Ontology-driven Exploration of RIKEN Bioresources via ChEBI Roles and Gene Ontology

Tatsuya Kushida[1], Daiki Usuda[1], Takatomo Fujisawa[2], Yasunori Yamamoto[3], Norio Kobayashi[4], and Hiroshi Masuya[1]

1 RIKEN BioResource Research Center (BRC), Japan
2 Research Organization of Information and Systems, NIG, Japan
3 Database Center for Life Science, ROIS, Japan
4 Information R&D and Strategy Headquarters, RIKEN, Japan

About RIKEN BioResource Research Center (BRC)

Mission

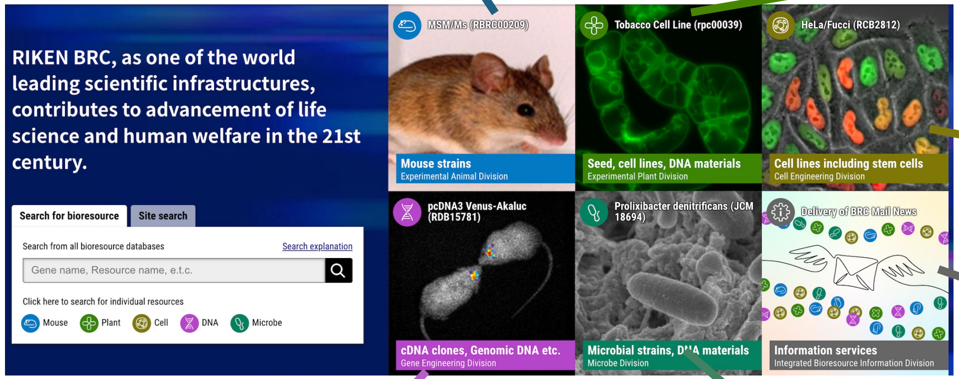
Contribute to the development of human health, medical science research, breeding, and production of useful chemical compounds through the RIKEN bioresources.

Core competence

Simultaneously provide the different types of bioresources.

Experimental animals (>6,000, *e.g.*, gene-modified mice)

Experimental plants (>610,000, *e.g.*, Arabidopsis c-DNA clones)



RIKEN BRC, as one of the world leading scientific infrastructures, contributes to advancement of life science and human welfare in the 21st century.

Search for bioresource | Site search

Search from all bioresource databases

Search explanation

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

Click here to search for individual resources

Mouse Plant Cell DNA Microbe

MSM/Ms (RBRC00209)
Mouse strains
Experimental Animal Division

Tobacco Cell Line (rpc00039)
Seed, cell lines, DNA materials
Experimental Plant Division

HeLa/Fucci (RCB2812)
Cell lines including stem cells
Cell Engineering Division

pcDNA3 Venus-Akaluc (RDB15781)
cDNA clones, Genomic DNA etc.
Gene Engineering Division

Prolixibacter denitrificans (JCM 18694)
Microbial strains, DNA materials
Microbe Division

Delivery of BRC Mail News
Information services
Integrated Bioresource Information Division

Cell materials (>9,000, *e.g.*, iPS cells derived from patients

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

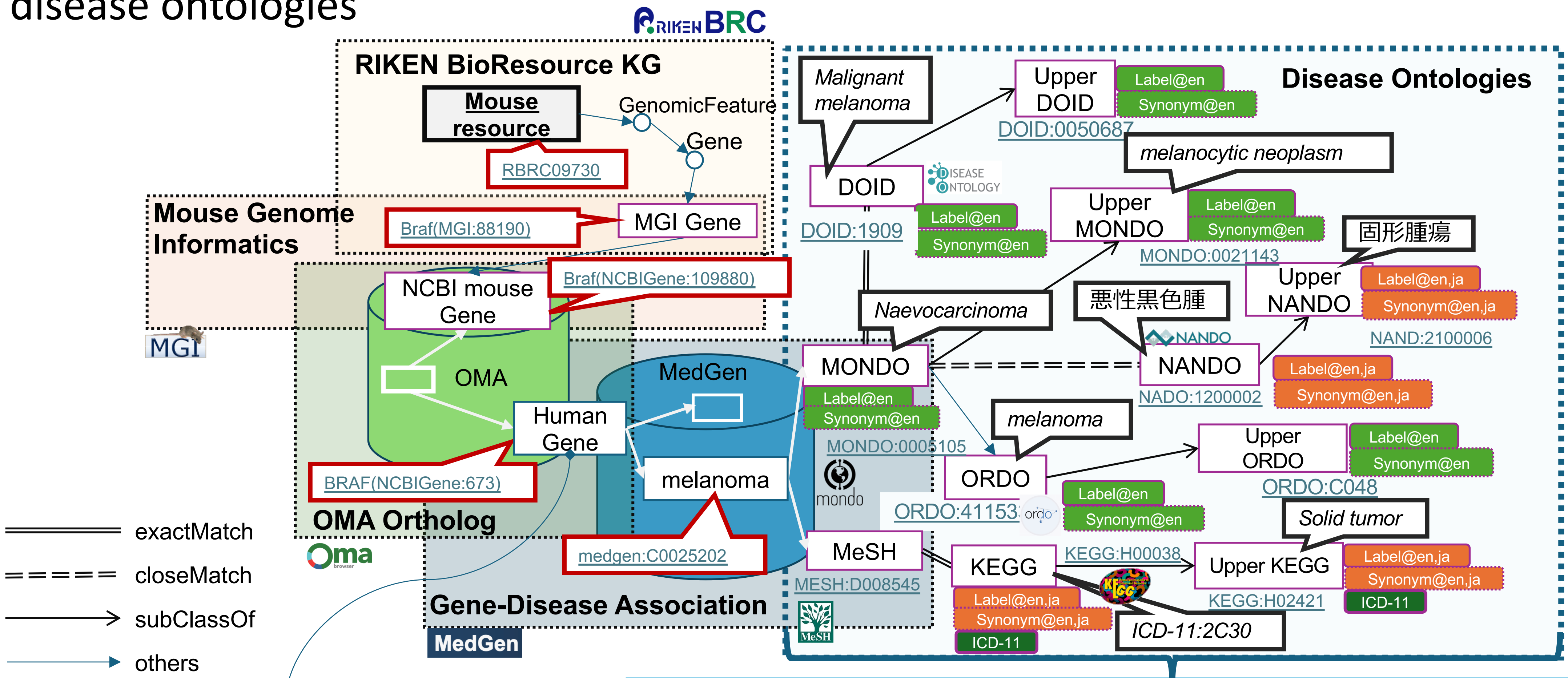
Microbe materials (>19,000, *e.g.*, Yeast strains)

Our Lab

Activities of Information division

1. Development of the **RIKEN BioResource Knowledge Graph (KG)**.
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.

Data schema of RIKEN BioResource KG integrated with external RDF data and disease ontologies



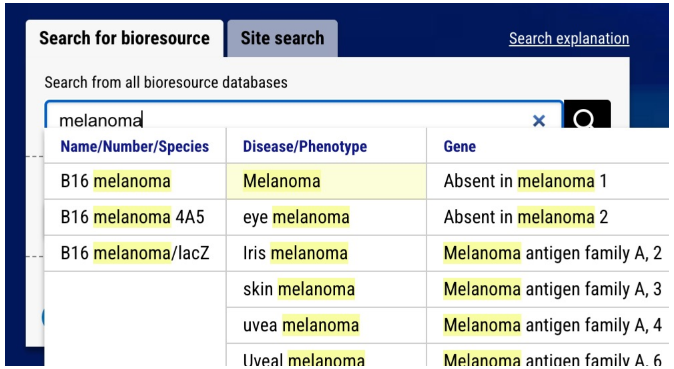
Neighboring graphs connect through the co-existing URIs. (e.g., https://www.ncbi.nlm.nih.gov/gene/673)

1. **Curate disease identifiers and labels (EN/JA):** Collect disease IDs and their labels and synonyms in both English and Japanese from the integrated KG.
2. **Link bioresources to diseases:** Create direct associations (**shortened KGs**) between each bioresource and the relevant disease IDs/labels (e.g., RBRC09730 and "melanoma").
3. **Integrate the shortened KGs into keyword retrieval system:** Expose these links in the keyword-based retrieval system so users can search bioresources by disease names or IDs.

Expansion of the retrieval function for exploration of KGs

RIKEN BRC
RIKEN BioResource Research Center

So far we have successfully implemented a keyword search system for identifying bioresources using disease IDs/labels (E/J).

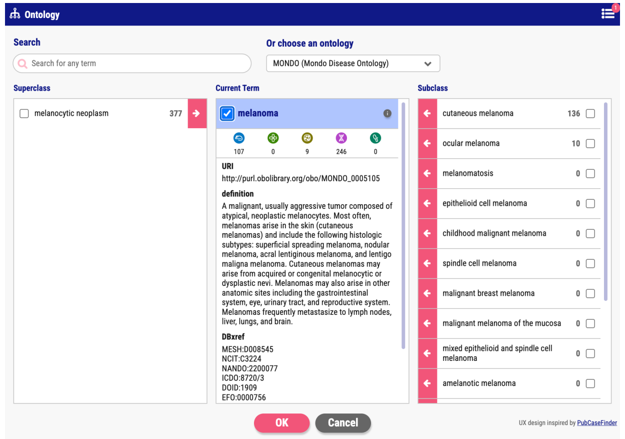


Search from all bioresource databases

melanoma

Name/Number/Species	Disease/Phenotype	Gene
B16 melanoma	Melanoma	Absent in melanoma 1
B16 melanoma 4A5	eye melanoma	Absent in melanoma 2
B16 melanoma/lacZ	Iris melanoma	Melanoma antigen family A, 2
	skin melanoma	Melanoma antigen family A, 3
	uvea melanoma	Melanoma antigen family A, 4
	Uveal melanoma	Melanoma antigen family A, 6

Search for bioresource



MONDO (Mondo Disease Ontology)

Superclass: □ melanocytic neoplasm 377

Current Term: ☑ melanoma

URI
http://purl.obolibrary.org/obo/MONDO_0005105

definition
A malignant, usually aggressive tumor composed of atypical, neoplastic melanocytes. Most often, melanomas arise in the skin (cutaneous melanomas) and include the following histologic subtypes: superficial spreading melanoma, nodular melanoma, acral lentiginous melanoma, and lentigo maligna melanoma. Cutaneous melanomas may arise from acquired or congenital melanocytic or dysplastic nevi. Melanomas may also arise in other anatomic sites including the gastrointestinal system, eye, urinary tract, and reproductive system. Melanomas frequently metastasize to lymph nodes, liver, lungs, and brain.

DBxref
MESH:D008545
NCIT:C3224
NANDO:2200077
ICDO:8720/3
DOID:1909
EFO:0000756

Subclass		
cutaneous melanoma	136	□
ocular melanoma	10	□
melanomatosis	0	□
epithelioid cell melanoma	0	□
childhood malignant melanoma	0	□
spindle cell melanoma	0	□
malignant breast melanoma	0	□
malignant melanoma of the mucosa	0	□
mixed epithelioid and spindle cell melanoma	0	□
amelanotic melanoma	0	□

UX design inspired by PubCaseFinder

Bioresource Advanced Search

In addition to diseases, we attempted to expand the bioresource KG, and enhance the retrieval function to explore bioresources related to chemical roles (e.g., anti-inflammatory agent) or Gene Ontology (GO) terms (e.g., cuticle development).

How do chemical roles/GO terms connect to bio-resources?

- Our solution
  - Create connections from **ChEBI roles** / **GO terms** to orderable bioresources by combining these ontologies and datasets:
    - **ChEBI ontology** (ChEBI role ---> ChEBI compound)
    - **Rhea RDF** (ChEBI compound ---> Reaction -> Enzyme)
    - **UniProt RDF** (Enzyme -> Protein, Protein -> Gene, Protein -> GO)
    - Bioresource KG (Gene -> Bioresource)

- Monarch KG
  - Large-scale, interoperability-oriented KG (Biolink-compliant).
  - Includes **ChEBI roles** (e.g., CHEBI:35474 - anxiolytic drug).
  - The Public stack currently does not expose a full biochemical reaction-participant chain linking EC (enzyme) ⇄ Rhea (reaction) ⇄ ChEBI (compound).

Our Rhea-based traversals are explicitly modeled and explorable in our KG.

Objectives

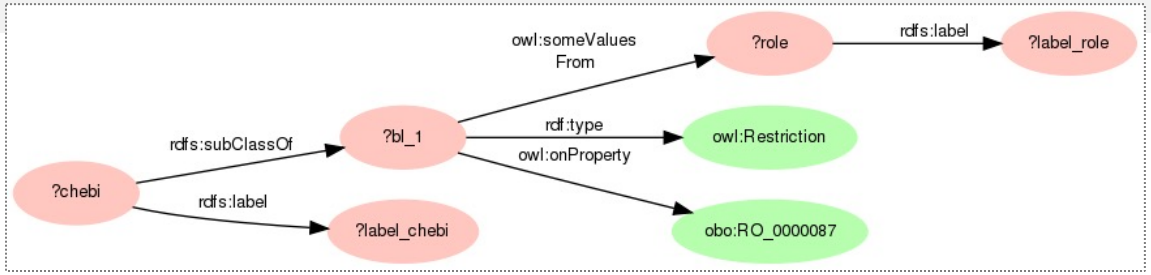
- Create paths reaching from ChEBI roles to bioresources by connecting multiple KGs through biochemical reactions, enzymes, proteins, and genes.
- And create paths reaching from GO terms to bioresources by connecting UniProt RDF through proteins, and genes.
- Explore paths by executing SPARQL queries.
- Measure the reachability (coverage) from the ChEBI role and GO to the bioresources.
    - To validate **practical effectiveness** of the ontology-driven paths at catalog scale.
    - To identify **bottlenecks** in the ontology-to-resource path.
    - To enable **benchmarking** across releases.

- **Contribution**: Operationalizing exploration paths, API-enabling, and making the coverage transparent

Our approach to data integration

1. Use typed edges (with defined domain and range) and ontology-backed link

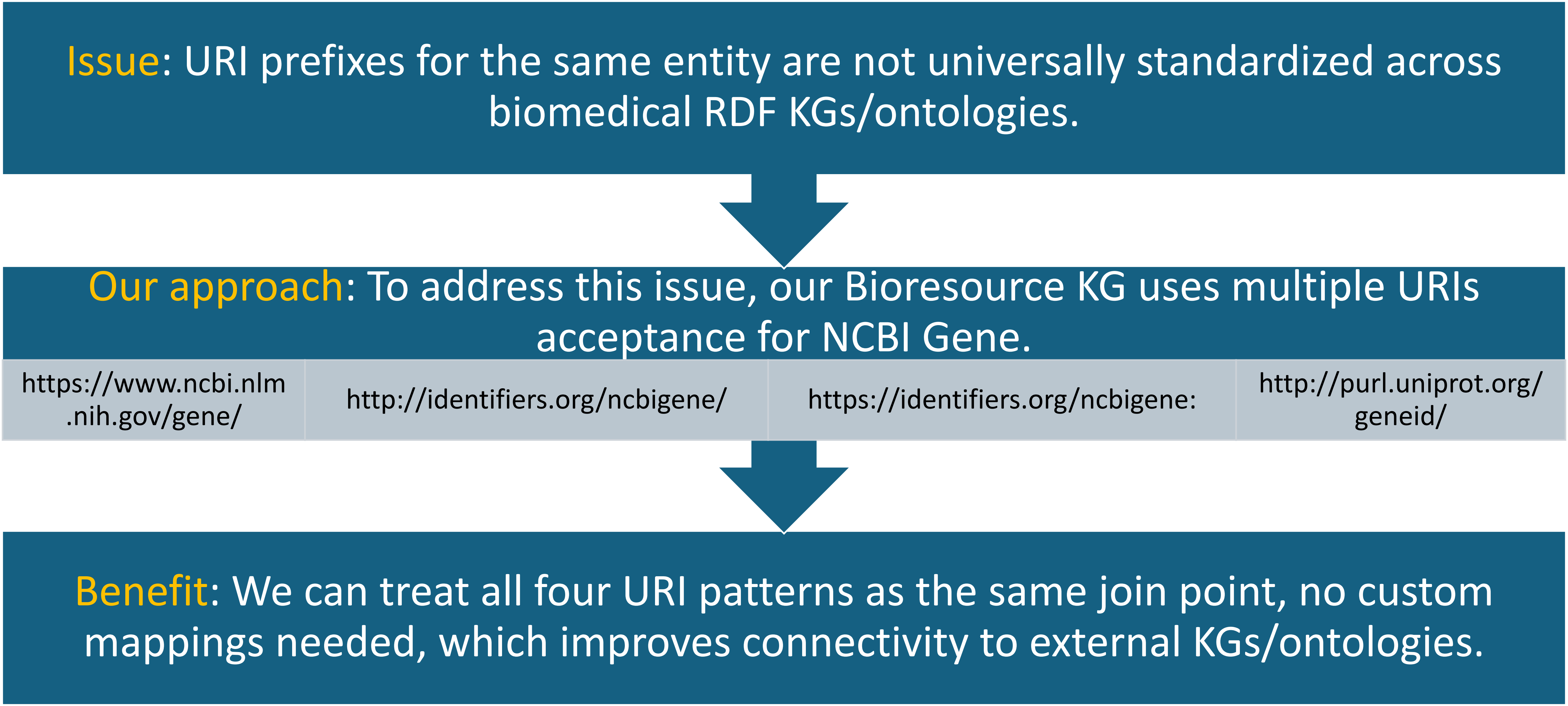
- UniProt RDF
  - Protein —**up:classifiedWith**→ GO term
  - Protein —**up:enzyme**→ EC class (Enzyme)
- Rhea RDF
  - Reaction —**rhea:ec**→ Enzyme
  - Reaction —**rhea:side** —Side —**rhea:contains** —Participant —**rhea:compound** —Compound —**rhea:chebi** → ChEBI compound
- ChEBI
  - “has role” is modeled as an OWL restriction (e.g., RO:0000087 some chebi:67079 #anti-inflammatory agent)



Ensure that every step is semantically valid and reproducible.

Our approach to data integration

2. URI prefix normalization



Issue: URI prefixes for the same entity are not universally standardized across biomedical RDF KGs/ontologies.

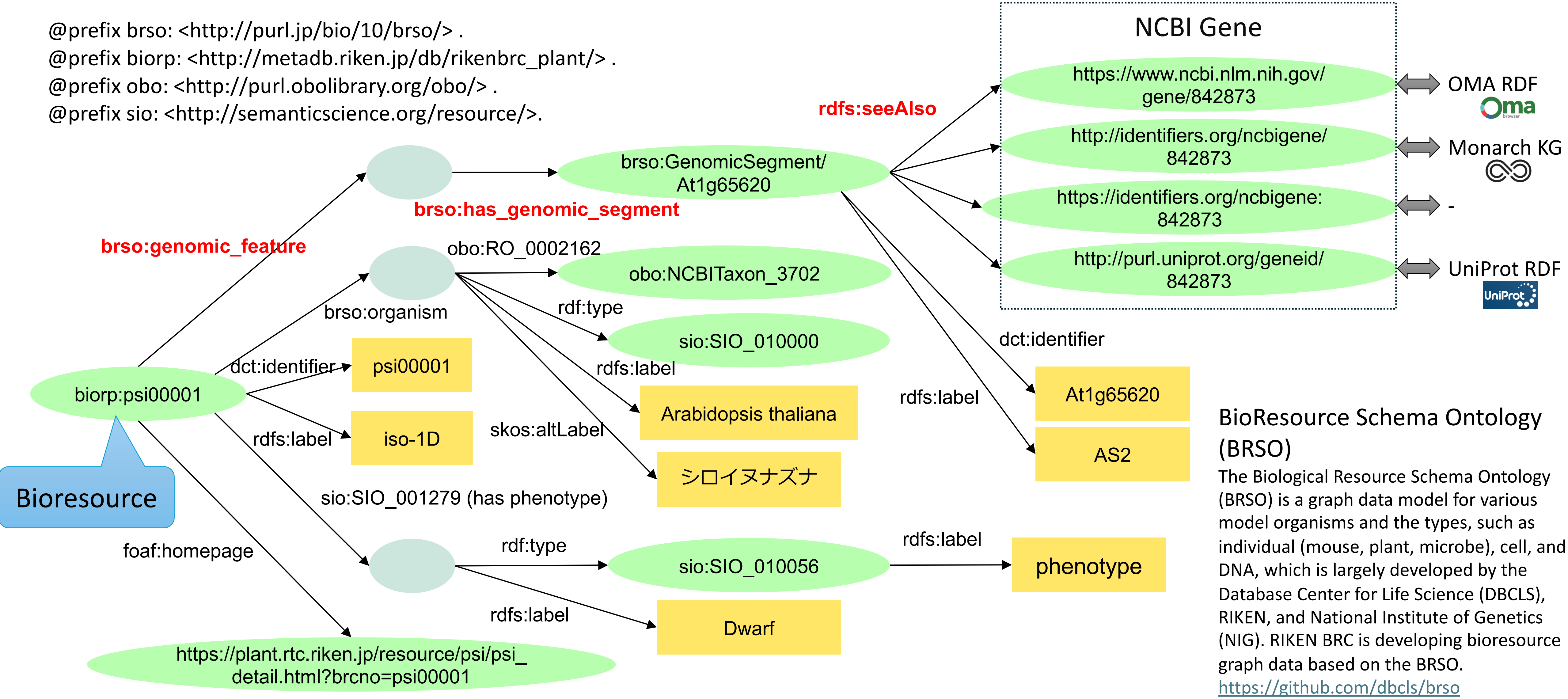
Our approach: To address this issue, our Bioresource KG uses multiple URIs acceptance for NCBI Gene.

https://www.ncbi.nlm.nih.gov/gene/	http://identifiers.org/ncbigene/	https://identifiers.org/ncbigene:	http://purl.uniprot.org/geneid/

Benefit: We can treat all four URI patterns as the same join point, no custom mappings needed, which improves connectivity to external KGs/ontologies.

Data structure of Plant DNA material using BRSO (extracted)

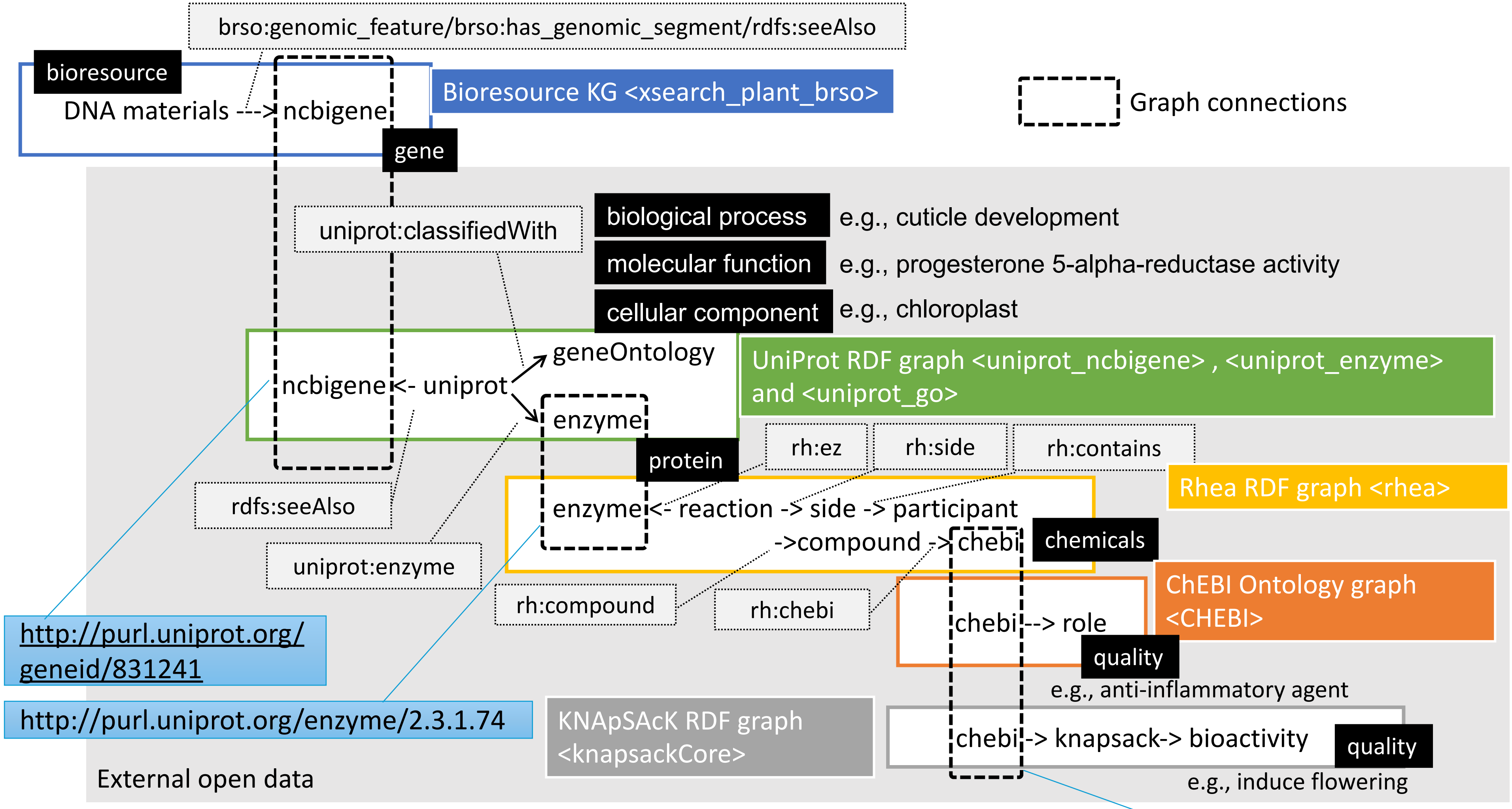
@prefix brso: <http://purl.jp/bio/10/brso/> .
@prefix biorp: <http://metadb.riken.jp/db/rikenbrc_plant/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix sio: <http://semanticscience.org/resource/>.



BioResource Schema Ontology (BRSO)

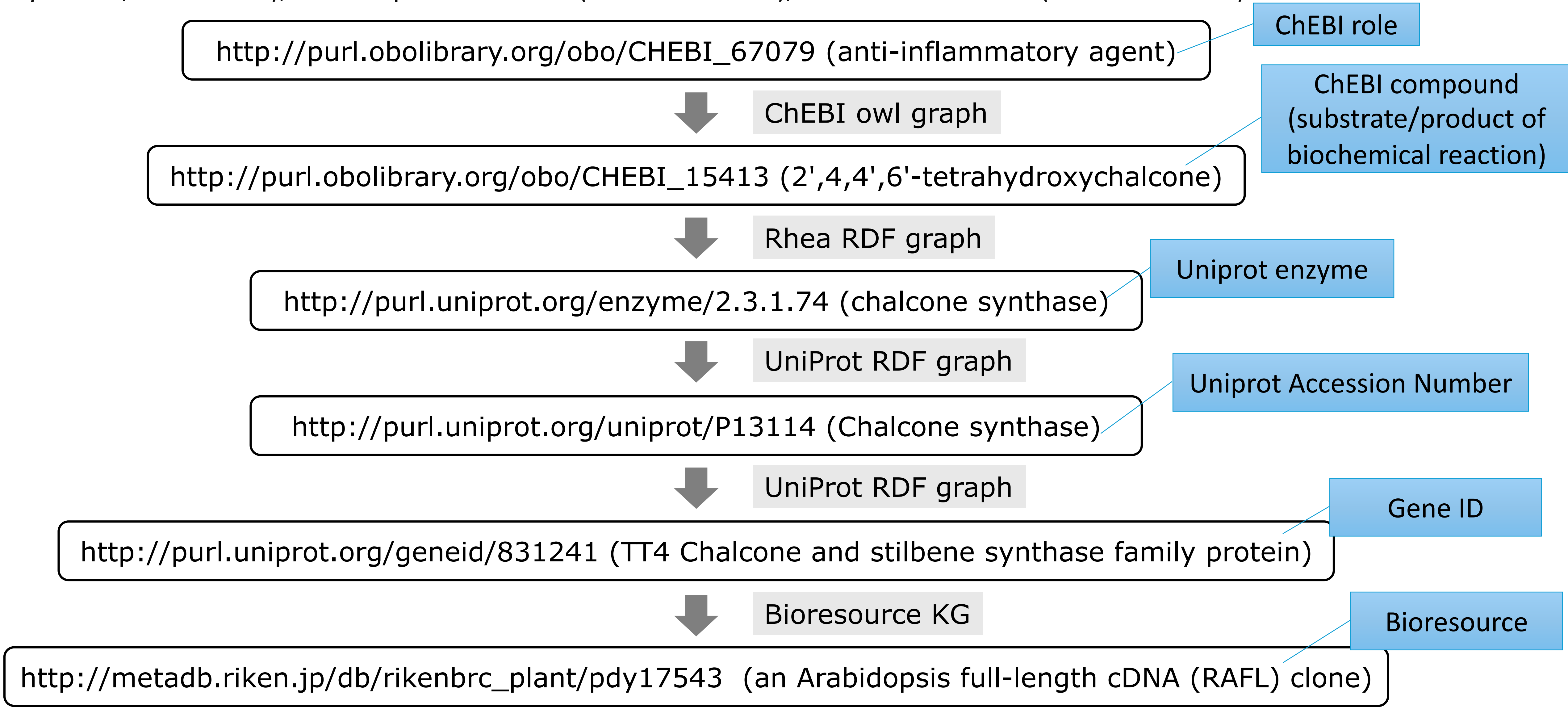
The Biological Resource Schema Ontology (BRSO) is a graph data model for various model organisms and the types, such as individual (mouse, plant, microbe), cell, and DNA, which is largely developed by the Database Center for Life Science (DBCLS), RIKEN, and National Institute of Genetics (NIG). RIKEN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

Semantic network of the integrated Bioresource KG with external KGs/ontologies


brso:genomic_feature/brso:has_genomic_segment/rdfs:seeAlso
bioresource
DNA materials ---> ncbigene
gene
Bioresource KG <xsearch_plant_brso>
Graph connections
uniprot:classifiedWith
biological process
e.g., cuticle development
molecular function
e.g., progesterone 5-alpha-reductase activity
cellular component
e.g., chloroplast
geneOntology
ncbigene <- uniprot
enzyme
UniProt RDF graph <uniprot_ncbigene> , <uniprot_enzyme> and <uniprot_go>
protein
rh:ez
rh:side
rh:contains
Rhea RDF graph <rhea>
rdfs:seeAlso
enzyme <- reaction -> side -> participant
->compound -> chebi
chemicals
uniprot:enzyme
rh:compound
rh:chebi
ChEBI Ontology graph <CHEBI>
chebi --> role
quality
e.g., anti-inflammatory agent
http://purl.uniprot.org/geneid/831241
http://purl.uniprot.org/enzyme/2.3.1.74
KNApSAcK RDF graph <knapsackCore>
chebi -> knapsack-> bioactivity
quality
e.g., induce flowering
External open data
http://purl.obolibrary.org/obo/CHEBI_15379

Example traversal starting from a ChEBI role ("anti-inflammatory agent", CHEBI:67079) to a RIKEN Arabidopsis cDNA clone via a Rhea reaction and its catalyzing enzyme (chalcone synthase, EC 2.3.1.74). This path represents ontology-level associations connecting ChEBI, Rhea, UniProtKB, and NCBI Gene identifiers.

Exploration path from the ChEBI role anti-inflammatory agent (CHEBI:67079) to an Arabidopsis DNA resource (pdy17543) via a chemical entity (2′,4,4′,6′-tetrahydroxychalcone; CHEBI:15413), an enzyme class (chalcone synthase; EC 2.3.1.74), UniProt protein entries (UniProt:P13114), and NCBI Gene TT4 (GeneID:831241).



http://purl.obolibrary.org/obo/CHEBI_67079 (anti-inflammatory agent)

ChEBI role

ChEBI owl graph

http://purl.obolibrary.org/obo/CHEBI_15413 (2',4,4',6'-tetrahydroxychalcone)

ChEBI compound (substrate/product of biochemical reaction)

Rhea RDF graph

http://purl.uniprot.org/enzyme/2.3.1.74 (chalcone synthase)

Uniprot enzyme

UniProt RDF graph

http://purl.uniprot.org/uniprot/P13114 (Chalcone synthase)

Uniprot Accession Number

UniProt RDF graph

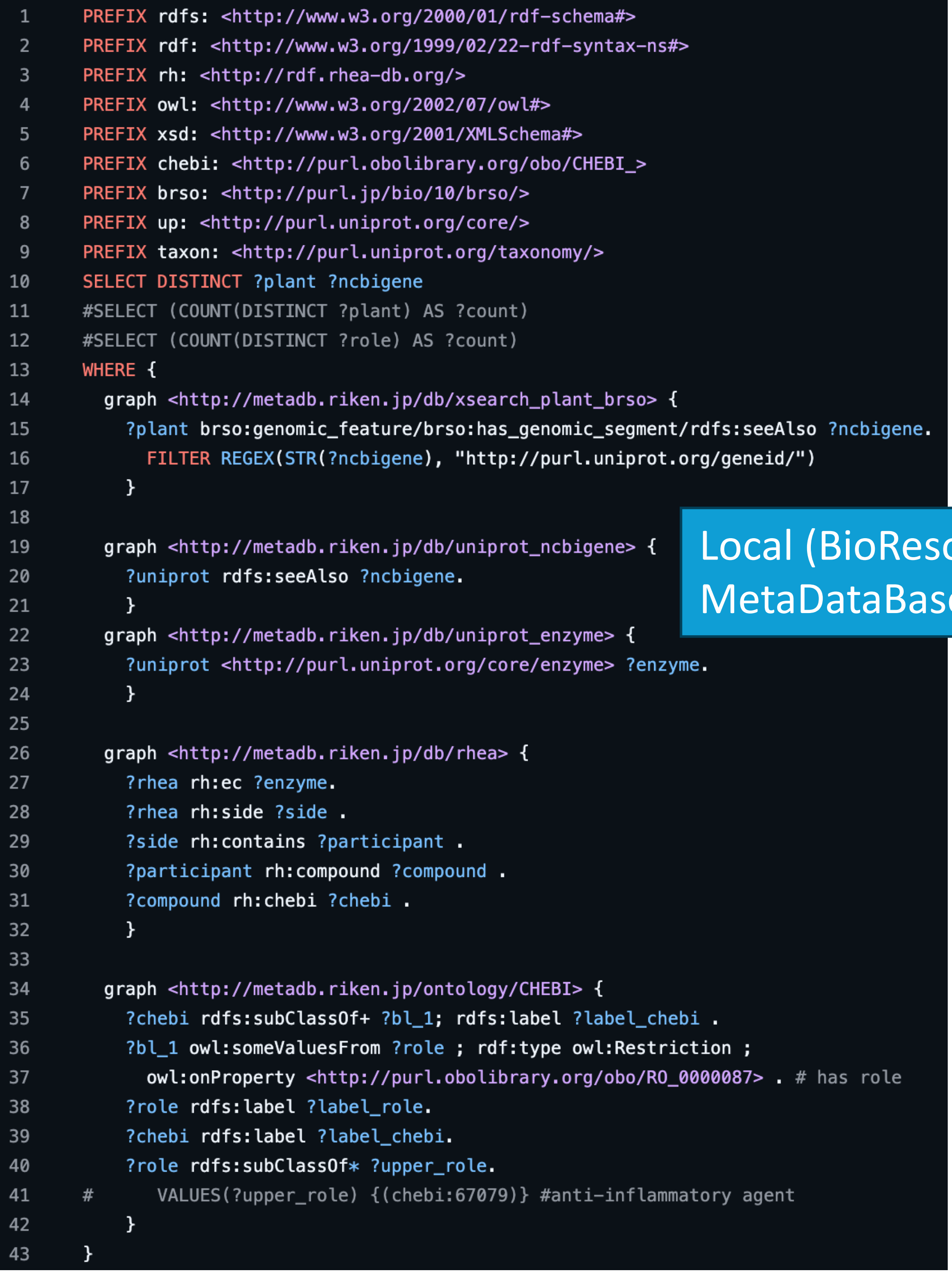
http://purl.uniprot.org/geneid/831241 (TT4 Chalcone and stilbene synthase family protein)

Gene ID

Bioresource KG

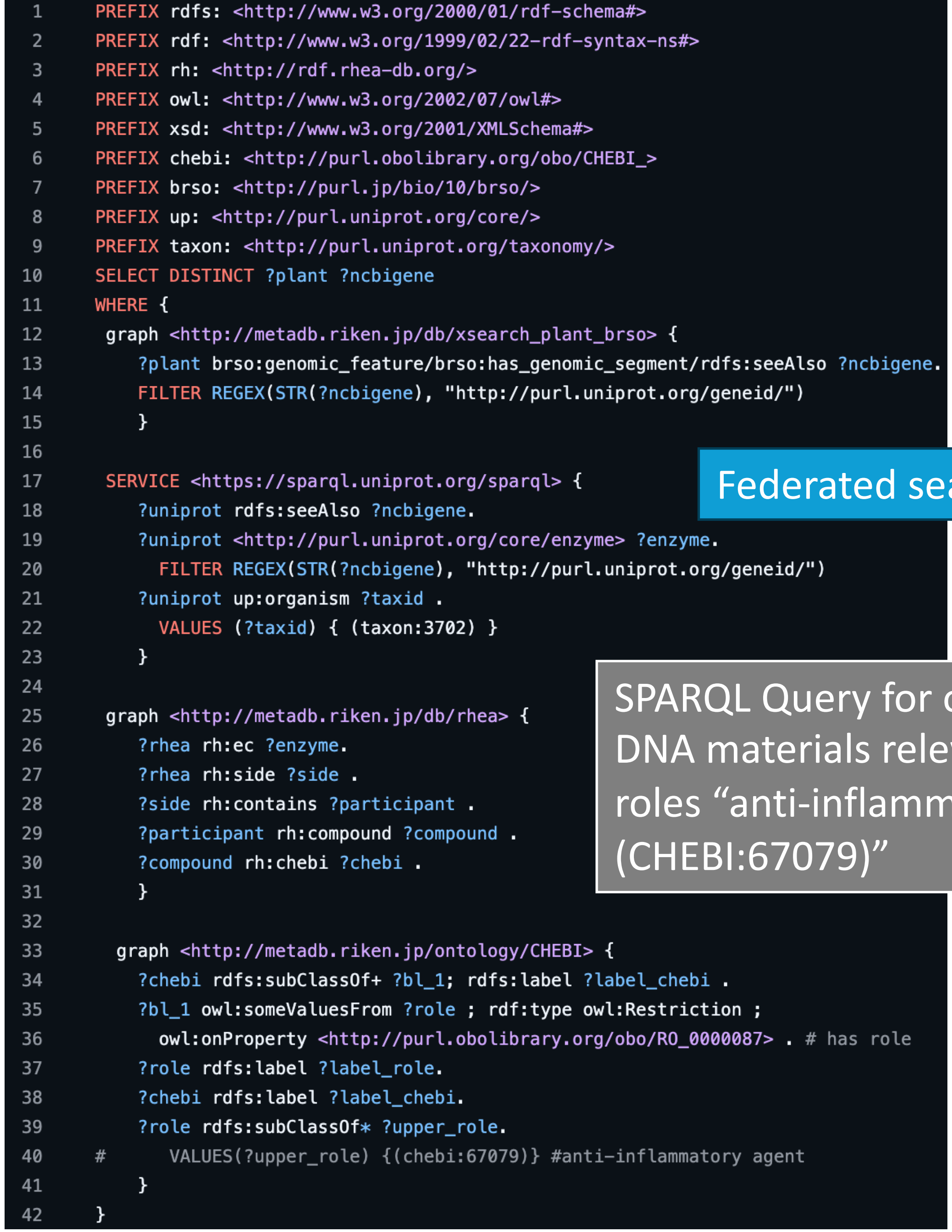
http://metadb.riken.jp/db/rikenbrc_plant/pdy17543 (an Arabidopsis full-length cDNA (RAFL) clone)

Bioresource



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rh: <http://rdf.rhea-db.org/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX chebi: <http://purl.obolibrary.org/obo/CHEBI_>
PREFIX brso: <http://purl.jp/bio/10/brso/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
SELECT DISTINCT ?plant ?ncbigene
#SELECT (COUNT(DISTINCT ?plant) AS ?count)
#SELECT (COUNT(DISTINCT ?role) AS ?count)
WHERE {
  graph <http://metadb.riken.jp/db/xsearch_plant_brso> {
    ?plant brso:genomic_feature/brso:has_genomic_segment/rdfs:seeAlso ?ncbigene.
      FILTER REGEX(STR(?ncbigene), "http://purl.uniprot.org/geneid/")
    }

  graph <http://metadb.riken.jp/db/uniprot_ncbigene> {
    ?uniprot rdfs:seeAlso ?ncbigene.
    }
  graph <http://metadb.riken.jp/db/uniprot_enzyme> {
    ?uniprot <http://purl.uniprot.org/core/enzyme> ?enzyme.
    }

  graph <http://metadb.riken.jp/db/rhea> {
    ?rhea rh:ec ?enzyme.
    ?rhea rh:side ?side .
    ?side rh:contains ?participant .
    ?participant rh:compound ?compound .
    ?compound rh:chebi ?chebi .
    }

  graph <http://metadb.riken.jp/ontology/CHEBI> {
    ?chebi rdfs:subClassOf+ ?bl_1; rdfs:label ?label_chebi .
    ?bl_1 owl:someValuesFrom ?role ; rdf:type owl:Restriction ;
      owl:onProperty <http://purl.obolibrary.org/obo/RO_0000087> . # has role
    ?role rdfs:label ?label_role.
    ?chebi rdfs:label ?label_chebi.
    ?role rdfs:subClassOf* ?upper_role.
#      VALUES(?upper_role) {(chebi:67079)} #anti-inflammatory agent
    }
}
```

Local (BioResource MetaDataBase)

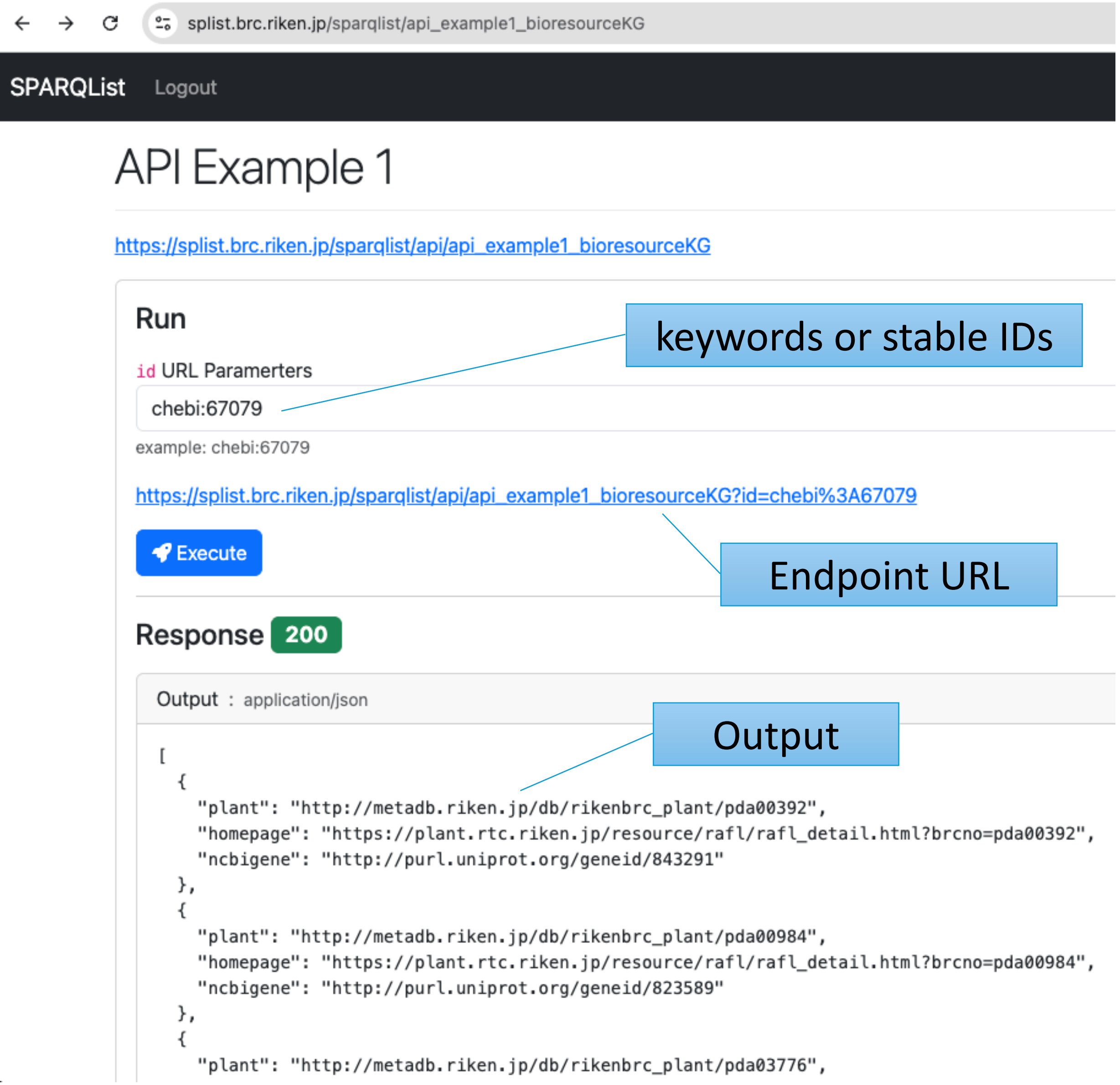


```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rh: <http://rdf.rhea-db.org/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX chebi: <http://purl.obolibrary.org/obo/CHEBI_>
PREFIX brso: <http://purl.jp/bio/10/brso/>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
SELECT DISTINCT ?plant ?ncbigene
WHERE {
 graph <http://metadb.riken.jp/db/xsearch_plant_brso> {
    ?plant brso:genomic_feature/brso:has_genomic_segment/rdfs:seeAlso ?ncbigene.
    FILTER REGEX(STR(?ncbigene), "http://purl.uniprot.org/geneid/")
    }

 SERVICE <https://sparql.uniprot.org/sparql> {
    ?uniprot rdfs:seeAlso ?ncbigene.
    ?uniprot <http://purl.uniprot.org/core/enzyme> ?enzyme.
      FILTER REGEX(STR(?ncbigene), "http://purl.uniprot.org/geneid/")
    ?uniprot up:organism ?taxid .
      VALUES (?taxid) { (taxon:3702) }
    }

 graph <http://metadb.riken.jp/db/rhea> {
    ?rhea rh:ec ?enzyme.
    ?rhea rh:side ?side .
    ?side rh:contains ?participant .
    ?participant rh:compound ?compound .
    ?compound rh:chebi ?chebi .
    }

  graph <http://metadb.riken.jp/ontology/CHEBI> {
    ?chebi rdfs:subClassOf+ ?bl_1; rdfs:label ?label_chebi .
    ?bl_1 owl:someValuesFrom ?role ; rdf:type owl:Restriction ;
      owl:onProperty <http://purl.obolibrary.org/obo/RO_0000087> . # has role
    ?role rdfs:label ?label_role.
    ?chebi rdfs:label ?label_chebi.
    ?role rdfs:subClassOf* ?upper_role.
#      VALUES(?upper_role) {(chebi:67079)} #anti-inflammatory agent
    }
}
```

Federated search (SERVICE)

SPARQL Query for obtaining plant DNA materials relevant to ChEBI roles “anti-inflammatory agent (CHEBI:67079)”

REST API for exploring the integrated KG



API Example 1

https://splist.brc.riken.jp/sparqlist/api/api_example1_bioresourceKG

Run

id URL Paramerters

chebi:67079

example: chebi:67079

https://splist.brc.riken.jp/sparqlist/api/api_example1_bioresourceKG?id=chebi%3A67079

Response 200

Output : application/json

```
[
  {
    "plant": "http://metadb.riken.jp/db/rikenbrc_plant/pda00392",
    "homepage": "https://plant.rtc.riken.jp/resource/rafl/rafl_detail.html?brcno=pda00392",
    "ncbigene": "http://purl.uniprot.org/geneid/843291"
  },
  {
    "plant": "http://metadb.riken.jp/db/rikenbrc_plant/pda00984",
    "homepage": "https://plant.rtc.riken.jp/resource/rafl/rafl_detail.html?brcno=pda00984",
    "ncbigene": "http://purl.uniprot.org/geneid/823589"
  },
  {
    "plant": "http://metadb.riken.jp/db/rikenbrc_plant/pda03776",
```

- Approach
  - Users submit **keywords or stable IDs** (EN/JA) → API expands to **representative SPARQL** → returns **JSON** (ID, label, linked GO/ChEBI/Rhea/EC/UniProtKB/NCBI Gene).

- Query examples
  - ChEBI role → plant / human DNA materials
    - Input: *“Anti-inflammatory agent”* or **CHEBI:67079**
  - GO terms → plant / human DNA materials
    - Input: “cuticle development” or GO:0042335

Coverage summary (as of 2025-10-15; all computed locally in MetaDB, no federation)

	Total bioresources	Bioresources with ≥1 NCBI Gene	Bioresources whose linked genes resolve to ≥1 UniProt protein	Bioresources reaching ≥1 ChEBI role	Bioresources with ≥1 GO term
DNA material (e.g., *Homo sapiens, Mus musculus*).	169,107	136,989 (81.0% = 136,989/169,107)	127,257 (75.3% = 127,257/169,107)	19,809 (**11.7%** = 19,809/169,107)	108,589 (**64.2%** = 108,589/169,107)
Plant DNA material (*Arabidopsis thaliana*)	612,129	267,613 (43.7% = 267,613/612,129)	266,167 (43.5% = 266,167/612,129)	61,499 (**10.0%** = 61,499/612,129)	254,082 (**41.5%** = 254,082/612,129)

Notes. “Bioresources whose linked genes resolve to ≥1 UniProtKB protein” means: at least one gene attached to the resource could be resolved to a UniProtKB protein (gene -> protein). NCBI Gene IRIs are normalized across four accepted patterns before DISTINCT counting. External graphs (UniProtKB, Rhea, ChEBI, GO) are hosted locally in MetaDB; no SERVICE federation was used.

	NCBI Gene related to bioresources	UniProt protein related to bioresources	ChEBI role related to bioresources	GO term related to bioresources
DNA material	36,060	49,670	166(**12.2%** = 166/1364)	18,273(**85.6%** = 18,273/21,352)
Plant DNA material	27,841	54,982	163(**12.0%** = 163/1364)	6,459(**30.3%** = 6,459/21,352)

Summary and future work

- KG integration enables **ChEBI/GO-driven retrieval** to the right bioresources, via a verifiable path from ontology terms to testable interventions.
- Implement **REST API** for exploring bioresources.
- Reachability (coverage) from ChEBI/GO to bioresources was quantified to assess practical effectiveness.
  - **ChEBI role–based coverage** to bioresources **:** *limited but valuable* (**11.7%** for DNA materials, e.g., human), offering a unique entry point via chemical/pharmacological roles.
  - **GO-based coverage** to bioresources : practically effective (**64.2%** for DNA materials, e.g., human).

- Future work
  - Release new UI, continue query optimization, re-evaluate cost-based federation with fallbacks, and use evidence tags (e.g., GO evidence codes).

Acknowledgements

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP), “Technologies for Smart Bio-industry and Agriculture” (funding agency: Bio-oriented Technology Research Advancement Institution).

Statistics of the integrated ontologies/graphs

Integrated ontologies/RDF graphs	Data format of original data or conversion methods to N-triples	Data acquisition date	License	No. of triples (distinct)	No. of properties (distinct)
ChEBI (ver. 244)	OWL	09/09/2025	CC BY 4.0	916,983	39
Gene Ontology (ver. 2025-07-22)	OWL	10/09/2025		241,191	50
UniProt-NCBI Gene (ver. 2025-06-18)	Conversion from TSV data to N-triples	30/07/2025		12,625,915	1
UniProt-Enzyme	Execution of the SPARQL query and acquisition of N-triple	30/07/2025		36,460,186	1
UniProt-GO		30/07/2025		349,538	2
Rhea	RDF/XML	10/03/2024		211,004	67
KNApSAcK	Turtle	23/08/2023		779,925	33
Bioresource KG (DNA material, e.g., *Homo sapiens, Mus musculus*))	Turtle	04/07/2025		2,185,956	31
Bioresource KG (Plant DNA material, e.g., *Arabidopsis thaliana*))	Turtle	29/06/2025		2,596,497	27

Comparison of **local** and **federated** SPARQL execution results

RIKEN BRC
RIKEN BioResource Research Center

Item	**Local (MetaDB)**	**Federated (SERVICE)**
Data location	Bioresource KG + external RDF hosted locally	Query remote official endpoints (e.g., UniProt)
Freshness	Snapshot-based (on ingest date)	Near real-time at source
Maintenance	Needs periodic mirroring/updates	Lower local maintenance
Typical runtime	Completed	>10 min or HTTP 502, often no result
Result of query *1 or *2	957 bioresources returned (*1)	0 (timeouts (*2) / 502 Proxy errors)
Practicality	Practical (completed; consistent results)	Impractical (timeouts/502; no results)
Next steps	Query/subquery tuning; store config; server scaling	Query/subquery tuning; Evaluate cost-based federated planners; selective federation

*1: [N**on-federated search, Local**]: SPARQL Query for obtaining plant DNA materials relevant to ChEBI roles “anti-inflammatory agent (CHEBI:67079)” (non-federated search)

*2 [**Federated search (SERVICE)]:** the same query