ライフサイエンス統合データベースセンター：オープンサイエンス推進のためのデータ統合

五斗　進

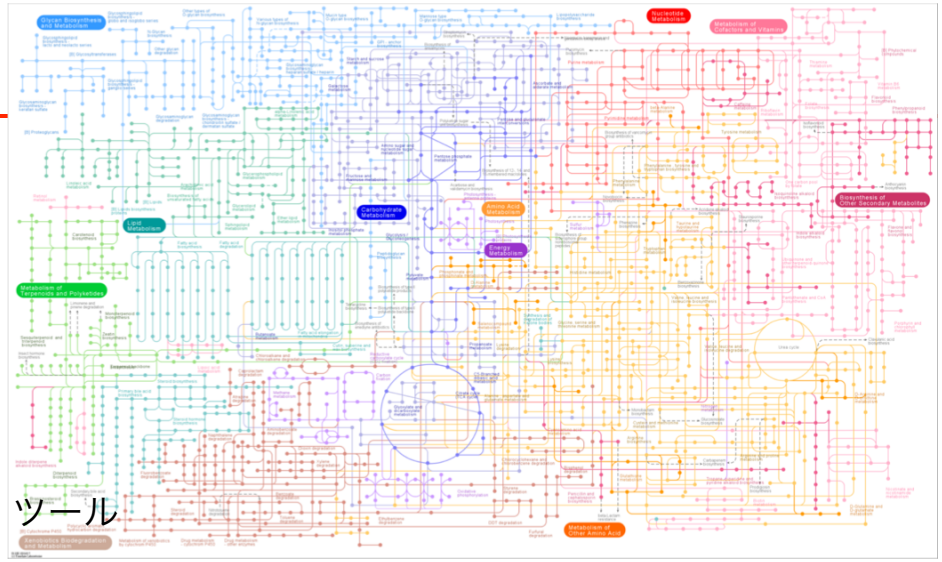
情報・システム研究機構 (ROIS)
データサイエンス共同利用基盤施設 (DS)
ライフサイエンス統合データベースセンター (DBCLS)

第５回オープンサイエンスデータ推進ワークショップ
2018年3月1日　京都大学理学研究科セミナーハウス

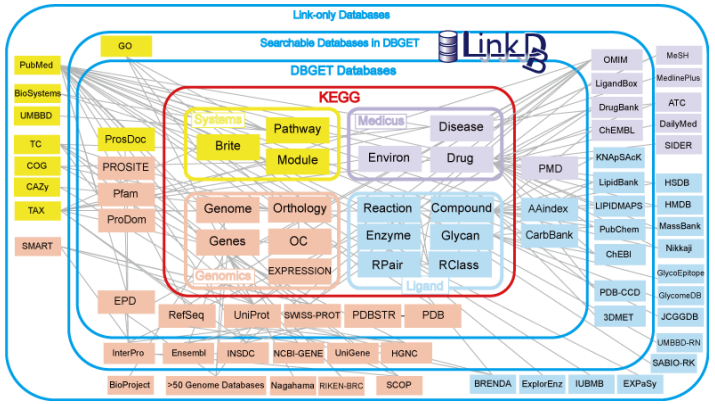
大学共同利用機関法人 情報・システム研究機構
データサイエンス共同利用基盤施設
Joint Support-Center for Data Science Research (DS)

大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

自己紹介

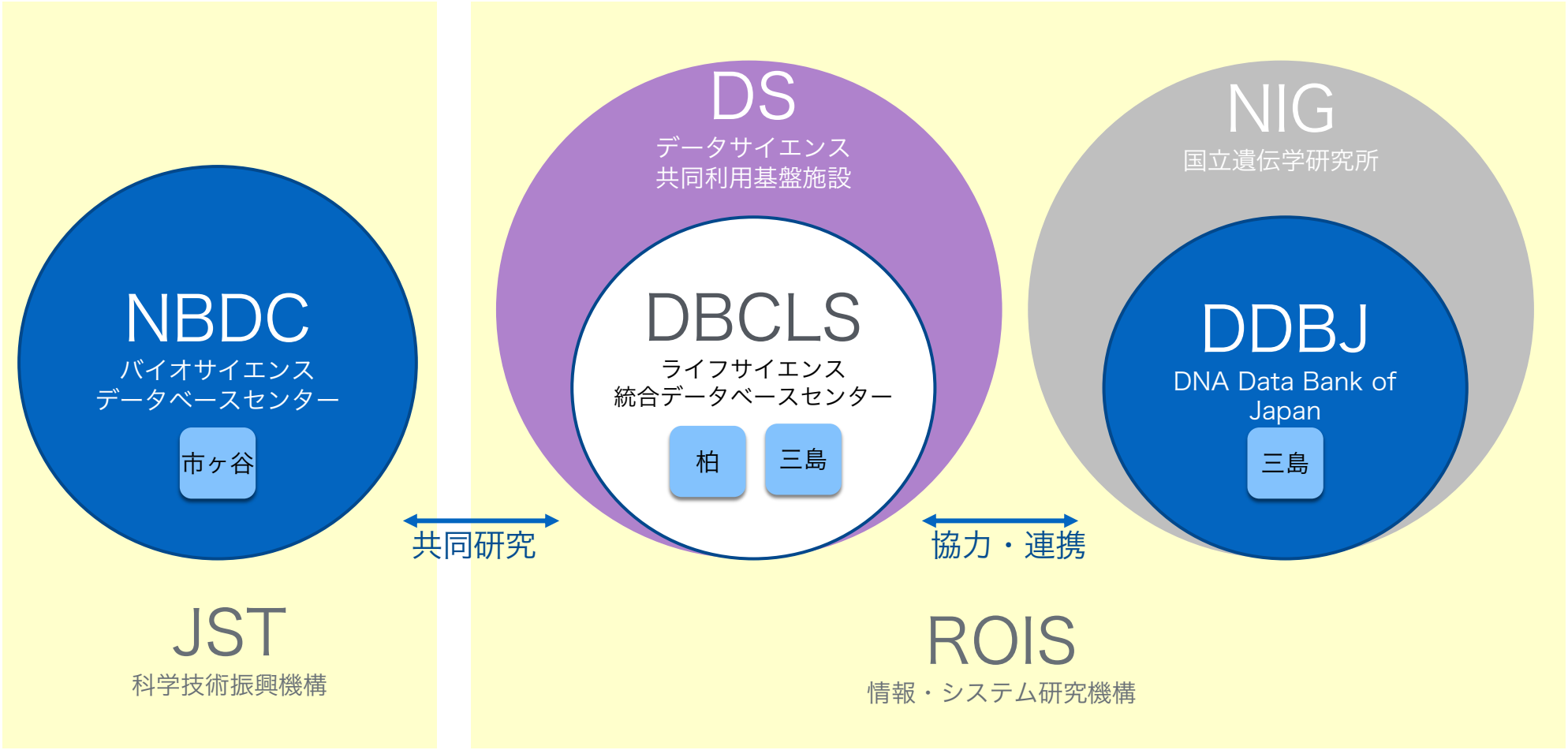


- 1994年〜2016年
  - 京都大学化学研究所
  - GenomeNet
    - バイオインフォマティクスに関するデータベース・ツール
  - KEGG: Kyoto Encyclopedia of Genes and Genomes
    - 遺伝子と化合物のネットワーク（代謝系やシグナル伝達系）
  - LinkDB
    - データベース間のリンク情報をまとめたデータベース



- 2017年〜
  - ライフサイエンス統合データベースセンター（DBCLS）
  - セマンティックウェブ（RDF）によるデータベース統合
  - 統合データベースの利活用

DBCLS と関連機関


NBDC
バイオサイエンス
データベースセンター
市ヶ谷
JST
科学技術振興機構
共同研究
DS
データサイエンス
共同利用基盤施設
DBCLS
ライフサイエンス
統合データベースセンター
柏
三島
協力・連携
NIG
国立遺伝学研究所
DDBJ
DNA Data Bank of
Japan
三島
ROIS
情報・システム研究機構

DBCLSの歴史

データベース統合化推進事業

H19

第0段階

H23

第1段階

H26

第2段階

H28

H29

データベースカタログ、横断検索、アーカイブ、
統合検索

FAIR principle

Web サービスを利用した DB 統合
計算機に適したデータアクセス

NBDC設立

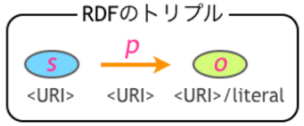
公募研究

統合検索の基盤技術開発

RDF データによる DB 統合
RDF = 共通データ形式とグローバルな ID

共同研究

- RDF: Resource Description Framework
  - 主語 (*Subject*) - 述語 (*Predicate*) - 目的語 (*Object*) からなるデータモデル
    - **主語** - モノの ID (URI)
    - **述語** - オントロジーで定義された属性 (URI)
    - **目的語** - 別のモノのID(URI) または 値 (literal)

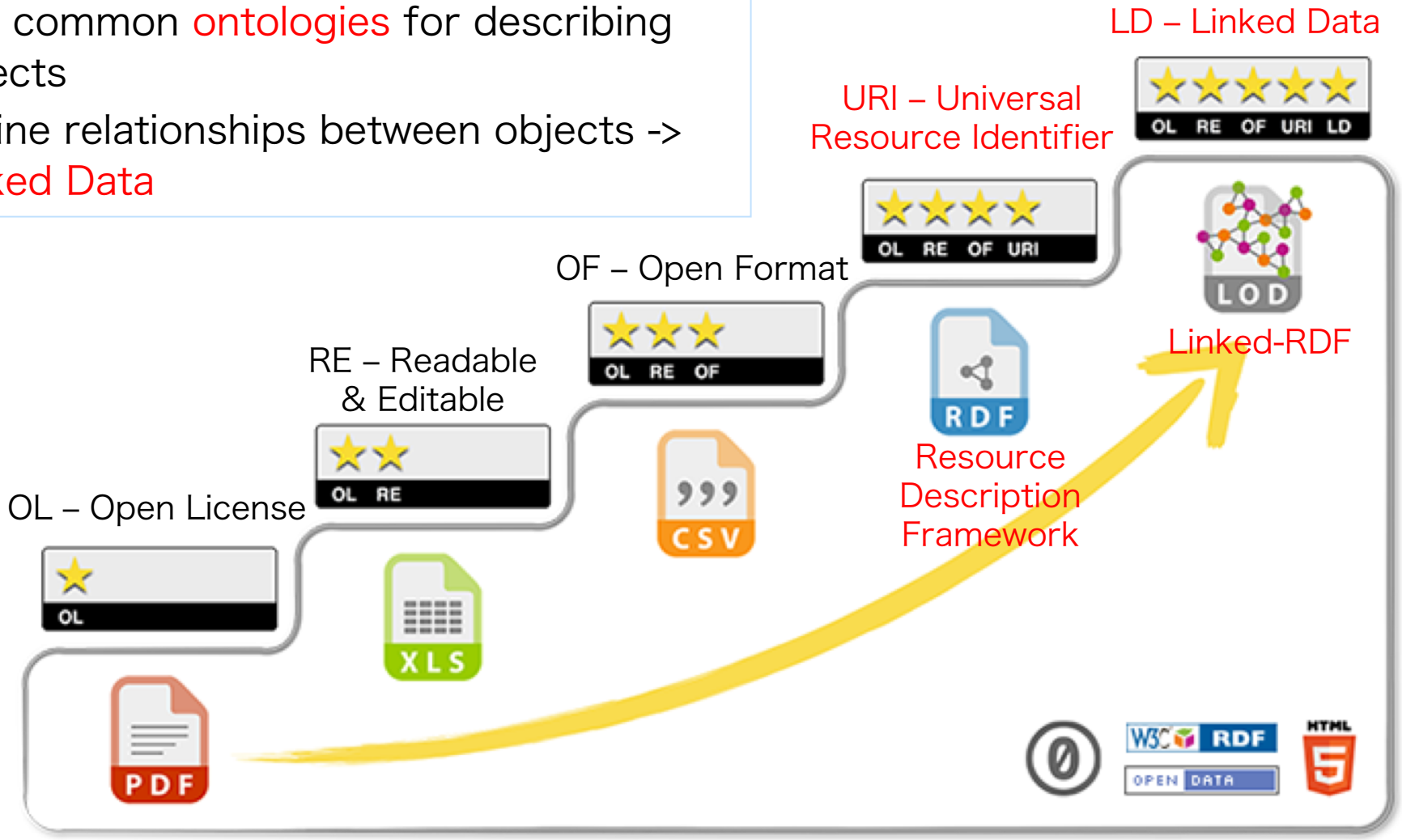


セマンティクスを重視した DB 統合
RDF + オントロジー = 実用的な分散知識ベース

5 ★ Linked Open Data

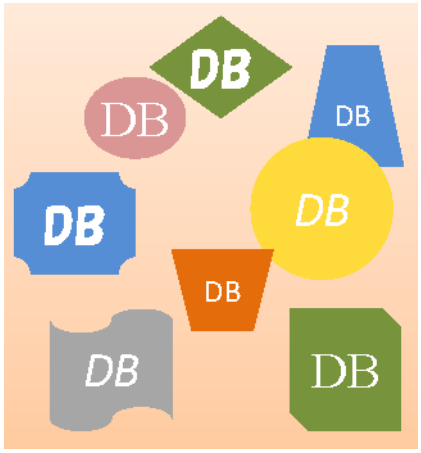

Tim Berners-Lee

- To give a unique ID to every object -> URI
- Use common ontologies for describing objects
- Define relationships between objects -> Linked Data



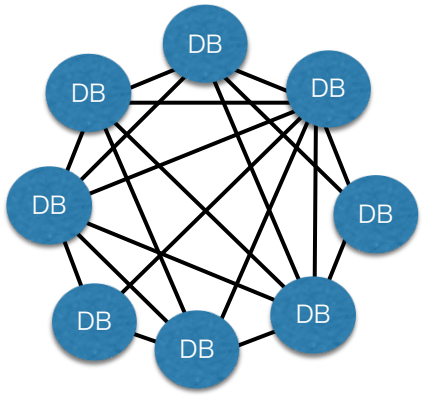
http://5stardata.info/en/

Database Integration @ DBCLS



Ontology

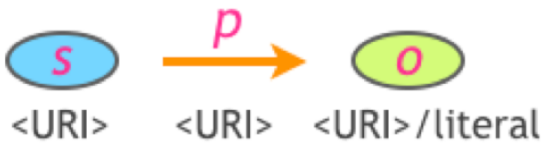
RDF

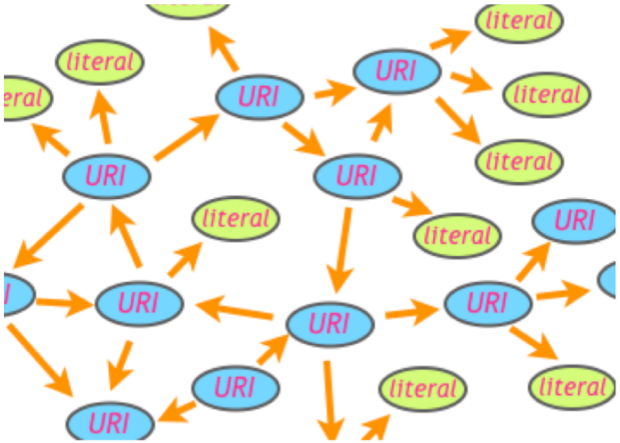


Highly heterogeneous databases using their own terms and formats

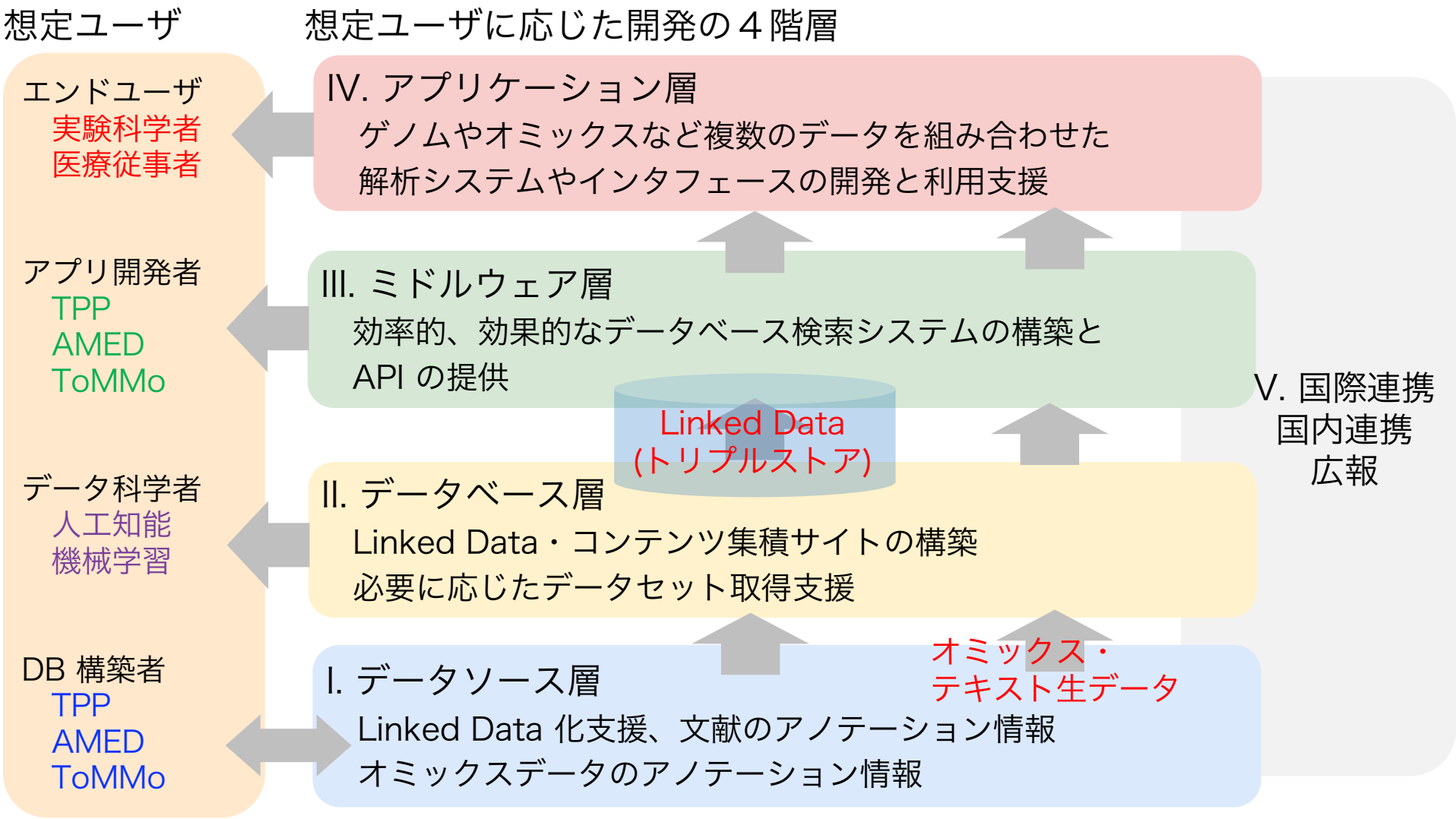
Databases integration for seamless access and knowledge mining

- RDF: Resource Description Framework
- Triples consisting of Subject, Predicate and Object
  - Subject: ID (URI) for an object
  - Predicate: Attribute (URI) defined by an ontology
  - Object: ID (URI) or value (literal) for another object





DBCLS データコア構想



NBDC RDF Portal

- SPARQLthon や BioHackathon を通じて、統合化推進プログラムのRDF化をサポート。そこでの知見を基にRDF化のガイドラインを作成。
  - http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline
- ガイドラインに沿って作られたRDFデータを一覧できるポータルサイトをNBDCと構築。RDFデータの利用促進および国際的認知度の向上。
- 20データセット（JST統合化推進プログラム9データセット）、約450億トリプルが登録されている（平成29年11月現在）。
- 微生物ゲノム、転写開始点、植物オーソログ、タンパク質立体構造、糖鎖構造、菌株・細胞情報、遺伝子発現など。

http://integbio.jp/rdf/

ライフサイエンス分野の RDF データ

タイプ	RDFデータセット	タイプ	RDFデータセット
遺伝子	DDBJ	遺伝子オーソログ	MBGD, PGDBj Orthology
ゲノム	Ensembl	タンパク質相互作用	IntAct, *Instruct, HINT*
メタゲノム	MicrobeDB.jp	パスウェイ	REACTOME, WikiPathway
エピゲノム	KERO, *ChIP-Atlas, iMETHYL*	システムバイオロジー	BioModels, SSBD
ゲノム変異	Linked ICGC, *ClinVar, ExAC*	バイオアッセイ	ChEMBL, PubChem
タンパク質	UniProt	病気	PAConto, GGDonto, DisGeNet, *ClinVar, MedGen*
タンパク質立体構造	wwPDB, BMRB, FAMSBASE	用語集	MeSH, Allie, LSD
糖鎖	GlyTouCan, GlycoEpitope, WURCS	トランスクリプトーム	ExpressionAtlas, RefEx, KERO, Open TG-GATEs
化学化合物	PubChem, Nikkaji	プロテオーム	neXtProt, The Human Protein Atlas, *jPOSTdb*
メタ情報	Quanto, integbio DB catalog, Colil, 新着論文レビュー	メタボローム	*MassBank, metabolonote*
サンプル	BioSamples, JCM	オントロジー	BioProtal, OLS

赤：NBDC RDF portal または DBCLS で作成　　*斜体：進行中*

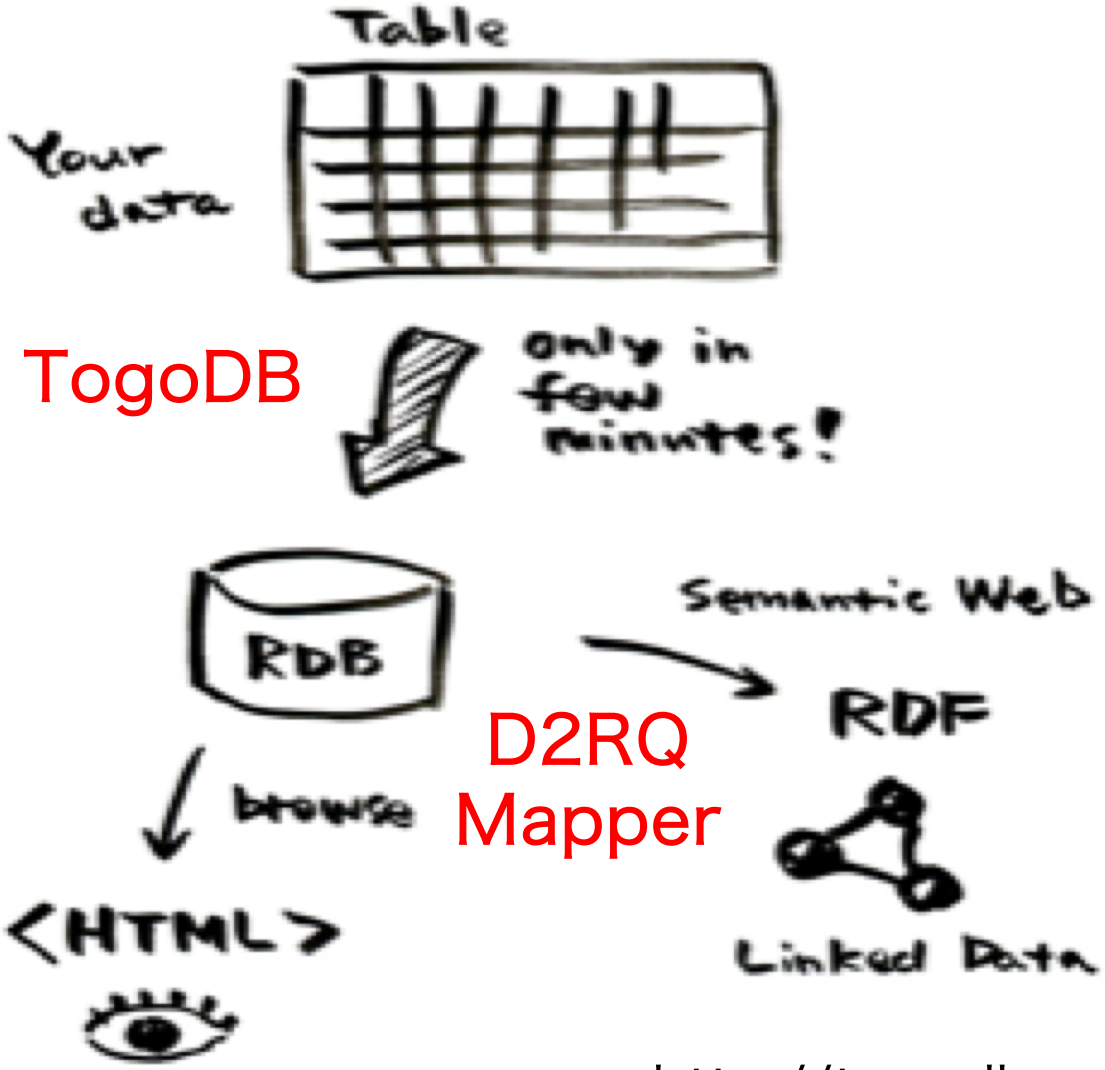
SPARQLthon

- 平成24年10月より、セマンティック・ウェブ技術による生命科学DBの統合をテーマに、毎月DBCLS主催で開催しているハッカソン。
- 平成30年2月までの開催回数66回、参加延べ人数約1400人、ユニークな参加者数140名超、所属機関数45（15大学、13研究機関、17企業）。
- 平成26年度からは、統合化推進プログラムから多くの参加者があり、データのRDF化に関する意識共有・技術情報共有・コラボレーションが著しく進んだ。また、ここで多くのRDFデータやオントロジーが作られてきた。

Biohackathon

- 最先端のデータ統合技術やその応用についての世界の技術動向を調査・把握しテーマの決定。先端技術を用いてシステムやプログラム開発を行っている国内外の現場の研究者が参集。
- 合宿形式で分野横断的に問題解決を行い、密度の高い情報交換と生産性の高い技術開発が会議の場でリアルタイムに進行。
- DBCLS はこの分野で世界をリードするハブとして高く評価され、継続的にリーダーシップを発揮。
- DBCLSの開発する統合データベースの基盤技術として活用。
- 研究者コミュニティの共有資産となり論文発表も行われている。

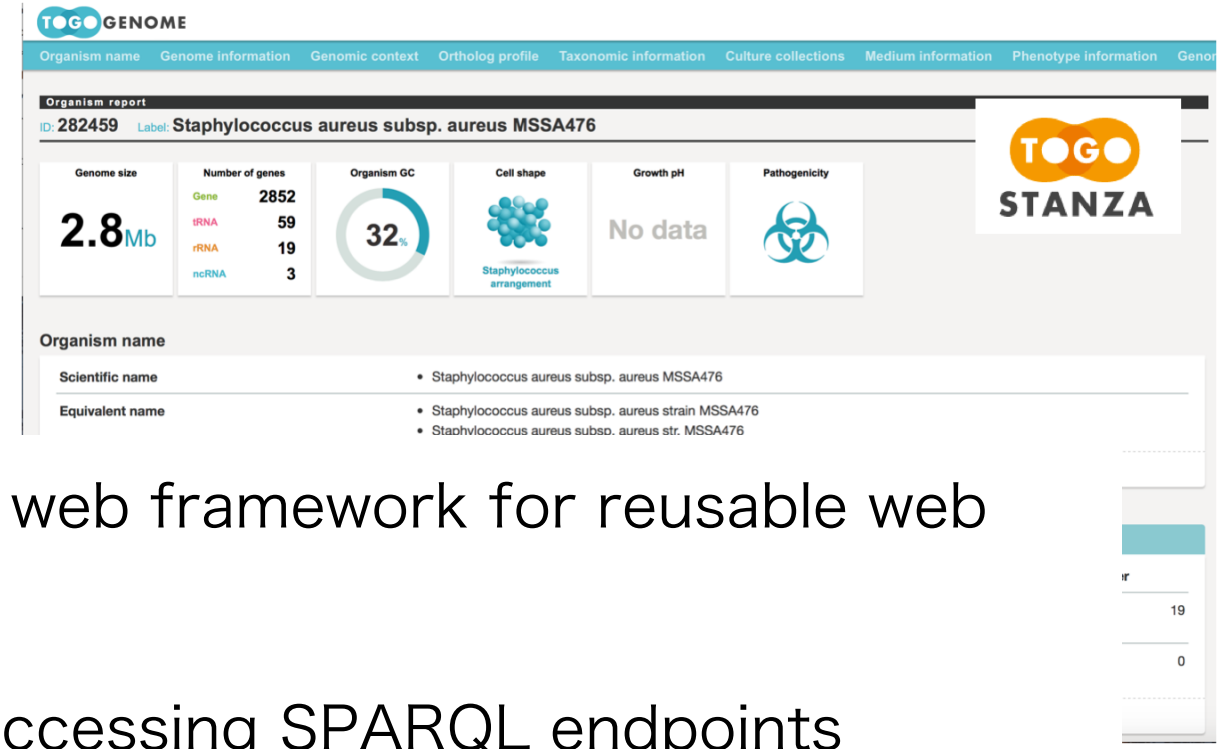
表データや RDB の RDF 化支援ツール開発



http://togodb.org/

ミドルウェアの開発

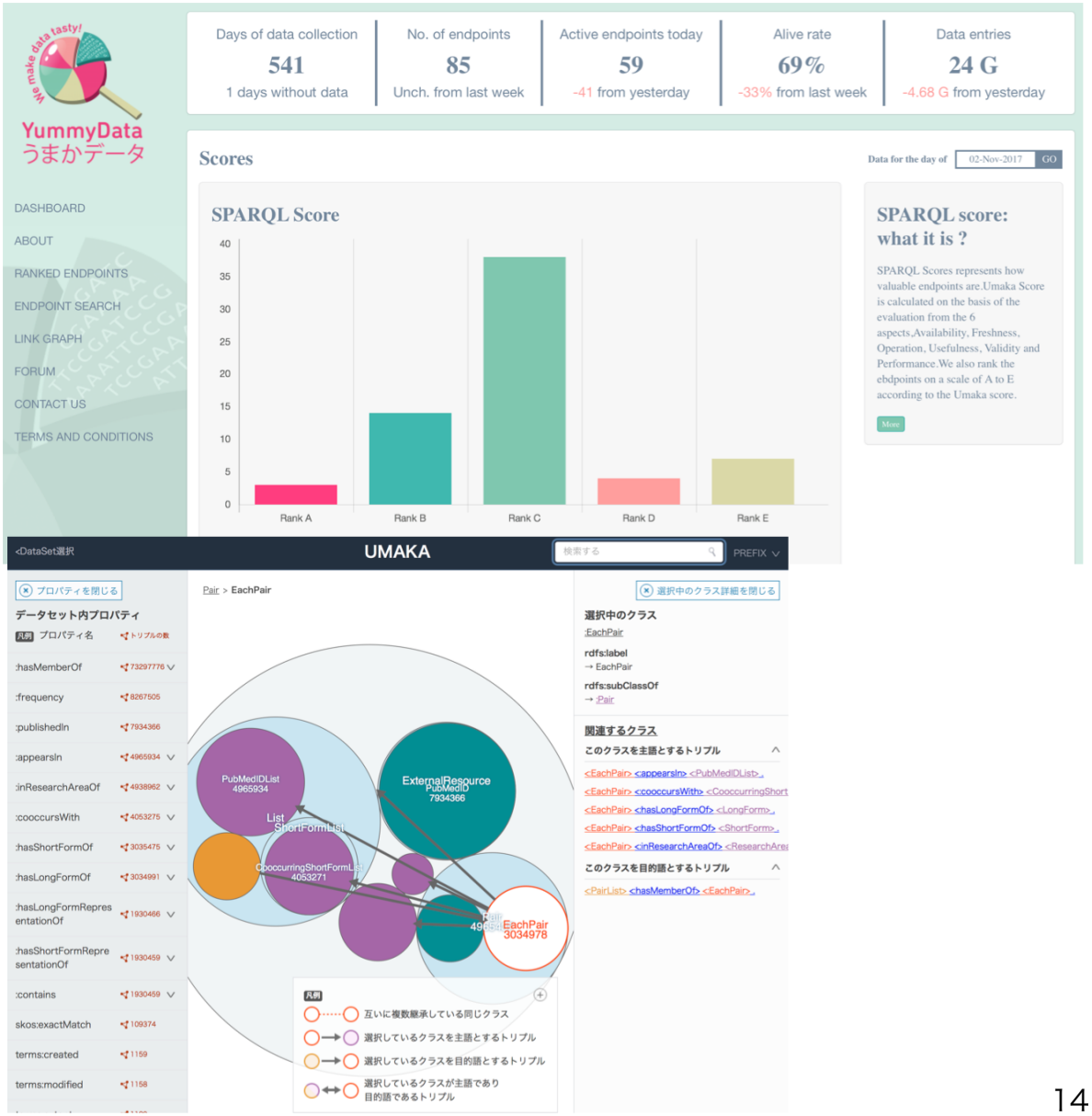
SPARQLエンドポイントへ効率よくアクセスするためのツール群



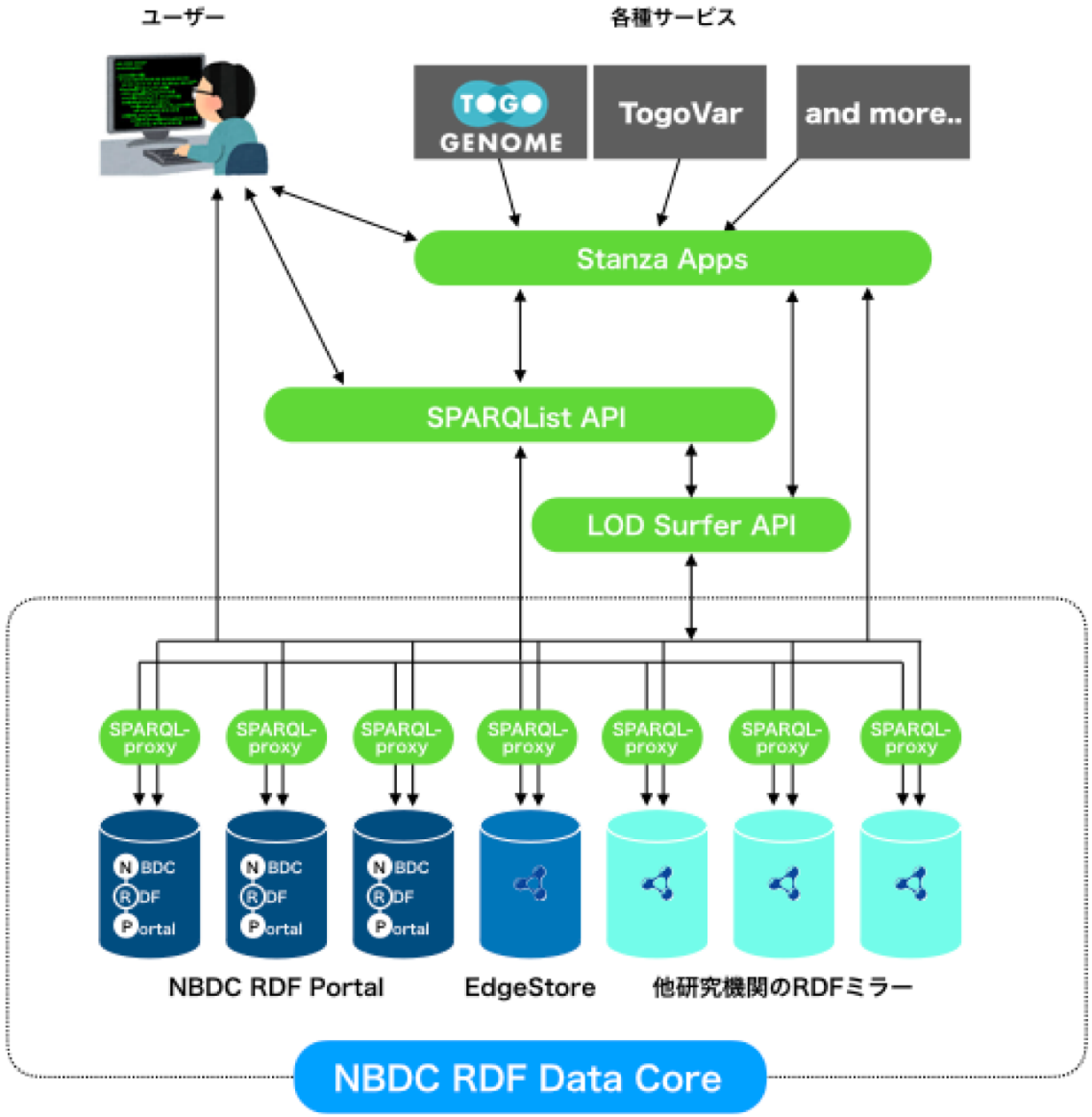
- **TogoStanza**: generic web framework for reusable web components
- **SPARQList**: API for accessing SPARQL endpoints
- **SPARQL support, SPARQL builder**: web interface to support building SPARQL queries
- **UmakaData**: listing and monitoring SPARQL endpoints

SPARQLエンドポイントの情報取得ツール

- UmakaData による適切なエンドポイントの選択
- UmakaViewer によるデータ構造の可視化



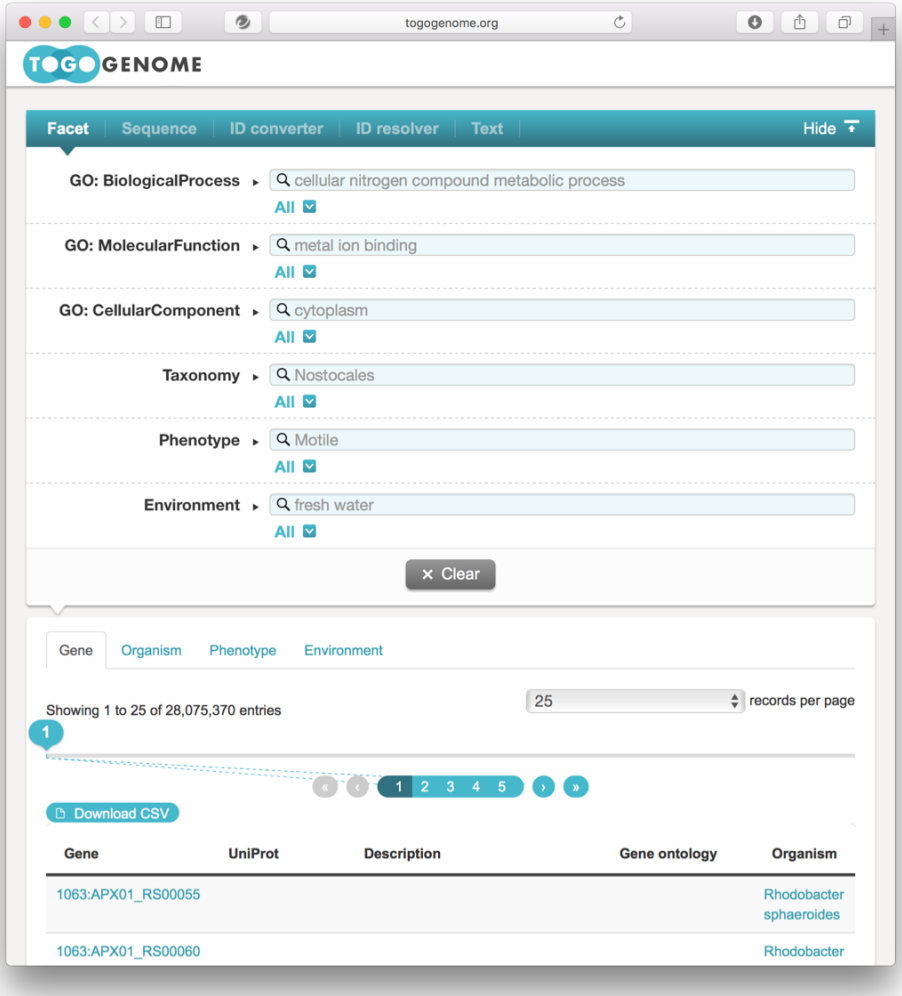
NBDC RDF データコアと関連ツール



Application: TogoGenome

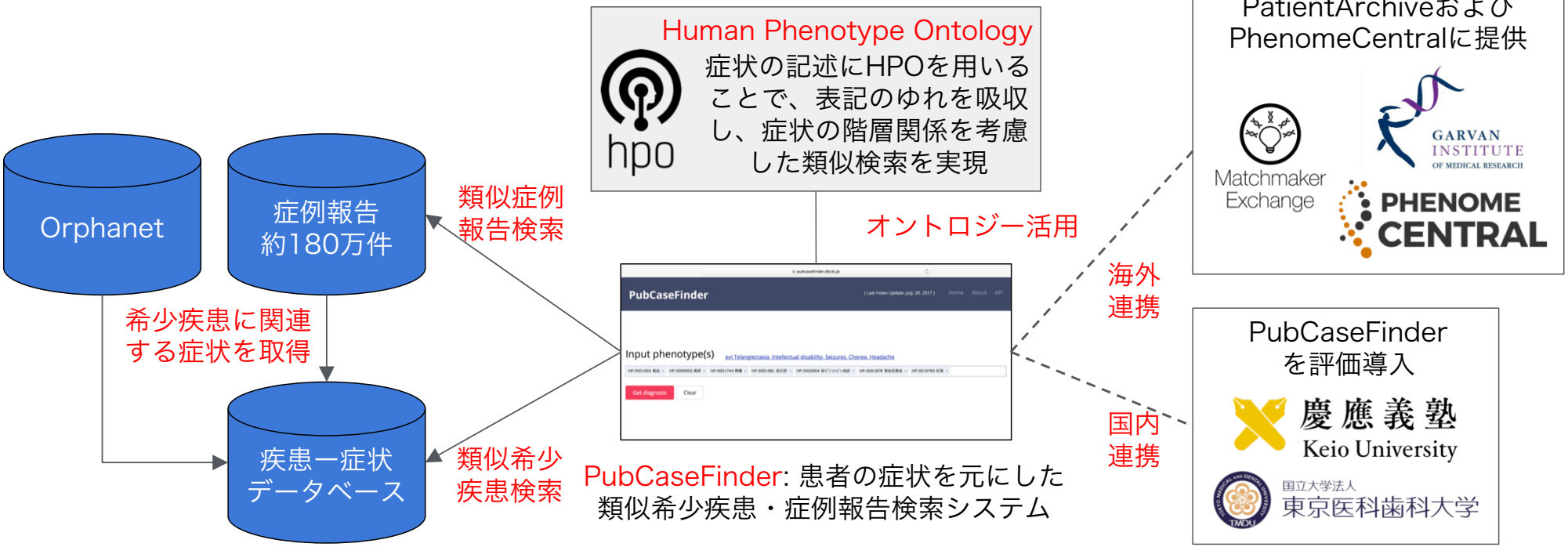
- セマンティック・ウェブ技術を利用したゲノムデータベース。RDFデータストアのみで実装されている点は、世界的にもユニーク。
- 昨年度から、原核生物種に加えて真核生物種も含めた生物種をカバー。
- 生物種数は真核生物360種を含む、計10,000以上。RDFデータ数では、10億以上。

→ ヒト変異・疾患情報



希少疾患診断支援システム PubCaseFinder

- 希少疾患診断支援を目的とした、症例報告からの希少疾患関連症状の収集、および患者の症状を元にした類似希少疾患・症例報告検索システム
  - 患者の症状と関連性の高い希少疾患を素早く検索
  - 症例報告を効率よく検索
  - 希少疾患と症状の関係をわかりやすく表示



https://pubcasefinder.dbcls.jp/

オミックスデータへのアクセスツール

1. Exhaustive, but functional index for public raw data repository

DBCLS SRA

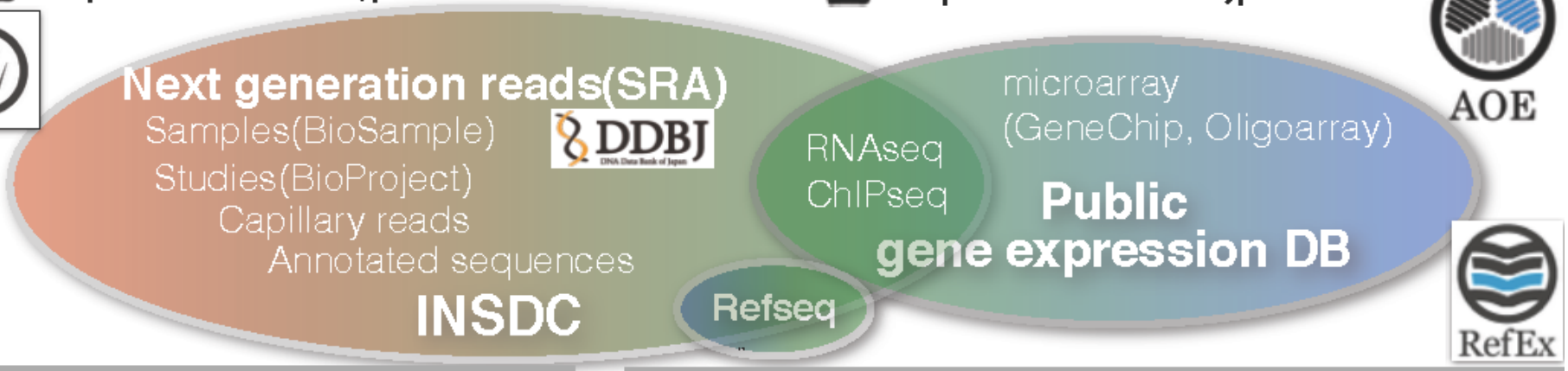
Yellow pages for Sequence Read Archive(SRA)

☑ http://SRA.dbcls.jp/

AOE(All Of gene Expression)

Graph shortcut for gene expression data

☑ http://AOE.dbcls.jp/



AOE

RefEx

4. Sequence analysis tools for nucleotides

☑ http://ggrna.dbcls.jp/

☑ http://gggenome.dbcls.jp/

統合遺伝子検索 GGRNA

Ultrafast sequence search GGGenome

2. Curated dataset for functional analysis

→ Reference transcriptome data

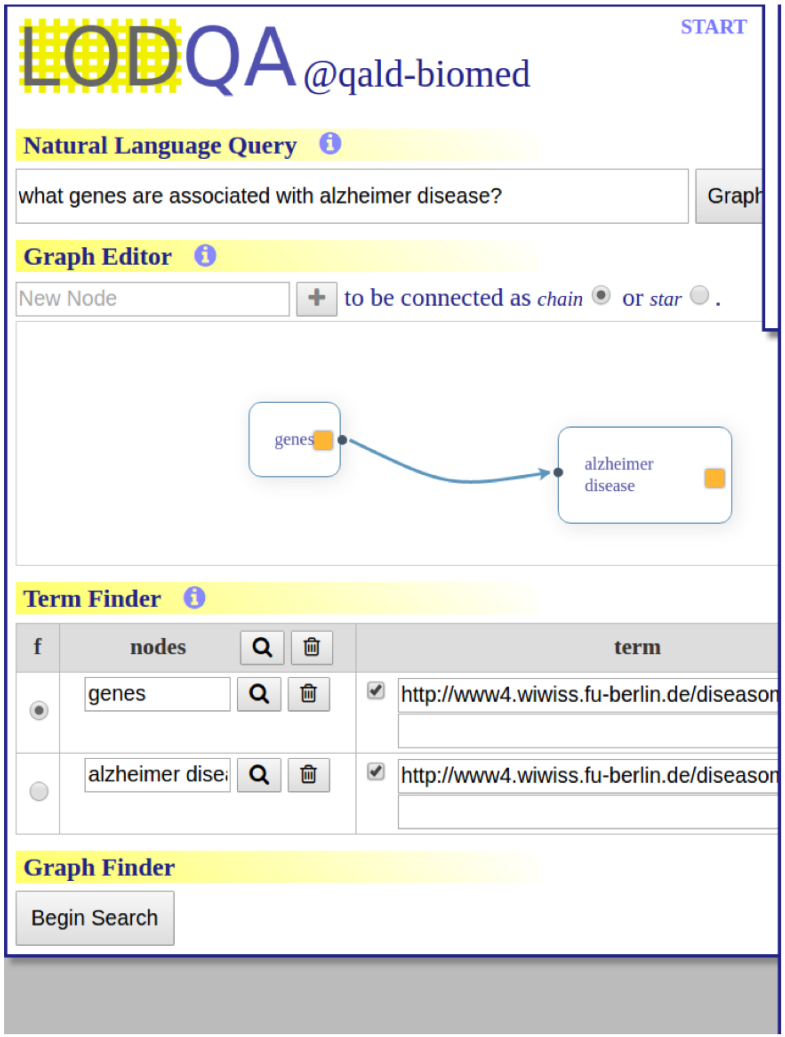
☑ http://RefEx.dbcls.jp/

→ Curation and visualization of public ChIP-seq data

☑ http://chip-atlas.org/

KYUSHU UNIVERSITY

自然言語Q&Aインタフェース

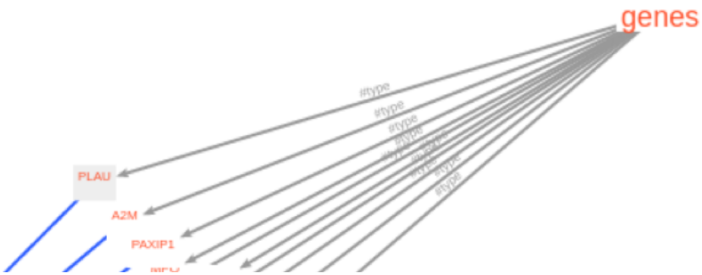


sparql	answer
SELECT ?it1 ?st1 ?p01 WHERE {?it1 ?st1 <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/genes> . ?it1 ?p01 <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseases/74> . FILTER (isIRI(?it1)) FILTER (str(?p01) NOT IN ("http://www.w3.org/1999/02/22-rdf-syntax-ns#type", "http://www.w3.org/2000/01/rdf-schema#subClassOf")) FILTER (str(?st1) IN ("http://www.w3.org/1999/02/22-rdf-syntax-ns#type", "http://www.w3.org/2000/01/rdf-schema#subClassOf"))} LIMIT 10	

sparql	answer
SELECT ?it1 ?st1 ?p01 WHERE {?it1 ?st1 <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/genes> . <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseases/74> ?p01 ?it1 . FILTER (isIRI(?it1)) FILTER (str(?p01) NOT IN ("http://www.w3.org/1999/02/22-rdf-syntax-ns#type", "http://www.w3.org/2000/01/rdf-schema#subClassOf")) FILTER (str(?st1) IN ("http://www.w3.org/1999/02/22-rdf-syntax-ns#type", "http://www.w3.org/2000/01/rdf-schema#subClassOf"))} LIMIT 10	A2M ACE APBB2 APOE APP BLMH MPO NOS3 PAXIP1 PLAU

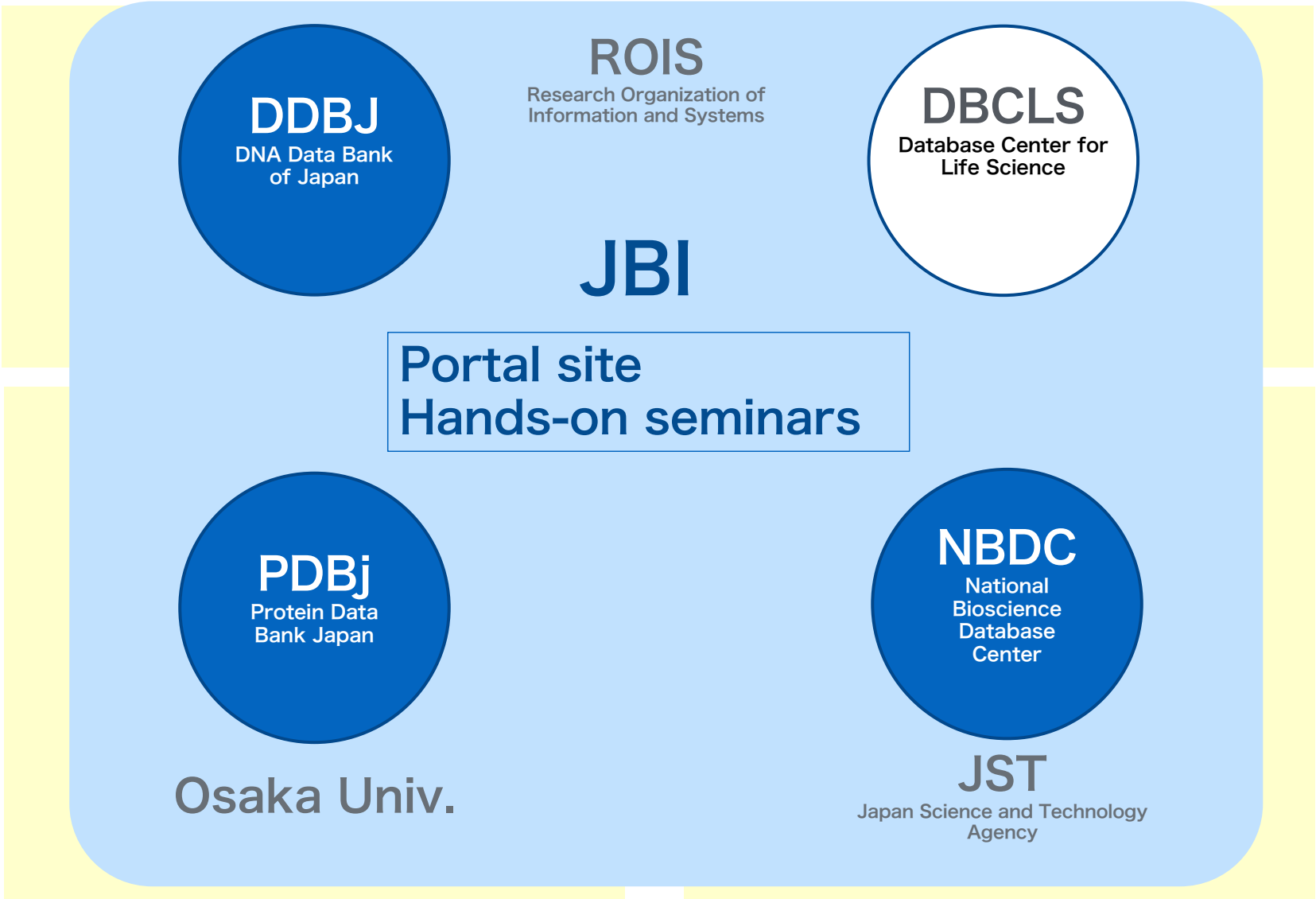
Show solutions in table

http://www4.wiwiss.fu-berlin.de/diseasome/resource/genes/APOE



http://lodqa.org

Japan alliance for Bioscience Information (JBI)



まとめ

- セマンティック・ウェブ技術を用いたデータベース統合
  - RDF, Linked Open Data
  - RDF Portal と変換ツール
- 統合データベースを利用するためのツール群
  - http://dbcls.jp/services
- コミュニティによる開発と利用促進
  - Biohackathon
  - SPARQLthon
  - 講習会, TogoTV：ツールの利用法など

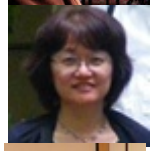
Acknowledgements

Director
KOHARA, Yuji

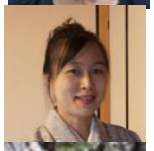
KAWANO, Shin

KIM, Jin-Dong

BONO, Hidemasa



MINOWA, Mari



YAMAGUCHI, Atsuko

YAMAMOTO, Yasunori

ONO, Hiromasa

KATAYAM, Toshiaki

KAWASHIMA, Shuichi

CHIBA, Hirokazu

NAITO, Yuki

NAKAZATO, Takeru

MORIYA, Yuki

IIDA, Keisuke

OHTA, Tazro

FUJIWARA, Toyofumi

WANG, Yue

OKUBO, Kousaku

KAWAMOTO, Shoko

OKAMOTO, Shinobu

National Bioscience Database Center

Japan Science and Technology Agency