A RDF-Based Portal of Biological Phenotype Data Created in Japan

Terue Takatsuki[1], Mikako Saito[1], Sadahiro Kumagai[2], Eiki Takayama[1], Kazuya Ohshima[1], Nozomu Ohshiro[1], Kai Lenz[3], Nobuhiko Tanaka[1], Norio Kobayashi[3,1] and Hiroshi Masuya[1,3],

[1] RIKEN BioResource Center, 3-1-1 Kouyadai, Tsukuba, Japan
[2]Hitachi Ltd., Omori Bellport B Bldg., Shinagawa, Tokyo, Japan
[3]Advanced Center for Computing and Communication, RIKEN, 2-1, Hirosawa, Wako, Japan
{hmasuya, takatter, e_takayama, kazuya22, ohshiro-n , nobtanak}@brc.riken.jp
{ kai.lenz, norio.kobayashi}@riken.jp
sadahiro.kumagai.jj@hitachi.com

Abstract. We developed RDF-based databases of the phenotypes and animal strains produced in Japan and a portal site termed as "J-Phenome". By the application of common schema, these databases can be retrieved by the same SPARQL query across graphs. In the operation of these databases, RDF represented multiple advantages such as improvement of comprehensive search, data integration using ontologies and public data, reuse of data and wider dissemination of phenotype data compared to conventional systems.

Keywords: Biological phenotype, data integration, RDF

1 Introduction

In life science, "Phenotype", a biological characteristic, which an organism shows as a result of interaction of genes and environment, is critical information for researchers and the scientific community in order to choose appropriate experimental materials for their studies.

In this context, the recent sharing of phenotype data is performed by various databases using phenotype ontologies. For example, Mammalian Phenotype ontology (MP) [1] and Human Phenotype ontology (HP) [2] are often used as standardized vocabularies of phenotypes and symptoms. Equivalent links between MP and HP terms, candidates of disease model animals can be shown through the disease-phenotype relationship across human, rodents, fish, worms and flies [3,4]. These data integrations are performed by Semantic Web technologies. Therefore, dissemination of phenotype-related information, RDF technology seems to be one of the best solutions.

We introduced a data integration project, "J-Phenome" (http://jphenome.info/), for wider dissemination of phenotype data produced in Japan using the RIKEN MetaDatabase (http://metadb.riken.jp/) as an infrastructure of the RDF data handling. In this paper, we overview the development of RDF datasets and discuss advantages of RDF for sharing phenotype data in Japan and worldwide.

2 Results

2.1 Collection of the phenotype data

For integration of the phenotype data available in the Japan databases datasets were collected from the original databases and are summarized in Table 1.

2.2 Conversion to the RDF data

For the conversion of phenotype and animal strain data, we used a RDF-based data schema, Bioresource Schema (BRS: http://metadb.riken.jp/metadb/db/bioresource_

Table 1. List of databases in J-Phenome.

Database project in J-phenome	Upper: URL of original site Lower: URI of graph (URL of database project)	Explanation of database	Number of triples	Ontologies used*
IMPC RDF data	http://www.mousephenotype.org http://metadb.riken.jp/db/IMPC_RDF	RDF version of phenotype data produced by International Mouse Phenotyping Consortium (IMPC)	53,902,370	MP, UO
NBRP Medaka Phenotype Metadata	https://shigen.nig.ac.jp/medaka/ http://metadb.riken.jp/db/NBRP_medaka	This database provides metadata of NBRP medaka database.	12,473	NCBITaxon, ZP, UO, PATO, RO
Metadata of NBRP Rat	http://www.anim.med.kyoto-u.ac.jp/nbr/ http://metadb.riken.jp/db/NBRP_rat	This database provides metadata of NBRP Rat Phenotype Database.	655,347	MP, RS, CL, GO, IAO, MA, NCBITaxon, UO, PATO, PR, RO
NIG Mouse Phenotype Database Metadata	http://molossinus.lab.nig.ac.jp/phenotype/ http://metadb.riken.jp/db/Nig_consomic_mouse	This database provides metadata of NIG Mouse Phenotype Database.	26,058	MP, CL, FMA, GO, IAO, MA, NCBITaxon, CMO, UO, PATO, RO
Metadata of BRC cell resources	http://cell.brc.riken.jp/ http://metadb.riken.jp/db/rikenbrc_cell	Database of cell lines available in the RIKEN BioResouce Center.	203,562	CLO, ORDO, CL, GO, IAO, NCBITaxon, UO, PATO, RO
Metadata of JCM resources	http://jcm.brc.riken.jp/ http://metadb.riken.jp/db/rikenbrc_jcm_microbe	Database of microbial strains available in Japan Collection of Microorganisms (JCM) in RIKEN BioResouce Center.	618,938	NCBITaxon, MCCV, MEO, CSSO, PDO, UO, RO
Metadata of BRC mouse resources and phenotypes	http://mus.brc.riken.jp/ http://metadb.riken.jp/db/rikenbrc_mouse	Database of mouse strains available in RIKEN BioResouce Center.	513,509	MP, CL, FMA, GO, IAO, MA, NCBITaxon, UO, PATO, PR, RO, UBERON, CHEBI, NBO
Metadata of Functional Glycomics with KO mice database	http://jcggdb.jp http://metadb.riken.jp/db/Glycomics_mouse	RDF based mata-database of "Functional Glycomics with KO mice database" in Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB).	1,462	MP, CL, GO, IAO, MA, NCBITaxon, PATO, PR, RO,

*Abbreviations: CLO: Cell line ontology, RS: Rat Strain Ontology, MP: Mammalian Phenotype Ontology, ORDO: Orphanet Rare Disease Ontology, CHEBI: Chemical Entities of Biological Interest Ontology, CL: Cell Ontology, FMA: Foundational Model of Anatomy, GO: Gene Ontology, IAO: Information Artifact Ontology, MA: Mouse Adult Gross Anatomy Ontology, MPATH: Mouse Pathology Ontology, NBO: Neuro Behavior Ontology, NCBITaxon, PATO: Phenotypic Quality Ontology, PR: Protein Ontology, RO: Relations Ontology, UBERON (Uber Anatomy Ontology, UO: Units of Measurement Ontology, ZP: Zebrafish phenotypes, CMO: Clinical Measurement Ontology, MCCV: Microbial Culture Collection Vocabulary, MEO: Metagenome and Microbes Environmental Ontology, CSSO: Clinical Signs and Symptoms Ontology, PDO: Pathogenic Disease Ontology

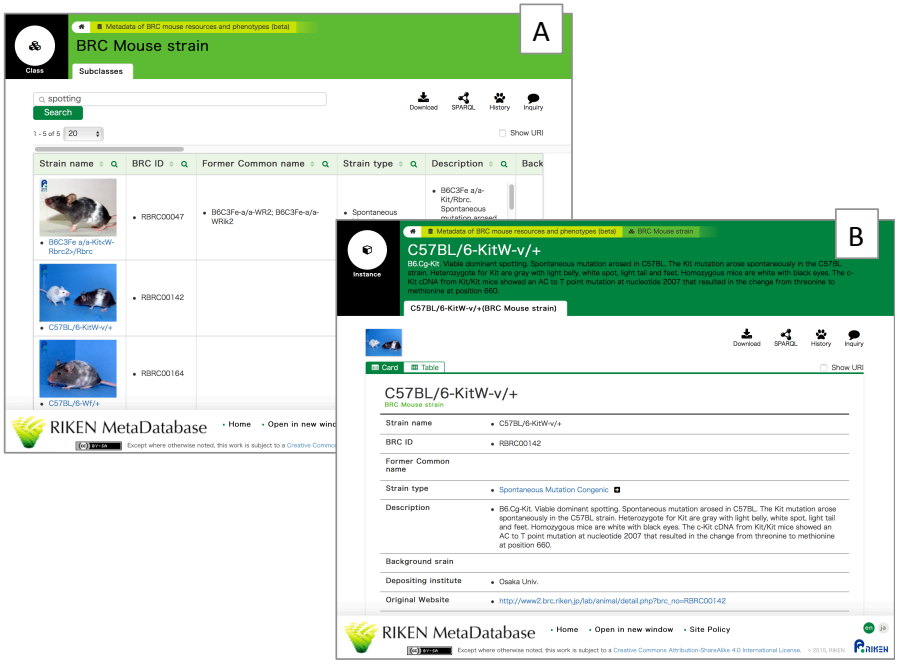

Fig. 1. Browsing of animal strain data in the RIKEN MetaDatabase. The panel (A) shows tabular window lists of the mouse strains. The panel (B) shows a card-view of one strain as an example.

schema). BRS provides standardized properties to describe attributes of instances. BRS-based data conversion was performed using the data-conversion function of RIKEN MetaDatabase [5]. As a result, we made eight RDF graphs for individual databases. As “common vocabularies”, Open Biomedical Ontologies were used for the annotations (Table 1).

2.3 Visualization and provision of the RDF data

The RDF data of phenotypes are visualized by RIKEN MetaDatabase, which provides functions of Web-based data browsing, (Fig. 1) downloading and SPARQL endpoint (http://metadb.riken.jp/sparql) that can process a query across graphs (databases).

3 Discussion

In the preceding section, we described the outline of RDF-based data integration in the J-Phenome project. Using the GUIs of the RIKEN MetaDatabase, biologists who are inexperienced in RDF data can easily browse and explore links of the RDF (Fig. 1).

1) Datasets of J-Phenome will contribute to the data integration in RIKEN MetaDatabase through use and provide common URIs for genes and bio-resources [5].

2) Comprehensiveness in the data retrieving. As a result of coordination of data by the common schema, especially utilizing common properties, cross-graph (cross-database) search using same or similar query(ies) were realized. Particularly, with the SPARQL endpoint, a single query can be applicable for phenotype data search of phenotype data in different species.
3) For the cooperation with external database(s) to share RDF datasets, we imported interrelationship data from Monarch Initiatives (https://monarchinitiative.org) [2], which easily expands the J-Phenome to show candidate diseases related with phenotypes. This expansion contributes to add "values" of Japanese animal strains as a "disease model". In addition, data import of the "opposite direction" is also useful.

We are currently planning to export the Japanese phenotype data to Monarch Initiative for wider dissemination of the Japanese phenotype data. J-Phenome data will be updated independently from Monarch Initiative's data, and will require frequent synchronizations to provide the latest data to users. However, we currently do not apply federated query because of problems on the performance. Improvement of performance on the federated query of SPARQL-related technology is expected to expand inter-database cooperation with the RDF.

Acknowledgements

We thank Drs. K. Naruse and H. Kaneko in National Institute for Basic Biology, T. Kuramoto in Kyoto Univ., T. Mashimo in Osaka Univ. T. Takada and K. Kawakami in National Institute of Genetics for giving useful advices and phenotype annotation in the conversion of original phenotype data into RDF. We also thank for H. Mori in Tokyo Institute of Technology, S. Kawashima in Database Center for Life Science and S. Carbon in Lawrence Berkeley National Laboratory for useful discussion for interoperability between databases. This work is partially supported by Database Integration Coordination Program (DICP) of National Bioscience Database Center (NBDC) / Japan Science and Technology Agency (JST).

References

1. Hoehndorf R, Schofield PN, Gkoutos GV.: PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Res. 18, e119 (2011).
2. Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PN, Mungall CJ.: Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. F1000Res. (2013).
3. Smith CL1, Goldsmith CA, Eppig JT.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 6, R7, (2005)
4. Köhler S, Doelken SC, Mungall CJ et al.: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 42(Database issue):D966--974. (2014).
5. Lenz K., Masuya H., Kobayashi N. RIKEN MetaDatabase: a database publication platform for RIKENs life-science researchers that promotes research collaborations over dierent research area. The 15th International Semantic Web Conference (ISWC2016), poster.