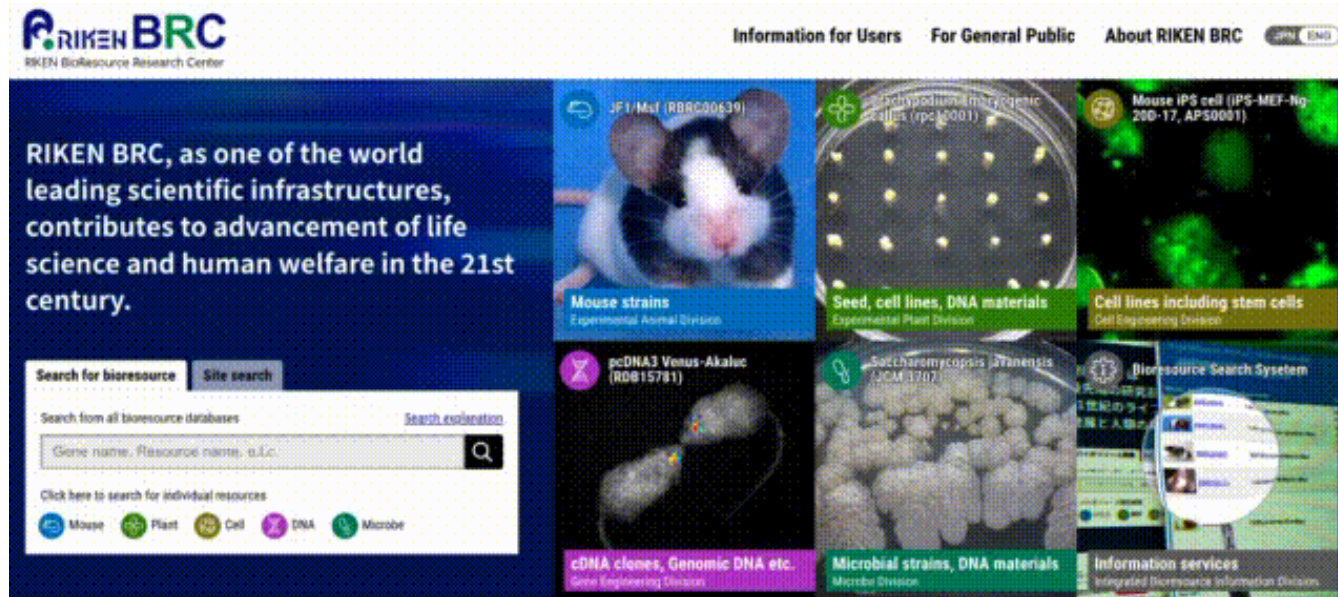


○ 檜田 達矢（理研）、臼田 大輝（理研）、桝屋 啓志（理研）
理化学研究所 バイオリソース研究センター（BRC）



理化学研究所 バイオリソース研究センター（BRC）

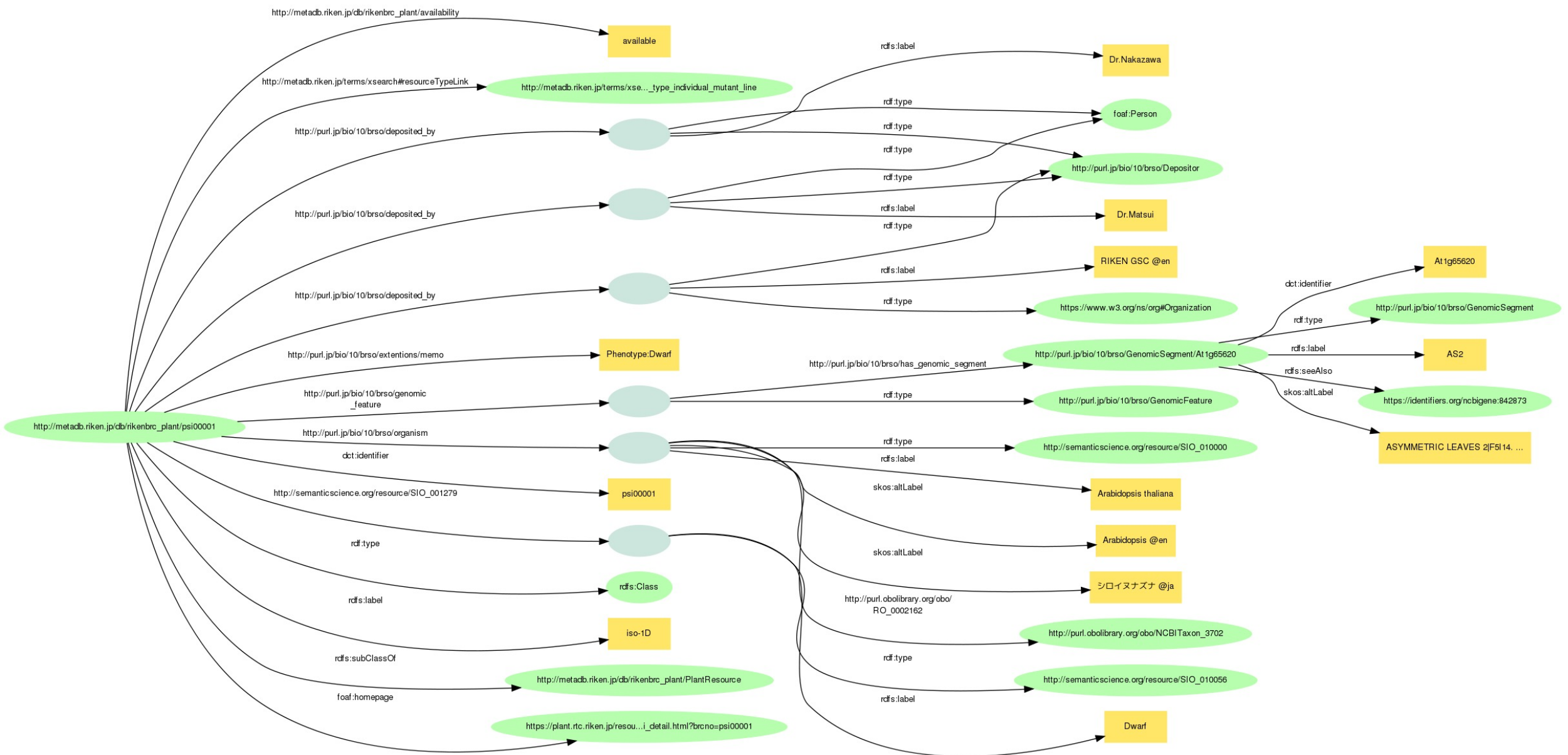
Outline

- 理研BRCでは、1) 実験動物（マウス）、2) **実験植物**、3) 細胞材料(例、iPS cells)、4) 遺伝子材料(例、ヒトcDNAクローン)、5) 微生物株を取り扱っている。
- 実験植物には、**植物遺伝子材料**（例、植物cDNAクローン[9植物種]）、種子[2植物種]及び細胞材料[24植物種]を含む。
- 理研BRCのミッションの1つは、理研BRCの植物材料を通して、生命科学の発展、育種や有用物質生産に貢献することである。
- 現在、理研BRCではバイオリソースRDFデータの開発を行い、外部データセットとの統合を進めている。
- 本報告では、抗炎症剤（anti-inflammatory agent）、開花誘導（induce flowering）、クチクラ形成（cuticle development）など**ChEBI**のロール、**KNApSack**の生物活性、**Gene Ontology**の各用語を用いて、目的とする実験植物材料の発見を可能にしたデータ整備の状況を紹介する。

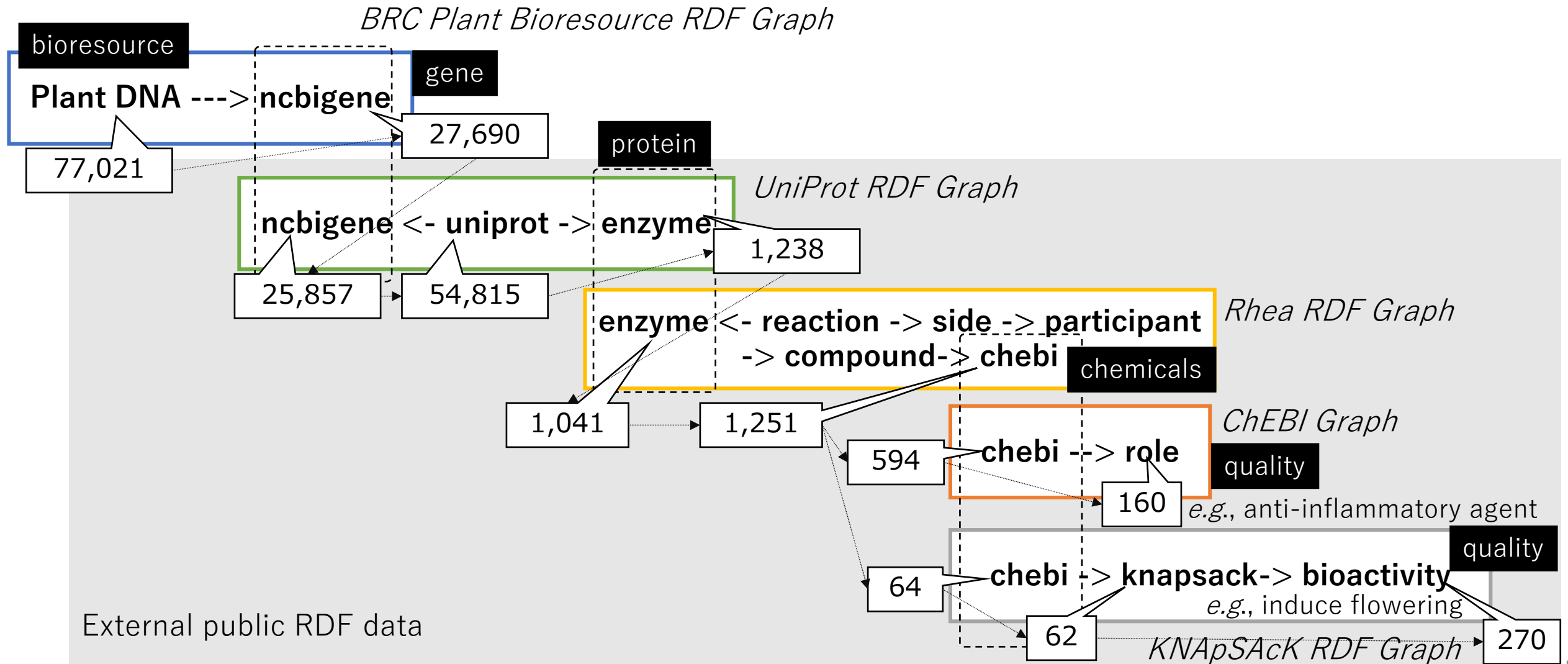
理研BRCが取り扱う実験植物

リソースのタイプ	植物種	種数 (合計:31種)	NCBI Geneの リンクを持つ実験植物 [遺伝子数]
遺伝子, 種子, 培養細胞	シロイヌナズナ	1 種	1 種 [27,690] (塩基配列あり)
遺伝子, 種子	ミナトカモジグサ	1 種	0 種 (塩基配列あり)
遺伝子, 培養細胞	タバコ	1 種	0 種 (塩基配列あり)
遺伝子	ヒメツリガネゴケ, キャッサバ, ハロフィラ (モデル 塩生植物, シロイヌナズナ近縁種), ストライガ, セ イヨウハコヤナギ, ハクサイ	6 種	0 種 (塩基配列あり)
培養細胞	ノラニンジン, ヨーロッパブドウ, ヨウシュヤマゴボ ウ, ホウレンソウ, ソメイヨシノ, ハチク, ツルニチ ニチソウ, ニチニチソウ, アスパラガス, ヒャクニチ ソウ, ミヤコグサ, ダイズ, イネ, オウレン, トマト, ゴマ, ヘチマ, ムラサキ, マダケ, キダチタバコ, ウ コン, ヨウシュハッカ	22種	0 種 (塩基配列なし)

BRC植物材料（シロイヌナズナcDNAクローン）のRDFスキーマ

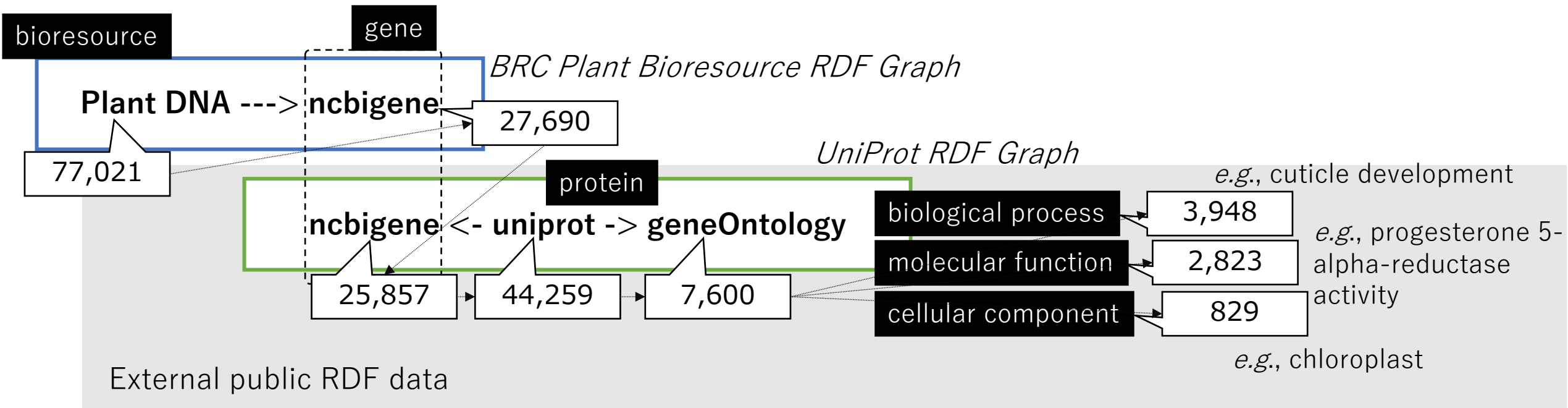


BRC実験植物と外部RDFデータの統合例 1/2



5

BRC実験植物と外部RDFデータの統合例 2/2



抗炎症剤の生成と関係が示唆されるBRC実験植物（遺伝子材料）を見つけるSPARQLクエリ

```
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_>
SELECT DISTINCT ?plant ?rhea ?chebi
WHERE {
```

BRC Bioresource RDF

```
graph <http://metadb.riken.jp/db/plant_ncbigene> {?plant rdfs:seeAlso ?ncbigene.}
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.}
```

UniProt RDF

```
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 .}
```

Rhea RDF

```
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.
    VALUES(?role) {(chebi:67079)} } #anti-inflammatory agent
```

ChEBI

```
}
```

外部データのトリプルストアへの実装状況

*: Total data size of the UniProt RDF data is 1.9 TB and more.
 **: The size of human, mouse, and Arabidopsis thaliana only.

Integrated RDF data	Status	Graph (in Bioresource MetaDB)	Data source	RDFization (RDF data conversion)	File size (ntriples)
UniProt	impleme nted	http://metadb.riken.jp/db/uniprot_ncbigene	idmapping_selected.tab.gz (from the FTP site)	Conversion from .tab to .ttl using a Python script	4.75 GB /1.9 TB*
	impleme nted	http://metadb.riken.jp/db/uniprot_enzyme	UniProt SPARQL Endpoint	Execution of SPARQL query and conversion of the query result to .ttl using a Python script	3.77 GB /1.9 TB*
	impleme nted	http://metadb.riken.jp/db/uniprot_go	UniProt SPARQL Endpoint	Execution of SPARQL query and conversion of the query result to .ttl using a Python script	203.3 MB** /1.9 TB*
Rhea	impleme nted	http://metadb.riken.jp/db/rhea	https://ftp.expasy.org/databases/rhea/rdf/	NOT necessary	189.3 MB
ChEBI	impleme nted	http://metadb.riken.jp/ontology/CHEBI	https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/	NOT necessary	586.4 MB (owl)
KNAPSAcK	NOT impleme nted	http://metadb.riken.jp/db/knapsackCore	https://github.com/ddbj/mb_knapsack	NOT necessary	612.6 MB



結論

- 77,021件の理研BRCの**実験植物材料**のRDFデータを, NCBI GENE ID使ってUniport, Rhea, ChEBI, KNApSack, Gene Ontologyなどの外部データセットと統合した.
- その結果, 160種類の**ChEBI**のロール (例, anti-inflammatory agent) に対して17,435件の実験植物材料が, 270種類の**KNApSack**の生物活性用語 (例, induce flowering) に対して1,921件の実験植物材料が, 7,600種類の**Gene Ontology**用語 (例, cuticle development) に対して71,746件の実験植物材料が発見された.
- 一方, NCBI GENE IDを持たない522,006件の実験植物材料 (例, 植物培養細胞材料) については, Uniport, Rhea, ChEBI, KNApSack, Gene Ontologyなどの外部データセットと結合はできなかった.
- UniProtは, 多数の外部データのURIやID (例, NCBI Gene, EC number, and Gene Ontology) を持つため, ライフサイエンスデータの統合におけるハブとして有用である.
- しかし, オリジナルのUniProtはデータサイズが大きく (1.9 TB以上), それをそのままの形でRDFデータのデータベース (トリプルストア) で取り扱う場合, 検索速度の低下などトラブル発生が懸念される.
- これに対応するため, UniProtオリジナルデータ (例, ttl, csv形式) を適当なサイズに分割, フォーマット変換し, トリプルストアに格納, 統合, 利活用した.

要旨

理化学研究所バイオリソース研究センター（BRC）は、総合的なバイオリソースセンターとして、シロイヌナズナなど31植物種の種子、培養細胞、遺伝子材料の開発、収集、保存、提供などの業務を行っている。当センターでは、これらバイオリソースの利用拡大に向けて、セマンティックウェブの標準書式であるResource Description Framework (RDF)形式によるデータ整備、検索システムの開発、運用を進めている。RDFデータは、バイオリソースの汎用的なデータスキーマであるBiological Resource Schema Ontology (BRSO)に準拠し、NCBI Taxonomy（例、[obo:NCBITaxon/3702](https://ncbi.nlm.nih.gov/taxonomy/obo:NCBITaxon/3702)）、Gene（例、[ncbigene:820887](https://ncbi.nlm.nih.gov/gene/ncbigene/820887)）、Nucleotide（例、[nucleotide:AF37539](https://ncbi.nlm.nih.gov/nucl/nucleotide/AF37539)）やTAIR Locus（例、[ensembl.plant:AT3G16400.1](https://tair.org/locus/ensembl.plant/AT3G16400.1)）などのURIを用いて構築している。現在、NCBI GeneとUniProt間のリンク情報の追加、リテラルで記述されている表現型の用語のPlant Trait Ontologyなどへのマッピングや関連付けを進めており、今後、Plant GARDEN, Orthologous Matrix (OMA), Rhea, KNApSack, Reactomeなどの外部機関が提供するデータセットとの統合を行う予定である。これら理研BRCのバイオリソースデータは、バイオリソースメタデータベースのSPARQLエンドポイント（<https://knowledge.brc.riken.jp/sparql>）および、理研BRCのトップページ（<https://brc.riken.jp/>）から検索可能である。

謝辞：本研究の一部は、内閣府 戦略的イノベーション創造プログラム（S I P）「スマートバイオ産業・農業基盤技術」（管理法人：農研機構生研支援センター）によって実施されました。