A Semantic Web-Based Infrastructure for Purpose-Driven Retrieval of Life Science Bioresources

Tatsuya Kushida[1], Daiki Usuda[1], Masanobu Yamagata[1], Norio Kobayashi[2, 1], Shoichiro Shindo[1], Tatsuya Yamada[1], Yuki Yamagata[2, 1] and Hiroshi Masuya[1, *]

[1] *BioResource Research Center, RIKEN, Koyadai 3-1-1, Tsukuba, Ibaraki, Japan*
[2] *RIKEN Information R&D and Strategy Headquarters, 2-1 Hirosawa, Wako, Saitama, Japan*

Abstract

In the life sciences, the shared use of research materials is essential for ensuring experimental reproducibility. These materials are commonly referred to as biological resources. To support life science research, biological resource centers have been established worldwide as institutional platforms for providing such resources. One of the core functions of these centers is the dissemination of information about available materials. The RIKEN BioResource Research Center, one of the major bioresource centers in Japan, has been offering a knowledge-based search system for life scientists since 2018. This system leverages Semantic Web technologies to provide detailed biological characteristics of the resources it manages. By integrating bioresource data, public life science datasets, and ontologies through a SPARQL endpoint backend, the system enables researchers to explore relevant research materials from diverse scientific perspectives via a dedicated search interface. Furthermore, the use of Semantic Web technologies contributes to sustainable and scalable system operation. This report outlines the usefulness of the system based on several years of operation, as well as a new search system developed to address the shortcomings that had been identified.

Keywords

biological resource, experimental material, bioinformatics, semantic web, ontology

1. Introduction

Life science research heavily relies on the biological materials used in experiments. Due to genetic variation across and within species, experimental outcomes are strongly influenced by the specific materials employed. Ensuring reproducibility therefore requires the preservation, maintenance, and shared use of these biological resources. To support this need, biological resource centers have been established worldwide as repositories for research materials. These centers play a vital role in providing information that helps researchers discover and select suitable resources.

A major challenge in information dissemination by these centers lies in addressing two kinds of diversity: the diversity of the biological resources themselves, and the diversity of research needs in the life sciences. Life science research spans basic biology to applied fields such as

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

[*] Corresponding author.

tatsuya.kushida@riken.jp (T. Kushida); hiroshi.masuya@riken.jp (H. Masuya)
0000-0002-0784-4113 (T. Kushida); 000-0002-3392-466X (H. Masuya)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

medicine, environmental science, and energy. It employs a wide range of methodologies, from macro-level studies of whole organisms to micro-level molecular analyses. Researchers examine biological functions from multiple perspectives to uncover fundamental mechanisms and develop new applications.

To support such research, biological resources—including organisms, cells, and DNA—must be accompanied by integrated information. This information needs to be presented in a way that is accessible and meaningful to life scientists. The RIKEN BioResource Research Center (RIKEN BRC) serves as a global repository of experimental materials such as mice, plants, cells, DNA, and microorganisms [1–4]. Since its founding in 2001, RIKEN BRC has offered online catalogs based on relational databases. In response to evolving research needs—particularly the need to convey biological characteristics and to support cross-type resource searches—the center launched a new system in 2018 that utilizes Resource Description Framework (RDF) and Semantic Web technologies to achieve enhanced data integration [5]. Since its launch, this system has been continuously operated, undergoing data updates and enhancements while also improving and adding search software according to evolving needs.

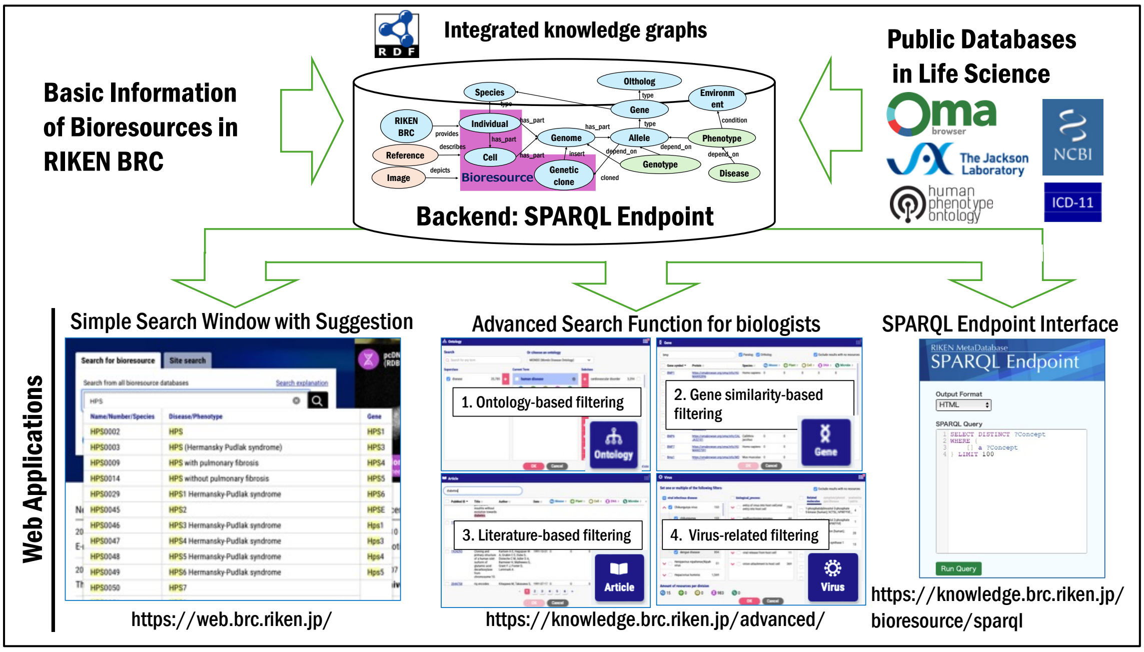
Integrated knowledge graphs
RDF
Basic Information of Bioresources in RIKEN BRC
Public Databases in Life Science
Species
Oltholog
Environment
Gene
RIKEN BRC
Individual
Genome
Allele
Phenotype
Reference
Cell
Genotype
Disease
Image
Bioresource
Genetic clone
type
has_part
provides
describes
depicts
insert
depend_on
condition
cloned
Backend: SPARQL Endpoint
Oma browser
The Jackson Laboratory
NCBI
human phenotype ontology
ICD-11
Web Applications
Simple Search Window with Suggestion
Advanced Search Function for biologists
SPARQL Endpoint Interface
1. Ontology-based filtering
Ontology
2. Gene similarity-based filtering
Gene
3. Literature-based filtering
Article
4. Virus-related filtering
Virus
SPARQL Endpoint
Run Query
https://web.brc.riken.jp/
https://knowledge.brc.riken.jp/advanced/
https://knowledge.brc.riken.jp/bioresource/sparql

Figure 1: Overview of the system configuration of the knowledge base for bioresources.

2. Overview of the knowledge base and query system

The bioresource search system at RIKEN BRC consists of multiple web-based search applications supported by an RDF repository implemented using Virtuoso (OpenLink). These web applications are integrated into the official RIKEN BRC website (https://web.brc.riken.jp), providing seamless access to various search functionalities. The backend system and several web applications are interconnected via APIs based on the SPARQLIST framework [6]. A public SPARQL endpoint is also available, allowing users to directly query the RDF repository (https://knowledge.brc.riken.jp/bioresource/sparql).

To meet the diverse needs of life science research, the RDF repository integrates not only the core metadata of RIKEN BRC's bioresources but also public life science datasets and ontologies [7–14] (Table 1). These data encompass both intrinsic genetic information and observable phenotypic traits inferred from genotypes. The data model follows the structural principles established by the OBO Foundry [15] and other relevant ontological frameworks [16, 17], ensuring semantic consistency and extensibility. The system employs tailored query patterns for each data source to enable practical integration of heterogeneous datasets.

Each dataset is managed as a named graph within the repository and is updated regularly. The integrated graphs include both datasets originally published in RDF and those converted in-house. The use of RDF technologies promotes interoperability across datasets via shared URIs, contributing to the long-term sustainability and cost-effectiveness of the knowledge base infrastructure.

The current knowledge base has grown significantly and now encompasses 350 named graphs and a total of 6,899,700,302 triples in total, which are continuously expanded and updated. These figures demonstrate the robustness and scalability of our system in integrating large-scale knowledge within the life science domain.

Table 1: List of integrated public ontologies and RDF datasets.

Category	Datasets	License
Genome/Proteome	Mouse Genome Informatics (MGI) RDF	Original
	UniProt RDF (including Gene Ontology)	CC BY 4.0
Ortholog	Orthologous Matrix (OMA) RDF	CC BY-SA 2.5
Gene-disease	DisGeNET RDF	CC BY-NC-SA 4.0
	MedGen RDF	Original
	KEGG MEDICUS RDF	CC BY-SA 4.0
Disease	Nanbyo Disease Ontology (NANDO)	CC BY 4.0
	Human Disease Ontology (DOID)	CC0 1.0
	Mondo Disease Ontology (MONDO)	CC BY 4.0
	Orphanet Rare Disease Ontology (ORDO)	CC BY 4.0
Phenotype	Mammalian Phenotype Ontology (MP)	CC BY 4.0
	Human Phenotype Ontology (HP)	Original
Chemicals	Chemical Entities of Biological Interest (ChEBI)	CC BY 4.0
Biochemical reaction	Knowledgebase of biochemical reactions (Rhea)	CC BY 4.0
Gene expression	Bgee: gene expression data in animals	CC0 1.0

3. Search Interfaces

To enhance usability for bioresource users, we have developed two types of search interfaces.

The first is a simple search function embedded in the top page of the RIKEN BRC website (https://web.brc.riken.jp). This interface provides a single search box that allows users to search for bioresources using keywords such as gene names or human disease terms. A suggestion list, generated by crawling the RDF repository, assists users in formulating their queries and retrieving relevant results. Search results are presented in a unified format across five different categories of bioresources (e.g., a sample of search result for “diabetes”).

A major technical challenge in implementing this function was the poor SPARQL performance when executing deep queries across multiple graphs. To overcome this, we created a simplified RDF graph by crawling the deeper knowledge graph at the time of each data update. The search interface operates on this optimized graph to ensure fast and responsive performance. Specifically, queries involving complex inferences or numerous joins tended to exhibit long response times.
To address this technical challenge, we construct a "shortened knowledge graph" optimized for simple keyword searches and basic filtering functionalities. It is created by crawling the original deep knowledge graph and extracting/aggregating only entities that are linked to BRC’s bioresources (e.g., genes, phenotypes, and diseases) and their primary associated properties (e.g., http://purl.obolibrary.org/obo/RO_0002200 (has phenotype)). Specifically, it is a reconstructed graph centered on essential URIs along with their labels, IDs, and classification information. This graph is specialized for particular search requirements, significantly reducing data volume to ensure fast and responsive performance for the most frequently utilized search patterns, such as keyword searches. This allows users to obtain search results quickly, with the option to execute more complex detailed queries against the original deep knowledge graph if needed. Over several years of operation this strategy has proven highly effective in improving system responsiveness and scalability, offering a practical solution to performance challenges in real-world applications of Semantic Web technologies.

The utilization of bioresources spans a wide range of fields. Over several years of operation, the need for a search function linked to more detailed conditions and research outcomes has been identified. To address this need, in addition to the simple search, we have newly implemented an advanced search interface (https://knowledge.brc.riken.jp/advanced/en/) for users who wish to conduct more specific or complex queries. This advanced interface offers four filtering options:

1. Ontology-based filtering, which uses hierarchical structures from multiple ontologies, including Gene Ontology (GO), NCBI Taxonomy, Chemical Entities of Biological Interest (ChEBI), and Mammalian Phenotype (MP) ontologies [10–13] (Tutorial 1).
2. Gene similarity-based filtering, which enables users to identify bioresources related to genes based on sequence or evolutionary similarity, such as orthologs and paralogs (Tutorial 2).
3. Literature-based filtering, which allows users to search for bioresources mentioned in scientific publications (Tutorial 3).
4. Virus-related filtering, which supports searches based on associations with infection-related processes, such as those observed in viral diseases like COVID-19 (Tutorial 4).

By combining these filters, users can generate tailored lists of bioresources that meet complex research criteria (Tutorial 5).

4. Future Challenges

Over the past six years, we have developed and operated a public database that provides information on bioresources—fundamental assets in life science research—based on RDF-based Semantic Web technologies. Because this database aims to deliver information grounded in domain-specific knowledge, the use of RDF has proven effective in reducing operational costs, supporting sustainable system maintenance, and ensuring alignment with the FAIR principles.

One persistent challenge, however, is the limited performance of RDF repositories when processing complex queries over deeply nested knowledge graphs. While we have mitigated this issue by constructing a shortened knowledge graph to improve responsiveness, several limitations remain apparent—particularly in search functionalities that are standard in general-purpose systems, such as partial keyword matching and relevance-based ranking. These features are not natively supported by graph-based RDF technologies.

To overcome these challenges, we plan to integrate a full-text search engine to enhance search capabilities. Additionally, the application of large language models (LLMs) has recently attracted attention as a means to further improve usability. In particular, the adoption of Domain-Expert-Guided Large Language Models (DEG-LLMs)—a class of LLMs fine-tuned using expert-curated knowledge—is anticipated to play a key role in future development.

Acknowledgements

This work is funded by the Management Expenses Grant for RIKEN BioResource Research Center, MEXT (http://www.mext.go.jp/), and ROIS-DS-JOINT (045RP2024).

Declaration on Generative AI

We used a large language model (LLM) to proofread and refine the English expression of this paper. The content and core ideas of the paper were entirely developed by the authors.

References

[1] Yokoyama KK, Murata T, Pan J, Nakade K, Kishikawa S, Ugai H, Kimura M, Kujime Y, Hirose M, Masuzaki S, Yamasaki T, Kurihara C, Okubo M, Nakano Y, Kusa Y, Yoshikawa A, Inabe K, Ueno K, Obata Y. Genetic materials at the gene engineering division, RIKEN BioResource Center. Exp Anim. 2010;59(2):115-24. doi: 10.1538/expanim.59.115.

[2] Yoshiki A, Ike F, Mekada K, Kitaura Y, Nakata H, Hiraiwa N, Mochida K, Ijuin M, Kadota M, Murakami A, Ogura A, Abe K, Moriwaki K, Obata Y. The mouse resources at the RIKEN BioResource center. Exp Anim. 2009 Apr;58(2):85-96. doi: 10.1538/expanim.58.85.

[3] Nakamura Y. Bio-resource of human and animal-derived cell materials. Exp Anim. 2010;59(1):1-7. doi: 10.1538/expanim.59.1.

[4] Mizuno-Iijima S, Nakashiba T, Ayabe S, Nakata H, Ike F, Hiraiwa N, Mochida K, Ogura A, Masuya H, Kawamoto S, Tamura M, Obata Y, Shiroishi T, Yoshiki A. Mouse resources at the RIKEN BioResource Research Center and the National BioResource Project core facility in Japan. Mamm Genome. 2022 Mar;33(1):181-191. doi: 10.1007/s00335-021-09916-x.

[5] Masuya H, Usuda D, Nakata H, Yuhara N, Kurihara K, Namiki Y, Iwase S, Takada T, Tanaka N, Suzuki K, Yamagata Y, Kobayashi N, Yoshiki A, Kushida T. Establishment and

application of information resource of mutant mice in RIKEN BioResource Research Center. Lab Anim Res. 2021 Jan 18;37(1):6. doi: 10.1186/s42826-020-00068-8.

[6] SPARQList URL: https://dbcls.rois.ac.jp/services-en.html#SPARQList

[7] Kushida T, Farias TM, Sima AC, Dessimoz C, Chiba H, Bastian FB, Masuya H. Exploring Disease Model Mouse Using Knowledge Graphs: Combining Gene Expression, Orthology, and Disease Dataset. BioRxiv, August 31, 2023. https://doi.org/10.1101/2023.08.30.555283

[8] Kushida T, Farias T, Sima A, Dessimoz C, Chiba H, Bastian F, and Masuya H. Federated SPARQL query performance evaluation for exploring disease model mouse: combining gene expression, orthology, and disease knowledge graphs. 2025 May 16;25(Suppl 1):189. doi: 10.1186/s12911-025-03013-8

[9] Adrian M Altenhoff, Alex Warwick Vesztrocy, Charles Bernard, Clement-Marie Train, Alina Nicheperovich, Silvia Prieto Baños, Irene Julca, David Moi, Yannis Nevers, Sina Majidian, Christophe Dessimoz, Natasha M Glover, OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem, Nucleic Acids Research, 2024, 52(D1):D513–D521

[10] Louden DN. MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. Med Ref Serv Q. 2020 Apr-Jun;39(2):183-191. doi: 10.1080/02763869.2020.1726152.

[11] Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-56. doi: 10.1093/nar/gku1179.

[12] Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012 Jan;40(Database issue):D136-43. doi: 10.1093/nar/gkr1178.

[13] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D456-63. doi: 10.1093/nar/gks1146.

[14] Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med. 2009 Nov-Dec;1(3):390-399. doi: 10.1002/wsbm.44.

[15] Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, Carbon S, Courtot M, Diehl AD, Dooley DM, Duncan WD, Harris NL, Haendel MA, Lewis SE, Natale DA, Osumi-Sutherland D, Ruttenberg A, Schriml LM, Smith B, Stoeckert CJ Jr, Vasilevsky NA, Walls RL, Zheng J, Mungall CJ, Peters B. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database (Oxford). 2021 Oct 26;2021:baab069. doi: 10.1093/database/baab069.

[16] Yamagata Y, Kushida T, Onami S and Masuya H. Homeostasis imbalance process ontology: a study on COVID-19 infectious processes. BMC Medical Informatics and Decision Making, 23 Supplement 4, May 22 2024. doi: 10.1186/s12911-024-02516-0.

[17] Yamagata Y, Fukuyama T, Onami S and Masuya H. Prototyping an Ontological Framework for Cellular Senescence Mechanisms: A Homeostasis Imbalance Perspective. Scientific Data, 11, Article number: 485, 2024. doi: 10.1038/s41597-024-03331-y.