**International Journal on Semantic Web and Information Systems**
Volume 14 • Issue 1 • January-March 2018

# RIKEN MetaDatabase:
## A Database Platform for Health Care and Life Sciences as a Microcosm of Linked Open Data Cloud

Norio Kobayashi, Advanced Center for Computing and Communication (ACCC), BioResource Center, RIKEN CLST-JEOL Collaboration Center, RIKEN, Japan

Satoshi Kume, RIKEN Center for Life Science Technologies, RIKEN CLST-JEOL Collaboration Center, RIKEN Compass to Healthy Life Research Complex Program, RIKEN, Japan

Kai Lenz, Advanced Center for Computing and Communication (ACCC), RIKEN, Japan

Hiroshi Masuya, BioResource Center, Advanced Center for Computing and Communication (ACCC), RIKEN, Japan

### ABSTRACT

Recently, the number and heterogeneity of life science datasets published on the Web have increased significantly. However, biomedical scientists face numerous serious difficulties finding, using and publishing useful databases. To address these issues, the authors developed a Resource Description Framework-based database platform, called the RIKEN MetaDatabase (http://metadb.riken.jp), that allows biologists to develop, publish and integrate multiple databases easily. The platform manages the metadata of both research and individual data described using standardised vocabularies and ontologies, and has a simple browser-based graphical user interface to view data including tabular and graphical forms. The platform was released in April 2015, and 113 databases, including mammalian, plant, bioresource and image databases, with 26 ontologies have been published using this platform as of January 2017. This paper describes the technical knowledge obtained through the development and operation of the RIKEN MetaDatabase to accelerate life science data distribution.

### KEYWORDS

Bioresource, Database Cloud Platform, Database Integration, Electron Microscopy Imaging Database, Imaging Data Integration, Life Sciences, Linked Open Data (LOD), Ontology, Open Microscopy Environment (OME), Phenotype, Resource Description Framework (RDF), Semantic Web, Web Ontology Language (OWL)

### 1. INTRODUCTION

Life sciences have developed rapidly and have been subdivided into specialised fields. Thus, the study of life sciences has generated numerous heterogeneous datasets, thereby making it difficult for researchers to find and use data appropriately in their research and publish data in a way that is useful for other researchers. Considering these difficulties, two major issues arise. The first is realising rich and useful data integration in a sustainable way. In the life sciences, many databases are still being created and the ongoing operational costs are a big problem. Since the database system is integrated

DOI: 10.4018/IJSWIS.2018010106

with the application program that browses data, it can no longer be operated continuously when the operating environment including the operating system gets older, mainly due to security issues. RDF and SPARQL technologies provide a standardized data representation and API. These technologies realise linking data, integrating data systematically using standardised vocabularies, representing semantics, and publishing data locations. They are expected to make database operation sustainable. The second issue involves realising easy, flexible and low-cost operation that allows many data developers and biologists to participate in the data integration process.

These difficulties also occur within a research institute. RIKEN is the largest Japanese comprehensive science institute, encompassing a network of research centres and institutes, and aims to lead high-quality research across a diverse range of scientific disciplines. As an umbrella organisation for research projects, RIKEN should be committed to widely disseminating the results of scientific research and technological developments, promoting excellent collaborative research across multiple scientific study fields, and integrating the activities of research centres. In the life sciences, RIKEN also encompasses both large-scale research centres and many small-scale laboratories that generate large-scale life science datasets in various fields. The institute is facing issues regarding the promotion of collaborative research across different fields, such as the share of large-scale genome sequence data, analysis of the molecular pathways of cells, functional analysis of biological tissues using imaging approaches, and provision of metadata for the bioresources (biological materials such as cells and tissues) used in each analysis. Therefore, a database infrastructure is required for the publication and promotion of RIKEN's research results. This situation can be presented as a microcosm of a linked open data cloud for life sciences.

The authors consider RIKEN's problem as a case study of data utilisation in life sciences. To address this problem, the authors developed the RIKEN MetaDatabase (http://metadb.riken.jp), a database platform based on the Resource Description Framework (RDF) that realises low-cost metadata management, systematic data integration and global publication on the Web. The RIKEN MetaDatabase was published in April 2015 with RIKEN's original databases, external databases associated with these databases and ontologies. Here we discuss the advantages of the RDF to solve life science data distribution and future issues, focusing on the RIKEN MetaDatabase implementation, data integration and comparison with other cases.

The novel results of this research include a database integration workflow that allows biologists to participate directly and an RDF-based data publication platform for multiple communities of different research projects in the life sciences. Furthermore, the authors have successfully achieved integration between actual research communities using these technologies.

The remainder of this paper is organised as follows. Section 2 discusses the requirement specifications for the database platform. Sections 3 and 4 discuss the functional design issues involved in realising the data integration platform, which has a data publication workflow in which experimental biologists can directly participate. Section 5 discusses the implementation of these platforms. Sections 6 and 7 introduce available databases and comprehensively review the RIKEN MetaDatabase by introducing concrete database projects. Section 8 describes future research directions and Section 9 concludes this paper.

## 2. REQUIREMENT SPECIFICATIONS FOR THE LIFE SCIENCE DATABASE PLATFORM

In the development of the RIKEN MetaDatabase, the authors were requested to expand data utilisation while lowering the cost of operation. In order to realise these requests, we recognised that both cloud-based life science databases and data integration functions were necessary. The details will be discussed below.

### 2.1. Requirements for Cloud-Based Life Science Databases

As an in-house database platform for a comprehensive research institute, the RIKEN MetaDatabase should support different types and sizes of datasets generated by research projects of all sizes, as shown in Figure 1. Each RIKEN research centre studies multiple specialised life science fields and develops individual databases as research results. The RIKEN MetaDatabase is designed to convert these databases to RDF format to realise internal and global data integration covering various research fields.

In addition, the RIKEN MetaDatabase should form a uniform knowledge base that includes internal RIKEN datasets and global datasets interlinked on the web. In addition, it should provide a simple operational workflow whereby biologists can easily participate in global data integration without requiring specialised data integration skills.

Ideally, both biologists and informaticians should cooperate closely for data integration and mining through the database platform, as shown in Figure 2. Therefore, the authors consider that the RIKEN MetaDatabase platform must provide well-coordinated datasets, namely, datasets described using URIs and properties used globally on the web for easy integration with other datasets, that can easily be integrated into global datasets and used by data scientists. Furthermore, the data publication workflow should be simple.

As illustrated in Figure 2, RIKEN has various databases produced by large-scale research centres and small laboratories that should be integrated (Existing databases). To publish integrated versions of these datasets, both the 'editing metadata' and 'publishing metadata' processes are crucial. In the 'editing metadata' process, prior to generating RDF data, building collaboration(s) amongst biologists and informaticians is desired. Because RIKEN deals with cutting-edge research results and a wide variety of research areas even within the life sciences, metadata creation including the selection of ontology terms is very difficult. Therefore, after mutual discussion among informaticians, ontology experts, experimental biologists, and life science data experts, these researchers settled on a set of descriptive metadata based on the detailed contents of the database. In the 'publishing metadata' process, multiple features, such as simple common viewers to represent data and their integration,

**Figure 1. Conceptual overview of the RIKEN MetaDatabase**

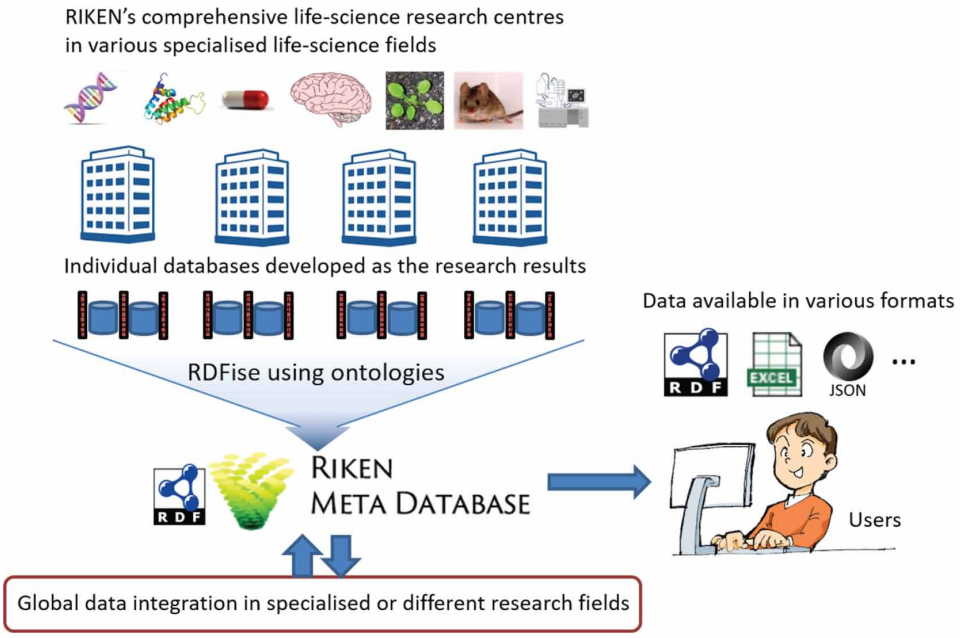
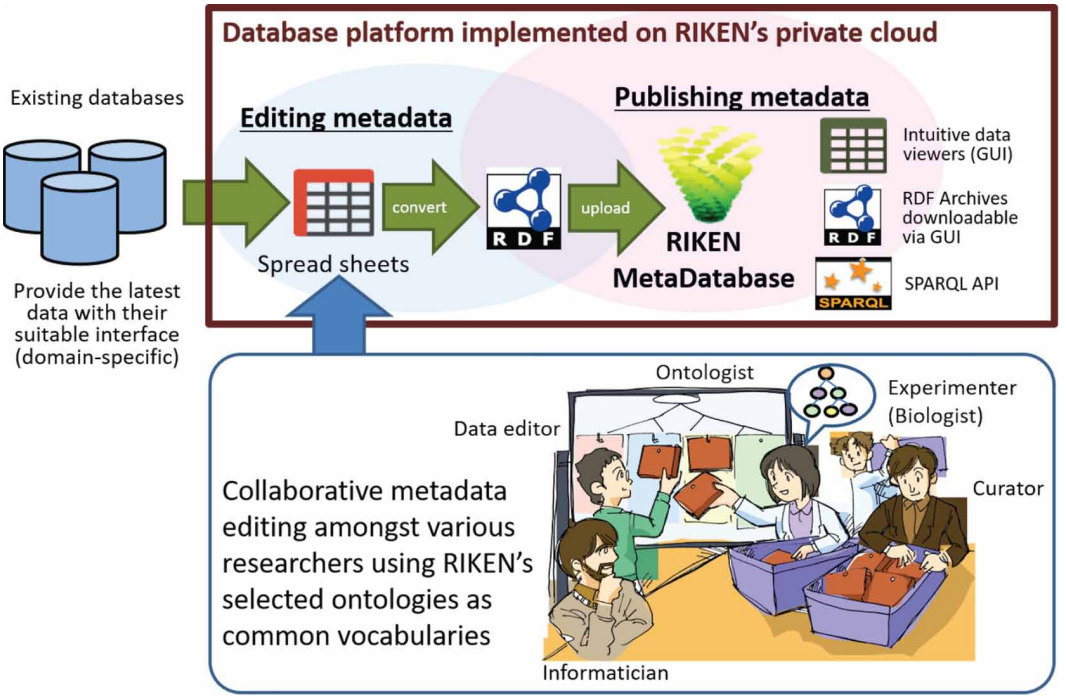


**Figure 2. Ideal workflow for data publication**




downloading RDF data and a SPARQL endpoint to output data in a user-specified format, are required. The requirements come from actual users' needs. The simple common viewer was introduced to satisfy the need for a data browser that could easily be used by non-data scientists, including biologists, with only a small development cost. The RDF data downloading function was introduced because there were users who wanted to download RDF archives and analyse them in their own local environments. A SPARQL endpoint was necessary for existing users who wanted to edit and integrate the RDF data with other RDF data, as shown in Section 7.

To satisfy these requirements, the authors have designed a cloud-based platform that allows many database developers to deploy data without management hardware and to ensure significant and flexible computational resources. The authors also adopted Semantic Web technologies including RDF, SPARQL, and ontologies, as described above. Simple interfaces for data generation and publication were also designed. To the best of the authors' knowledge, no other database platform realises cloud-based data storage and database publication features that satisfy both data scalability and data portability based on standardised data formats and datatypes. In other words, this platform meets the previously-described needs of database developers in an organisation or community in a cost-effective manner, providing a suitably scalable database platform and reducing the cost of data management.

## 2.2. Data Integration

In recent bioinformatics research, both comprehensive data coverage for given data types and integration between heterogeneous datasets are necessary. In the development of the RIKEN MetaDatabase, we aimed to realise both data integration within a particular specialised research field (Case 1) and data integration between different research fields (Case 2):

1. **Data integration in a specialised research field:** Case 1 refers to data integration to realise a comprehensive dataset across research projects (research organisations or international consortiums) to form a unified database. In this case, we assume that research projects generate non-redundant datasets that may belong to the same data class. To enable unified data handling and management, research projects should share the data structure defined using their class rather than sharing data entities (or instances). For example, in the field of micro-organism research, it is desired that basic metadata for microbial strains (culture collections) are broadly shared across databases. For efficient data integration within the micro-organism research field, a common data schema, the Microbial Culture Collection Vocabulary (MCCV; http://bioportal.bioontology.org/ontologies/MCCV), was designed. The MCCV helps with data integration between databases hosted in the RIKEN MetaDatabase and outside it (see Section 8.4);
2. **Data integration amongst different research fields:** Case 2 refers to data integration amongst different research fields or co-operable research to realise mutual data links across datasets. In this case, projects provide related datasets that belong to different data classes. Here, some data entities may act as links between the different datasets. In Section 7.2, data integration across databases via sharing of common URIs (data items and ontology terms) is described. In addition, in Section 7.3, we describe a trial of developing common schema for image metadata that models interlinks among different data, images, experiments, and bioresources.

Case 1 can be achieved by introducing common or standardised data schemata and ontologies wherein each data entity is typically described as an instance of a class or an ontology term. Therefore, data integration is achieved by sharing common classes and semantic links (RDF properties) that define the data structure rather than by providing a direct link between data entities. In case 2, data entities from different datasets are directly connected by semantic links and each dataset is described by different specialised data schemata. The data entities are links that allow the expansive combination of different communities.

The Semantic Web with the RDF satisfies these two types of integration simultaneously. Case 1 can be realised using the RDF scheme and OWL, and a data linking mechanism can be applied in case 2. However, as explained in Section 2, most platforms do not satisfy both cases. Therefore, the authors propose a practical approach to solve this problem.

## 3. RDF-BASED RIKEN METADATABASE PLATFORM DESIGN

As mentioned in Section 3, the authors decided to develop the RIKEN MetaDatabase as a consolidated database based on RDF-related technologies. This section discusses how the authors designed the RIKEN database platform using RDF technologies.

### 3.1. RDF Data Structure Suitable for Life Science Data Integration

First, we describe the data structures used in RIKEN databases. Prior to functional design, we first reviewed the data structures of databases published from RIKEN. As a result, we found that most databases (datasets) were represented in tabular form hosted by a relational database system.

In a relational database, extending data schema is not easy. In the RIKEN MetaDatabase, new databases created for research datasets must be added and hosted from time to time. Therefore, if a relational database had been used for the RIKEN MetaDatabase, the development and deployment costs would have been extremely large. With RDF, there is no significant distinction between the schema and data record definitions and it is easy to host various types of data. Moreover, because relationships between data items are defined by the data itself, maintenance of relationships is not necessary. Moreover, RDF promotes the distribution of data as linked open data and data integration across databases. Due to these advantages, RDF meets the requirements for the RIKEN MetaDatabase.

Therefore, to realise a simple and user-friendly database infrastructure system, the RDF data handled by the RIKEN MetaDatabase is restricted to tabular-type database data and hierarchical ontology data.

Tabular-type database data can be easily generated and browsed. The tree form used to describe ontology data represents concepts and data classes with their conceptual hierarchy. Using these data forms, the RIKEN MetaDatabase aims to build a single integrated RDF dataset by managing multiple tabular and hierarchical ontology data individually.

## 3.2. Tabular Data Model

We introduce a tabular data model to describe RDF data in which all RDF resources are associated with an RDF class. A table is generated for each class of subject instances of RDF triplets. Figure 3 shows the RDF data structure in tabular form. The table is divided into two parts, one for RDF schema definitions and the other for data. The first four rows are RDF schema definitions. The fifth and subsequent rows show data, which are instances of the datatype given in the fourth row.

### *3.2.1. Rows*

In Figure 3, the RDF scheme definition is presented in the top four rows. The first and second rows give the English and Japanese column names. These are displayed in the graphical user interface (GUI) but are not included in RDF triplets. The third and fourth rows describe the properties and classes of the objects of the triplets used to convert the tabular data to RDF format, respectively. The fifth and subsequent rows show the data.

### *3.2.2. Columns*

The first column is a comment column, which is not converted to RDF.

The second column shows instances (resources) of the common class that is the subject of all triplets described in the table, i.e. a list of subject instances. Using the table coordinates $(r,c)$ to locate the data points, where $r$ is the row and $c$ is the column, $(4,2)$ contains the data class, $(3,2)$ is empty and $(\mathrm{m},2)$ for $\mathrm{m} \geq 5$ are instances of the $(4,2)$ class.

The third and subsequent columns describe the properties and objects for the subjects listed in the second column. $(3,n)$ is a property and $(4,n)$ is a class or data type of the instances or the literals listed as $(m,n)$, respectively, where $m \geq 5$ and $\mathrm{n} \geq 3$. Here, the triplet $(m,2),(3,n),(m,n)$ is equivalent to the following set of RDF triplets:

**Figure 3. Spreadsheet describing RDF data for the Background strain class (http://metadb.riken.jp/db/rikenbrc_mouse/animal_0000004) in tabular form. The second column shows instances of the Background strain class, the third column shows literal rdf:langString values and the fourth column shows Taxon classes as instances of the owl:Class.**

| | 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 |

$(m,2)$ $(3,n)$ $(m,n)$
$(m,2)$ rdf:type $(4,2)$
$(m,n)$ rdf:type $(4,n)$

where $(4,n)$ is an RDF class, or:

$(m,2)$ $(3,n)$ $(m,n)$
$(m,2)$ rdf:type $(4,2)$

where $(4,n)$ is a data type and $(m,n)$ is a literal denoted as the form of data type $(4,n)$.

Moreover, each pair of a property and object class (or data type) in the third and subsequent columns can appear multiple times to describe multiple triplets sharing a common subject, property and object class (or data type).

### 3.3. Correspondence with the RDF Scheme

To manage multiple RDF datasets as databases or ontologies in the RIKEN MetaDatabase, we introduce a specialised data category that corresponds to the existing RDF scheme elements, as shown in Table 1.

A database is an RDF dataset with tabular data that comprises an individual database and corresponds to an RDF named graph. An ontology is an OWL ontology managed as an RDF named graph. A property and a class are equivalent to an RDF property and an RDF class as an instance of rdf:Property and rdfs:Class, respectively. An instance is limited to an instance $i$ of rdf:Resource explicitly described as triplet $i$ rdf:type $c$, where $c$ is an RDF class. We introduce limited instances to establish data reusability. When a class is specified, the instances of that class can be obtained accurately without orphan instances that are not associated with any class.

### 3.4. RDF Data Generation and Publication

The authors designed a procedure by which users can generate and publish their RDF data. Tree ontology data, usually described in OWL, that can be downloaded from public repositories or generated by an existing ontology editor can be uploaded directly to the RIKEN MetaDatabase platform and published immediately. On the other hand, for tabular data, we apply the following spreadsheet-based workflow:

**Table 1. Correspondence between the RIKEN MetaDatabase and the RDF scheme**

| RIKEN MetaDatabase | RDF Scheme | Description |
|---|---|---|
| Database | Named graph | An individual dataset with multiple classes |
| Ontology | Named graph | An individual ontology written in OWL |
| Property | Instance of rdf:Property | Equivalent to rdf:Property |
| Class | Instance of rdfs:Class | A concept or an rdf:Resource set |
| Instance | Instance of class | An instance typed by a class |

**Step 1:** Generating a spreadsheet

In this step, the user (database developer) builds a spreadsheet as a Microsoft Excel file or Tab-Separated Values (TSV) files that represent a tabular data model. Using multiple spreadsheets in Microsoft Excel or TSV files, the user can describe a complicated database in which multiple tables are linked in a relational database management system. The RDF resources are written as URIs.

**Step 2:** Generating an RDF dataset

The spreadsheet generated in the previous step is converted by the user to RDF format using the authors' application program. The program generates RDF data converted from raw data and a structure definition file that describes the order of the columns and the column names in RDF format.

**Step 3:** Uploading the RDF dataset

Both the database and the structure definition files are uploaded by a service administrator to the database platform. The uploaded data are published immediately.

### 3.5. User Interface for Data Input and Output

The RIKEN MetaDatabase employs both a GUI that works with the user's web browser and an application programming interface (API) for data input and output.

For data input, a registration interface for RDF-format tabular data and tree ontology data is implemented. Only service administrators can use this function because they should be able to check the uploaded data before publication.

The data publishing function is implemented in both the API and GUI. As an API, we use an interface that acts as a SPARQL endpoint accessible via the HTTP, which is a standardised RDF-data-access protocol. The GUI works on the user's web browser and displays RDF data in various formats, such as tabular and tree formats. In addition, it offers a list of RDF data archives for download and access to the SPARQL endpoint described above with a query editor and result display functions.

The data input and output interfaces are summarised in Figure 4. The data output interface in particular will be discussed in detail when we consider the data display functions in Section 4.

## 4. DATA DISPLAY FUNCTIONS

To demonstrate RIKEN managements of RDF data, several fixed display forms have been developed for each data category. By default, data are displayed with multilingual labels rather than Unified Resource Identifiers (URI); however, both labels and URIs can be shown to RDF experts.

The implemented views are summarised as follows:

- **Tabular view:** Lists instances of a specified class;
- **Card view:** Displays a selected instance;
- **List view:** Lists databases and ontologies, and includes a keyword search function to filter data;
- **Database view:** Shows the classes in the database and statistics, i.e. the numbers of triplets, classes and properties;
- **Tree view:** Shows OWL ontologies as trees based on subclass relationships;
- **Download view:** Used to download RDF data archives for each database;
- **SPARQL search view:** Supports editing queries and displaying results.

**Figure 4. Relationships amongst data views and data upload pages, starting from the top page of the RIKEN MetaDatabase**

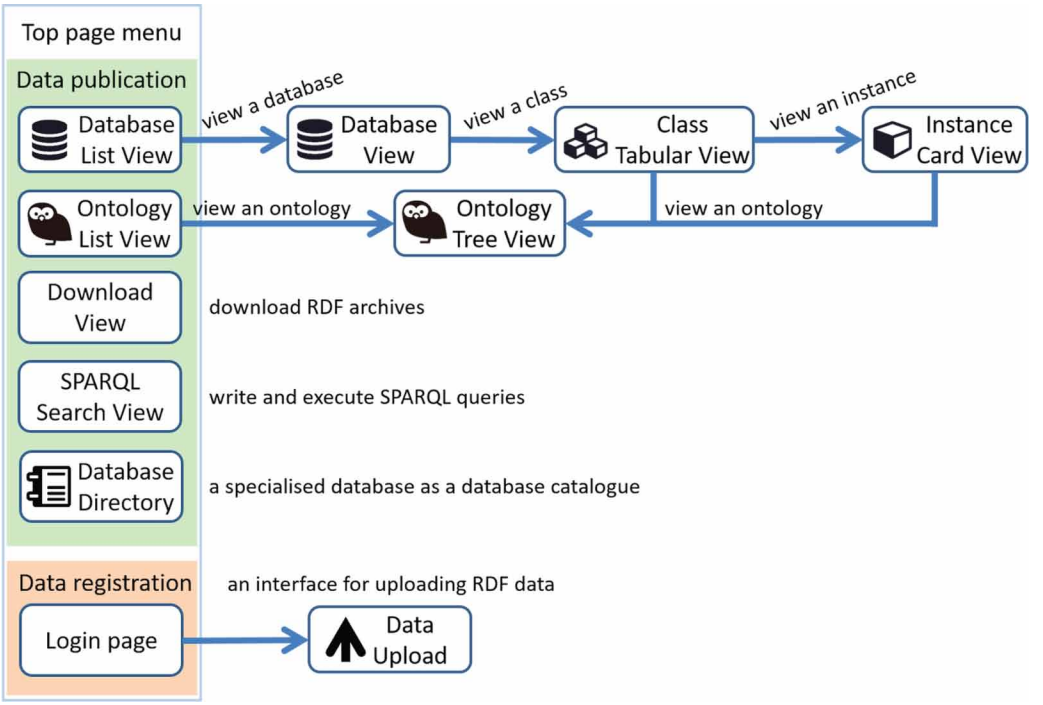



Figure 4 shows a typical data browsing flow starting from the top page of the RIKEN MetaDatabase. The top page traverses RDF categories from higher to lower levels, i.e. database → class → instance, by referencing the ontology tree if necessary. By default, the tabular and card views display RDF data. We describe these views in detail in the following.

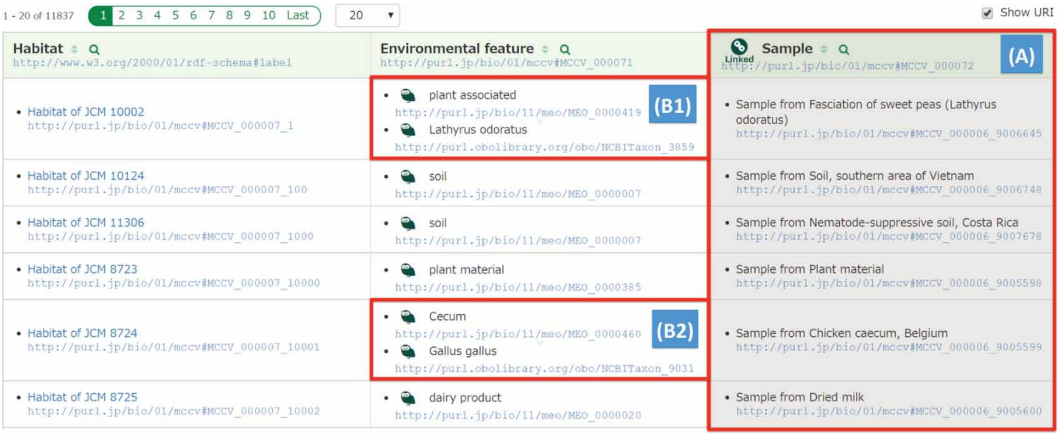
As shown in Figure 4, the top page comprises 'Database list', 'Ontology list', 'Download', 'SPARQL search' and 'Database directory' views. The 'Class tabular view' and 'Instance card view' display detailed content. The 'Ontology list view' is linked to the 'Ontology tree view' to show the detailed content of each ontology. The Ontology tree view is link to both Class tabular and Instance card views to represent usage of common vocabularies in each dataset. The Database directory lists the metadata of databases (datasets) published by RIKEN.

### 4.1. Tabular View

Tabular views display RDF graph data. These views can be generated for each class and show the name and description of the target class, all instances of the class, the triplets whose subject is one of the instances, and the triplets whose object is of the instances. The former triplets are called forward triplets, and the latter triplets are called reverse triplets. In addition, if the object of a forward triplet or the subject of a reverse triplet is an instance described in both the target database and at least one other database, then that triplet is shown so that data integration via instances can be realised. A selected RDF class with its instances can be shown in this view. However, the column names and column order can be customised using a structure definition file generated from a spreadsheet.

An example tabular view is shown in Figure 5. The first column shows instances of the class. The second and subsequent columns form sets, each of which is associated with a property and describes a list of objects of forward triplets or subjects of reverse triplets with that property. In more detail for displaying reverse triplets, each column is associated with a property reverse-linked to the instances in the first column.

**Figure 5. Tabular view of the Habitat of the Japan Collection of Microorganisms (JCM) class resource database (http://metadb.riken.jp/metadb/db/rikenbrc_jcm_microbe). (A) The third column (class sample) shows instances to link to the subject instances of the first column, i.e. reverse linked instances of the Sample class. (B1 and B2) Multiple objects with the same subject and predicate pairs can be displayed as a list in the corresponding cell.**




By default, the name of the first column is the class label and those of the second and subsequent columns are the corresponding properties labels. However, a structure definition file with different column names can be uploaded and the original column names can be overwritten.

The data in each row can be sorted in ascending or descending order as specified by the user. Furthermore, the row data can be filtered using full-text search for human readable metadata with user-specified keywords for each column.

### 4.2. Card View

The card view, as shown in Figure 6, is primarily used to show an instance and its triplets linked to other instances or reverse linked from other instances. In the card view, a user can view a long triplet path by traversing the connected triplets in a sequence from the corresponding instance.

By default, only triplets including that particular instance are shown. A user can select an instance connected via a triplet to show further triplets with the selected instance, and the new triplets are shown as a new nested card in the original card view.

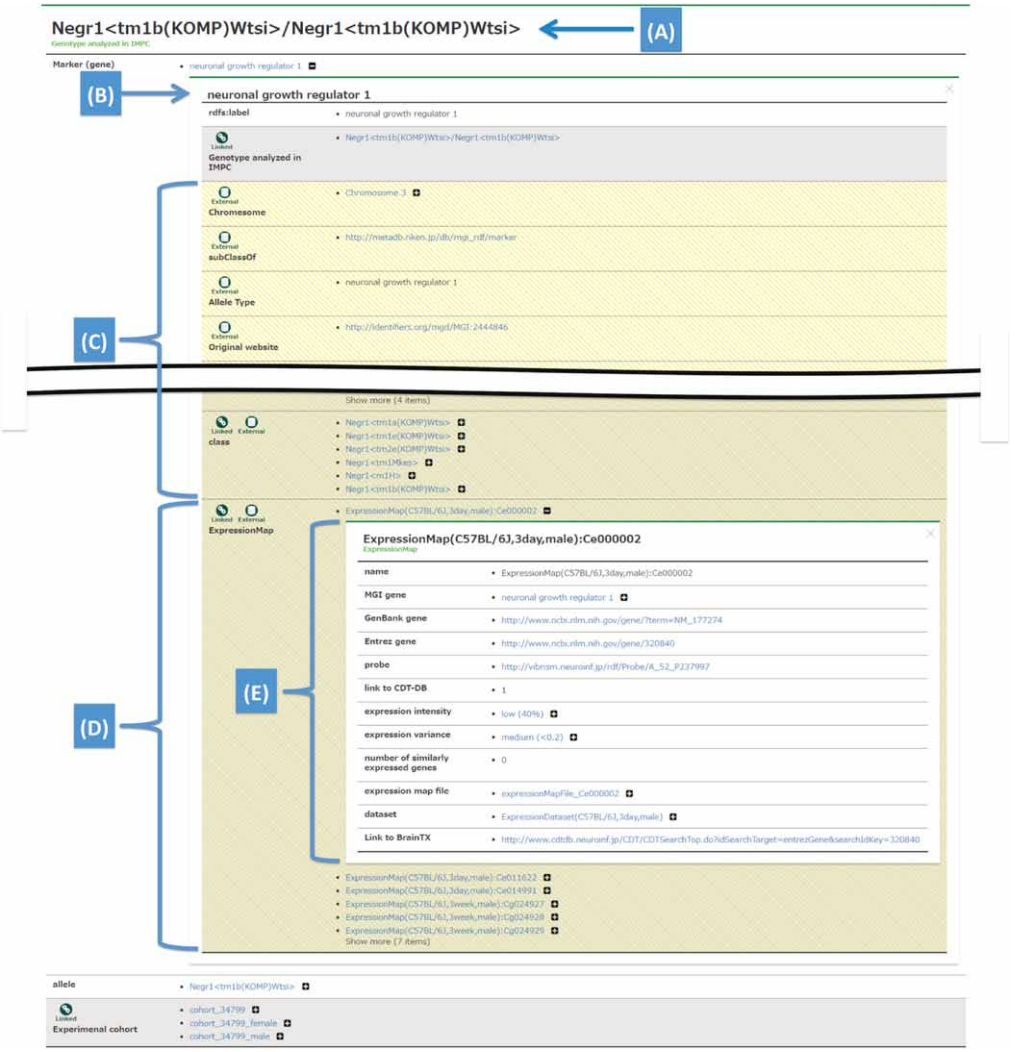
### 4.3. List Views

The RIKEN MetaDatabase is an integrated database platform that manages multiple databases and ontologies. Figure 7 shows list views for (A) databases and (B) ontologies reached from the top page of the RIKEN MetaDatabase. In the database list view, databases with their classes are listed and search functions, including faceted search, for selecting databases categorised with Integbio (Section 6.1) terms having species, themes, publishers and datatypes and keyword search against database names and descriptions of databases and classes. The ontology list view also displays a list of ontology names and descriptions and provides a keyword search function against names and ontology and ontology term (classes) descriptions.

### 4.4. Other Views

Other views are specialised views for data format and function, including database, ontology, downloadable archives and SPARQL search, as shown in Figure 8. The download view (A) shows the name and description of the corresponding database and a list of classes and statistical data, including the numbers of triplets, classes, properties, etc. The ontology tree view (B) shows ontology terms in

**Figure 6. Card view of an instance of an experimental cohort (http://metadb.riken.jp/db/IMPC_RDF/Cohort) used in the International Mouse Phenotype Consortium (IMPC) database (Section 7.2) that links to instances of other databases. (A) is an instance of the Cohort class of KO mice, (B) represents an allele, i.e. Cdh23-v (waltzer), carried by the cohort in the IMPC database and described in the URI of the Mouse Genome Informatics (MGI) RDF database, (C) is a list of triplets in the MGI RDF database, (D) is a list of links from the BioResource Center (BRC) Mouse Strain database and (E) is the detailed information of the mouse strain with multiple alleles including Cdh23-v in the BRC Mouse Strain database. The detailed RDF data structure shown in this figure can be explained as follows: Cdh23-v, shown in (B), is the object of the triple whose subject is the cohort_33938_female shown in (A) and whose predicate is 'Allele'. (B) shows a card view of the data related to Cdh23-v. (C) shows a part of (B), namely, the list of triples whose subjects are Cdh23-v, retrieved from an external database. For example, the first column of (C) is a triple whose subject, predicate, and object are Cdh23-v, 'Allele Name', and 'waltzer', respectively.**




tree format based on subclass relationships. In addition, detailed information about the selected term is shown with classes and instances linked to the term. The download view (C) shows the archived files of a database, an archive of whole RDF data of the database and a list of archives of classes of the database. Furthermore, the original spreadsheet file and database descriptions of the Health Care and Life Sciences (HCLS) Community Profile and SPARQL Builder Metadata (Section 6.1) can be downloaded from this view. The SPARQL search view (D) allows a user to write a SPARQL query and shows query results.

**Figure 7. (A) Database list view and (B) ontology list view reached from the top page of the RIKEN MetaDatabase**

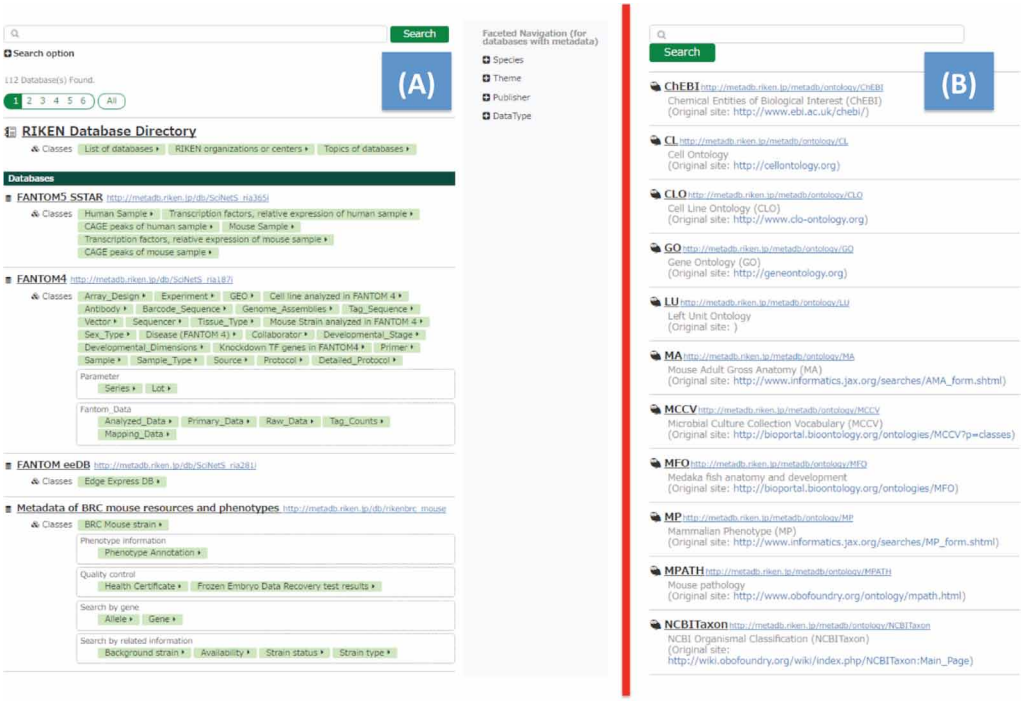
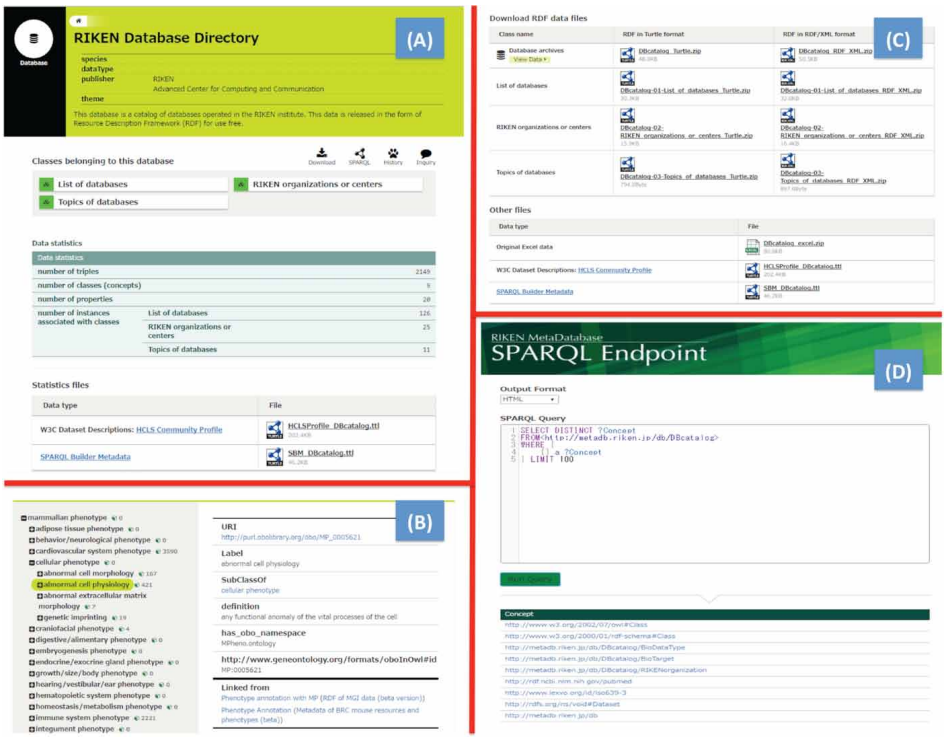



**Figure 8. Specialised views: (A) Database view; (B) Ontology tree view; (C) Download view and (D) SPARQL search view**



### 4.5. URI Display Mode

In the table and card views, the URI display mode can be turned on. In the normal mode, each data item is denoted using the label given by rdfs:label for biologists who are not interested in URIs. In the URI display mode, the URIs are also displayed below each of the labels. Therefore, when a data creator wants to link to data in the RIKEN MetaDatabase, the appropriate URI can easily be found while browsing the actual data.

## 5. IMPLEMENTATION

This section introduces the system architecture and deployment that uses a cloud platform and experimental implementation towards realising distributed databases for handling unpublished research datasets.

### 5.1. System Architecture and Cloud Deployment

Reducing both the development and operational costs is most important for realising persistent database services worldwide while ensuring service stability. In the authors' implementation of the RIKEN MetaDatabase, the authors adopt a simple architecture comprising two components: (a) a web server that provides a GUI and (b) an RDF triplet store. The web server provides web pages with data display functions through a data display view, as described previously. The RDF data displayed in a view are obtained from the RDF triplet store. The RDF triplet store manages both RDF data and structure definition data generated by the authors' application from spreadsheets, as described in Section 3.3. The structure definition data are referenced to generate a tabular view to obtain the data schema part of the tabular data model, including column names, properties and data types, and these data are also used to generate SPARQL queries to obtain the corresponding RDF data (instances). In addition, the web server functions as a SPARQL endpoint for submitting SPARQL queries generated by the web server.

In the authors' current platform, they employ OpenLink Software's Virtuoso Open-Source Edition version 7 (https://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/) as the RDF triplet store, and the web server is implemented as a Java servlet using Apache Tomcat version 8 (http://tomcat.apache.org/).

To ensure stability, portability and continuity of the platform, these software components are deployed on RIKEN's private cloud, called the RIKEN Cloud Service (http://cloudinfo.riken.jp), which provides multipurpose Linux-based virtual machines. The web server runs on a virtual machine connected to the global network. The RDF triplet store is deployed on a specialised virtual machine to realise fast SPARQL operations and is connected to a 1-TB flash memory storage via an InfiniBand network where the Virtuoso database directory is located.

### 5.2. Experimental Implementation Toward Realisation of Distributed Databases

In the actual operation of the RIKEN MetaDatabase, a single instance of the system is insufficient; thus, multiple distributed instances are necessary for the following reasons.

#### *5.2.1. Data Testing Before Publication*

Typically, users want to preview their data before publication. In the preview step, users check how their data are shown to avoid simple errors in data construction. For this purpose, a private mirror instance located on the intranet (private network) is required for previewing the data before publication.

#### *5.2.2. Exclusive System for Un-Published Research Data*

Database development and data creation are performed from the beginning of a research project, and these processes should remain private until a progress report is released. For this purpose, an individual

instance of the system is required for each research project. The instance can be closed within a project but can also be open when the project is finished and the developed databases are published.

RIKEN acts as an umbrella organisation for multiple research projects leading specific study fields. There were multiple requests from projects to have their own individual database 'instances'. Because the RIKEN MetaDatabase is lightweight in that it only requires two virtual machines (one as a web server and the other as an RDF triplet store, as shown in Section 5.1), multiple instances can be generated in the RIKEN Cloud Service. We also desire that these multiple distributed instances are able to act as a single integrated database for functions requiring database co-operation, such as SPARQL federated query searches over instances. Unfortunately, current standardised SPARQL federated query search (https://www.w3.org/TR/sparql11-federated-query/) is not efficient, and queries must be limited for a quick response. As an experimental implementation of such functions, specialised SPARQL federated querying is used for data display functions (Section 4). Currently, the public system described above and a specialised internal publication repository system have been implemented successfully. In addition, multiple instances of the RIKEN MetaDatabase are also used to distribute processing loads. The internal repository system manages MEDLINE data, including bibliographic information with abstracts, totalling approximately 0.93 billion triplets and accepts SPARQL queries from the public system against both data and structure definition data in RDF format so that the public system can generate tabular and card views containing publication information.

As part of the RIKEN Cloud Experimental Service, six individual private system instances for ongoing research projects, including mammalian and primates imaging, bioresource management and health informatics, have been generated with their own access control using firewalls.

## 6. AVAILABLE DATABASES

As of January 2017, 26 public ontologies, including the Gene Ontology (GO), Phenotypic Quality Ontology (PATO), NCBI Organismal Classification (NCBITaxon) and Semanticscience Integrated Ontology (SIO), have been selected and published as mirrors. These ontologies refer to 113 databases, including 62 original RIKEN databases. The remaining 51 databases are external databases that have been converted from originally non-RDF databases and linked from RIKEN's databases. In total, the RIKEN MetaDatabase has 161 million triplets, 2,238 classes, 3.18 million instances and 1,271 properties. The original databases are from various research fields, e.g. FANTOM (mammalian (The FANTOM Consortium & the RIKEN PMI & CLST (DGT), (2014))), FOX Hunting (plant (Ichikawa et al., 2006)), Heavy-atom Database System (protein (Sugahara et al., 2009)) and Metadata of BioResource Center (BRC) resources (bioresources (Yoshiki et al., 2009), (Nakamura, 2010) and (Yokoyama et al., 2010)).

The RIKEN MetaDatabase launched in April 2015 and has been operated stably to date, with the exception of scheduled power outages twice a year. In 2016, the average monthly data was as follows. The number of unique users was approximately 2,500, and the numbers of browser view and SPARQL accesses via programs were approximately 30,000 and 180,000, respectively.

### 6.1. Database Directory Service

The RIKEN MetaDatabase provides a specialised database and RDF datasets that provide easy data access. The specialised database is the RIKEN Database Directory, which is a catalogue of RIKEN's databases, including non-RDF databases. The catalogue data are designed to be compatible with the Integbio Database Catalog (http://integbio.jp/dbcatalog/?lang=en), which aims at inter-ministry integration of life science databases in Japan. In addition, W3C's HCLS Community Profile data (http://www.w3.org/TR/hcls-dataset/), including statistics data, are generated for each database and for entire datasets, and are published as RDF archives and via the SPARQL endpoint.

The RIKEN MetaDatabase also provides SPARQL Builder Metadata (http://sparqlbuilder.org/), which are generated and published for more intelligent SPARQL searches. The SPARQL Builder

Metadata is a profile of the SPARQL endpoint that describes the RDF graph structure. Thus, the SPARQL Builder tool (Yamaguchi et al., 2015) generates a SPARQL query that obtains triplet paths connecting two user-specified ontological concepts.

## 6.2. Integrated Databases

As described above, RDF data in the RIKEN MetaDatabase are designed and generated by editing a spreadsheet in collaboration amongst biologists and informaticians, including ontologists. The current platform does not support a data annotation function, which assists in selecting suitable ontology terms for data creators (biologists); however, the ontologists and curators carefully examine user data and select ontology terms from the introduced ontology sets in the RIKEN MetaDatabase.

Table 2 lists databases that share common class or instance URIs amongst other databases and common URIs for ontology terms amongst ontologies. The numbers of databases show the degree of database integration in which a corresponding database has direct links to or reverse links from the data entities of other databases that are not necessarily from the same research field. These relationships are visualised by the card view interface, as shown in Figure 6. On the other hand, the numbers of ontologies show the degree of database integration of the same research field of the corresponding database. The database list in Table 2 includes many bioresource databases, which are the bases of life sciences, and these databases are integrated in a wide range of other data.

Moreover, from an ontology-based integration perspective, Table 3 lists ontologies used in multiple databases in the RIKEN MetaDatabase. This shows that the ontologies successfully integrate amongst databases by narrowing the concepts by selecting a specific set of ontologies.

## 6.3. Ongoing Project for Microscopy Imaging Data Sharing

Imaging data is a very important fundamental in life sciences, and RIKEN has various imaging methods for experiments, such as optical microscopy (OM), electron microscopy (EM), magnetic resonance imaging (MRI) and positron emission tomography (PET), to obtain imaging data for a variety of biological samples. For example, ultra-microstructural imaging data, such as images obtained

**Table 2. Databases that integrate other databases and ontologies**

| Database | Number of Databases | Number of Ontologies |
|---|---|---|
| Metadata of BRC mouse resources and phenotypes | 14 | 15 |
| Metadata of Functional Glycomics with KO mice (ACGG-DB) | 13 | 11 |
| RIKEN Database Directory | 13 | 6 |
| Metadata of Functional Glycomics with KO mice (JCGGDB) | 13 | 11 |
| NIG Mouse Phenotype Database Metadata | 13 | 12 |
| Metadata of BRC cell resources | 13 | 10 |
| Metadata of NBRP Rat | 12 | 15 |
| Bioresource schema | 11 | 12 |
| NBRP Medaka Phenotype Metadata | 9 | 11 |
| NIG Zebrafish | 9 | 5 |
| IMPC RDF | 8 | 5 |
| Metadata of JCM resources | 8 | 9 |
| Metadata of quantitative data and datasets of microscopy images provided from SSBD database | 8 | 10 |
| Cell phenotype | 2 | 4 |

**Table 3. Ontologies referenced by multiple databases**

| Ontology | Number of Databases | Ontology | Number of Databases |
|---|---|---|---|
| Cell line ontology | 13 | Unit ontology | 7 |
| NCBI taxon | 11 | Mouse adult gross anatomy | 5 |
| Phenotypic quality | 10 | Uber anatomy ontology | 4 |
| Current procedural terminology | 10 | Left unit ontology | 3 |
| Cell ontology | 8 | Mouse pathology | 3 |
| OBO-relation ontology | 8 | Zebrafish anatomy and development | 3 |
| Gene ontology | 7 | Clinical measurement ontology | 2 |
| Mammalian phenotype | 7 | Metagenome and microbes environmental ontology | 2 |
| Statistics ontology | 7 | Semanticscience integrated ontology | 2 |

by scanning EM (SEM), provide evidence of detailed morphological phenomena in mammalian tissues and/or cells (Kasthuri *et al.*, 2015). Such comprehensive imaging studies form the basis of 'Morphomics', a field of omics research.

RIKEN is working on the production of the large-scale nanoscale microstructural imaging data of mammalian (mouse and rat) tissues using SEM in a collaborative study with JEOL Ltd. High-spec OM allows us to image multidimensional dynamics at sub-second intervals (Chen et al, 2014), and RIKEN is conducting time-lapse imaging experiments for living cells. In neuroimaging research, RIKEN is performing structural and functional imaging analysis of human and primate brains using MRI and PET, and brain imaging data has been accumulated (Mizuno et al., 2015 and Takahashi et al., 2015). However, a common procedure for data analysis of accumulating images is still not organised. To address this issue, the authors have been developing a common platform to systematically conduct an integrated analysis for these imaging data. Furthermore, using this platform, the authors aimed to carry out a data-driven research for contributing to medical science and health care.

### *6.3.1. Development of OWL Ontology as an Extension of OME Vocabulary*

To reference such image data from experimental results, the authors developed an OWL ontology that describes a variety of imaging metadata, including biological samples and experimental conditions for multiple devices such as OM and EM by extending the data schema of the Open Microscopy Environment (OME), as shown in Figure 9. OME is the de facto standard interoperability toolset for biological imaging data and has been proposed to manage multidimensional and heterogeneous imaging data for optical microscopy (Allan et al., 2012).

Translation of an XML-based OME data model into OWL/RDF format was performed previously (Little et al., 2004); however, the authors have extended the OME data model step by step (Kume et al., 2016), and the ontology OWL file is published at GitHub (https://github.com/imageMetadata/OME). Currently, EM, phenotype data, biological samples and imaging and experimental conditions for SEM can be described as an example of SEM imaging of rat liver tissue in graph structure (Figure 10). Using the microscopy ontology, the authors have generated 2,500 images and their metadata of rat liver tissue and have published the RIKEN CLST Multimodal Microstructure database (http://metadb.riken.jp/db/clstMultimodalMicrostruct) as a prototype. The authors will generate images and metadata for other tissues, such as brain and kidney tissue. This effort is performed in collaboration

**Figure 9. Overview of classes and relations extracted in the RDF-based OME data model and its extension. The circles represent rdfs:Class or owl:Class instances, while the arrows represent their relationships. The grey classes are directly translated from the OME data model and the white classes are the proposed extended classes. The dotted rectangles represent expedient groups of classes for describing the biological sample (biosample) as an object in the image, screening (a type of biological experiment for sorting substances), images, the experimenters who are 'authors' of the image, and the instruments used in imaging experiments. For example, the class ImagingCondition is linked from Image by the relationship, imagingCondition. The 'ome:' prefixes are omitted.**

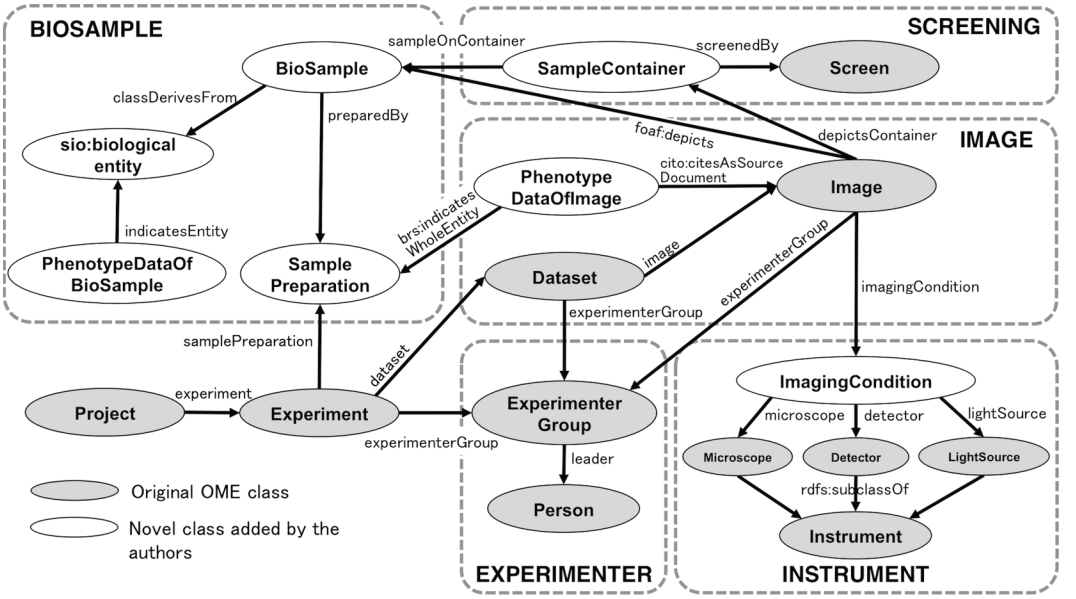
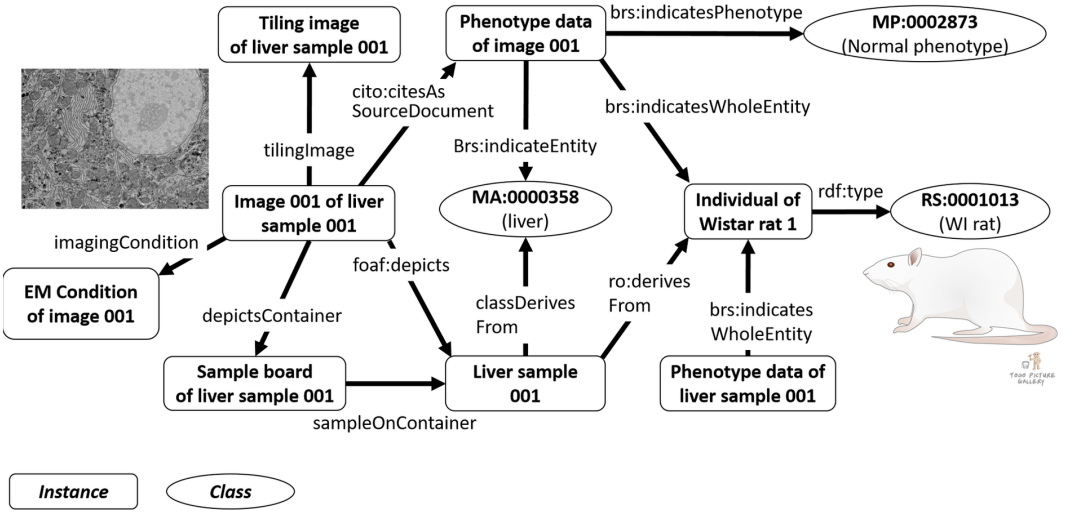



**Figure 10. An example of a graph structure that describes imaging data. The rounded rectangles are instances of classes and the ovals are classes. An electron microscope image (Image 001 of liver sample 001) depicts a liver sample (Liver sample 001) derived from a Wistar rat (Individual of Wistar rat 1). Referencing Image 001 of liver sample 001, phenotype data (Phenotype data of image 001) was produced. The graph structure can link the bioresource information in RIKEN to an external database from RS:0001013.**



with multiple RIKEN research centres. RIKEN plans to use this ontology as a common schema to describe imaging metadata across research centres.

## 7. DISCUSSION: CONTRIBUTIONS OF RIKEN METADATABASE

The authors developed the RIKEN MetaDatabase as a cloud-based database platform that realises Semantic-Web-based data integration with a simplified workflow, implemented in cooperation with biologists and informaticians. Here we comprehensively review our methodology, as well as the development process and operation of the RIKEN MetaDatabase from various perspectives, including: 1) related work, 2) contributions for different types of users, 3) open data promotion, 4) data coordination referencing common data resources, 5) scheme-level integration across databases, and 6) international collaboration.

### 7.1. Related Work

Here, we review existing RDF-based database platforms for life science data publication and integration that are related to the RIKEN MetaDatabase. The technical differences amongst those platforms are summarised in Table 4.

Harvard Catalyst (https://catalyst.harvard.edu) is an information resource-sharing platform for human health research that enables collaboration amongst researchers within a group of 31 institutes, including Harvard University. RDF-based data integration and federated search amongst distributed servers are used as the network's mining tools (Vasilevsky et al., 2012). However, the platform does not aim at hosting the researchers' databases for data integration purposes.

Bio2RDF (http://bio2rdf.org) provides major existing life science datasets by converting them to the RDF format (Belleau et al., 2008). In this case, the creators of the original data

**Table 4. RDF-based database platforms for life science data**

| | | RIKEN MetaDatabase | Harvard Catalyst | Bio2RDF | NCBO BioPortal | NBDC RDF Portal | EMBL-EBI RDF Portal | TogoDB | OpenRefine | RightField |
|---|---|---|---|---|---|---|---|---|---|---|
| Concept | Hosting data | Researcher's data from RIKEN | Data for shared resource | Public databases | Ontologies | Public data and researcher's data | Public databases | Researcher's data | Researcher's data | - |
| | Schema | User defines | Fixed | Host defines | User defines | User defines | Host defines | User defines | User defines | User defines |
| | Generating RDF data from user's data | Y | Y | N | Y | N | N | Y | Y | Y |
| Function | Table view | Y | Y | Y | N | N | N | Y | Y | Y |
| | Tree view | Y | N | N | Y | N | N | N | N | N |
| | SPARQL Endpoint | Y | Y | Y | Y | Y | Y | Y | Y | N |
| | Multiple tables in a database | Y | N | Y | N | N | N | Y | Y | Y |
| | Spreadsheet support for data input | Y | Y | N | N | N | N | Y | Y | Y |
| | Support for data annotation | N | Y | N | Y | N | N | Y | N | Y |

Notes: Researcher's data, in 'hosting data': data uploaded by a user (data creator and publisher).
User defines, in 'schema': The system allows users to define their own data schema.
Host defines, in 'schema': The data schema is fixed by the system and users cannot introduce their own data schema.

and the informaticians who convert the data to RDF format differ. In contrast, in the proposed approach, the original data creators participate in the data integration by converting the data to RDF format themselves.

Similar to the proposed platform, BioPortal (http://bioportal.bioontology.org), which is hosted by the National Center for Biomedical Ontology (NCBO), is a data federation platform based on the RDF and the Web Ontology Language (OWL) (Whetzel et al., 2011). However, the primary focus of BioPortal is data integration and coordination between ontologies, while the authors' approach focuses on directly inter-linking data items across research databases.

The RDF Portal (http://integbio.jp/rdf/), which is hosted by the National Bioscience Database Center (NBDC), and the proposed platform apply a common data integration concept that collects RDF datasets from various fields. The RDF Portal allows researchers from different institutes and universities to combine their RDF datasets. In addition, it provides SPARQL query interfaces for each dataset and across all datasets. In contrast, the proposed platform supports both the generation and collection of RDF data. Furthermore, the RIKEN MetaDatabase provides a data browser that can be used by non-RDF users.

The European Bioinformatics Institute (EBI) RDF Portal (https://www.ebi.ac.uk/rdf/) currently hosts six datasets produced by large-scale projects that many bioinformaticians find indispensable for data analysis and integration. In contrast, the proposed platform is intended to host both large-, middle- and small-scale projects and laboratories that want to contribute to life sciences through the publication of research-based data.

TogoDB (http://togodb.org) provides a simple database service that can generate data in RDF format. Users can deploy a database by simply uploading a CSV file. The TogoDB service is publicly available and users can release multiple data formats, such as CSV, JSON and RDF (XML and Turtle), and its table view is highly customisable. With TogoDB, users can define a multiple-table database using multiple spreadsheets as well as RIKEN MetaDatabase. In contrast, the RIKEN MetaDatabase provides various data views including a tree view.

To generate RDF data, the authors employ a spreadsheet to describe the raw data and convert the data to RDF format. Similarly, OpenRefine (http://openrefine.org/) can generate RDF data from various source files, including a spreadsheet. With OpenRefine, data to be converted to RDF format are defined outside the source files. In contrast, the authors' spreadsheet includes all data to be converted to RDF format, and an RDF expert can easily recognise the RDF data structure from the spreadsheet.

RightField (Wolstencroft et al., 2011) is a tool for editing life science data in spreadsheets by embedding ontology annotations. A RightField spreadsheet allows a user to select terms from a given ontology dataset that includes subclass relations, individuals and combinations.

## 7.2. Contributions for Different Types of Users

Here we discuss the contribution of the RIKEN MetaDatabase from three viewpoints or roles: the database publisher, the database user, and the RIKEN institute. A database publisher is an experimental researcher who publishes research data as a database, while a database user is a researcher, inside or outside RIKEN, who uses data from the RIKEN MetaDatabase for their own research. The RIKEN Institute is the organisation that manages the various research activities, wants to reveal inter-relationships between the research activities via an integrated database, and wants to promote cooperation between different research activities, thereby aiming to produce cutting edge studies.

### *7.2.1. Database Publishers*

A data publisher is a biologist who has research results, converts the data into RDF format and publishes the converted RDF data. The advantages of data generation using a spreadsheet are summarised as follows:

1. New columns can be added easily;
2. Text format data, such as TSV, can be imported easily;
3. The readability of tabular data is high.

Adding new columns is required to include a triplet that corresponds to a new RDF property. This feature is enabled because the RDF format is open for adding new data (open world assumption).

Importing text format data allows bioinformaticians to input data derived through existing techniques where life science data processing is often performed using script languages, such as Perl, and text data are often used for data exchange rather than the RDF graph format.

Finally, the tabular form is suitable for data typing and data confirmation prior to publication.

Especially for biologists, this methodology does not solve RDF-specific difficulties, such as the usage of URIs for data resource identification and selecting suitable vocabularies, including ontologies, properties and classes. However, these difficulties are mitigated by generating spreadsheet templates in collaboration with informaticians.

Data integration based on the RDF is also an advantage to publishers. Though the integration can be realised by creating semantic links from one publisher's data to another's as RDF triplets, a more attractive advantage is that data published later may be linked to existing data already integrated by the original publishers. Furthermore, the appropriate data to link new data can be discovered easily through the tabular view without SPARQL.

#### *7.2.2. Database Users*

Previously, database users had great difficulties discovering the types of databases available, where these were published and how to use them. The RIKEN MetaDatabase collects the metadata of databases published by RIKEN and functions as a single multipurpose database collection. The metadata are published as a database catalogue in the RIKEN Database Directory and the HCLS Community Profile using standardised vocabularies, which help users discover data.

Furthermore, by employing standards for metadata publication, such as the RDF and SPARQL, the RIKEN MetaDatabase provides a standardised API to access data as a SPARQL endpoint. In addition, for users who are unfamiliar with the RDF, the RIKEN MetaDatabase provides intuitive data views, such as the tabular data view, which is a popular form for biologists.

#### *7.2.3. The RIKEN Institute*

Since RIKEN has researchers in various fields, including genome, plant, animal, brain, medical, bioresource and informatics researches, RIKEN handles a wide range of metadata descriptions and biomedical concepts. The development of a novel ontology is required for new types of research data and concepts. RIKEN easily realises collaboration amongst various researchers for internal collaborative research. Consequently, the RIKEN's researchers are accomplishing the difficult task of ontology development. The authors propose that this collaborative metadata integration model should be used in an open environment.

### 7.3. Open Data Promotion

The development of the RIKEN MetaDatabase is a step toward open access to research data. The platform provides easy and interactive access to previously untapped data stored in laboratory records. In addition, the RIKEN MetaDatabase facilitates easy, rapid and cost-effective publication of databases by small laboratories. For example, in the case of ENU-induced Mutations in RIKEN Mutant Mouse Library (http://metadbdev.riken.jp/sandbox/db/BRC-ENU-inducedMutationsInRIKENMutantMouseLibrary), the data developer did not have sufficient expertise and hardware to develop a public database. Using the RIKEN MetaDatabase, they easily published their data on the web. Furthermore, through collaboration with us and RDF experts at the DNA Data Bank of Japan (DDBJ), their data were integrated with data within RIKEN and DDBJ by applying a common data scheme.

### 7.4. Data Coordination Referencing Common Data Resources

The third contribution is data coordination between the different databases hosted by the RIKEN MetaDatabase to share common instance URIs. To describe alleles and genes in mice, the authors applied common gene records (http://metadb.riken.jp/metadb/db/mgi_rdf) imported from the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org). The MGI project approved the publication of the RDF version of their mouse gene records. The authors have promoted the common use of MGI gene records in the RIKEN MetaDatabase. As a result, MGI records are used in multiple databases, such as Metadata of BRC mouse resources and phenotypes (http://metadb.riken.jp/metadb/db/rikenbrc_cell), Metadata of Functional Glycomics with KO mice database (http://metadb.riken.jp/metadb/db/Glycomics_mouse) and the International Mouse Phenotype Consortium (IMPC; http://www.mousephenotype.org) RDF data (http://metadb.riken.jp/metadb/db/IMPC_RDF). In these databases, the data items related to genes are linked to MGI allele or gene records. Through this association, integrated information, including public experimental material (mouse strain in this case), that corresponds to phenotype data published by the IMPC, can be obtained, as shown in Figure 6.

### 7.5. Scheme-Level Integration Across Databases

The fourth contribution is related to scheme-level integration.

#### *7.5.1. Case 1: RIKEN Mutant Mouse Library*

In the ENU-induced Mutations in the RIKEN Mutant Mouse Library, next-generation sequencing (NGS) metadata are described in the common RDF scheme, which is developed in cooperation with DDBJ and RIKEN, based on the broadly used XML scheme for NGS metadata. Using this scheme, RIKEN plans to develop a unified pipeline to publish NGS metadata on the web and deposit NGS data as public archives operated by DDBJ. It is expected that this pipeline will promote worldwide sharing of NGS data from RIKEN.

#### *7.5.2. Case 2: Metadata from the Japan Collection of Microorganism*

Metadata from the Japan Collection of Microorganisms (JCM) resources (http://metadb.riken.jp/metadb/db/rikenbrc_jcm_microbe) are described based on a common RDF scheme for microorganism strains, i.e. MCCV, which is used by the MicrobeDB.jp project (http://microbedb.jp/). The integrated database represents an encyclopaedia of microbes based on metagenome data. By applying the MCCV, basic information about microbe strains released by the JCM can be related to the metagenome data in the MicorbeDB.jp project.

#### *7.5.3. Case 3: Phenotype Data of Experimental Animals*

Phenotype data of experimental animals are also integrated by the J-phenome project (http://jphenome.info). J-phenome is a portal of phenotype databases hosted by the RIKEN MetaDatabase in which the RDF scheme for the description of animal phenotypes are unified using common phenotype ontologies, such as the Mammalian Phenotype Ontology and PATO. The unified scheme contributes to the development of a common application to determine cross-species relationships between phenotypes using an inter-ontology relationship library produced by machine reasoning (Hoehndorf, et al., 2011) (Robinson et al., 2011).

#### *7.5.4. Case 4: Imaging Metadata*

The ongoing project to share imaging metadata (Section 6.3) aims at scheme-level integration of various imaging data, including imaging and experimental conditions, biological samples and phenotypes. Currently, SEM imaging data can be described using the authors' ontology and the RIKEN MetaDatabase to successfully realise scheme-level integration for SEM. As a next step, integration of current imaging data with MRI imaging data of primate brains is planned.

In summary, scheme-level integration in the RIKEN MetaDatabase contributes to the common use of query, application and workflow pipelines to handle the same (or similar) data across multiple databases.

### 7.6. International Collaboration

The integration of multiple datasets in the RIKEN MetaDatabase contributes to international collaboration. The IMPC is an umbrella of comprehensive phenotyping of mouse mutants (Dickinson et al., 2016). Through cooperation with the IMPC, multiple research centres have released measurement data produced from the standardised phenotyping pipeline. As a member of the IMPC, RIKEN BRC has produced an RDF version of the IMPC phenotype data, including more than 50 million triplets, which is now hosted by the RIKEN MetaDatabase. Although the IMPC website provides a rich interface to visualise various phenotype data, the RDF version of the IMPC data in the RIKEN MetaDatabase can be used by data scientists who want to integrate these phenotype data with other datasets from different databases. For example, using the SPARQL endpoint of the RIKEN MetaDatabase (http://metadb.riken.jp/sparql), a data user can perform a federated query between RIKEN and the EBI RDF platform (https://www.ebi.ac.uk/rdf/) to retrieve what phenotype can be expressed when a specific biological pathway is inactivated by utilising a connection between the IMPC and Reactome (http://www.reactome.org) datasets.

In summary, using the RIKEN MetaDatabase, seamless data integration can be performed from an inner-research institute level to a worldwide level.

## 8. FUTURE RESEARCH DIRECTION

The RIKEN MetaDatabase is a simple database system and platform built on RIKEN's private cloud infrastructure. Data generation and publication costs for biologists are reduced because they do not need to prepare and operate their own servers. Since the system does not support data access control, it cannot handle private datasets or datasets that are under development. However, the system is lightweight and requires only two virtual machines. Therefore, the authors built multiple private instances of the system on the private cloud for ongoing research projects with their own access control using firewalls. However, federated SPARQL search amongst arbitrary instances of the system is not yet available. Future work will include the development of effective federation amongst such instances as an ideal database federation model on the web.

Since the publication of the RIKEN MetaDatabase in April 2015, efforts toward data dissemination have continued. Thus far, the authors have participated in international database projects, such as the IMPC for mouse phenotype databases and W3C's HCLS group for a database profile, such that RIKEN's published metadata can be easily linked to other published datasets.

Sharing a data schema is an important factor for data integration. The authors will continue to promote collaboration amongst researchers, including biological scientists and informaticians. To facilitate schema sharing, the authors are working on the development of an OME ontology, which is a common imaging metadata schema (Section 6.3). The authors will expand this ontology for wider use (e.g. MRI imaging). The authors also plan to set up a RIKEN working user group to discuss the application or construction of common data schema for various data items or concepts.

Currently, the RIKEN MetaDatabase does not provide support for finding the proper terms in the ontologies of other databases for data publishers in RIKEN. Discussion-based co-operation amongst database developers is currently being promoted by forming a working group of representatives from various RIKEN research centres. The co-operation includes deciding on guidelines for sharing ontology terms, common vocabularies, properties, and data schema. To promote the reuse of common URIs, data publishers want to implement an automatic ontology annotation function in the RIKEN MetaDatabase. To address this issue, the use of existing applications may be appropriate. For example,

RightField may prove useful in the extension of Excel spreadsheets, thereby allowing semi-automatic ontology annotation in the data construction workflow.

## 9. CONCLUSION

In this paper, we have discussed the requirements specifications, design, development and operation of the RIKEN MetaDatabase. One of the major difficulties is the practical co-localisation of an open data RDF and the development of simple data processing methods for biologists. To address these issues, the authors have developed a template spreadsheet for data creation, which is a GUI that realises intuitive data views including a tabular view. The database platform is deployed on RIKEN's private cloud infrastructure and multiple system instances can be generated. Thus far, data integration from different research fields, such as the IMPC, has been successfully realised using the RIKEN MetaDatabase.

Future work includes the realisation of practical federation amongst multiple public system instances to construct an integrated database that supports the authors' proposed data views. This will be accomplished by developing an individual database for each research project in a distributed environment and intelligent support systems to select suitable vocabularies for biologists.

## ACKNOWLEDGMENT

This paper is dedicated to RIKEN's centennial.

## REFERENCES

Allan, C., Burel, J. M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., & Swedlow, J. R. et al. (2012). OMERO: Flexible, model-driven data management for experimental biology. *Nature Methods*, *9*(3), 45–53. doi:10.1038/nmeth.1896 PMID:22373911

Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–716. doi:10.1016/j.jbi.2008.03.004 PMID:18472304

Chen, B. C., Legant, W. R., Wang, K., Shao, L., Milkie, D. E., Davidson, M. W., & Betzig, E. et al. (2014). Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science*, *346*(6208), 1257998. doi:10.1126/science.1257998 PMID:25342811

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., & Murray, S. A. et al. (2016). High-throughput discovery of novel developmental phenotypes. *Nature*, *537*(7621), 508–514. doi:10.1038/nature19356 PMID:27626380

Forrest, A. R. R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M. J. L., Haberle, V., & Hayashizaki, Y. et al.The FANTOM Consortium & the RIKEN PMI & CLST (DGT). (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. doi:10.1038/nature13182 PMID:24670764

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2011). PhenomeNET: A whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, *39*(18), e119. doi:10.1093/nar/gkr538 PMID:21737429

Ichikawa, T., Nakazawa, M., Kawashima, M., Iizumi, H., Kuroda, H., Kondou, Y., & Matsui, M. et al. (2006). The FOX hunting system: An alternative gain-of-function gene hunting technique. *The Plant Journal*, *45*(6), 974–985. doi:10.1111/j.1365-313X.2006.02924.x PMID:17227551

Kasthuri, N., Hayworth, K. J., Berger, D. R., Schalek, R. L., Conchello, J. A., Knowles-Barley, S., & Lichtman, J. W. et al. (2015). Saturated Reconstruction of a Volume of Neocortex. *Cell*, *162*(3), 648–661. doi:10.1016/j.cell.2015.06.054 PMID:26232230

Kume, S., Masuya, H., Kataoka, Y., & Kobayashi, N. (2016). Development of an Ontology for an Integrated Image Analysis Platform to enable Global Sharing of Microscopy Imaging Data. In *Proceeding of 15th International Semantic Web Conference, poster session*.

Little, S., & Hunter, J. (2004). Rules-By-Example a Novel Approach to Semantic Indexing and Querying of Images. In *Proceeding of 3rd International Semantic Web Conference*. doi:10.1007/978-3-540-30475-3_37

Mizuno, K., Kawatani, J., Tajima, K., Sasaki, A. T., Yoneda, T., Komi, M., & Watanabe, Y. et al. (2016). Less efficient and costly processes of frontal cortex in childhood chronic fatigue syndrome. *NeuroImage. Clinical*, *12*, 600–606. doi:10.1016/j.nicl.2016.09.016 PMID:27709065

Nakamura, Y. (2010). Bio-resource of human and animal-derived cell materials. *Experimental Animals*, *59*(1), 1–7. doi:10.1538/expanim.59.1 PMID:20224164

Robinson, P. N., Köhler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., & Smedley, D. et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, *24*(2), 340–348. doi:10.1101/gr.160325.113 PMID:24162188

Sugahara, M., Asada, Y., Shimada, H., Taka, H., & Kunishima, N. (2009). HATODAS II: Heavy-atom database system with potentiality scoring. *Journal of Applied Crystallography*, *42*(3), 540–544. doi:10.1107/S0021889809012370

Takahashi, K., Mizuno, K., Sasaki, A.T., Wada, Y., Tanaka, M., Ishii, A., Tajima K., Tsuyuguchi, N., Watanabe K., Zeki S., & Watanabe Y. (2015). Imaging the passionate stage of romantic love by dopamine dynamics. *Frontiers in Human Neuroscience, 09 April 2015*.

Vasilevsky, N., Johnson, T., Corday, K., Torniai, C., Brush, M., Segerdell, E., & Haendel, M. et al. (2012). Research resources: Curating the new eagle-i discovery system. *Database (Oxford)*, 20. PMID:22434835

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., & Musen, M.A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res. 39*(Web Server issue), W541-545.

Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J. L., & Goble, C. et al. (2011). RightField: Embedding ontology annotation in spreadsheets. *Bioinformatics (Oxford, England)*, *27*(14), 2021–2012. doi:10.1093/bioinformatics/btr312 PMID:21622664

Yamaguchi, A., Kozaki, K., Lenz, K., Wu, H., Yamamoto, Y., & Kobayashi, N. (2015). Efficiently finding paths between classes to build a SPARQL query for life-science databases. *The 5th Joint International Conference (JIST 2015), LNCS,* 9544, 321-330.

Yokoyama, K. K., Murata, T., Pan, J., Nakade, K., Kishikawa, S., Ugai, H., & Obata, Y. et al. (2010). Genetic materials at the gene engineering division, RIKEN BioResource Center. *Experimental Animals*, *59*(2), 115–124. doi:10.1538/expanim.59.115 PMID:20484845

Yoshiki, A., Ike, F., Mekada, K., Kitaura, Y., Nakata, H., Hiraiwa, N., & Obata, Y. et al. (2009). The mouse resources at the RIKEN BioResource center. *Experimental Animals*, *58*(2), 85–96. doi:10.1538/expanim.58.85 PMID:19448331