植物ゲノム統合データベースPlant GARDENにおける微生物、メタボローム統合データベースとの連携

○平川　英樹[1]、藤澤　貴智[2]、長崎　英樹[1]、森　宙史[2]、福島　敦史[3]、ジェルフィ　アンドレア[1]、市原　寿子[1]、
中村　保一[2]、金谷　重彦[4]、有田　正規[2]、黒川　顕[2]、磯部　祥子[1]、田畑　哲之[1]

1. かずさDNA研究所
2. 国立遺伝学研究所
3. 理化学研究所環境資源科学研究センター
4. 奈良先端科学技術大学院大学先端科学技術研究科

要旨

統合化推進プログラムにおいて、植物ではPlant GARDEN（https://plantgarden.jp）、微生物ではMicrobeDB.jp（https://microbedb.jp）、メタボロームではMetaboBank（https://www.ddbj.nig.ac.jp/metabobank/）が構築されており、それぞれが扱う以下の主要なデータセットについてRDF化が行なわれている。

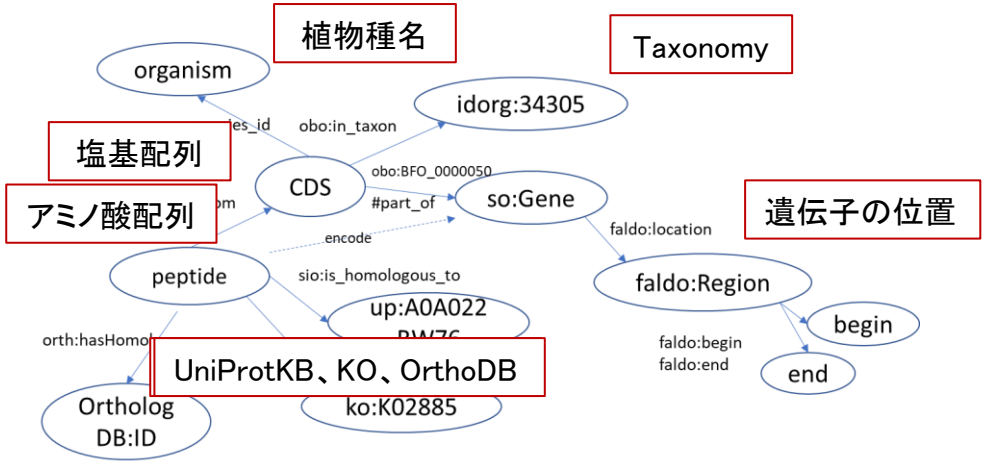
- Plant GARDEN：植物ゲノム、遺伝子、アノテーション、DNAマーカー、QTL
- MicrobeDB.jp：メタ16Sデータ、メタゲノム解析データ、オントロジーアノテーションされたサンプル属性情報
- MetaboBank：実験で得られた代謝産物に関する生データや実験に関するメタデータ

植物ゲノム統合では、PubTatorにより各植物種に関する病気や遺伝子、化合物などの情報について文献キュレーションを行い、DBCLSのTogoDBを用いて得られたデータに対してRDF化を行っている。現在、SPARQLエンドポイントを構築し、各統合化データベースの連携によって、更なるデータ統合を図っている。SPARQLを用いることで、植物種名や遺伝子、化合物（代謝産物）などでクエリ検索し、新たな知識発見ができるようになることを試みている。

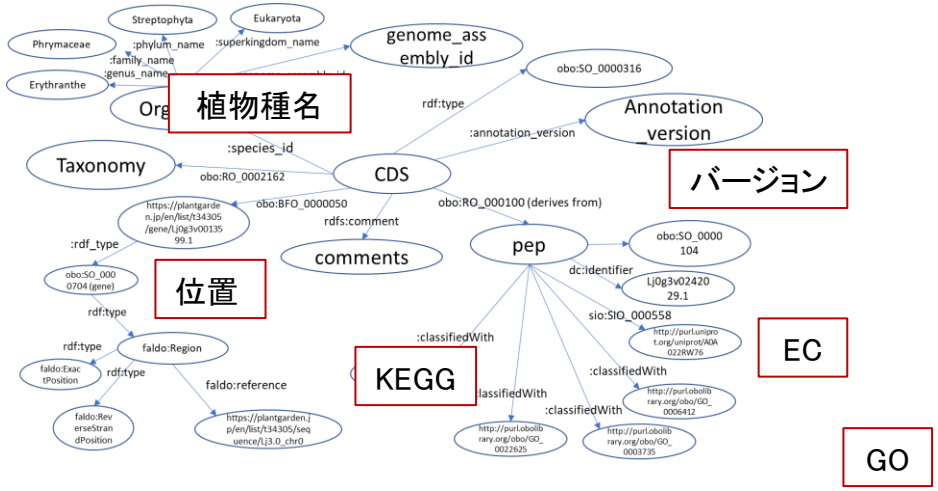
植物ゲノム関連データのRDF化

データ連携を行うため、Plant GARDENで公開している遺伝子、アノテーション、DNAマーカー、QTL情報のRDF化を行った。
国立遺伝学研究所とDBCLSからのご協力を頂いた。

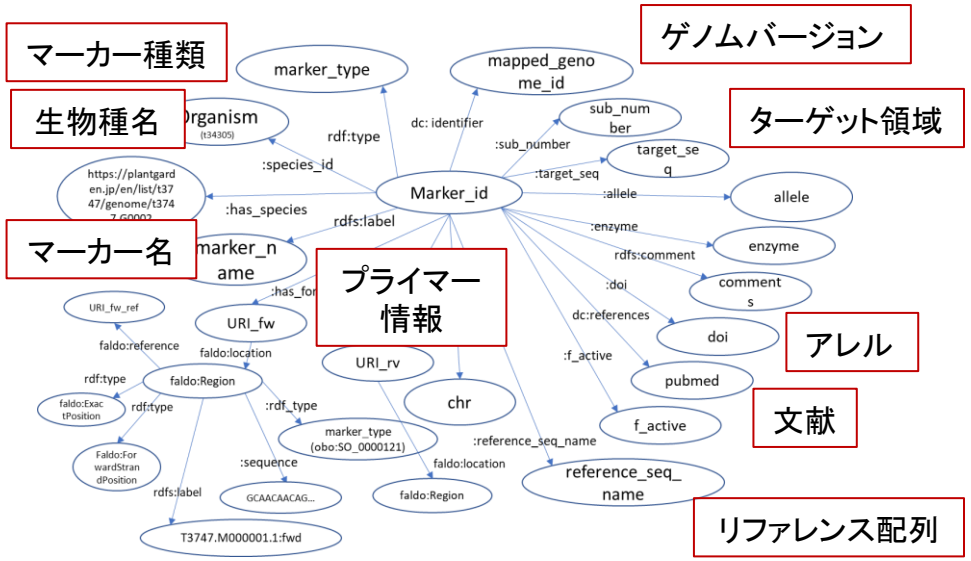
① 遺伝子情報（gff）



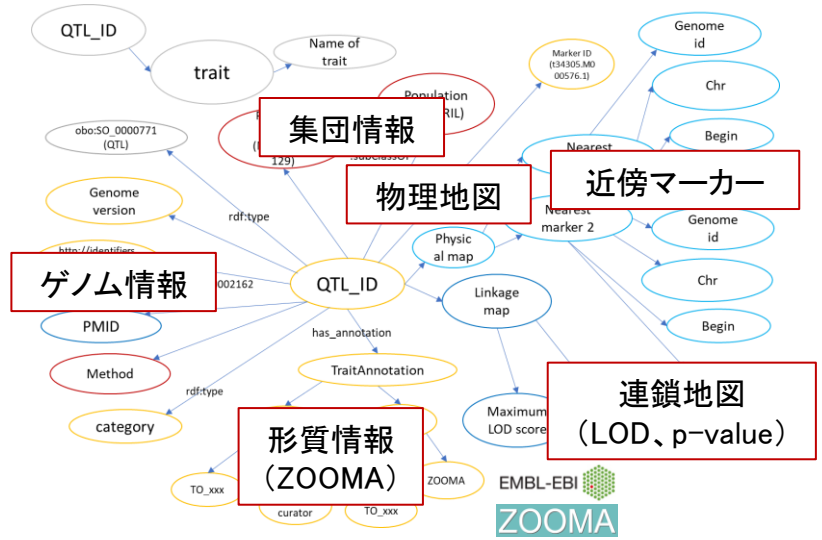
② アノテーション情報（Hayai-Annotation）



③ DNAマーカー情報



④ QTL情報



RDFを用いたデータ連携

植物は微生物と相互作用しており、様々な二次代謝産物を産生する。RDFを用いて植物、微生物、メタボローム統合データベースを繋ぐことを試みている。

Plant GARDEN

・ゲノム情報（134種）
　ゲノム配列、CDS、PEP、アノテーション
・DNAマーカー（38種）
・QTL（27種）

PubTatorを用いた文献キュレーションのRDF

・生物種、病気、遺伝子、化合物など
・植物と微生物の相互作用
　Plant Growth Promoting Bacteria（PGPB；共生、成長促進）、病原性細菌（細菌、カビ、ウイルス）
・代謝産物（MESH ID）

連携

SPARQL

MetaboBank

・実験で得られた代謝産物に関する生データや実験に関するメタデータ
・データ数
・KNApSAcKのRDF
（DBCLSにて実施）

KNApSAcK

連携

SPARQL

MicrobeDB.jp

・系統・遺伝子・環境の3つの軸に沿って整理・統合されたフルRDFのデータベース
・メタゲノム解析データ（メタ16S rRNA、遺伝子）
・オントロジーアノテーションされたサンプル属性情報
・単細胞の真菌類、藻類のゲノムデータ
・解析プロトコルの標準化
・解析パイプラインの開発

SPARQL

外部データベースのRDF

Rhea SIB
化合物（11,886個）
酵素反応（13,673個）

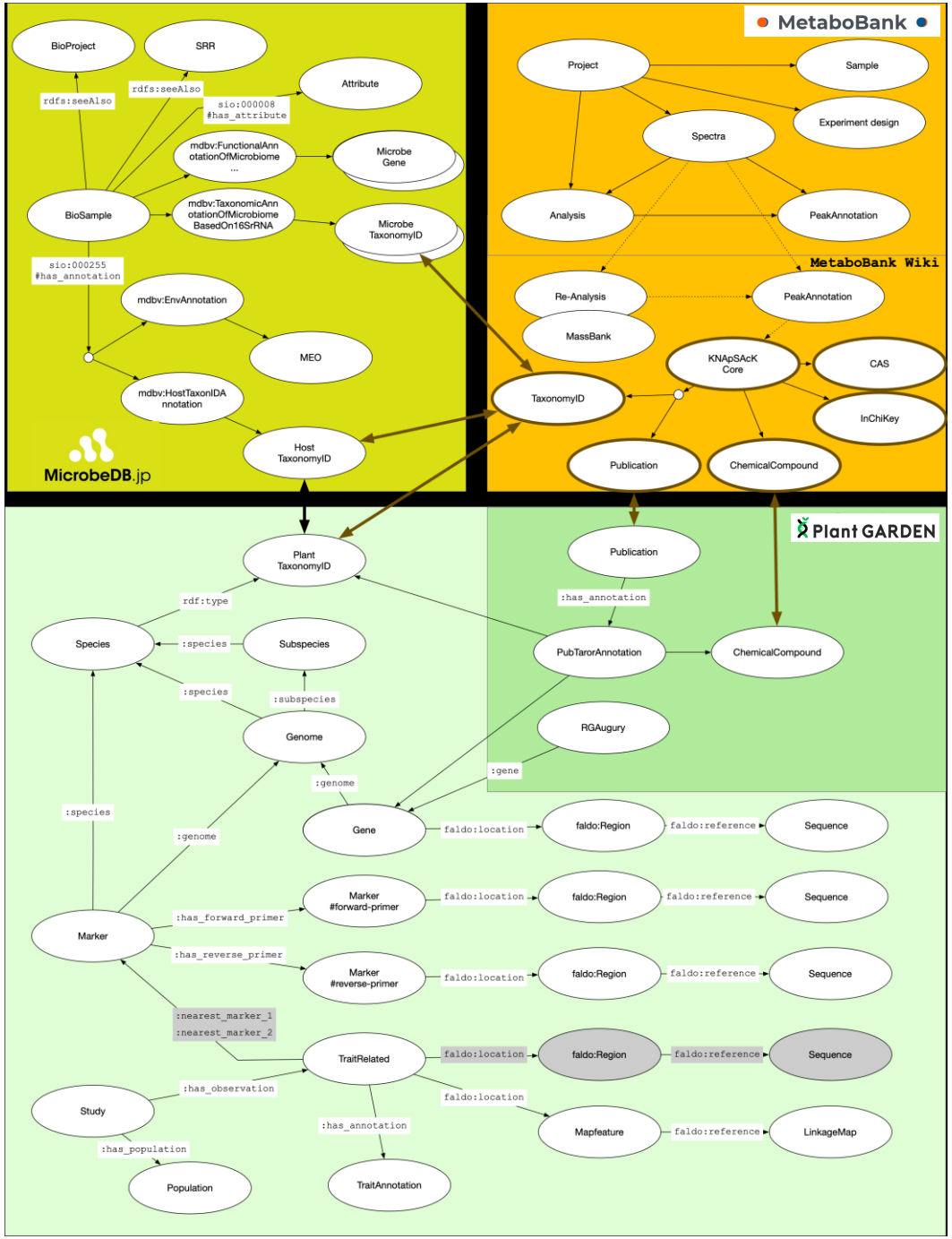
Medical Subject Headings RDF NIH

NBDC RDF Portal
NBDC National Bioscience Database Center
統合化推進プログラムの他のDB

Plant GARDEN、MicrobeDB.jp、MetaboBank間のRDFの関係図

3者間の統合データベースの連携に関するRDFスキーマ図を示した。スキーマ図の作成には国立遺伝学研究所の藤澤さんのご協力を頂いた。

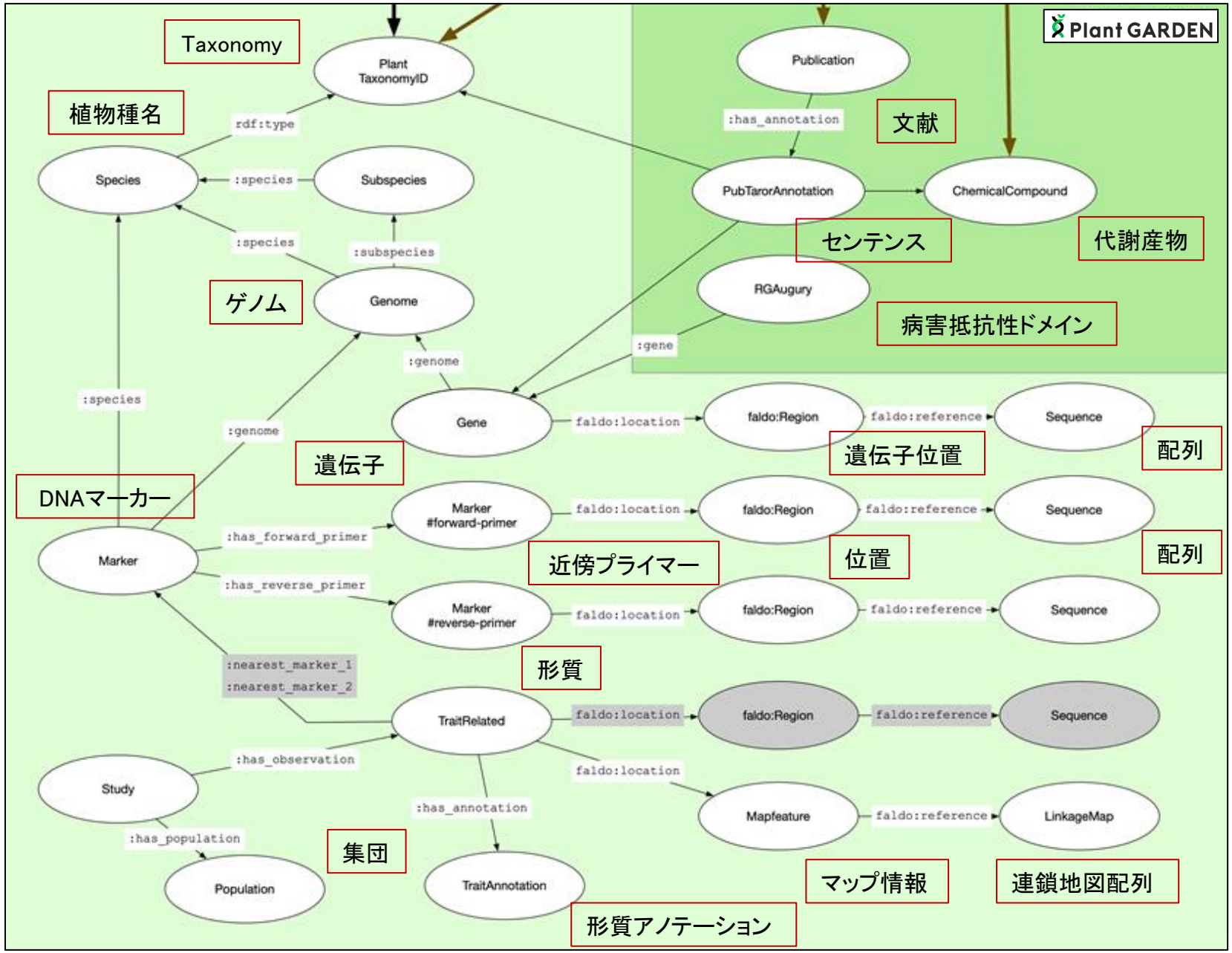
それぞれの統合データベースについて拡大したスキーマ図を次ページに示した。



Plant GARDENにおけるRDFの関係図

以下の項目についてRDF化を行った。

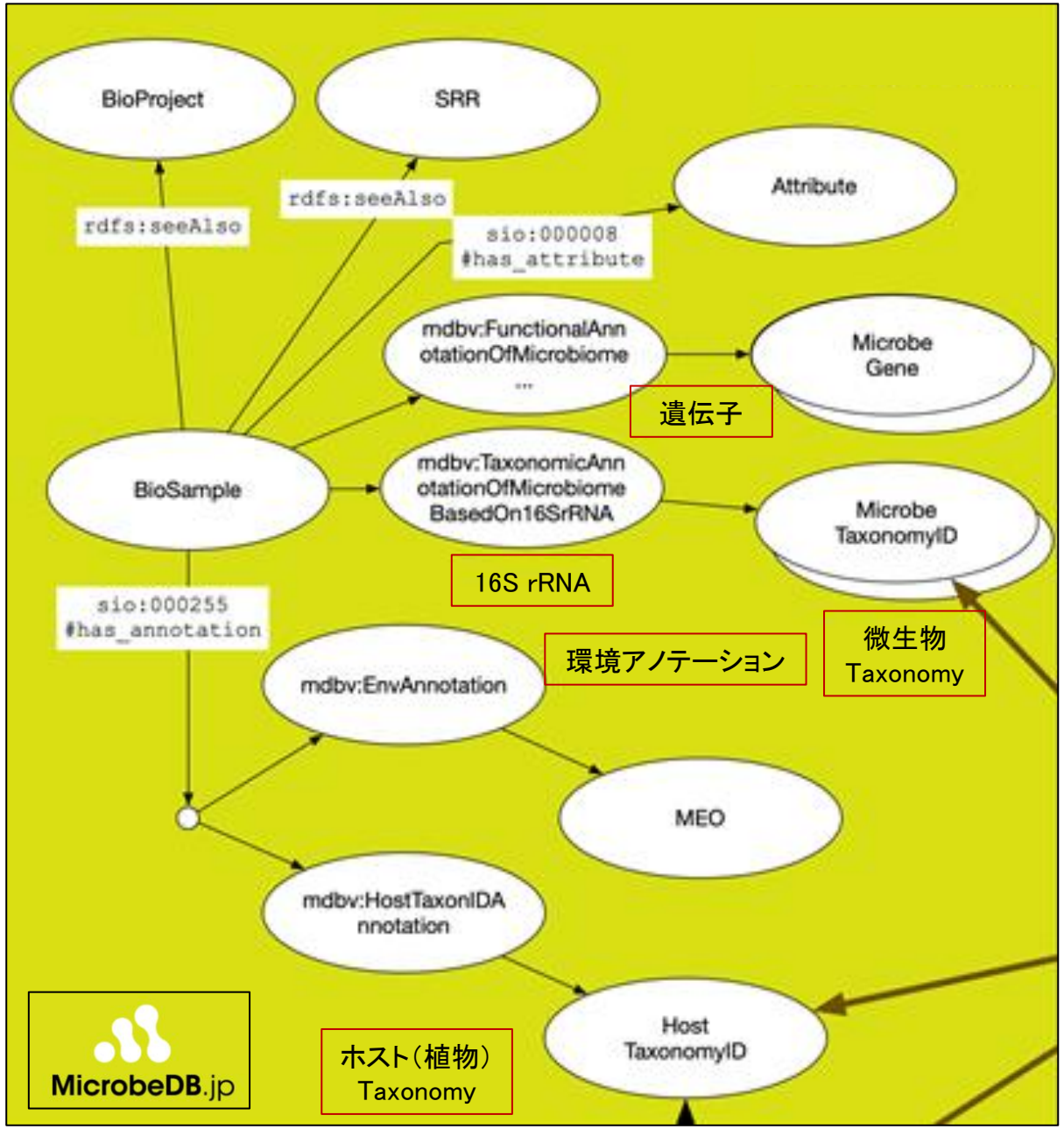
・植物種名
　Taxonomy ID
・植物ゲノム
　遺伝子の位置、配列
・DNAマーカー
　近傍マーカーの位置
　マーカー配列
・形質
　集団名、形質アノテーション、
　マップ情報、連鎖地図配列



MicrobeDB.jpにおけるRDFの関係図

以下の項目についてRDF化を行った。

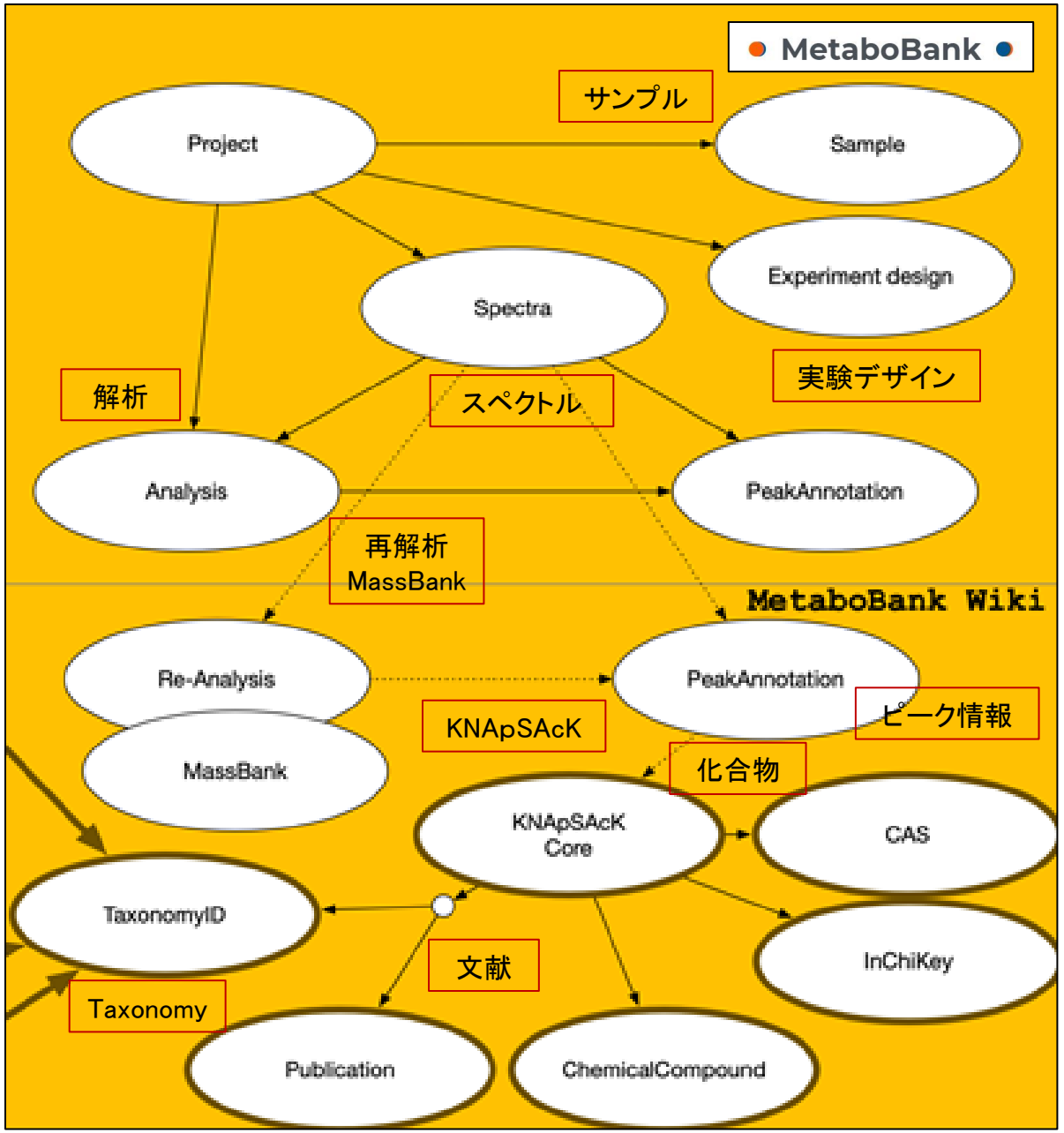
・ホスト（植物種）名
　Taxonomy ID
・環境アノテーション
　MEO
・BioSample
　遺伝子
　16S rRNA
　微生物名、Taxonomy ID
・BioProject
・SRR



MetaboBankにおけるRDFの関係図

以下の項目についてRDF化を行った。

・生物種名
  Taxonomy ID
・Project ID
  サンプル情報
  実験デザイン
  スペクトル
  解析内容
  ピークアノテーション
  再解析（MassBank）
  再解析されたpeakアノテーション
・KNApSAcK Core
  文献情報
  化合物
  CAS
  InChiKey



エンドポイントに格納したトリプル

各統合データベースにおいてエンドポイントを作成した。以下の情報についてのトリプルデータを格納した。

MetaboBank（メタボローム）
https://mbs1.ddbj.nig.ac.jp/sparql

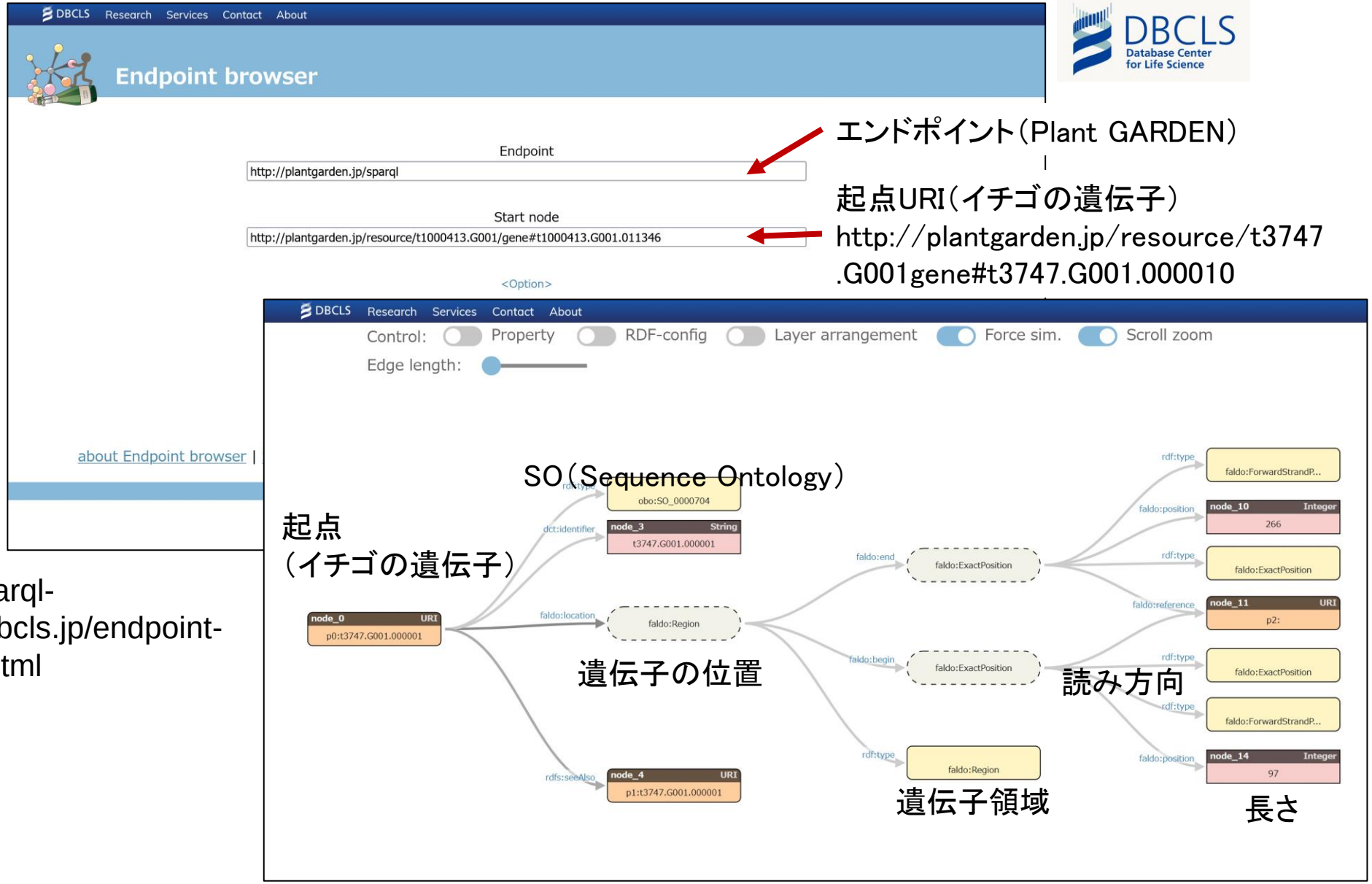
Plant GARDEN（植物）
http://plantgarden.jp/sparql

MicrobeDB.jp（微生物）
https://microbedb.jp/sparql

Abbreviate URIs	#triples	
http://plantgarden.jp/resource/gene	323,678,344	遺伝子
http://metadb.riken.jp/db/plantMetabolomics	33,893,537	MetaboWikiエントリー情報
http://ddbj.nig.ac.jp/ontologies/taxonomy	33,503,477	Taxonomy
http://plantgarden.jp/resource/pg_marker	12,435,112	DNAマーカー
http://mb-wiki.nig.ac.jp/resource	3,489,616	リソース
http://plantgarden.jp/resource/rgaugury	2,337,589	病害抵抗性遺伝子
http://plantgarden.jp/resource/pubtator	447,403	センテンス
http://plantgarden.jp/resource/genome	4,396	ゲノム
http://localhost:8890/DAV/	2,959	
http://www.openlinksw.com/schemas/virtrdf#	2,479	
http://plantgarden.jp/resource/species	1,792	植物種
http://plantgarden.jp/resource/subspecies	1,639	植物種（subspecies）
http://plantgarden.jp/resource/trait	1,378	形質
http://www.w3.org/2002/07/owl#	160	
http://localhost:8890/sparql	14	
http://www.w3.org/ns/ldp#	3	

格納されたデータの確認（Endpoint browser）

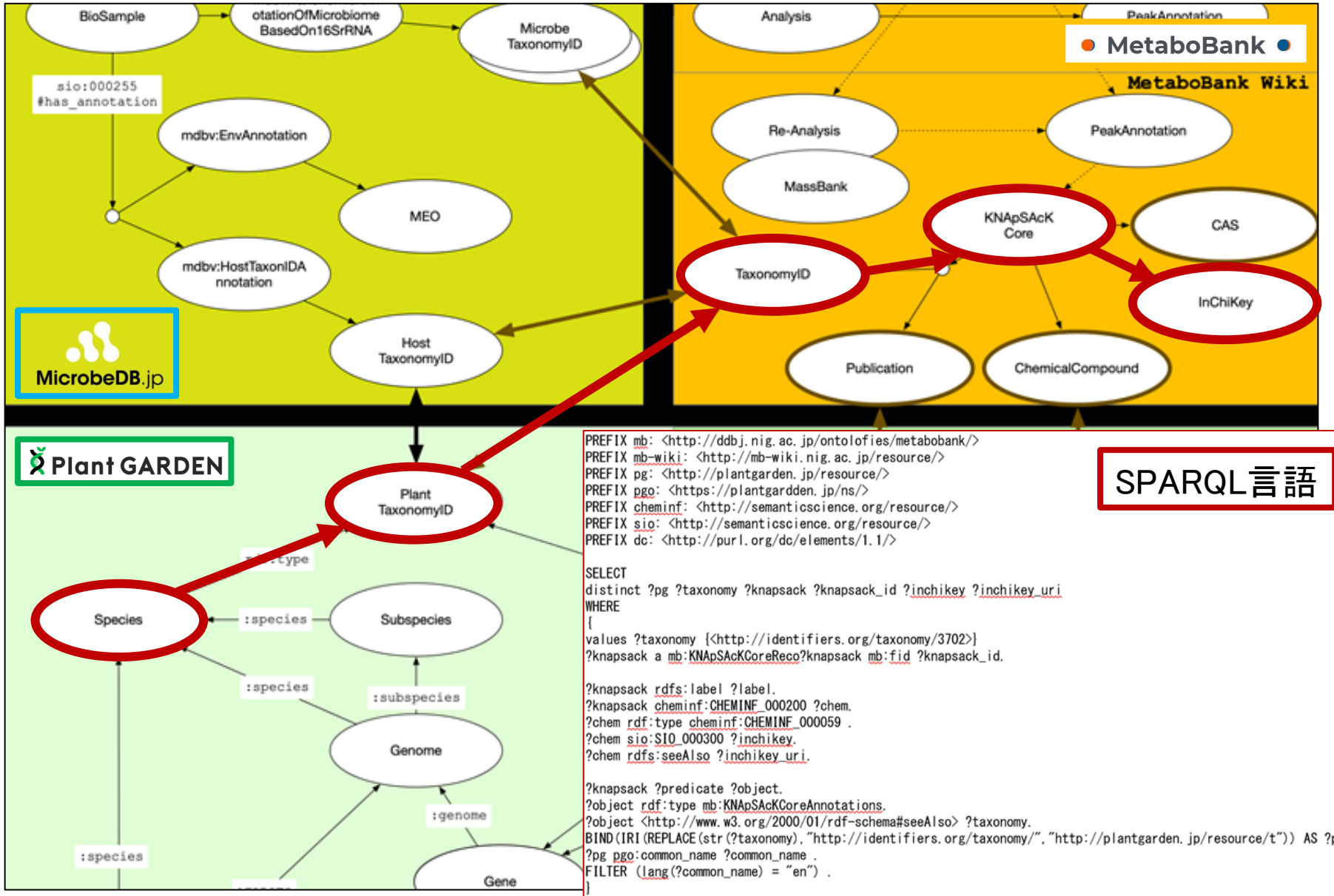
DBCLSから提供されているEndpoint browserを用いてデータが繋がっていることを確認した。



https://sparql-support.dbcls.jp/endpoint-browser.html

3者間におけるRDFの繋がり

データが繋がっていることを確認するため、植物種名を起点として、KNApSAcKの化合物情報を検索することを試みた。



TaxonomyからKNApSAcKを介したInChIKeyへの検索

シロイヌナズナ（*Arabidopsis thaliana*）のTaxonomy IDからKNApSAcKのIDを経て、InChIKeyのURIを取得し、ChemSpiderの化合物情報を取得した。

Virtuoso（SPARQL言語）

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX mb: <http://ddbj.nig.ac.jp/onto
PREFIX mb-wiki: <http://mb-wiki.nig.ac
PREFIX pg: <http://plantgarden.jp/reso
PREFIX pgo: <https://plantgardden.jp/r
PREFIX cheminf: <http://semanticscienc
PREFIX sio: <http://semanticscience.or
PREFIX dc: <http://purl.org/dc/element

SELECT
distinct ?pg ?taxonomy ?knapsack ?knap
WHERE
{
values ?taxonomy {<http://identifiers.
?knapsack a mb:KNApSAcKCoreReco?knapsa

?knapsack rdfs:label ?label.
?knapsack cheminf:CHEMINF_000200 ?chem
?chem rdf:type cheminf:CHEMINF_000059
?chem sio:SIO_000300 ?inchikey.
?chem rdfs:seeAlso ?inchikey_uri.

?knapsack ?predicate ?object.
?object rdf:type mb:KNApSAcKCoreAnnota
?object <http://www.w3.org/2000/01/rdf
BIND(IRI(REPLACE(str(?taxonomy),"http:
?pg pgo:common_name ?common_name .
FILTER (lang(?common_name) = "en") .
}
LIMIT 1000
```

pg	taxonomy	knapsack	knapsack_id	inchikey	inchikey_uri
http://plantgarden.jp/resource/t3702	http://identifiers.org/taxonomy/3702	http://mb-wiki.nig.ac.jp/resource/C00033848	"C6H1206_C00033848"	"BJHIKXHVCXFQLS-FBTDRCNSNA-N"	http://identifiers.org/inchikey/BJHIKXHVCXFQLS-FBTDRCNSNA-N
http://plantgarden.jp/resource/t3702	http://identifiers.org/taxonomy/	http://mb-wiki.nig.ac.jp			http://identifiers.org
http://plantgarden.jp/resource/t3702	http://identifiers.org/taxonomy/				
http://plantgarden.jp/resource/t3702	http://identifiers.org/taxonomy/				
http://plantgarden.jp/resource/t3702	http://identifiers.org/taxonomy/				
http://plantgarden.jp	http://identifiers.org				

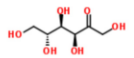
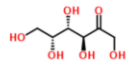
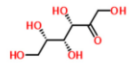
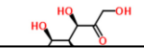
SPARQL検索の結果

ChemSpider

Found 32 results

Search term: **BJHIKXHVCXFQLS-FBTDRCNSNA-N** (Found by InChIKey (skeleton match))

1 2

ID	Structure	Molecular Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC
5764 - 3/3 defined		$C_6H_{12}O_6$	180.1559	111	347	20193	1944
83142 - 3/3 defined		$C_6H_{12}O_6$	180.1559	106	227	268	187
6638 - 3/3 defined		$C_6H_{12}O_6$	180.1559	84	208	537	658
96680		$C_6H_{12}O_6$	180.1559	77			

化合物情報の検索

統合化データベース連携検索システムの構築

SPARQL検索には言語に関する知識が必要であり、一般ユーザが容易に検索できるようにする必要性がある。

Plant GARDEN

・ゲノム情報（134種）
　ゲノム配列、CDS、PEP、アノテーション
・DNAマーカー（38種）
・QTL（27種）

PubTatorを用いた文献キュレーションのRDF

・生物種、病気、遺伝子、化合物など
・植物と微生物の相互作用
　Plant Growth Promoting Bacteria（PGPB；共生、成長促進）、病原性細菌（細菌、カビ、ウイルス）
・代謝産物（MESH ID）

連携　SPARQL

MetaboBank

・実験で得られた代謝産物に関する生データや実験に関するメタデータ
・データ数
・KNApSAcKのRDF
（DBCLSにて実施）

連携　SPARQL

MicrobeDB.jp

・系統・遺伝子・環境の3つの軸に沿って整理・統合されたフルRDFのデータベース
・菌叢メタゲノムデータ（メタ16S rRNA、遺伝子）
・ゲノム・ドラフトゲノムデータ
・単細胞の真菌類、藻類のゲノムデータ

SPARQL

外部データベースのRDF

Rhea SIB　化合物（11,886個）
酵素反応（13,673個）

Medical Subject Headings RDF NIH

NBDC RDF Portal　NBDC National Bioscience Database Center　統合化推進プログラムの他のDB

統合化データベース連携検索システムの構築

基礎研究から育種などの応用研究に至る幅広い分野での新たな知識発見を目指す

今後

3者の統合データベースのエンドポイントにトリプルデータを格納し、横断検索できることを確認した。
今後、以下の内容を実施し、データ統合化により有益な情報が得られるようになることを目指す。

RDF化
- ・有益な情報を得ることができる検索の経路、ステップを検討する
- ・スキーマの改変（必要があれば）
- ・センテンスキュレーション、病害抵抗性遺伝子との繋がり

連携（統合化データベース連携検索システムの構築）
- ・一般ユーザにとって使い易いシステムを構築する

植物、微生物、メタボローム、糖鎖やフェノタイプなど他の統合化データベース、UniProtなどRDFに対応した外部データベースに対する横断検索を可能にする。

検索の例）作物についてのDNAマーカー・形質に関わる遺伝子　→　産生される代謝産物の検索
→　菌叢において見られる微生物の検索　→　有益性または病原性の検索
→　作物の生育に適した土壌（環境）の推定に役立てる

3者間の統合データベースや外部データベースとデータを繋げることによって、様々なデータベースをボーダレスに活用することが可能となり、幅広い研究開発の現場でこれまでにない新たな知識発見が行なえるようになることを目指す。

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