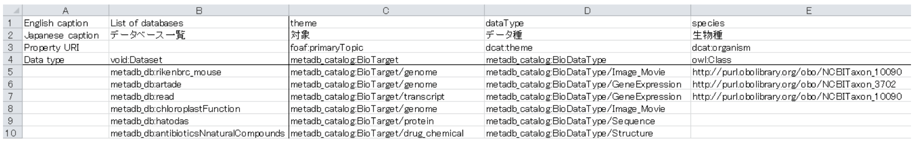
For data viewers, the requirements are as follows: 1) enabling users to find datasets via a simple graphical interface (GUI) and 2) showing the interrelations of data entities across databases.

For data publishers, RDF data generation should be handled by biologists who are unfamiliar with Semantic Web technologies. The platform supports a tabular form RDF data, which is frequently used in life sciences, to enable biologists to create RDF data more easily.

3 Implementation

3.1 Spreadsheet for RDF data generation

As described above, the core methodology that allows biologists to generate RDF data is the introduction of a tabular form that can be implemented using a spreadsheet application. Figure 1 shows an example of a Microsoft Excel spreadsheet.



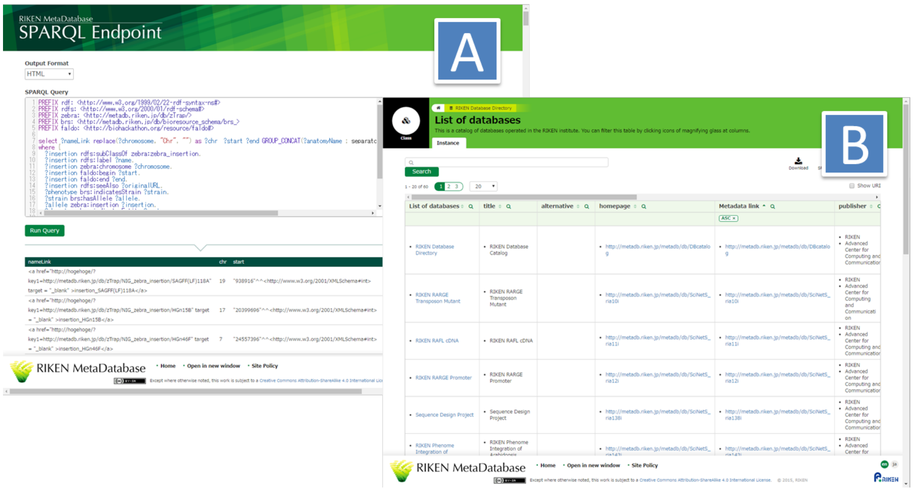
	A	B	C	D	E
1	English caption	List of databases	theme	dataType	species
2	Japanese caption	データベース一覧	対象	データ種	生物種
3	Property URI		foaf:primaryTopic	dcat:theme	dcat:organism
4	Data type	void:Dataset	metadb_catalog:BioTarget	metadb_catalog:BioDataType	owl:Class
5		metadb_db:rikenbrc_mouse	metadb_catalog:BioTarget/genome	metadb_catalog:BioDataType/Image_Movie	http://purl.obolibrary.org/obo/NCBITaxon_10090
6		metadb_db:artade	metadb_catalog:BioTarget/genome	metadb_catalog:BioDataType/GeneExpression	http://purl.obolibrary.org/obo/NCBITaxon_3702
7		metadb_db:read	metadb_catalog:BioTarget/transcript	metadb_catalog:BioDataType/GeneExpression	http://purl.obolibrary.org/obo/NCBITaxon_10090
8		metadb_db:chloroplastFunction	metadb_catalog:BioTarget/genome	metadb_catalog:BioDataType/Image_Movie	
9		metadb_db:hatodas	metadb_catalog:BioTarget/protein	metadb_catalog:BioDataType/Sequence	
10		metadb_db:antibioticsNnaturalCompounds	metadb_catalog:BioTarget/drug_chemical	metadb_catalog:BioDataType/Structure	

Fig. 1. Spreadsheet describing RIKEN Database Directory

In the spreadsheet, a table, column names, and data cells correspond to a class, properties and instances/literals, respectively, and the spreadsheet corresponds to a database that can be represented as an RDF graph. In a single Microsoft Excel workbook, users can define or update a graph (database) with multiple classes (tables), properties, instances, and triples.

3.2 Database management platform

The RIKEN MetaDatabase is implemented as a web server that connects a backend RDF triple store to a GUI accessed via a web browser. In addition, a SPARQL endpoint acts as an application programming interface (API). We employ two virtual machines(VMs) on our private cloud, i.e., the ”RIKEN Cloud Service”. One VM provides the GUI, and the other hosts Virtuoso as the backend RDF triple store. With cloud computing, database publishers using the RIKEN MetaDatabase do not require to configure hardware or use external software. The GUI reproduces the tabular list defined by the spreadsheet (Fig. 2).



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX zebra: <http://metadb.riken.jp/db/zTrap/>
PREFIX brs: <http://metadb.riken.jp/db/bioresource_schema/brs_>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>

select ?nameLink replace(?chromosome, "Chr", "") as ?chr  ?start ?end GROUP_CONCAT(?anatomyName ; separatc
where {
  ?insertion rdfs:subClassOf zebra:zebra_insertion.
  ?insertion rdfs:label ?name.
  ?insertion zebra:chromosome ?chromosome.
  ?insertion faldo:begin ?start.
  ?insertion faldo:end ?end.
  ?insertion rdfs:seeAlso ?originalURL.
  ?phenotype brs:indicatesStrain ?strain.
  ?strain brs:hasAllele ?allele.
  ?allele zebra:insertion ?insertion.
```

nameLink	chr	start
<a href="http://hogehoge/?key1=http://metadb.riken.jp/db/zTrap/NIG_zebra_insertion/SAGFF(LF)118A" target = "_blank" >insertion_SAGFF(LF)118A</a>	19	"938916"^^<http://www.w3.org/2001/XMLSchema#int>
<a href="http://hogehoge/?key1=http://metadb.riken.jp/db/zTrap/NIG_zebra_insertion/HGn15B" target = "_blank" >insertion_HGn15B</a>	17	"20399696"^^<http://www.w3.org/2001/XMLSchema#int>
<a href="http://hogehoge/?key1=http://metadb.riken.jp/db/zTrap/NIG_zebra_insertion/HGn46F" target = "_blank" >insertion_HGn46F</a>	7	"24557396"^^<http://www.w3.org/2001/XMLSchema#int>

Except where otherwise noted, this work is subject to a Creative Commons Attribution-ShareAlike 4.0 International Lic

List of databases

This is a catalog of databases operated in the RIKEN institute. You can filter this table by clicking icons of magnifying glass at columns.

List of databases	title	alternative	homepage	Metadata link	publisher
• RIKEN Database Directory	• RIKEN Database Catalog		• http://metadb.riken.jp/metadb/db/DBcatalog	• http://metadb.riken.jp/metadb/db/DBcatalog	• RIKEN • Advanced Center for Computing and Communication
• RIKEN RARGE Transposon Mutant	• RIKEN RARGE Transposon Mutant		• http://metadb.riken.jp/metadb/db/SciNetS_ria10i	• http://metadb.riken.jp/metadb/db/SciNetS_ria10i	• RIKEN • Advanced Center for Computing and Communication
• RIKEN RAFL cDNA	• RIKEN RAFL cDNA		• http://metadb.riken.jp/metadb/db/SciNetS_ria11i	• http://metadb.riken.jp/metadb/db/SciNetS_ria11i	• RIKEN • Advanced Center for Computing and Communication
• RIKEN RARGE Promoter	• RIKEN RARGE Promoter		• http://metadb.riken.jp/metadb/db/SciNetS_ria12i	• http://metadb.riken.jp/metadb/db/SciNetS_ria12i	• RIKEN • Advanced Center for Computing and Communication
• Sequence Design Project	• Sequence Design Project		• http://metadb.riken.jp/metadb/db/SciNetS_ria138i	• http://metadb.riken.jp/metadb/db/SciNetS_ria138i	• RIKEN • Advanced Center for Computing and Communication
• RIKEN Phenome Integration of	• RIKEN Phenome Integration of Arabidopsis		• http://metadb.riken.jp/metadb/db/SciNetS_ria143i	• http://metadb.riken.jp/metadb/db/SciNetS_ria143i	• RIKEN • Advanced Center for

Except where otherwise noted, this work is subject to a Creative Commons Attribution-ShareAlike 4.0 International License. © 2015, RIKEN

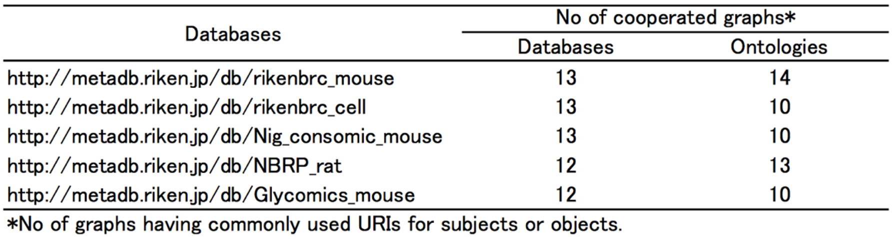
Fig. 2. RIKEN MetaDatabase graphical user interface. (A) SPARQL application programming interface and (B) tabular list

The RIKEN MetaDatabase has a ”built-in” database, i.e., the RIKEN Database Directory (http://metadb.riken.jp/metadb/db/DBcatalog), that list the metadata in order to easily find the RIKEN databases.

4 Activities

As of July 2016, there were 110 individual databases, 21 ontologies, and 152,784,620 RDF triples registered in the RIKEN MetaDatabase. These databases as well as the non-RDF-based databases published by RIKEN are also listed in the specialized RIKEN Database Directory. The RIKEN Database Directory is designed to be data-compatible with the IntegBio Database Catalog, which is a portal site for life sciences database integration by four ministries of Japan.

This system realizes data-mediated collaboration among the RIKEN database developers. Figure 3 shows the examples of the common uses of URIs across databases, where multiple databases have multiple inter-reletions with other datasets (databases and ontologies) via common URIs.



Databases	No of cooperated graphs*	
	Databases	Ontologies
http://metadb.riken.jp/db/rikenbrc_mouse	13	14
http://metadb.riken.jp/db/rikenbrc_cell	13	10
http://metadb.riken.jp/db/Nig_consomic_mouse	13	10
http://metadb.riken.jp/db/NBRP_rat	12	13
http://metadb.riken.jp/db/Glycomics_mouse	12	10

*No of graphs having commonly used URIs for subjects or objects.

Fig. 3. Cross-database cooperation in the RIKEN MetaDatabase

5 Discussion and Conclusions

Applying native RDF technologies, we have developed the RIKEN MetaDatabase, a database infrastructure for the publication and integration of data produced by RIKEN. The data-mediated collaboration shown in Fig. 3 benefits both the data developers as well as the data viewers, facilitating the integration of published data and a viewer's private data. For the further reuse of existing URIs, development of a retrieval function of public data realized in RightField[4] may be useful.

In another case, laboratories have initiated developing common schemata, including common upper-level classes and properties (e.g., mouse strains and cell lines as experimental materials or images), to standardize the data format of commonly used data entities. The standardization of such RDF schemata should be consistent with global activities such as Open Biomedical Ontology and BioSharing [5]. Therefore, in the future, we plan to organize a committee or a user group to discuss data coordination in RIKEN and globally. We expect that the RIKEN MetaDatabase will promote global data integration and collaboration across different life science fields.

References

1. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J. Biomed Inform., 5, 706-716. (2008)
2. Queralt-Rosinach, N., Piñero, J., Bravo, À., Sanz, F., Furlong, L.I.: DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. Bioinformatics. (2016) in press
3. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère. N., Parkinson. H., Birney. E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics, 9, 1338–1339. (2014)
4. Wolstencroft K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F., Goble, C.: RightField: embedding ontology annotation in spreadsheets.Bioinformatics 27(14), 2021–2012 (2011)
5. McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., Sansone, S.A.: BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences.Database 2016, 1–8 (2016)