

Development of a Knowledge Graph Based Integrated Database for Bioresources

Tatsuya Kushida, Daiki Usuda, Toyoyuki Takada, Hiroshi Masuya

tatsuya.kushida@riken.jp

Integrated BioResource Information Division,
RIKEN BioResource Research Center (BRC)

ANRRC 2024
September 4, 2024



About RIKEN BioResource Research Center (BRC)

Mission

Contribute to the development of human health, medical science research, breeding, and production of useful chemical compounds through the RIKEN bioresources.

Core competence

Simultaneously provide the different types of bioresources.



DNA materials (>160,000,
e.g., human cDNA clones)

Microbe materials (>19,000,
e.g., Yeast strains)

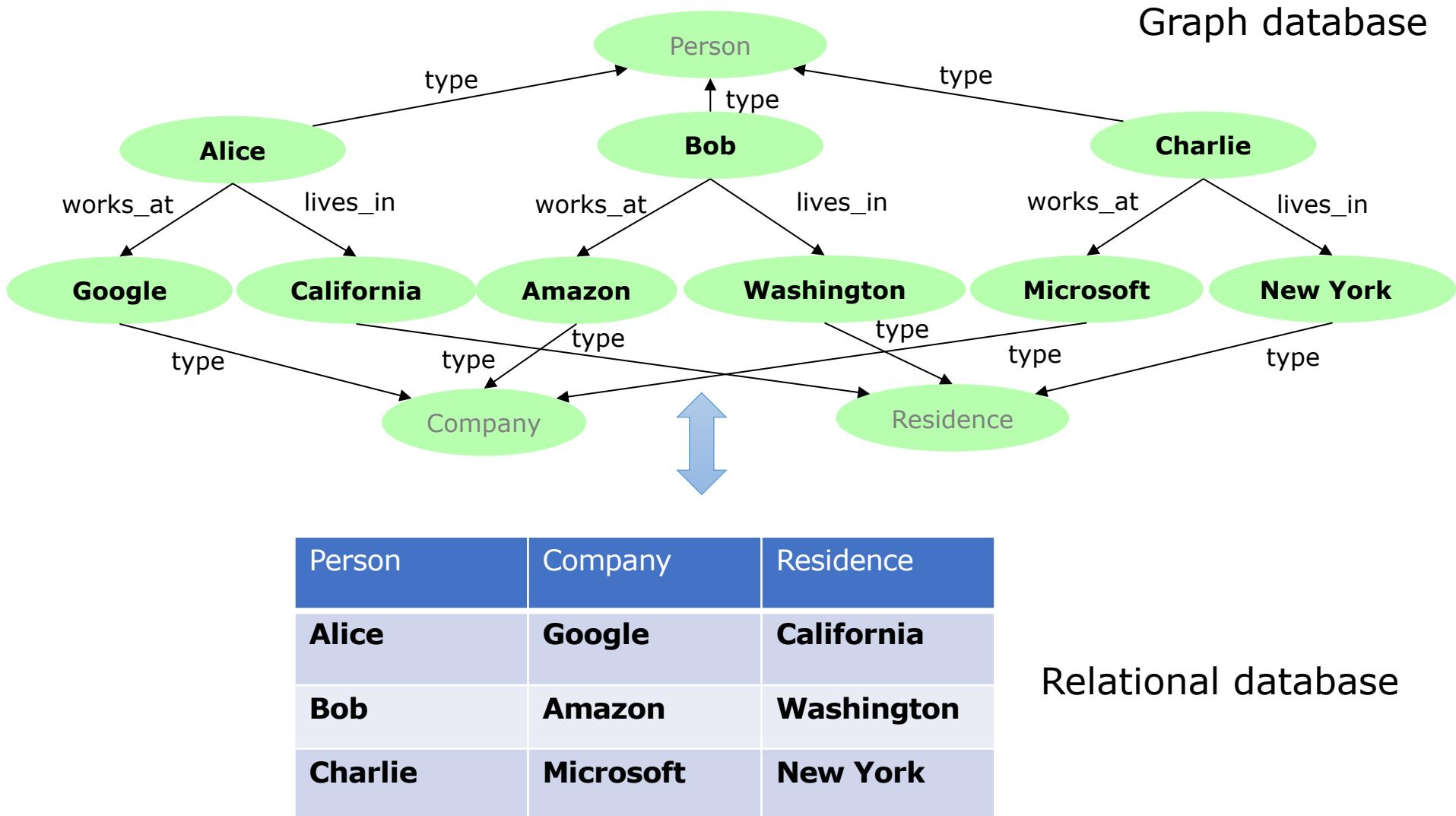
Activities of
Information division

1. Development of the **RIKEN BioResource Graph data** (Knowledge Graph).
2. **Integration** of this data with external public datasets (e.g., UniProt, ChEBI, DisGeNET, and Disease Ontology).
3. Storage of the graph data in a triple store, **BioResource MetaDatabase** (6.8 billion triples).
4. Implementation of a **bioresource retrieval system** to explore the bioresources available for life science research and industry.

Development of bioresource graph data (knowledge graph)



Data structure of Graph Database vs Relational Database



Graph Database vs Relational Database - Advantages of Graph Database (green)

Aspect	Graph Database	Relational Database
Schema Flexibility	No predefined schema needed; easily handles complex, evolving data structures.	Rigid schema required; difficult to handle complex relationships.
Ease of Data Insertion	New data and relationships can be added without modifying the schema.	Requires schema modification when adding new data types or relationships.
Query Performance	Optimized for queries involving relationships; efficient traversal of connected data.	Performance may degrade with complex joins or hierarchical data.
Use Cases	Ideal for social networks, recommendation engines, fraud detection, etc.	Best suited for structured data, like financial records, inventory management, etc.



Graph Database vs Relational Database - Disadvantages of Graph Database (**pink**)

Aspect	Graph Database	Relational Database
Maturity	Still relatively new; less mature than relational databases.	Highly mature and well-established technology.
Tooling and Ecosystem	Limited tools and ecosystem support.	Rich set of tools, libraries, and community support.
Performance on Simple Queries	May not perform as well on simple, flat queries.	Optimized for simple, structured queries.
Data Consistency	Ensuring consistency across distributed nodes can be challenging.	Strong consistency models, well-supported by most systems.



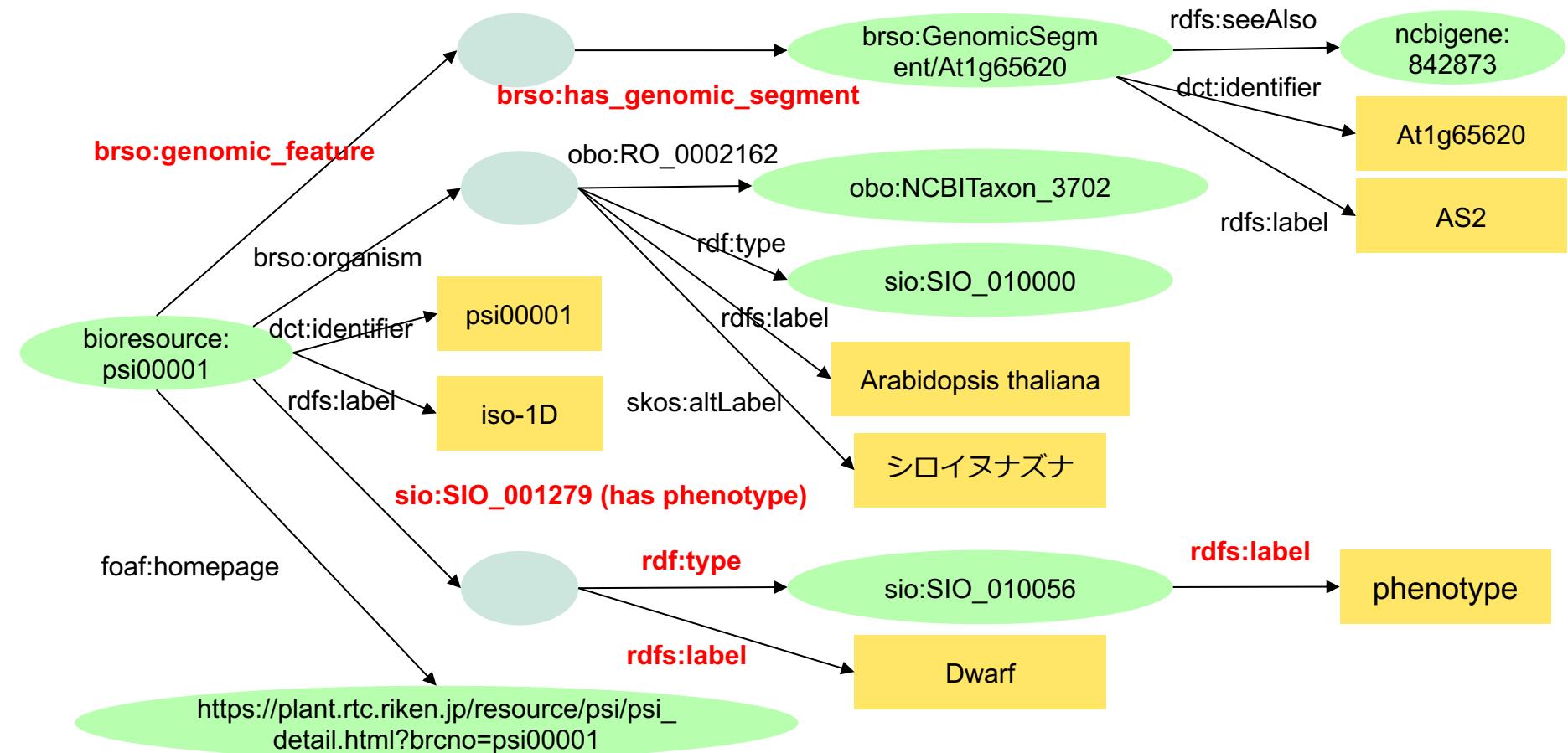
BioResource Schema Ontology (BRSO)

- The Biological Resource Schema Ontology (BRSO) is a graph data model for various model organisms and the types, such as individual (mouse, plant, microbe), cell, and DNA, which is largely developed by the Database Center for Life Science (DBCLS), RIKEN, and National Institute of Genetics (NIG). RIKEN BRC is developing bioresource graph data based on the BRSO.
- <https://github.com/dbcls/brso>

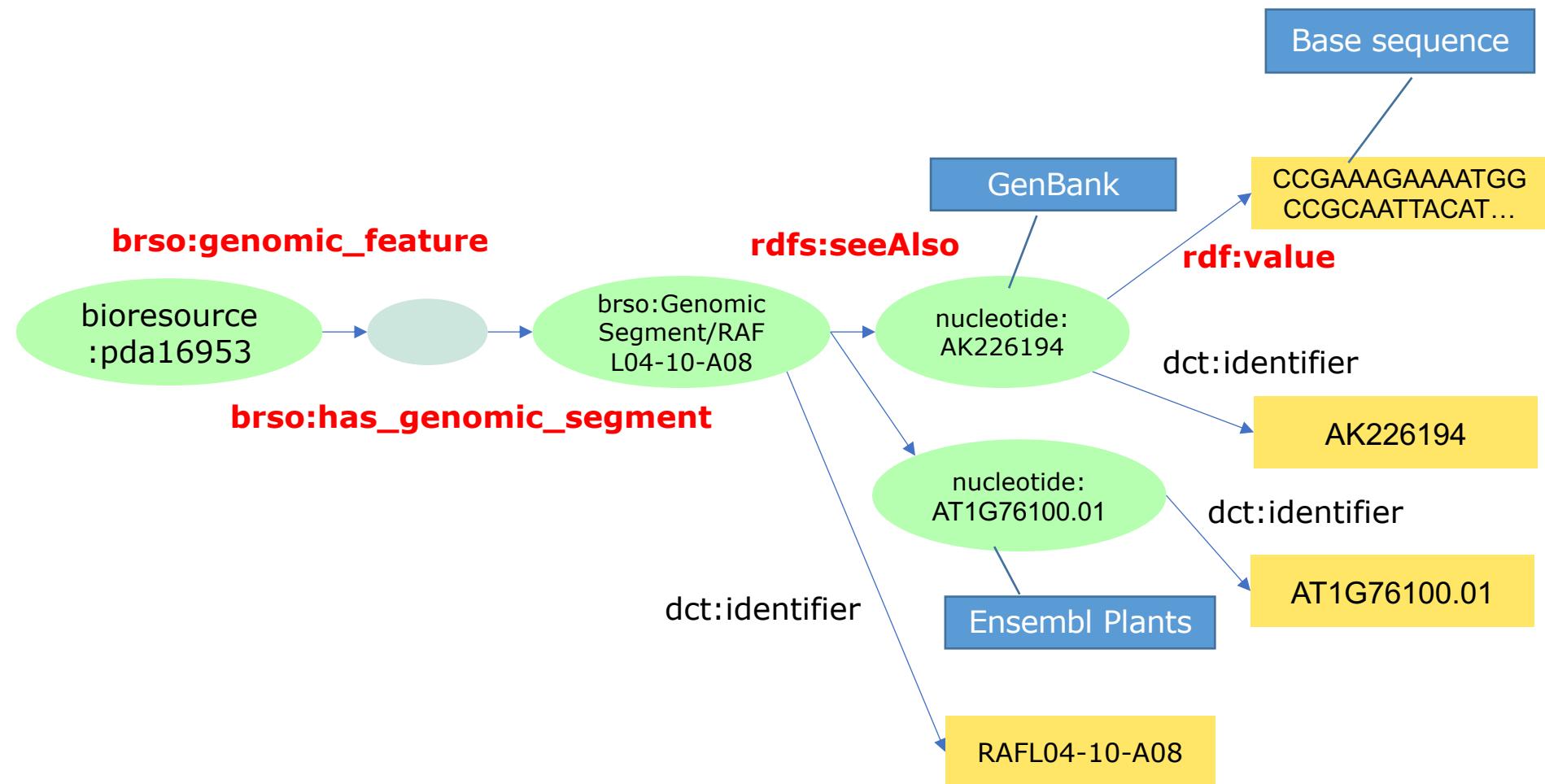
Examples of properties (edges) of BRSO

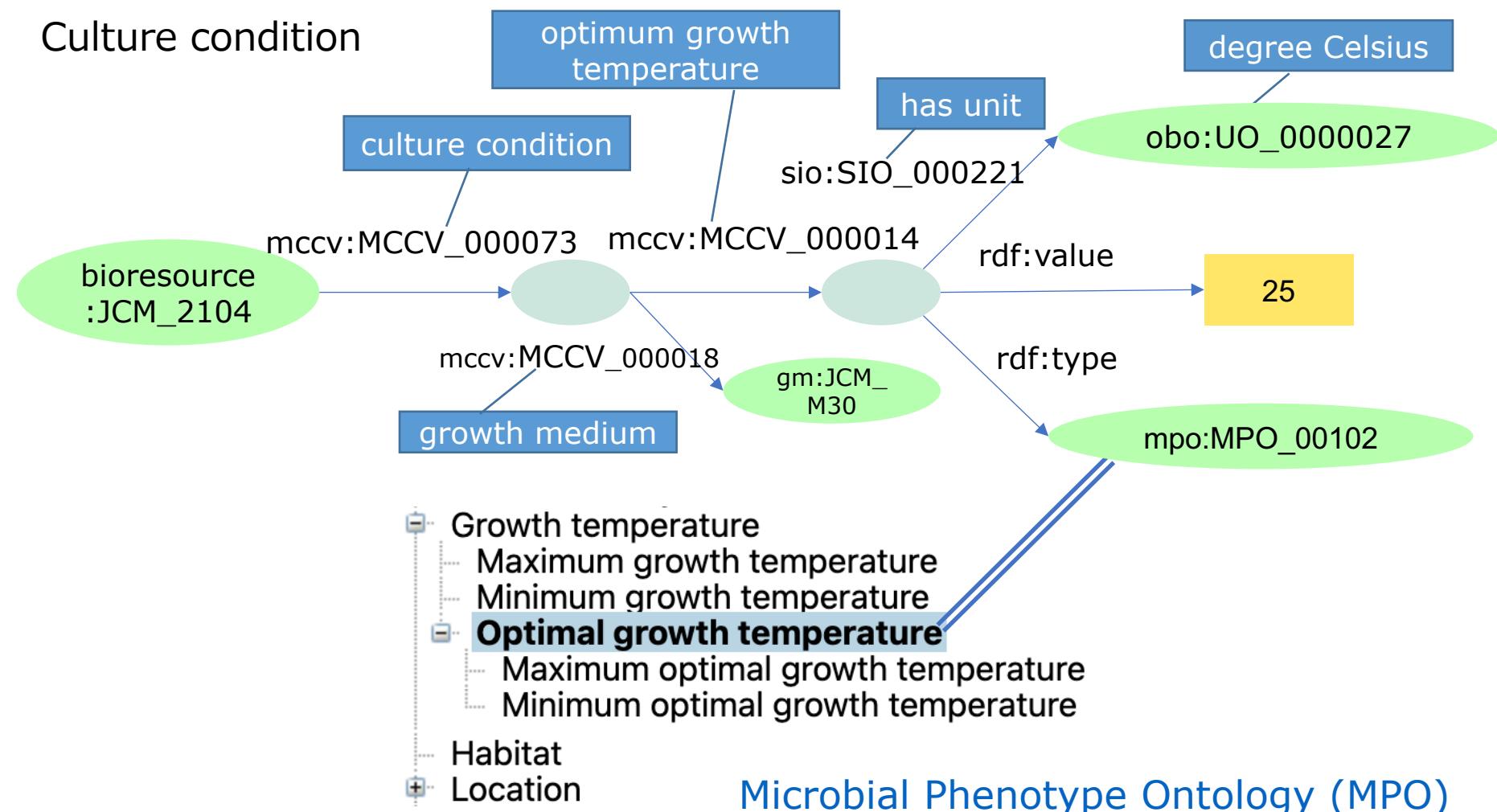
Property	Domain	Range
http://purl.jp/bio/10/brso/deposited_by	Bioresource	Depositor
http://purl.jp/bio/10/brso/sampled_from	Bioresource	Location
http://purl.jp/bio/10/brso/donor	Bioresource	donor
http://purl.jp/bio/10/brso/organism	Bioresource	organism
http://purl.jp/bio/10/brso/genomic_feature	Bioresource	genomic segment
http://purl.jp/bio/10/brso/derived_from	Bioresource	Anatomical part
http://purl.jp/bio/10/brso/has_genomic_segment	genomic segment	gene
http://purl.jp/bio/10/brso/has_allele	genomic segment	allele



Data structure of gene and phenotype of Experimental Plants using BRSO (extracted)

Data structure of DNA sequence data using BRSO (extracted)



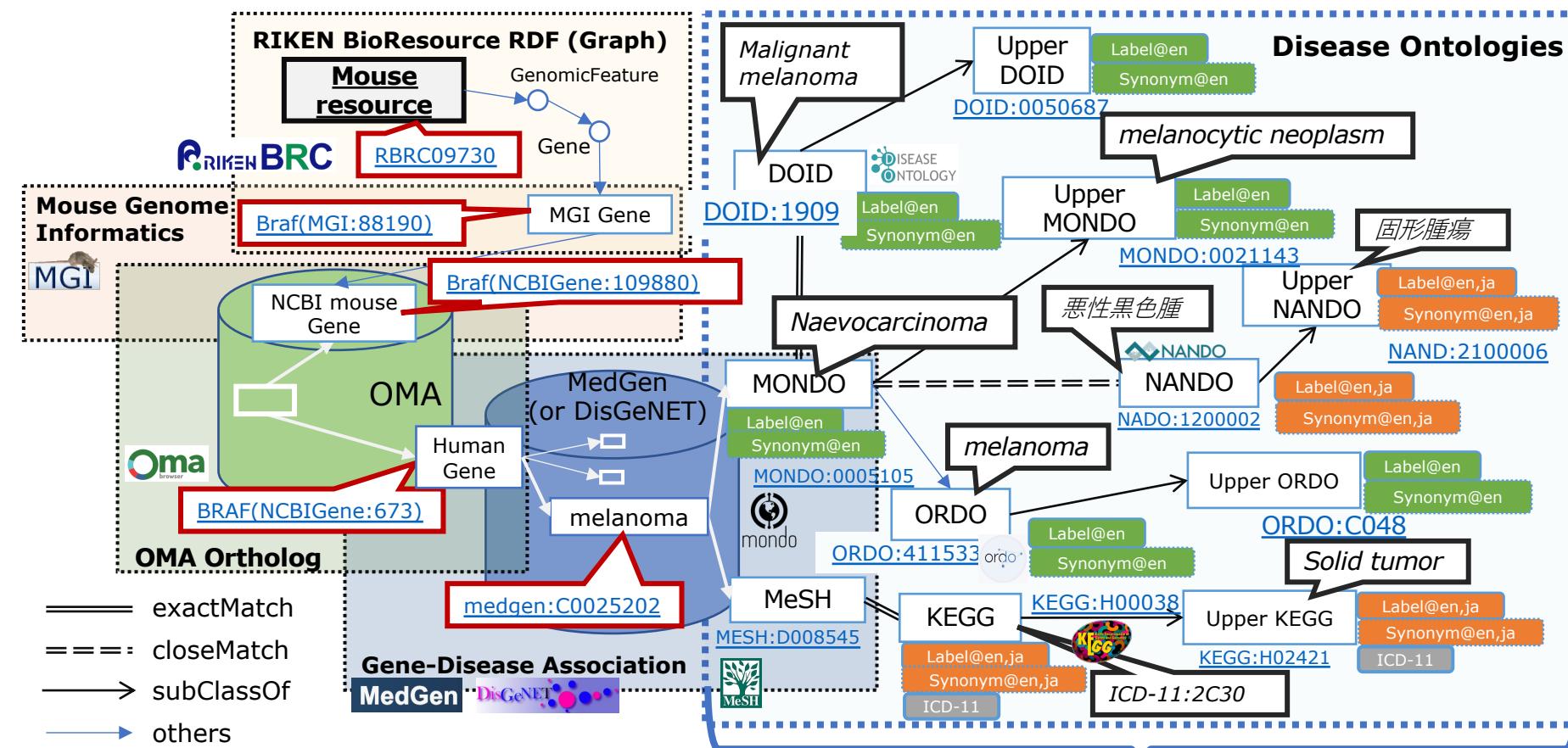
Data structure of **culture condition** for Microbe materials using BRSO (extracted)

Data integration



Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0)
©2024 Tatsuya Kushida (RIKEN BioResource Research Center (BRC))

Integration of RIKEN mice data with human disease ontologies



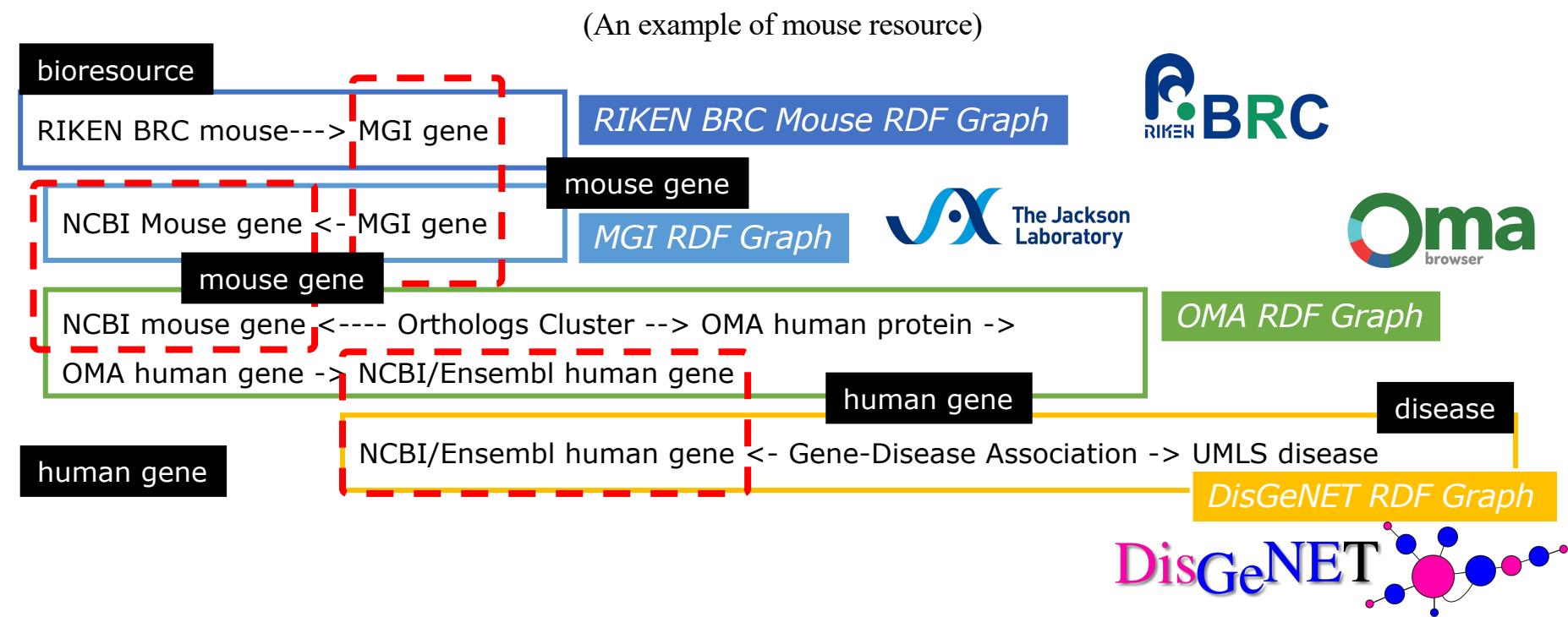
Keyword retrieval system to explore the bioresources relevant to human diseases

Reuse

Bioresource-disease associations

Collecting

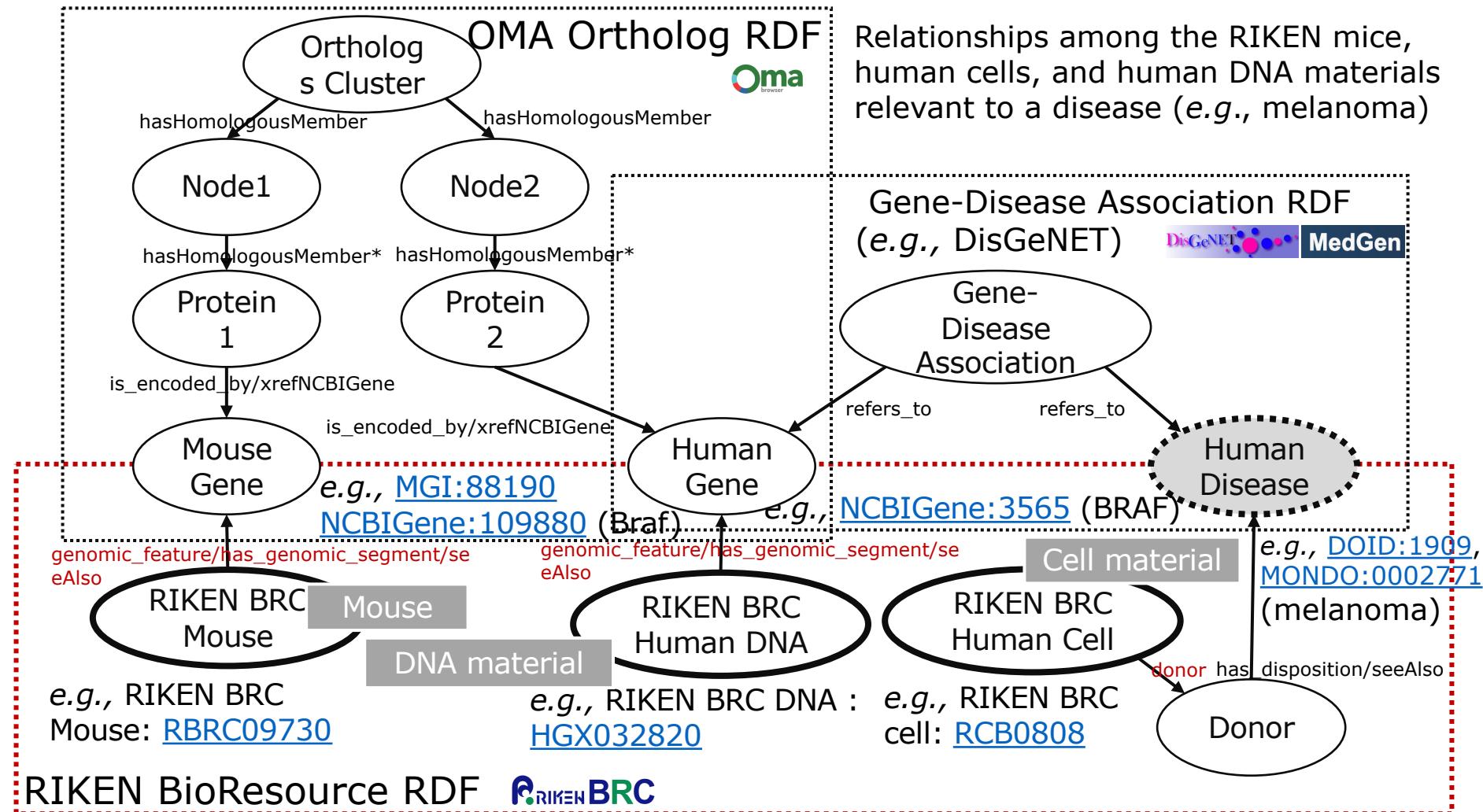
Linking a bioresource with **disease** using multiple databases (Animal)

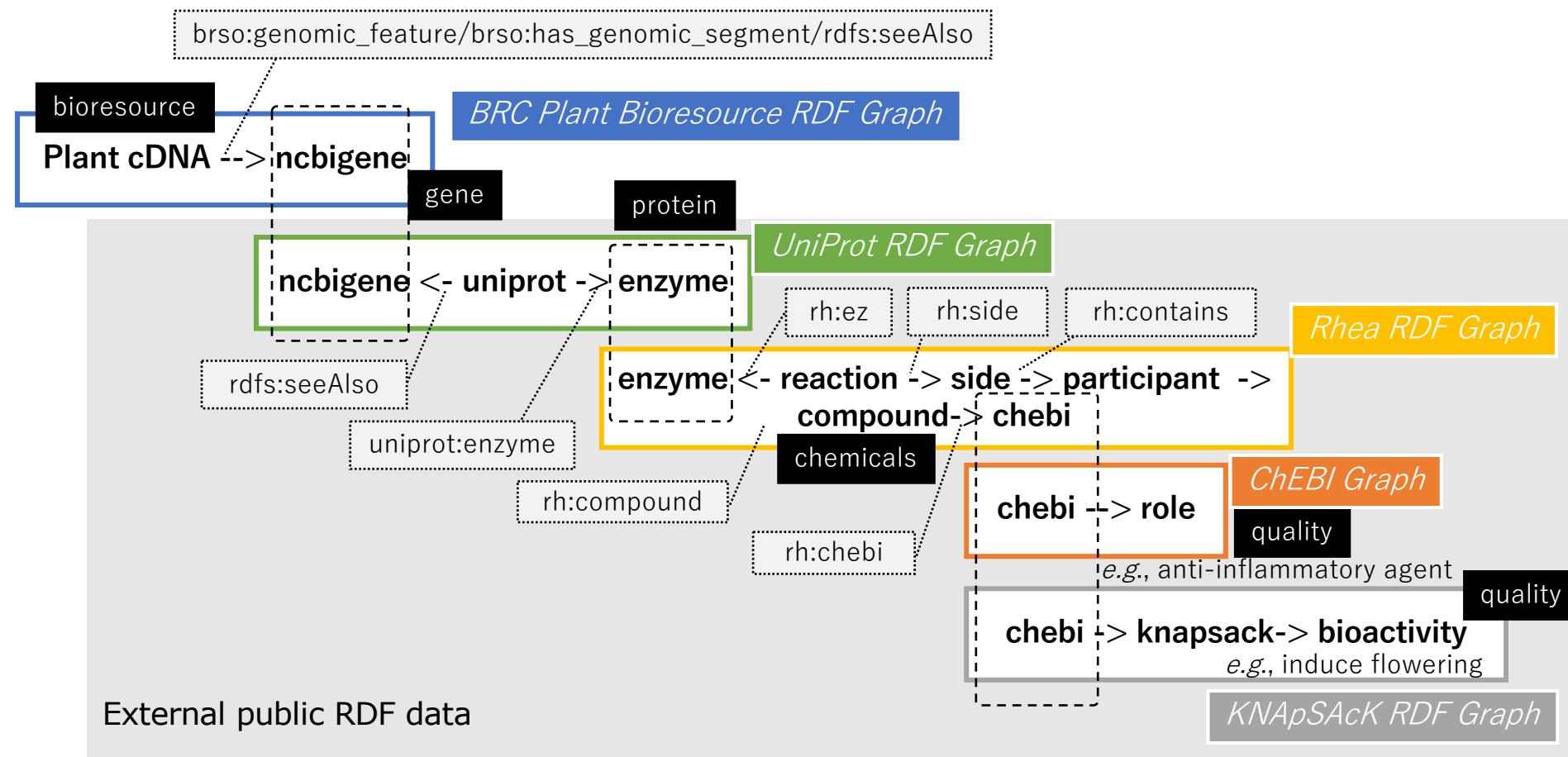


: Graph connection through URI

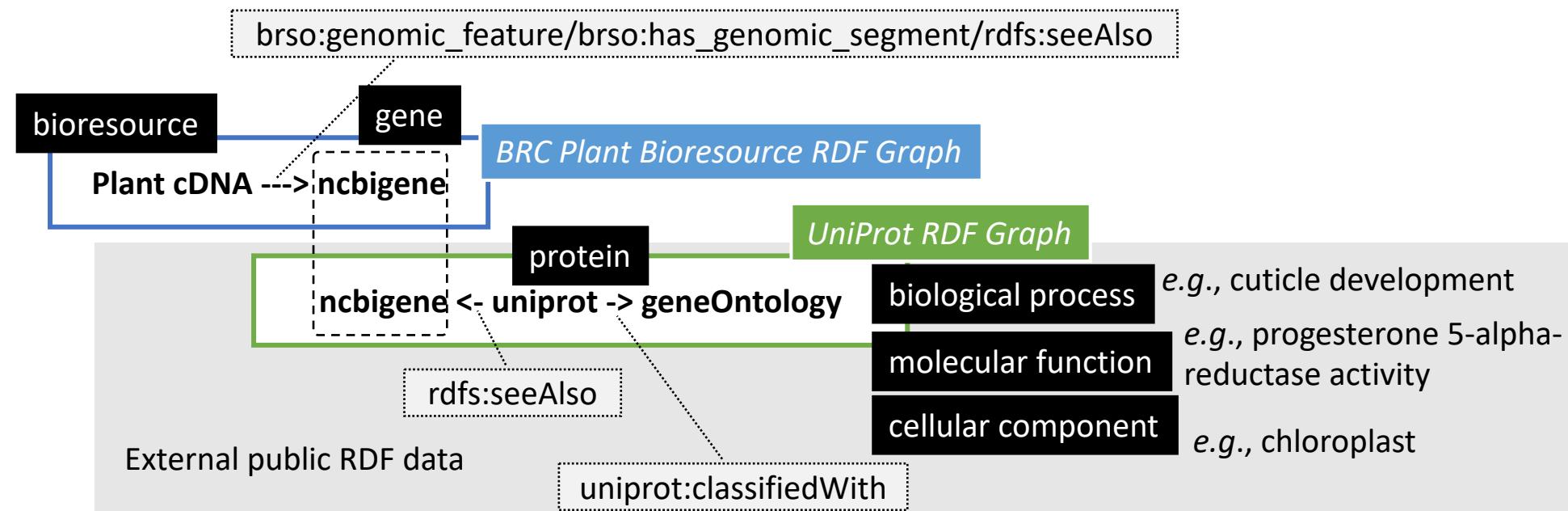
13

Relationships among mice, cell and DNA materials

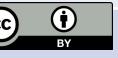


Linking a bioresource with **chemical data** using multiple databases (**Plants**) 1/2

Linking a bioresource with Gene Ontology using UniProt Graph (Plants) 1/2



External public data integrated with the bioresource graph data

Category	Datasets		Who created graph data (RDF/OWL)	License	
Genome /Proteome	1	Mouse Genome Informatics (MGI) RDF	RIKEN BRC	Original (... used for research and educational purposes only. Reproduction or use for commercial purposes without the express prior written permission of The Jackson Laboratory is prohibited.)	
	2	UniProt RDF (including Gene Ontology)	UniProt	CC BY 4.0	
Ortholog	3	Orthologous Matrix (OMA) RDF	OMA	CC BY-SA 2.5	
Gene-Disease interaction	4	DisGeNET RDF	DisGeNET	CC BY-NC-SA 4.0	
	5	MedGen RDF	DBCLS	NCBI Website and Data Usage Policies and Disclaimers	
	6	KEGG MEDICUS RDF	RIKEN BRC	CC BY-SA 4.0	
Disease	7	Nanbyo Disease Ontology (NANDO)	DBCLS	CC BY 4.0	
	8	Human Disease Ontology (DOID)	DOID	CC0 1.0	
	9	Mondo Disease Ontology (MONDO)	MONDO	CC BY 4.0	
	10	Orphanet Rare Disease Ontology (ORDO)	ORDO	CC BY 4.0	
Phenotype	11	Mammalian Phenotype Ontology (MP)	MP	CC BY 4.0	
	12	Human Phenotype Ontology (HP)	HP	Original (... vocabularies, annotation files, tools and documentation are freely available. ... is acknowledged and cited properly.	
Chemicals	13	Chemical Entities of Biological Interest (ChEBI)	ChEBI	CC BY 4.0	
Biochemical reactions	14	Knowledgebase of biochemical reactions (Rhea)	Rhea	CC BY 4.0	
Gene expression	15	Bgee: gene expression data in animals	Bgee	CC0 1.0	



STATISTICS (No. of bioresources, and triples by bioresource types)

Bioresource type (graph)	No. of resources	No. of resources relevant to human diseases*	No. of relevant human disease	No. of triples (set of a node, an edge, and a node)
Experimental animals (e.g., mice)	6,295	1,623	1,295	1,004,439
Experimental plants	612,129	N/A	N/A	3,772,663
Cell materials	9,634	215	607	766,279
DNA materials	168,387	2,470	2,333	4,915,475
Microbe materials	19,946	N/A	N/A	1,113,745
Total (RIKEN bioresource)	816,391	4,308	2,628	11,572,601
Total (including external datasets)	N/A	N/A	N/A	6,803,456,799**

*: Used datasets(DisGeNET-DOID)

**: Acquired from [YummyData](#)

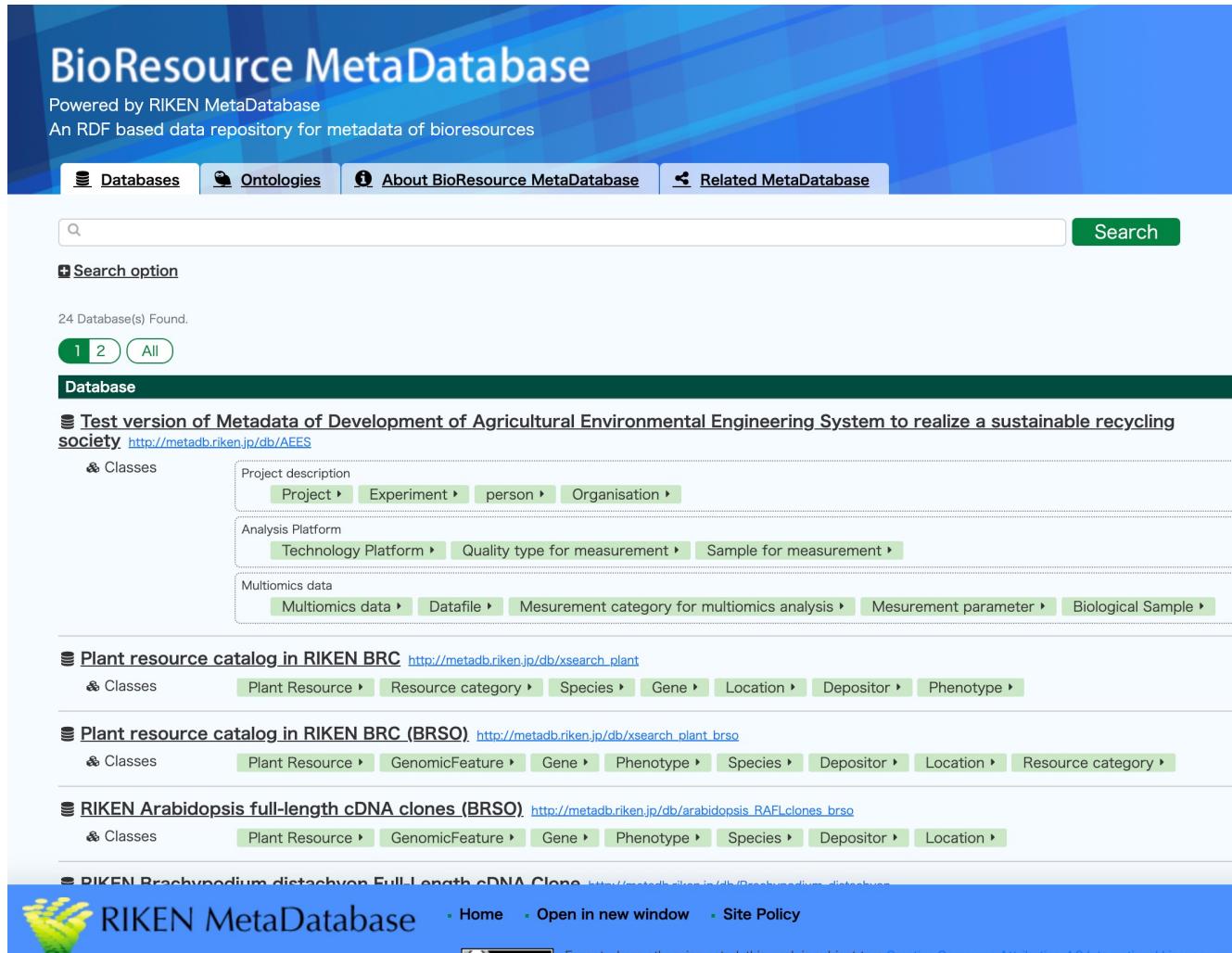


Open triplestore (Graph Database storage)



BioResource MetaDatabase top page

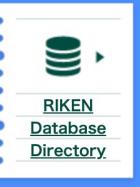
<https://knowledge.brc.riken.jp/>



The screenshot shows the BioResource MetaDatabase homepage. At the top, there's a navigation bar with links for Databases, Ontologies, About BioResource MetaDatabase, and Related MetaDatabase. Below the navigation is a search bar with a 'Search' button. A message indicates '24 Database(s) Found.' with page navigation buttons for 1, 2, and All. A green header bar labeled 'Database' contains a list of database entries. Each entry includes a title, a brief description, and a class hierarchy. For example, the first entry is 'Test version of Metadata of Development of Agricultural Environmental Engineering System to realize a sustainable recycling society' with the URL <http://metadb.riken.jp/db/AEFS>. The second entry is 'Plant resource catalog in RIKEN BRC' with the URL http://metadb.riken.jp/db/xsearch_plant. The third entry is 'Plant resource catalog in RIKEN BRC (BRSO)' with the URL http://metadb.riken.jp/db/xsearch_plant_brsq. The fourth entry is 'RIKEN Arabidopsis full-length cDNA clones (BRSO)' with the URL http://metadb.riken.jp/db/arabidopsis_RAFLclones_brsq. The fifth entry is 'RIKEN Brachypodium distachyon Full Length cDNA Clone' with the URL http://metadb.riken.jp/db/Brachypodiumdistachyon_fullLengthcDNA_Clone.



SPARQL endpoint



RIKEN Database Directory

[en](#) [ja](#)

Faceted Navigation (for databases with metadata)

- [Species](#)
- [Theme](#)
- [Publisher](#)
- [DataType](#)

Related Projects



RIKEN Open Science Portal



RIKEN HACKATHON 2020



RIKEN Open Life Science Platform

Related MetaDatabase



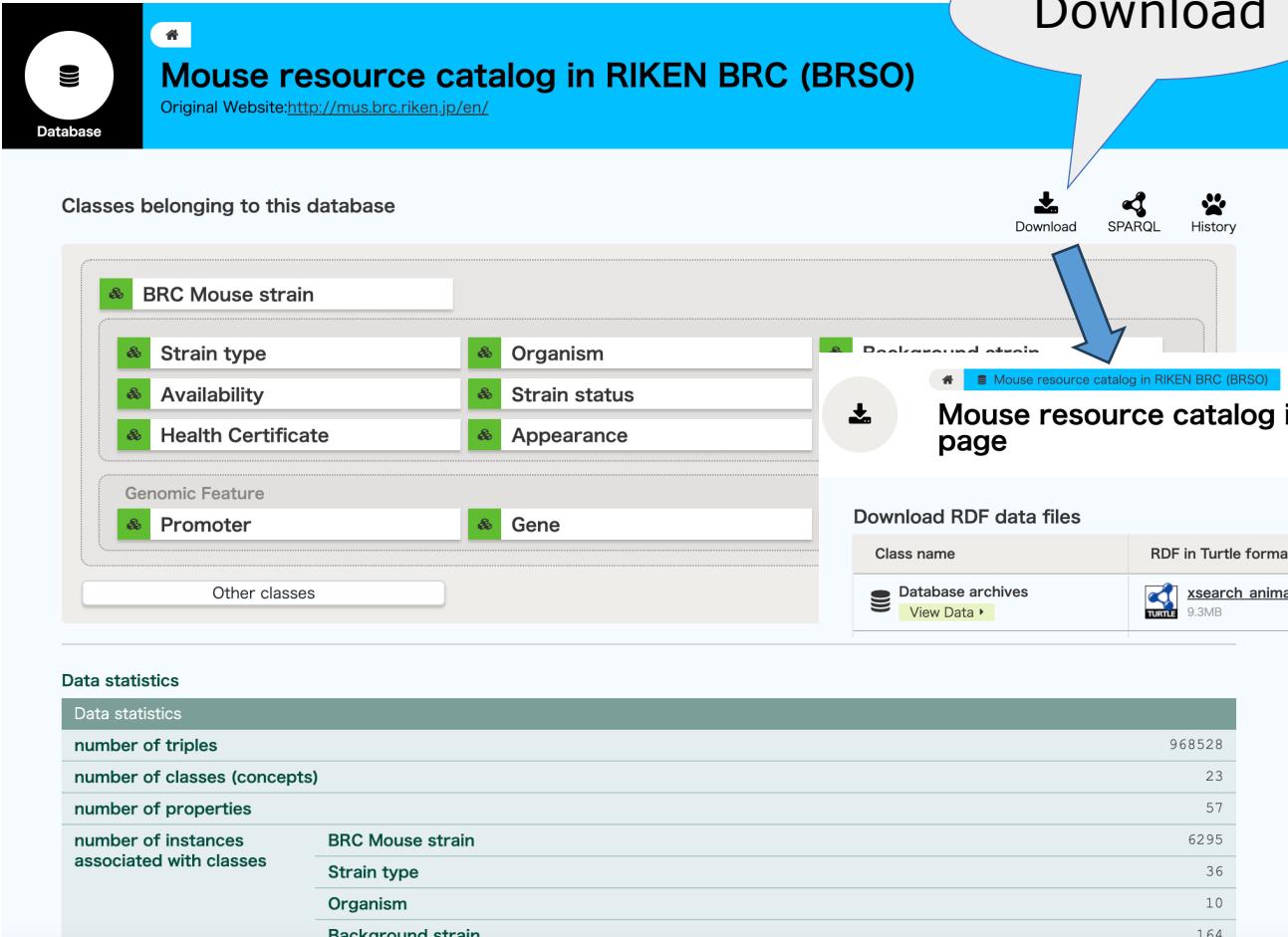
RIKEN Microstructural Imaging MetaDatabase



RIKEN Plant Metabolome MetaDatabase

Mouse resource download page

https://knowledge.brc.riken.jp/bioresource/db/xsearch_animal_brso



Download

Classes belonging to this database

BRC Mouse strain		
Strain type	Organism	
Availability	Strain status	
Health Certificate	Appearance	
Genomic Feature		
Promoter	Gene	
Other classes		

Data statistics

Data statistics									
number of triples	968528								
number of classes (concepts)	23								
number of properties	57								
number of instances associated with classes	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td>BRC Mouse strain</td> <td>6295</td> </tr> <tr> <td>Strain type</td> <td>36</td> </tr> <tr> <td>Organism</td> <td>10</td> </tr> <tr> <td>Background strain</td> <td>164</td> </tr> </table>	BRC Mouse strain	6295	Strain type	36	Organism	10	Background strain	164
BRC Mouse strain	6295								
Strain type	36								
Organism	10								
Background strain	164								

Download

Download SPARQL History

Download RDF data files

Class name	RDF in Turtle format	RDF in RDF/XML format
Database archives	 xsearch animal brso Turtle.zip <small>9.3MB</small>	

Graph data



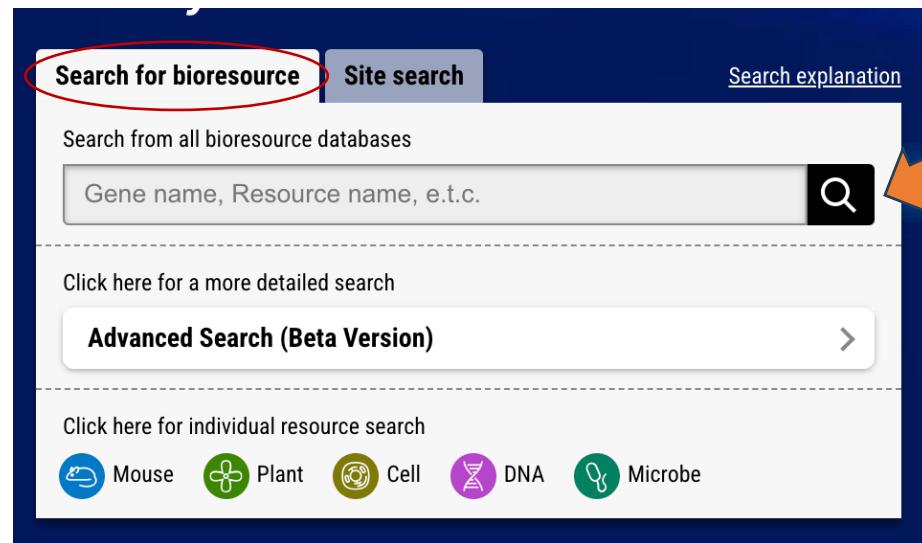
Implementation of two search systems for <https://web.brc.riken.jp/>



1. Cross-search system

A simple **keyword retrieval system** for all RIKEN's bioresources reusing graph data.

at <https://web.brc.riken.jp>



Search for bioresource Site search

Search explanation

Search from all bioresource databases

Gene name, Resource name, e.t.c.

Click here for a more detailed search

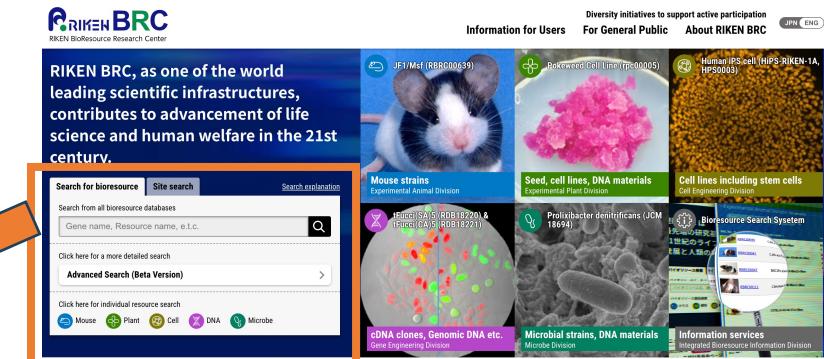
Advanced Search (Beta Version) >

Click here for individual resource search

 Mouse  Plant  Cell  DNA  Microbe

Cross-search window

As the user types, the system presents the resource name, species, disease, phenotype, and gene as possible search terms that match the text.

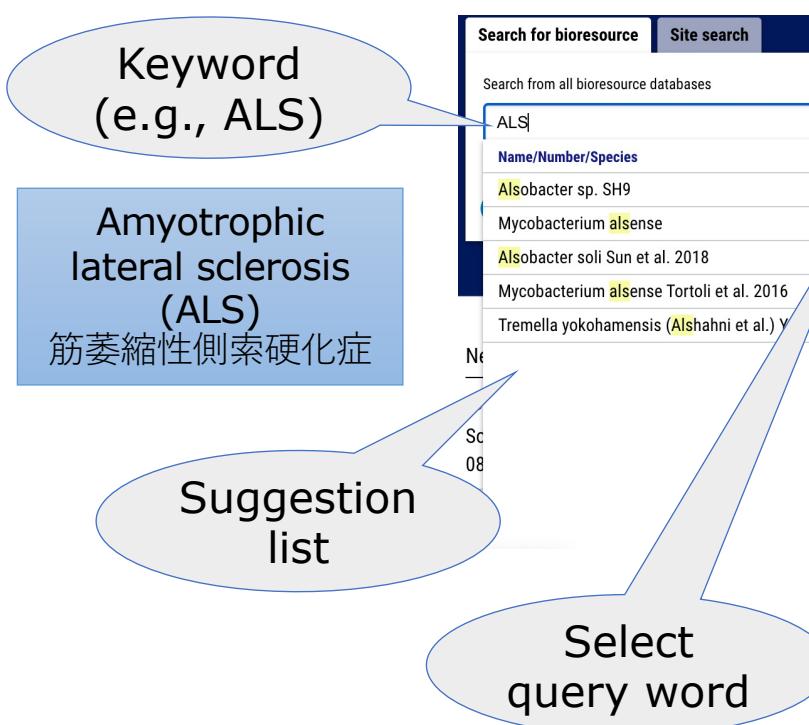


- 6,295 mice, 612,129 plants, 9,634 cells, 168,387 DNAs and 19,946 microbes can be retrieved at once.
- The knowledge base has biological names for bioresources, species, genes, phenotypes and diseases to support bioresource users for retrieve.
- The system is integrated in the BRC top page with search window.

1. Cross-search system

Function

- Perform a **keyword search** using disease names, gene symbols, phenotypes, resource IDs, and so on.
- Support for keyword entry via **suggestion lists**.

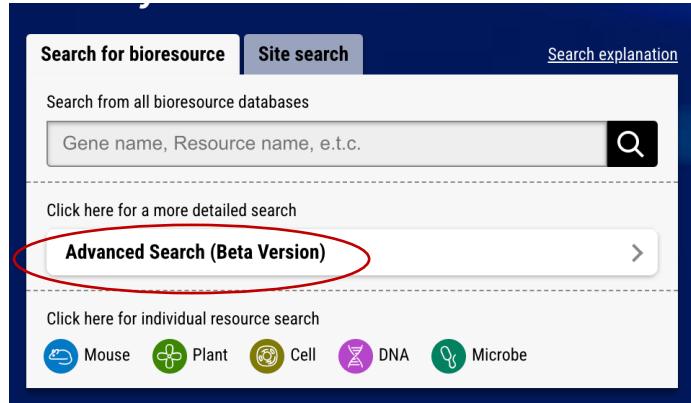


Display search results by resource type, such as mice, cell and DNA materials.

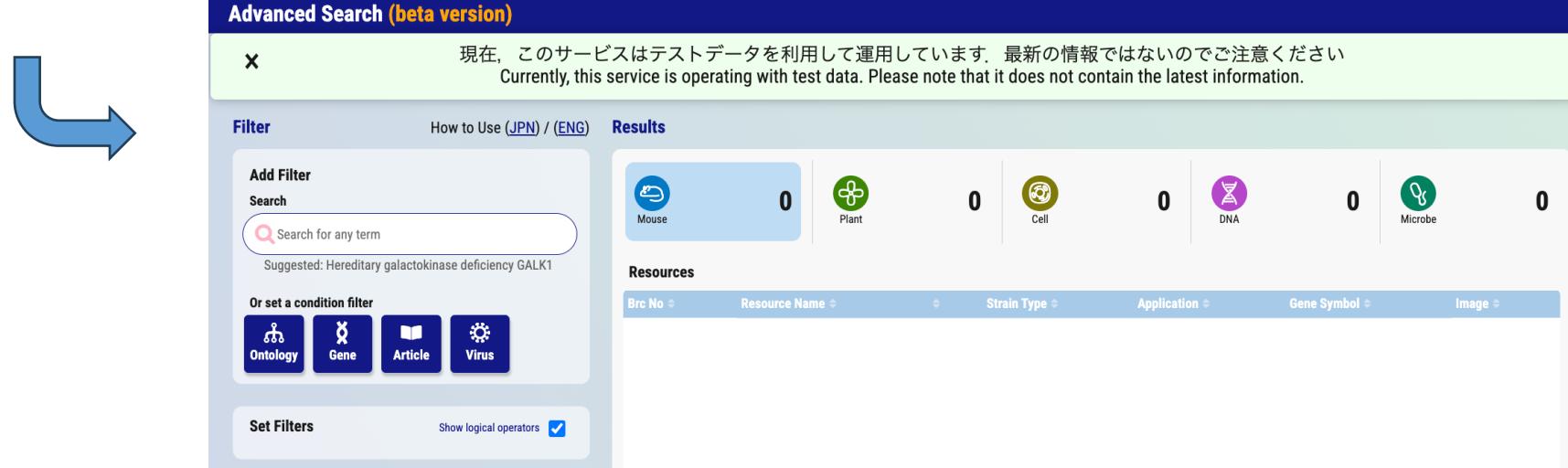
2. Advanced search system (Beta version)

↳ A **novel retrieval system** that is functionally enhanced.

at <https://web.brc.riken.jp> (RIKEN BRC top page)



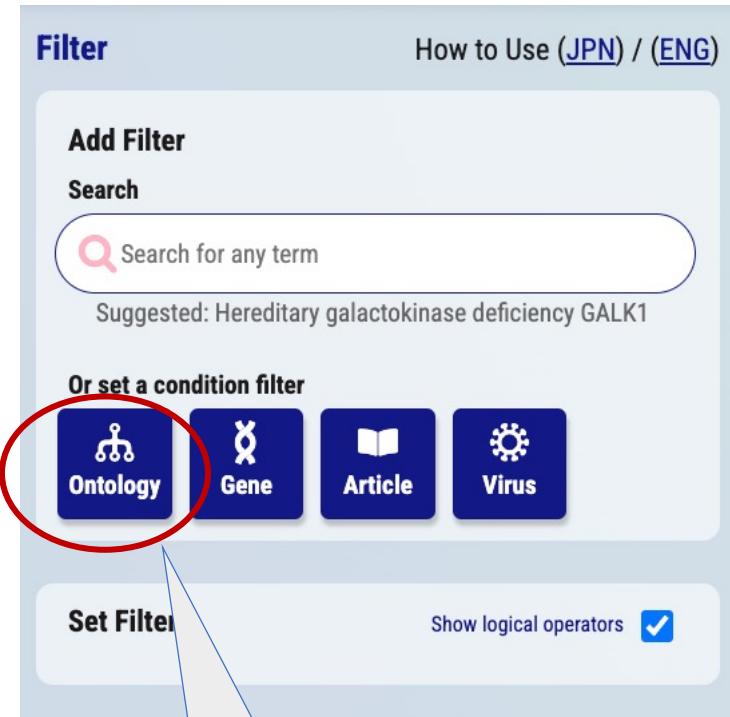
Interface of the Advanced search system



Features of Advanced Search

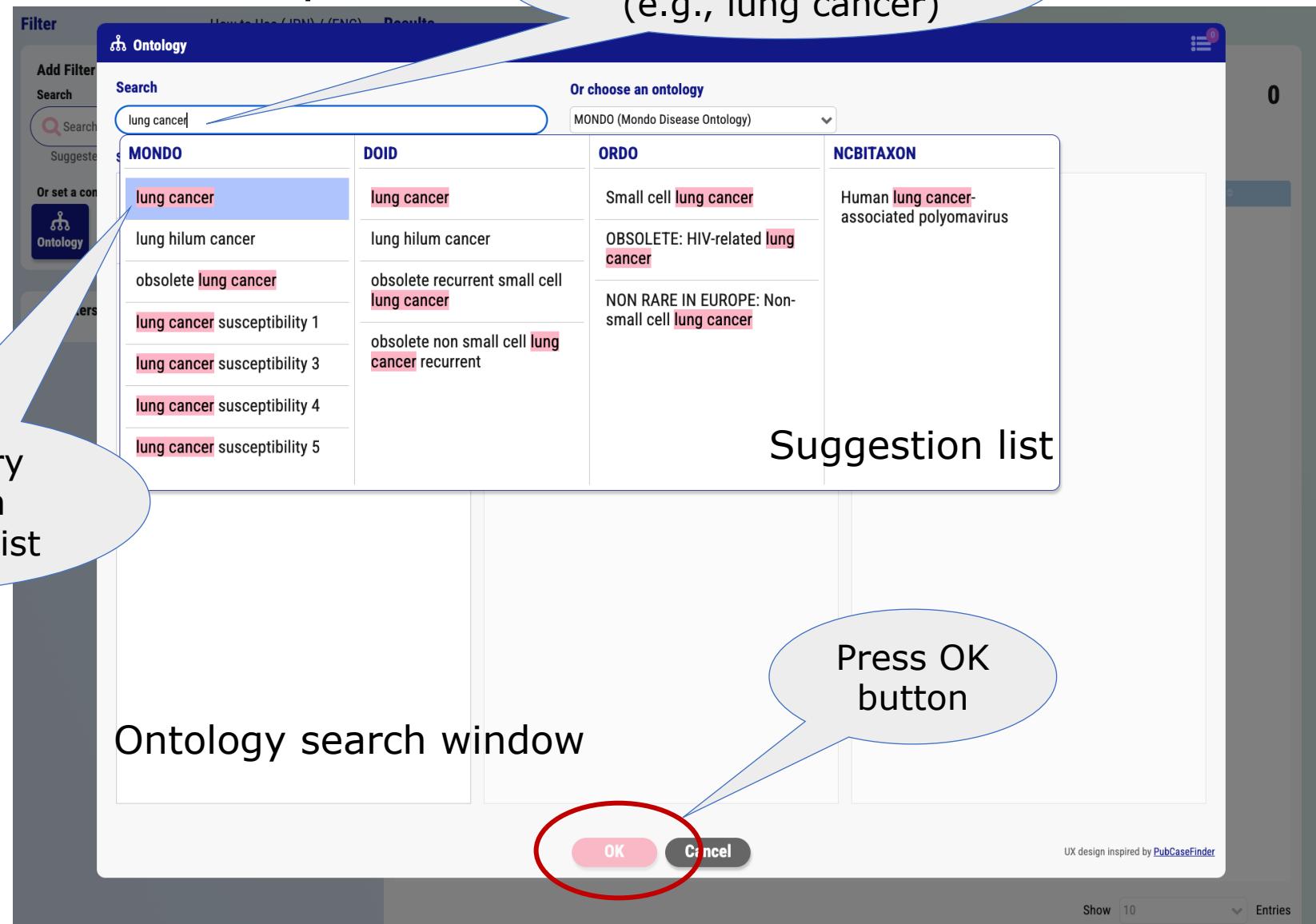
- Users can search by category in addition to keyword search for all fields. Currently, categories include **ontology**, **gene**, **article** (PubMed ID), and **virus**.
- In the case of an **ontology search**, users select multiple ontology terms from the ontology hierarchy and find the relevant bioresources.
- During the search process, users can confirm **the number of hit bioresources** and proceed with the search.
- Users can combine multiple search conditions (perform **OR/AND search**).

Advanced search window

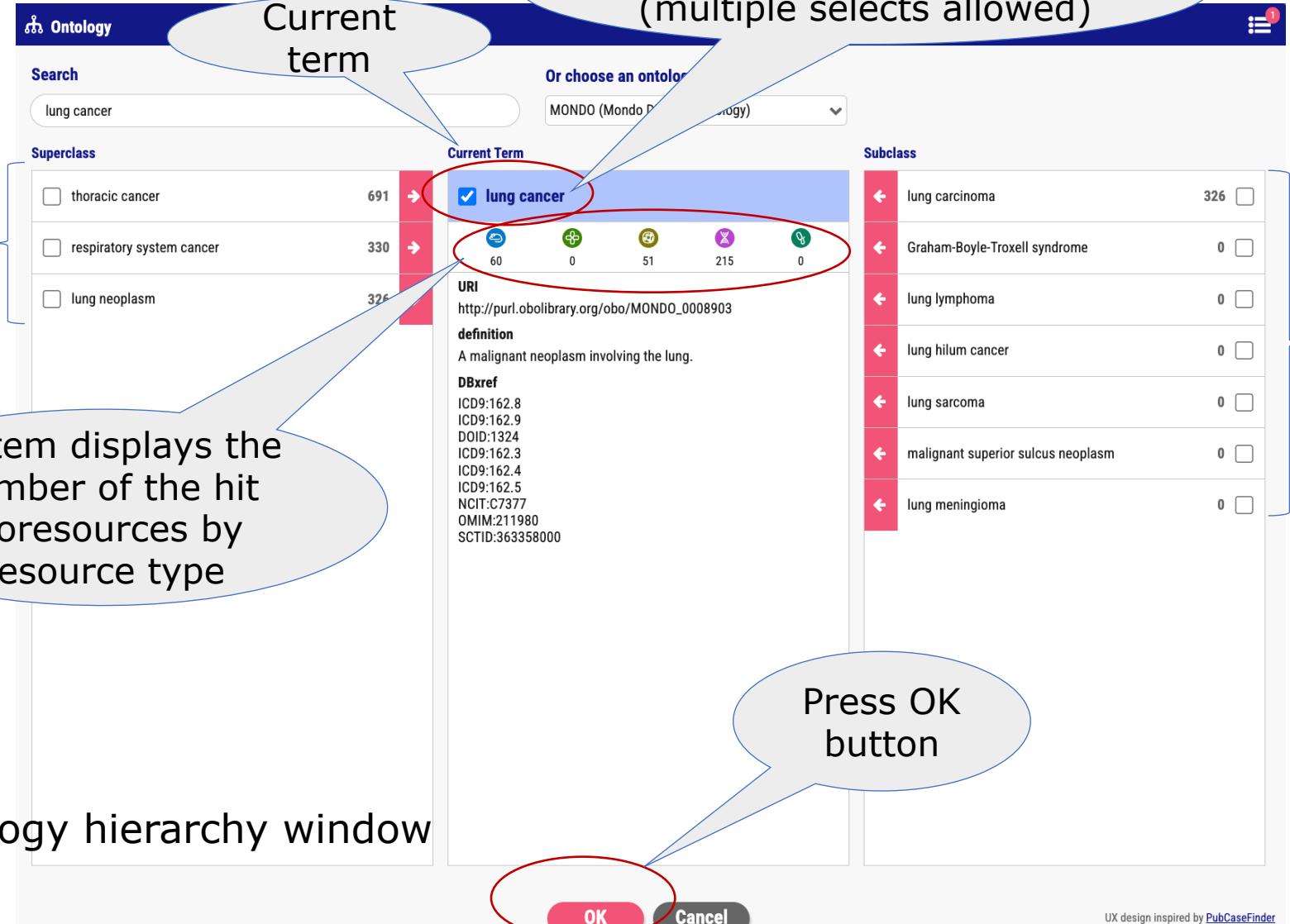


Select Ontology search

Ontology search examples



Ontology search examples



The screenshot shows the RIKEN BRC Advanced search system interface for an ontology search. The search term 'lung cancer' is entered in the search bar. The results are displayed in three main sections: Superclass, Current term, and Subclass.

Superclass: Shows categories like thoracic cancer (691 hits), respiratory system cancer (330 hits), and lung neoplasm (326 hits). A callout points to the number of hits for each category.

Current term: The central panel displays the selected term 'lung cancer' with a checkmark. It includes the following details:

- URI:** http://purl.obolibrary.org/obo/MONDO_0008903
- definition:** A malignant neoplasm involving the lung.
- DBxref:** ICD9:162.8, ICD9:162.9, DOID:1324, ICD9:162.3, ICD9:162.4, ICD9:162.5, NCIT:C7377, OMIM:211980, SCTID:363358000

Subclass: Shows specific subtypes of lung cancer, such as lung carcinoma (326 hits), Graham-Boyle-Troxell syndrome, lung lymphoma, lung hilum cancer, lung sarcoma, malignant superior sulcus neoplasm, and lung meningioma. A callout points to the number of hits for each subtype.

A large blue oval at the bottom right points to the 'OK' button, with the text 'Press OK button' inside it. A callout also points to the 'Current term' section with the text 'Check query words (multiple selects allowed)'.

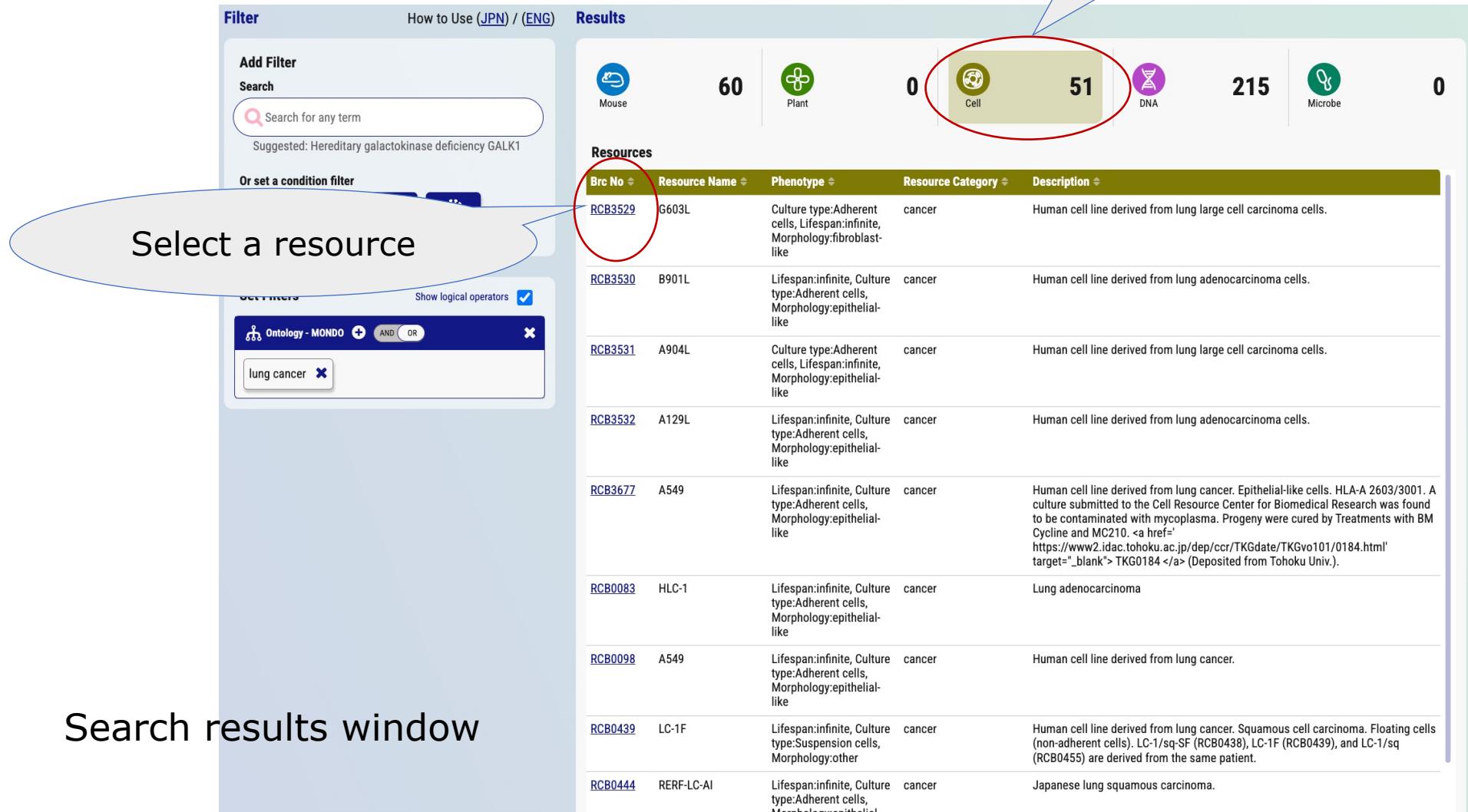


Ontology search examples

Select a resource

Search results window

Select resource type



The screenshot illustrates the RIKEN BioResource Research Center (BR) search interface, demonstrating how to search for resources using ontologies.

Filter: Shows a search bar with the term "lung cancer" and a dropdown menu for "Ontology - MONDO".

Results: A grid of resources categorized by type: Mouse (60), Plant (0), Cell (51), DNA (0), Microbe (215). The "Cell" category is highlighted with a red oval.

Brc No	Resource Name	Phenotype	Resource Category	Description
RCB3529	G603L	Culture type:Adherent cells, Lifespan:infinite, Morphology:fibroblast-like	cancer	Human cell line derived from lung large cell carcinoma cells.
RCB3530	B901L	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Human cell line derived from lung adenocarcinoma cells.
RCB3531	A904L	Culture type:Adherent cells, Lifespan:infinite, Morphology:epithelial-like	cancer	Human cell line derived from lung large cell carcinoma cells.
RCB3532	A129L	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Human cell line derived from lung adenocarcinoma cells.
RCB3677	A549	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Human cell line derived from lung cancer. Epithelial-like cells. HLA-A 2603/3001. A culture submitted to the Cell Resource Center for Biomedical Research was found to be contaminated with mycoplasma. Progeny were cured by Treatments with BM Cycline and MC210. TKG0184 (Deposited from Tohoku Univ.).
RCB0083	HLC-1	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Lung adenocarcinoma
RCB0098	A549	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Human cell line derived from lung cancer.
RCB0439	LC-1F	Lifespan:infinite, Culture type:Suspension cells, Morphology:other	cancer	Human cell line derived from lung cancer. Squamous cell carcinoma. Floating cells (non-adherent cells). LC-1/sq-SF (RCB0438), LC-1F (RCB0439), and LC-1/sq (RCB0455) are derived from the same patient.
RCB0444	RERF-LC-AI	Lifespan:infinite, Culture type:Adherent cells, Morphology:epithelial-like	cancer	Japanese lung squamous carcinoma.

Ontology search examples

[CELL SEARCH SYSTEM](#)
[CELL BANK Website](#)
[Back](#)
Cell No. : Cell Name

RCB3529 : G603L

update : 2023/10/16

Comment

Human cell line derived from lung large cell carcinoma cells.

Comment from the depositor
Terms and conditions

In publishing research results obtained by the use of the BIOLOGICAL RESOURCE, a citation of the literature ref. (Cancer Sci, 2005;96:882-8) designated by the DEPOSITOR is required.

Remarks
[Order Form\(C-0005.pdf\)](#), [MTA\(C-0007.pdf\)](#), [MTA\(C-0007p.pdf\)](#)
Order Form

 Regarding MTA between user institutions and RIKEN BRC, there are two kinds of MTA, not-for-profit academic purpose (C-XXXX) and for-profit research purpose (C-XXXXp), depending on the sort of user institutions and the purposes of use. Please use an appropriate MTA([to see](#)). In relation to commercial use and use for patent filing, first of all Please contact RIKEN BRC (cellbank.brc@riken.jp).

Basic information	Depositor	HANAGIRI, Takeshi
	Originator	TAKENOYAMA, Mitsuhiro
	Year of deposit	2010
	Animal	_human < Mammals
	Genus	Homo
	Species	sapiens
	Race	Japanese
	Gender	Male
	Age at sampling	75 years
	Tissue	lung
	Disease name	large cell carcinoma
	Metastatic ability	Unknown
	Classification	cancer
	Year of origin	2002
	Lifespan	infinite
	Morphology	fibroblast-like
	Cellosaurus(Expasy)	CVCL U091
	deposit info	lot info

A resource page



Trial of AI search



Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0)
©2024 Tatsuya Kushida (RIKEN BioResource Research Center (BRC))

Trying to use LLM to search for bioresources

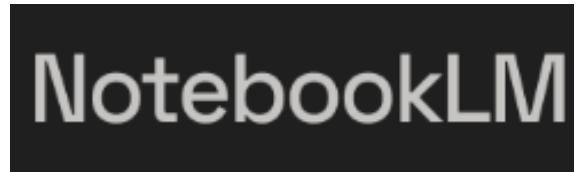
Large Language Models

- **Objectives:**

- Implement the LLM in the bioresource retrieval system to enable **fuzzy search** capabilities.

- **Expected Results:**

- Enable users to perform more intuitive and flexible queries.
- Enhance the overall user experience.
- Prevent failures in information retrieval due to user errors, such as typos or misspellings.
- Improve search accuracy, with a particular focus on enhancing recall.



Examples of fuzzy search for RIKEN microorganisms using LLM

```
The file has been successfully read after handling some formatting issues. The data includes the following columns:

- `jcm`: The JCM Number (Microorganism ID)
- `organism`: The species name
- `meo`: The MEO ID (Habitat ID)
- `label_meo`: The label for the habitat

I'll now search the data for the term `[kw]` provided, which is "forst," as well as related terms like "forest" (considering possible misspellings). I will extract relevant information and prepare a table with the requested details. [x]  
  
Here are the details for the first 30 organisms found in the dataset related to the keyword "forest" and its related terms:  
  
Total Results: 246

| JCM Number | Species Name                   | MEO ID      | Label of MEO      | Relation |
|------------|--------------------------------|-------------|-------------------|----------|
| JCM_12113  | Agromyces aurantiacus          | MEO_0000126 | forest soil       | related  |
| JCM_12733  | Aspergillus ornatus            | MEO_0000126 | forest soil       | related  |
| JCM_12733  | Sclerocleista ornata           | MEO_0000126 | forest soil       | related  |
| JCM_12833  | Micromonospora pataeolongensis | MEO_0000126 | forest soil       | related  |
| JCM_1334   | Arthrobacter ramosus           | MEO_0000193 | beech forest soil | related  |
| JCM_39028  | Mortierella parvispora         | MEO_0000126 | forest soil       | related  |

JCM Show the next 30 forest-related organisms. Can you analyze organisms found in soil? related
```

• Methodology

- *LLM*: ChatGPT vs NotebookLM
- *Source data to upload to LLMs*: RIKEN microorganism habitat data derived from RIKEN graph data
- *Prompt** : [A part of the prompt] Search for microorganisms found not only in forests, and but also in habitats relevant to forests including parts of forests, the subterms, and the terms associated with the forests, using only "forest" as a query word.

Search statement that provide an instruction to LLMs to analyze or generate specified target data.

The source data is used by the LLM to analyze or generate the specified target data, in combination with its original background knowledge.

The prompt to perform a fuzzy search for the term 'forest'.

*: See this link if you want to know the full prompt.

Results of fuzzy search using LLMs

	Preparation of source (upload data*)	No. of LLM suggesting microbes found in forests, or the relevant habitats	No. of LLM suggesting habitats relevant to forests (and examples)	Examples of relations between forests and the habitats
ChatGPT (GPT 4o)	Easy	258	5 (forest soil, beech forest soil, tropical forest soil, tropical rainforest soil, an woodland)	part of (242 habitats) related (16 habitats)
NotebookLM (Gemini 1.5 Pro)	Easy	36	12 (forest soil, wood, leaf, twig, bark, culm, stem, branch, forest temperate grasslands, savannas, and shrubland biome)	part of (28 habitats) subclass (2 habitats) related (1 habitat)
Dify (text-embedding-3- large + gpt-4o-mini)	Error** occurred in the embedding process	N/A	N/A	N/A
ChatGPT API (text- embedding-3-large)	Error** occurred in the embedding process	N/A	N/A	N/A

*: The size is about 1.2 M.

**: Rate Limit Exceeded Error. As we have uploaded other datasets, this may be the reason.



Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0)

©2024 Tatsuya Kushida (RIKEN BioResource Research Center (BRC))

Summary

1. Development of **graph data** of various bioresources.
 - Experimental animals, plants, cells, DNA and microbe materials
2. **Integration of datasets** provided by external institutes
 - Disease ontologies (DOID, MONDO, ORDO, NANDO), orthologs (OMA), genes-diseases (MedGen, DisGeNET), UniProt (including GO), biochemical reactions (Rhea), and chemical compounds (ChEBI).
3. Service provided by triple store **BioResource MetaDatabase**
 - Data release (more than 6.8 billion triples (bioresources + external data))
4. Implementation of a **keyword retrieval, and an advanced search system (Beta version)**
 - Reusing bioresource graph data
 - Cover misspellings, and types by using suggestion lists for experimental researchers.

