Life Science Database Integration Using Linked Data

Susumu Goto

Database Center for Life Science (DBCLS)
Joint Support-Center for Data Science Research (DS)
Research Organization of Information and Systems (ROIS)

International Life Science Integration Workshop
2018 / 3 / 6 @ Nakano Sunplaza, Tokyo, Japan

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)

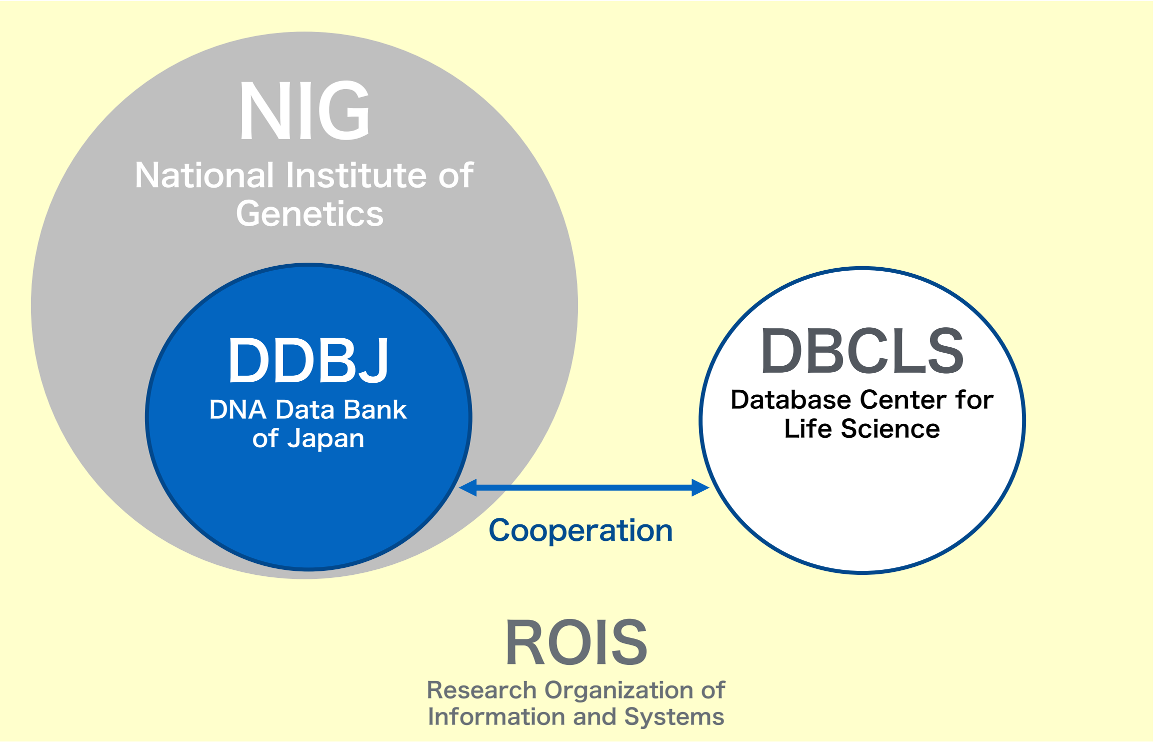
大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

Database Center for Life Science

- 2008-
  - Database integration based on web application
- 2011-
  - Funded by JST National Bioscience Database Center for the database integration with the FAIR principle

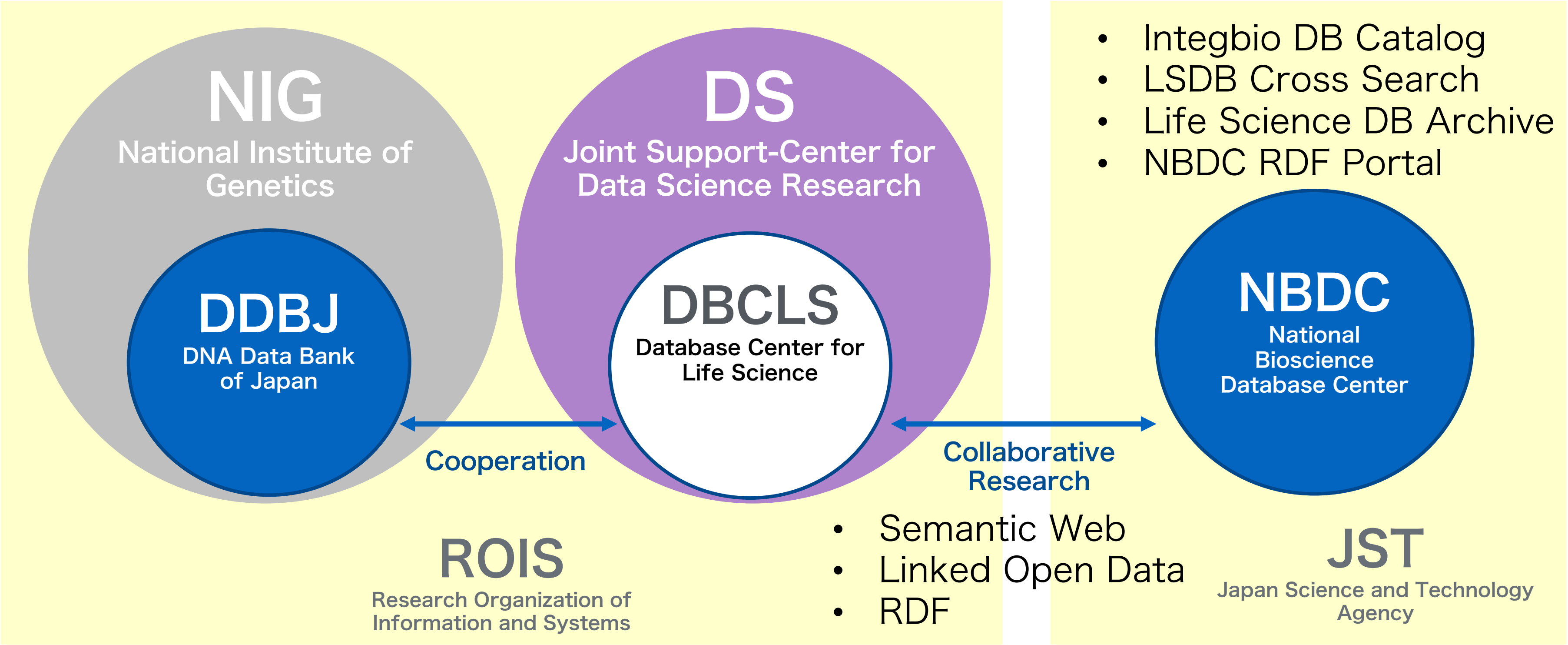
- Integbio DB Catalog
- LSDB Cross Search
- Life Science DB Archive
- Technology development

FAIR
- Findable
- Accessible
- Interoperable
- Reusable


NIG
National Institute of Genetics
DDBJ
DNA Data Bank of Japan
Cooperation
DBCLS
Database Center for Life Science
ROIS
Research Organization of Information and Systems

Database Center for Life Science

- 2008-
  - Database integration based on web application
- 2011-
  - Funded by JST National Bioscience Database Center for the database integration with the FAIR principle



NIG
National Institute of Genetics
DDBJ
DNA Data Bank of Japan
DS
Joint Support-Center for Data Science Research
DBCLS
Database Center for Life Science
Cooperation
Collaborative Research
NBDC
National Bioscience Database Center
ROIS
Research Organization of Information and Systems
JST
Japan Science and Technology Agency

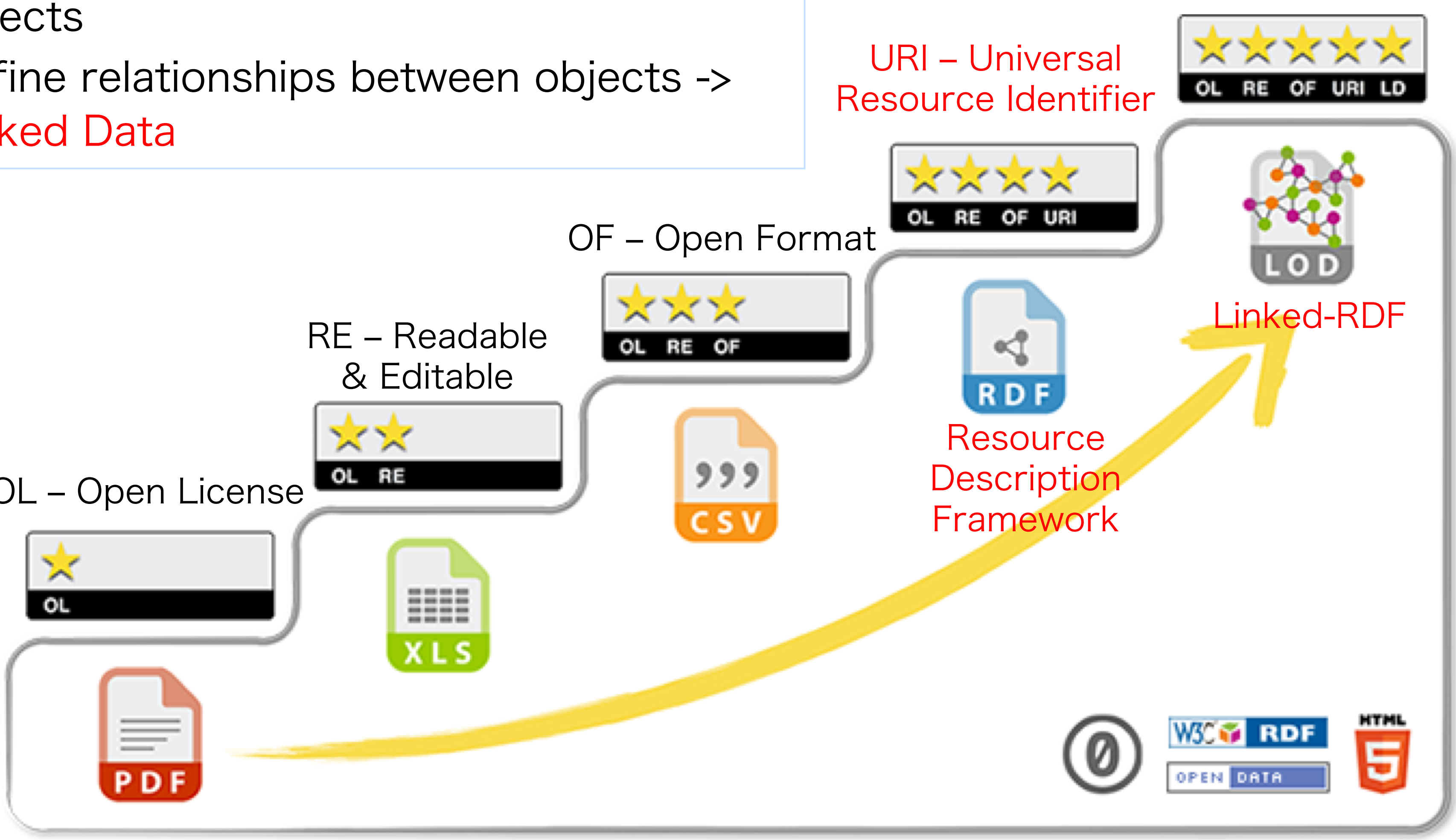
- Integbio DB Catalog
- LSDB Cross Search
- Life Science DB Archive
- NBDC RDF Portal

- Semantic Web
- Linked Open Data
- RDF

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



OL – Open License

RE – Readable & Editable

OF – Open Format

URI – Universal Resource Identifier

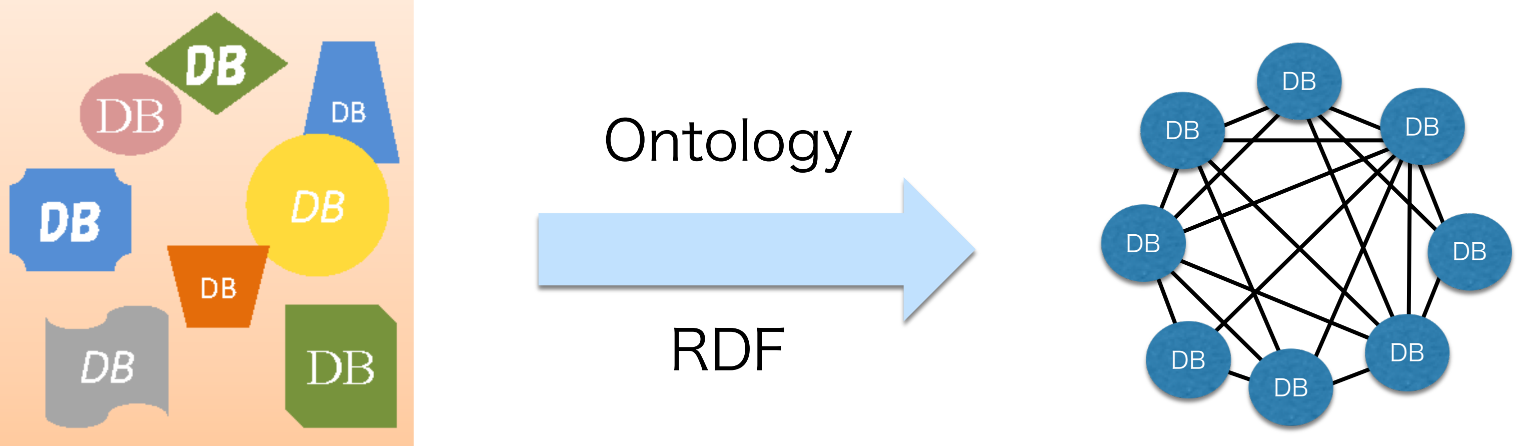
LD – Linked Data

Linked-RDF

Resource Description Framework

http://5stardata.info/en/

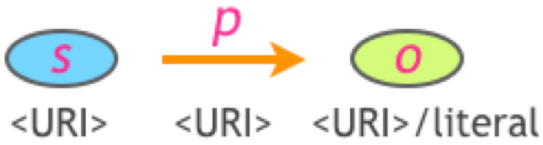
Database Integration @ DBCLS

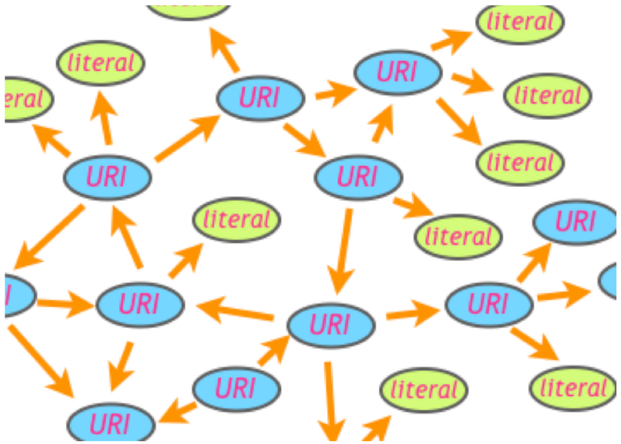


Highly heterogeneous databases
using their own terms and formats

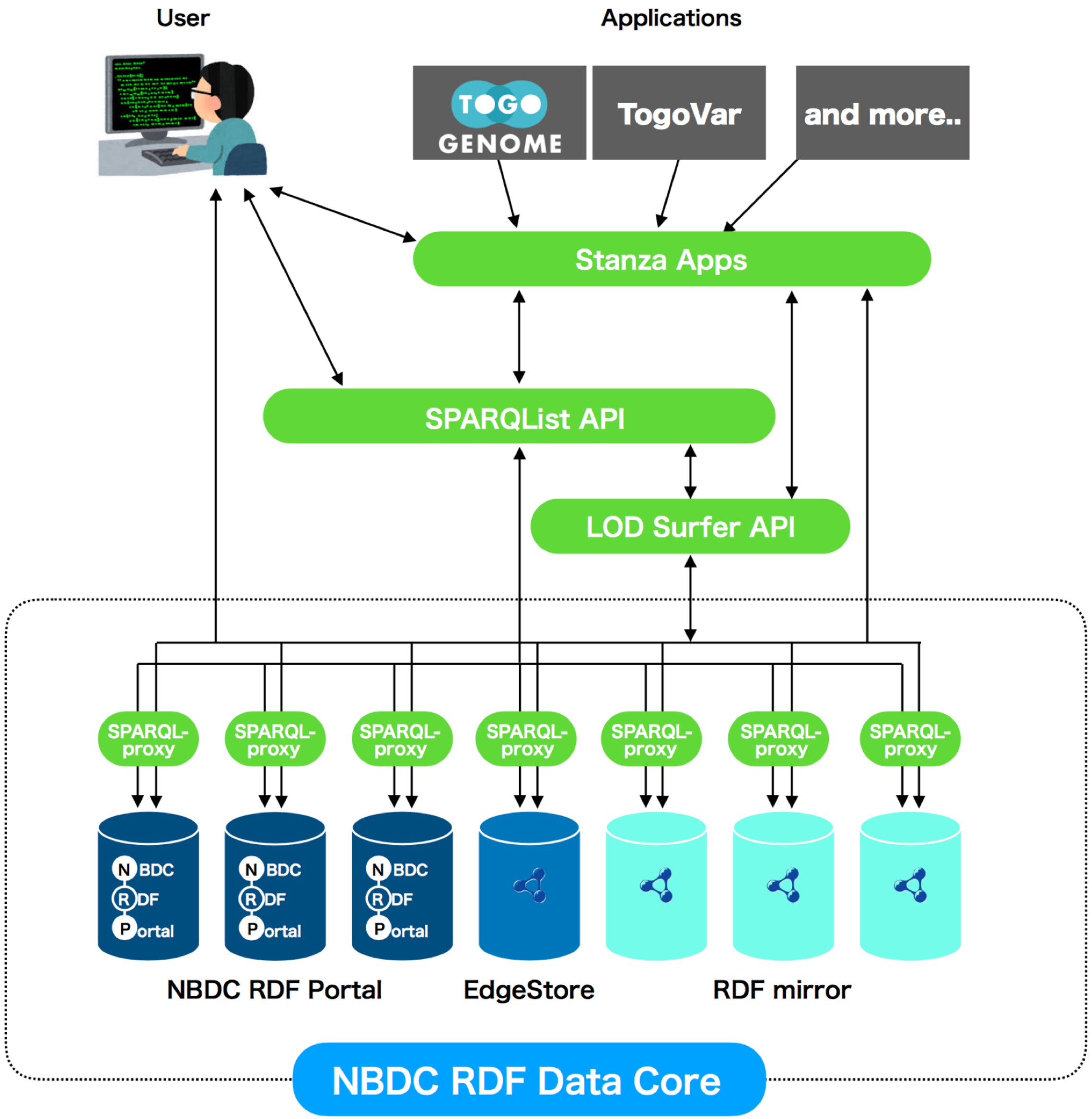
Databases integration for seamless
access and knowledge mining

- 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





Database Integration @ DBCLS



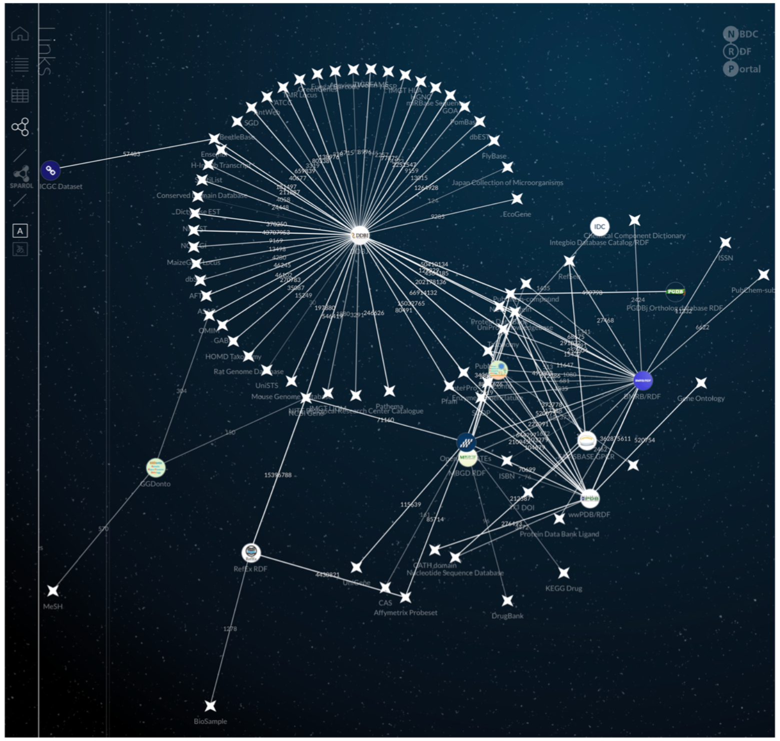
Application Development

Middleware Development

Database Development

NBDC RDF Portal

- Portal site for RDF data from research groups in Japan
- 20 data sets including nine from NBDC funded databases comprising 45 billion triples (as of Nov. 2017)
- Microbial genomes, protein 3D structures, glycan structures, …
- RDF file download, SPARQL endpoints, Statistics, Metadata, …



http://integbio.jp/rdf/
Network of Databases

Two important topics

- RDFyzing database guideline
  - http://wiki.lifesciencedb.jp/mw/BH14.14/RDFizingDatabaseGuidelineEnglishDraft0.1
- BioHackathons and SPARQLthons

SPARQLthon

- Two days hackathon held every month from 2012 October.
- Theme: Life science database integration by semantic web technologies.
- >60 times in total and 1,328 (138 unique) participants from 45 institutes (15 universities, 13 research institutes, 17 private companies).
- From 2014, researchers from integrated database project funded by NBDC have attended and collaborated for creating RDF data and ontologies.

Biohackathon

- International hackathon hosted by DBCLS/NBDC once a year in Japan from 2008
- Discuss and develop up-to-date technologies and systems for database integration and its applications
- One week intense development by international collaboration
- Summary papers have been published
- FAIR principle paper acknowledges biohackathon

Currently Available RDF Data

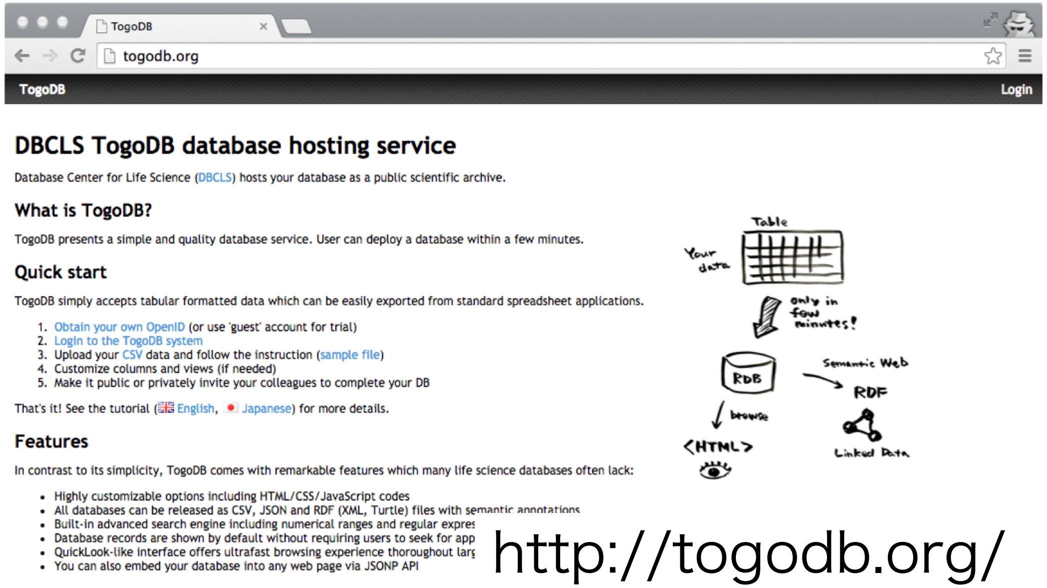
Type	RDF Data Set	*Type*	RDF Data Set
Gene	DDBJ	*Ortholog*	MBGD, PGDBj Orthology
Genome	Ensembl	*Protein interaction*	IntAct, Instruct, HINT
Metagenome	MicrobeDB.jp	*Pathway*	REACTOME, WikiPathway
Epigenome	KERO, ChIP-Atlas, iMETHYL	*Systems biology*	BioModels, SSBD
Genome variation	Linked ICGC, ClinVar, ExAC	*Bioassay*	ChEMBL, PubChem
Protein	UniProt	*Disease*	PAConto, GGDonto, DisGeNet, ClinVar, MedGen
Protein structure	wwPDB, BMRB, FAMSBASE	*Dictionary*	MeSH, Allie, LSD
Glycan	GlyTouCan, GlycoEpitope, WURCS	*Transcriptome*	ExpressionAtlas, RefEx, KERO, Open TG-GATEs
Chemical compound	PubChem, Nikkaji	*Proteome*	neXtProt, The Human Protein Atlas, jPOSTdb
Meta data	Quanto, integbio DB catalog, Colil, First Authors	*Metabolome*	MassBank, metabolonote
Sample	BioSamples, JCM	*Ontology*	BioProtal, OLS

Red: RDF Portal、Blue: On-going

Tools for RDFyzing Data

TogoDB

Converting table data to RDB / RDF



DBCLS TogoDB database hosting service

Database Center for Life Science (DBCLS) hosts your database as a public scientific archive.

What is TogoDB?

TogoDB presents a simple and quality database service. User can deploy a database within a few minutes.

Quick start

TogoDB simply accepts tabular formatted data which can be easily exported from standard spreadsheet applications.

1. Obtain your own OpenID (or use 'guest' account for trial)
2. Login to the TogoDB system
3. Upload your CSV data and follow the instruction (sample file)
4. Customize columns and views (if needed)
5. Make it public or privately invite your colleagues to complete your DB

That's it! See the tutorial (English, Japanese) for more details.

Features

In contrast to its simplicity, TogoDB comes with remarkable features which many life science databases often lack:

- Highly customizable options including HTML/CSS/JavaScript codes
- All databases can be released as CSV, JSON and RDF (XML, Turtle) files with semantic annotations
- Built-in advanced search engine including numerical ranges and regular expres
- Database records are shown by default without requiring users to seek for app
- QuickLook-like interface offers ultrafast browsing experience thoroughout larg
- You can also embed your database into any web page via JSONP API

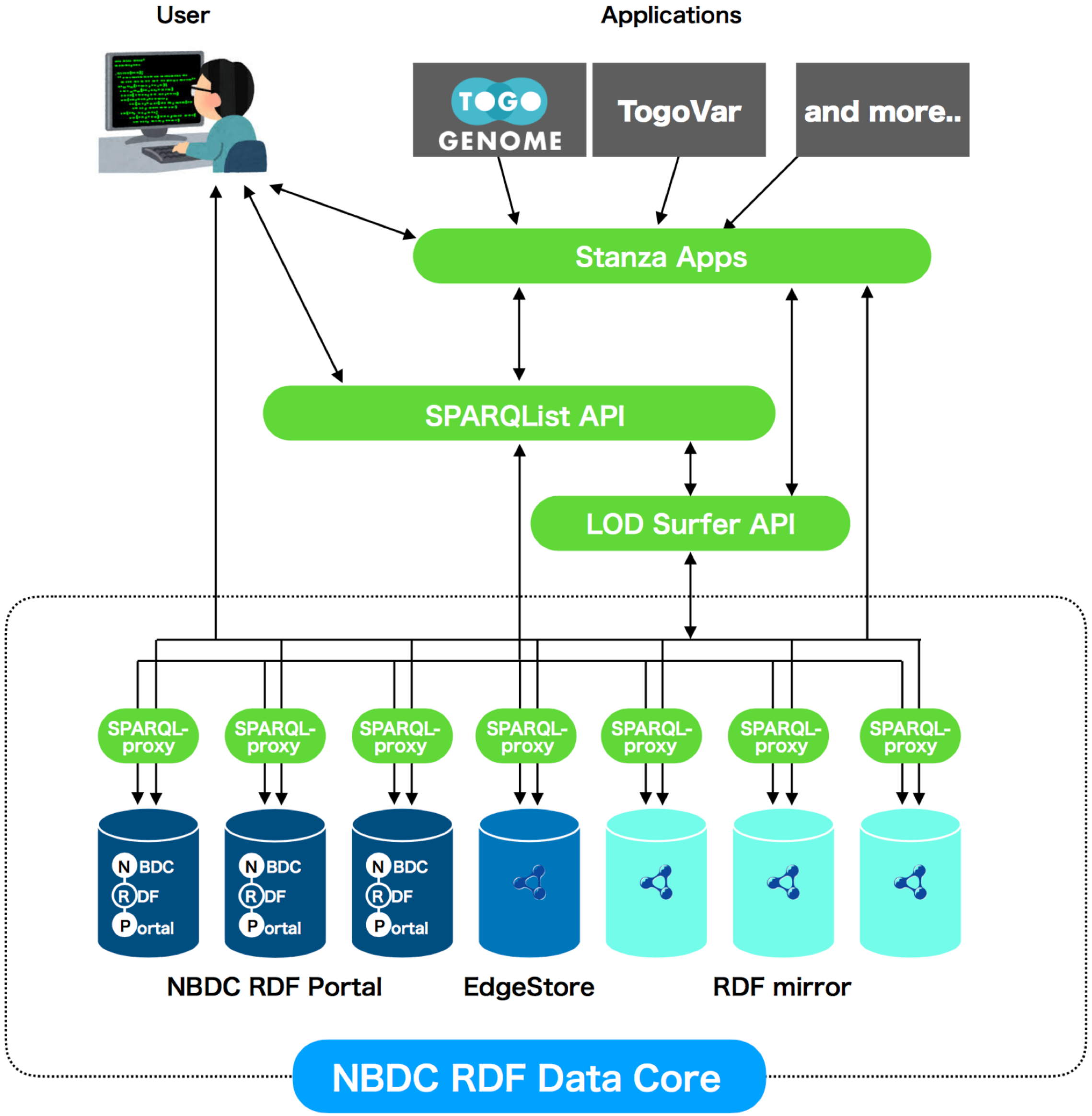
http://togodb.org/

D2RQ Mapper

Converting RDB to RDF

http://d2rq.dbcls.jp/

Database Integration @ DBCLS



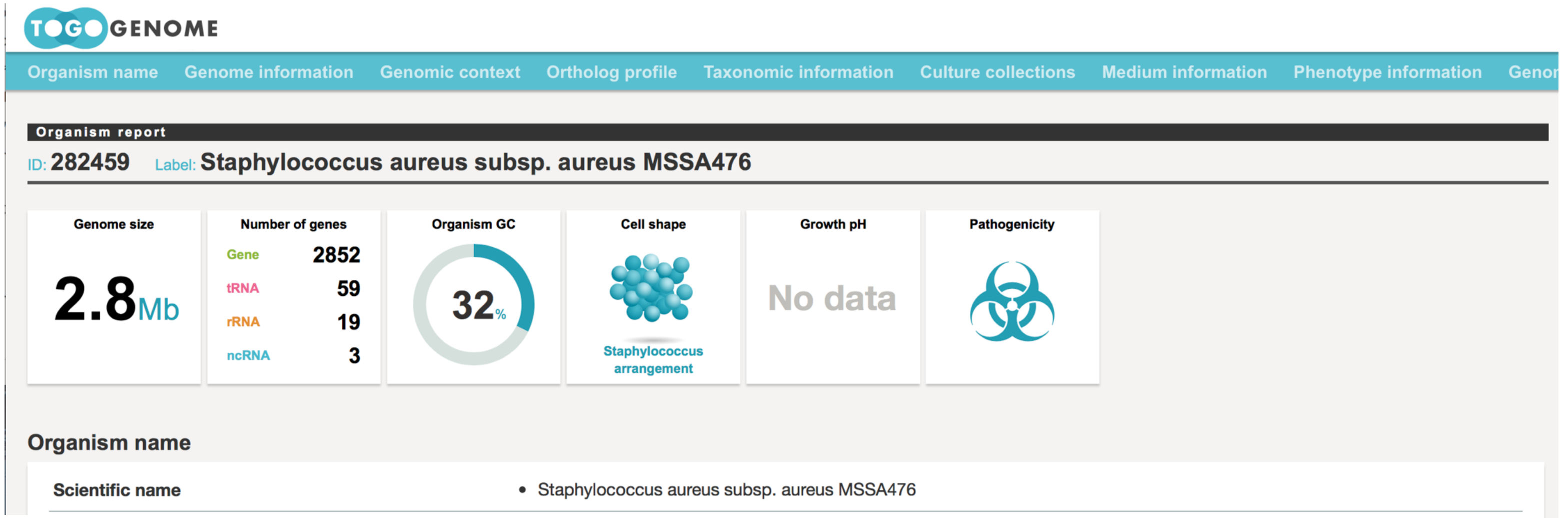
Application Development

Middleware Development

Database Development

Middleware: Accessing SPARQL EPs

TOGO STANZA



Organism report

ID: 282459 Label: Staphylococcus aureus subsp. aureus MSSA476

Genome size	Number of genes		Organism GC	Cell shape	Growth pH	Pathogenicity
2.8Mb	Gene	2852	32%	Staphylococcus arrangement	No data	
	tRNA	59				
	rRNA	19				
	ncRNA	3				

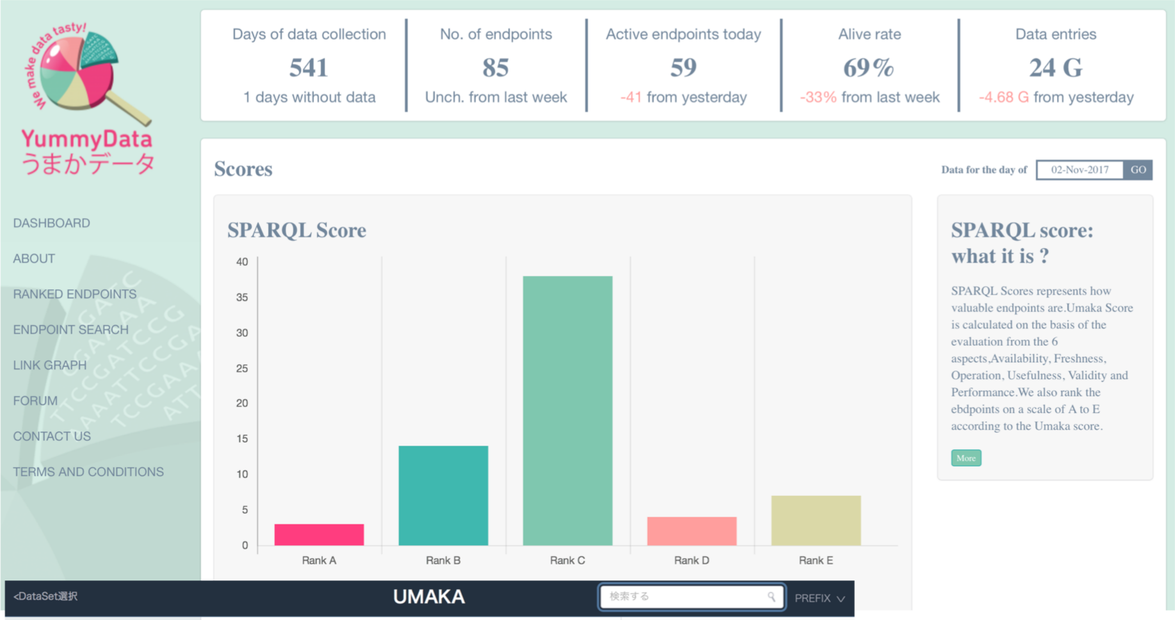
Organism name

Scientific name
- Staphylococcus aureus subsp. aureus MSSA476

- **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
- **YummyData**: listing and monitoring SPARQL endpoints

YummyData: Information for SPARQL endpoint

- YummyData for endpoint information
- YummyViewer for visualization of class relationships



We make data tasty!
YummyData
うまかデータ

DASHBOARD
ABOUT
RANKED ENDPOINTS
ENDPOINT SEARCH
LINK GRAPH
FORUM
CONTACT US
TERMS AND CONDITIONS

Days of data collection	No. of endpoints	Active endpoints today	Alive rate	Data entries
541	85	59	69%	24 G
1 days without data	Unch. from last week	-41 from yesterday	-33% from last week	-4.68 G from yesterday

Scores

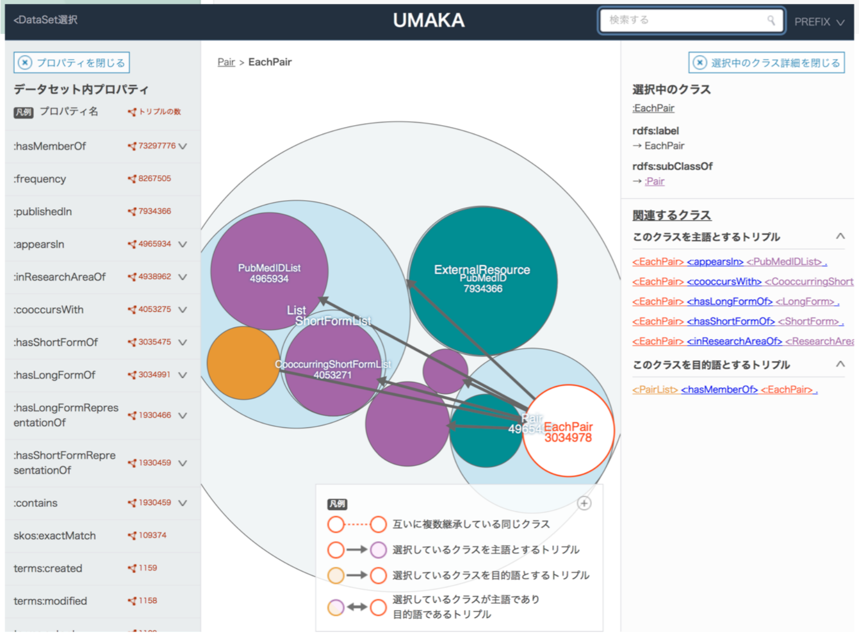
Data for the day of 02-Nov-2017 GO

SPARQL Score

SPARQL score: what it is ?

SPARQL Scores represents how valuable endpoints are.Umaka Score is calculated on the basis of the evaluation from the 6 aspects,Availability, Freshness, Operation, Usefulness, Validity and Performance.We also rank the ebdpoints on a scale of A to E according to the Umaka score.

More



<DataSet選択 UMAKA 検索する PREFIX

ⓧ プロパティを閉じる

データセット内プロパティ

凡例 プロパティ名	トリプルの数
:hasMemberOf	73297776
:frequency	8267505
:publishedIn	7934366
:appearsIn	4965934
:inResearchAreaOf	4938962
:cooccursWith	4053275
:hasShortFormOf	3035475
:hasLongFormOf	3034991
:hasLongFormRepresentationOf	1930466
:hasShortFormRepresentationOf	1930459
:contains	1930459
skos:exactMatch	109374
terms:created	1159
terms:modified	1158

Pair > EachPair

凡例
互いに複数継承している同じクラス
選択しているクラスを主語とするトリプル
選択しているクラスを目的語とするトリプル
選択しているクラスが主語であり目的語であるトリプル

ⓧ 選択中のクラス詳細を閉じる

選択中のクラス

:EachPair

rdfs:label
→ EachPair

rdfs:subClassOf
→ :Pair

関連するクラス

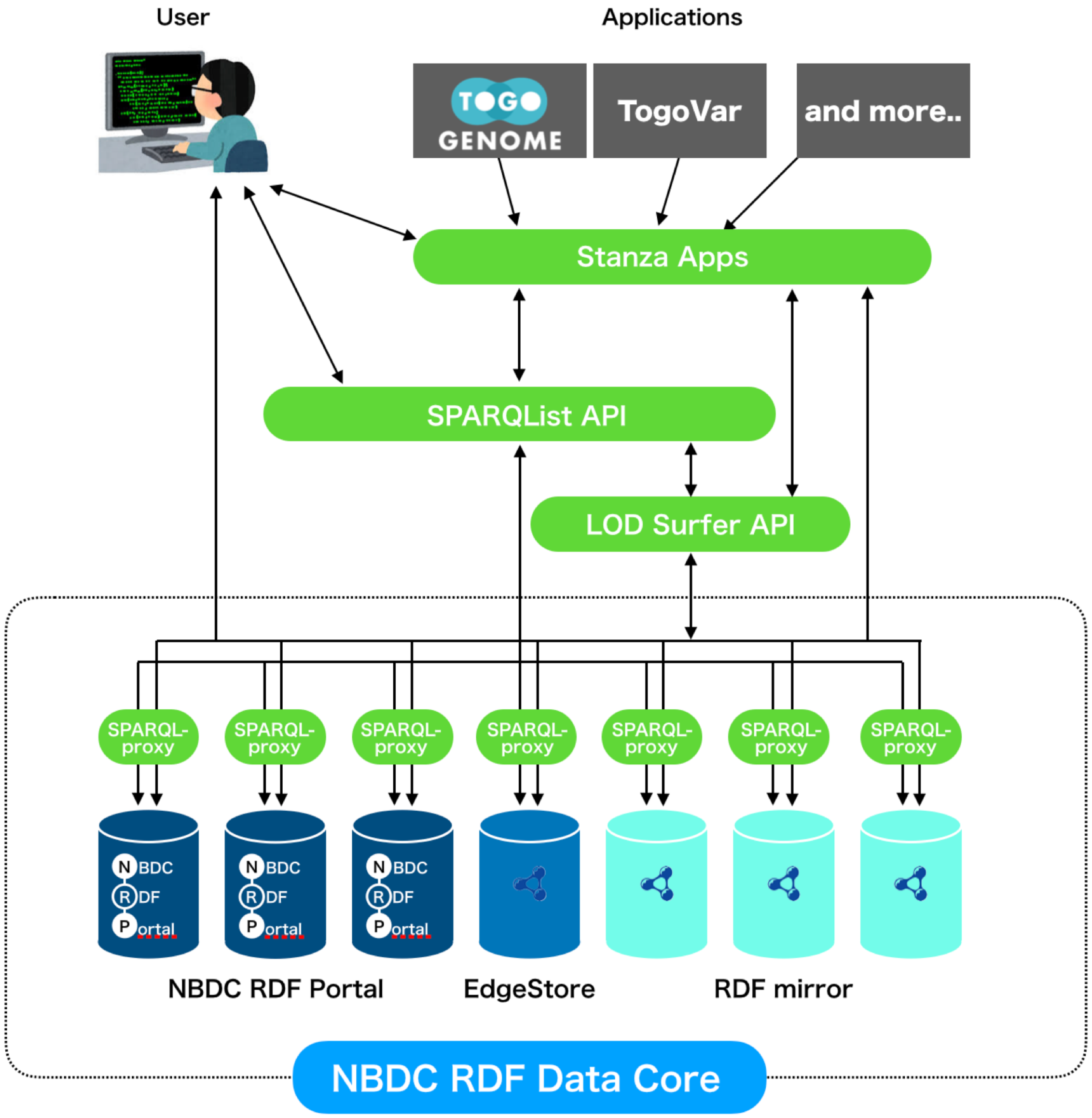
このクラスを主語とするトリプル

<:EachPair> <:appearsIn> <:PubMedIDList> .
<:EachPair> <:cooccursWith> <:CooccurringShort
<:EachPair> <:hasLongFormOf> <:LongForm> .
<:EachPair> <:hasShortFormOf> <:ShortForm> .
<:EachPair> <:inResearchAreaOf> <:ResearchArea

このクラスを目的語とするトリプル

<:PairList> <:hasMemberOf> <:EachPair> .

Database Integration @ DBCLS



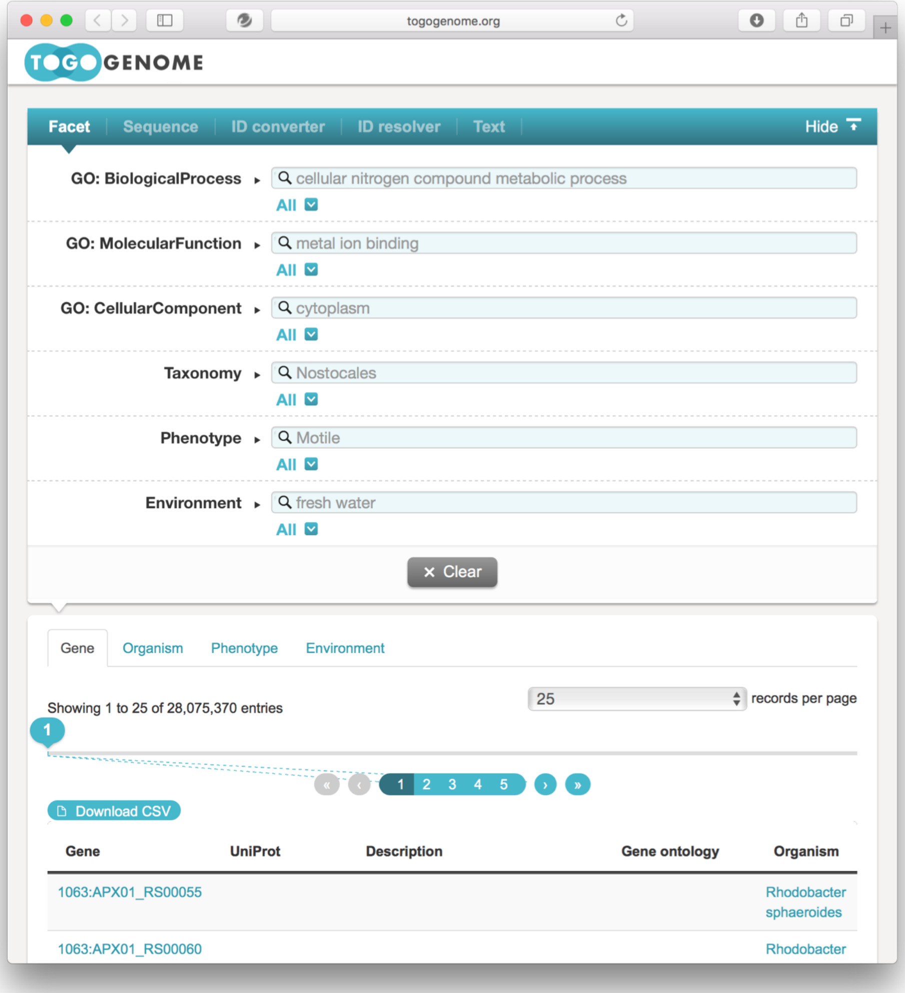
Application Development

Middleware Development

Database Development

Application: TogoGenome

- Genome database based on semantic web technology.
- Unique: implemented only by RDF data stores.
- >10,000 species including 360 eukaryotes.
- > 1 billion triples
- Genes and genomes, environmental and growth conditions, links to other DBs



togogenome.org

TOGOGENOME

Facet | Sequence | ID converter | ID resolver | Text | Hide

GO: BiologicalProcess ▸	cellular nitrogen compound metabolic process All ☑
GO: MolecularFunction ▸	metal ion binding All ☑
GO: CellularComponent ▸	cytoplasm All ☑
Taxonomy ▸	Nostocales All ☑
Phenotype ▸	Motile All ☑
Environment ▸	fresh water All ☑

× Clear

Gene | Organism | Phenotype | Environment

Showing 1 to 25 of 28,075,370 entries

25 records per page

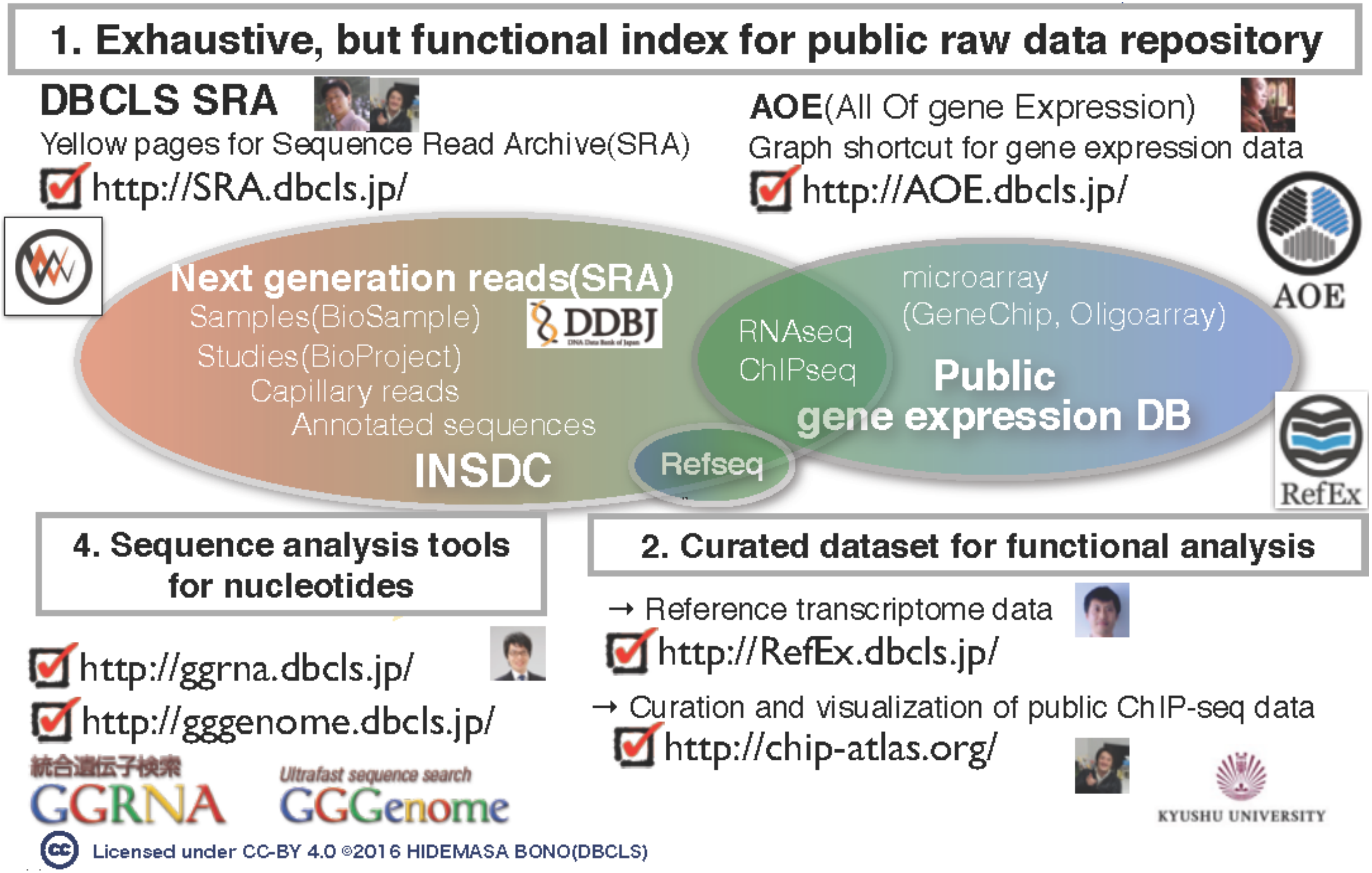
« ‹ 1 2 3 4 5 › »

Download CSV

Gene	UniProt	Description	Gene ontology	Organism
1063:APX01_RS00055				Rhodobacter sphaeroides
1063:APX01_RS00060				Rhodobacter

Variation data → TogoVar

Application: Easy access to omics data



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/

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

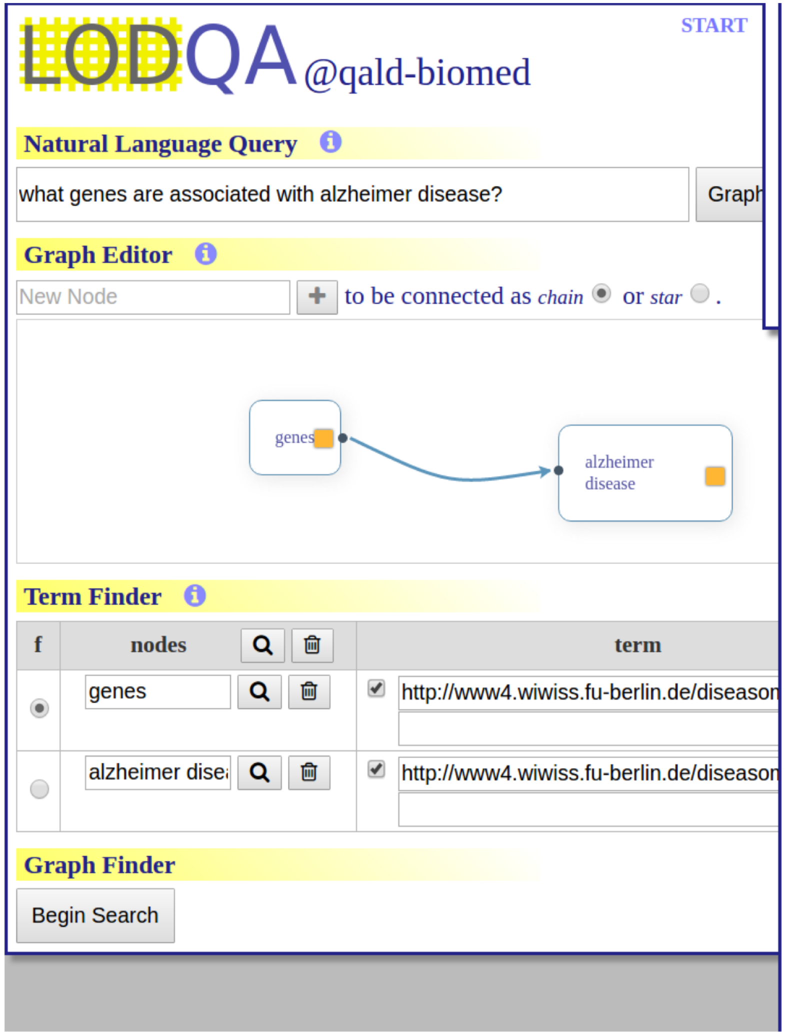
☑ http://RefEx.dbcls.jp/

→ Curation and visualization of public ChIP-seq data

☑ http://chip-atlas.org/

KYUSHU UNIVERSITY

Application: Natural language Q&A



LODQA @qald-biomed

Natural Language Query

what genes are associated with alzheimer disease?

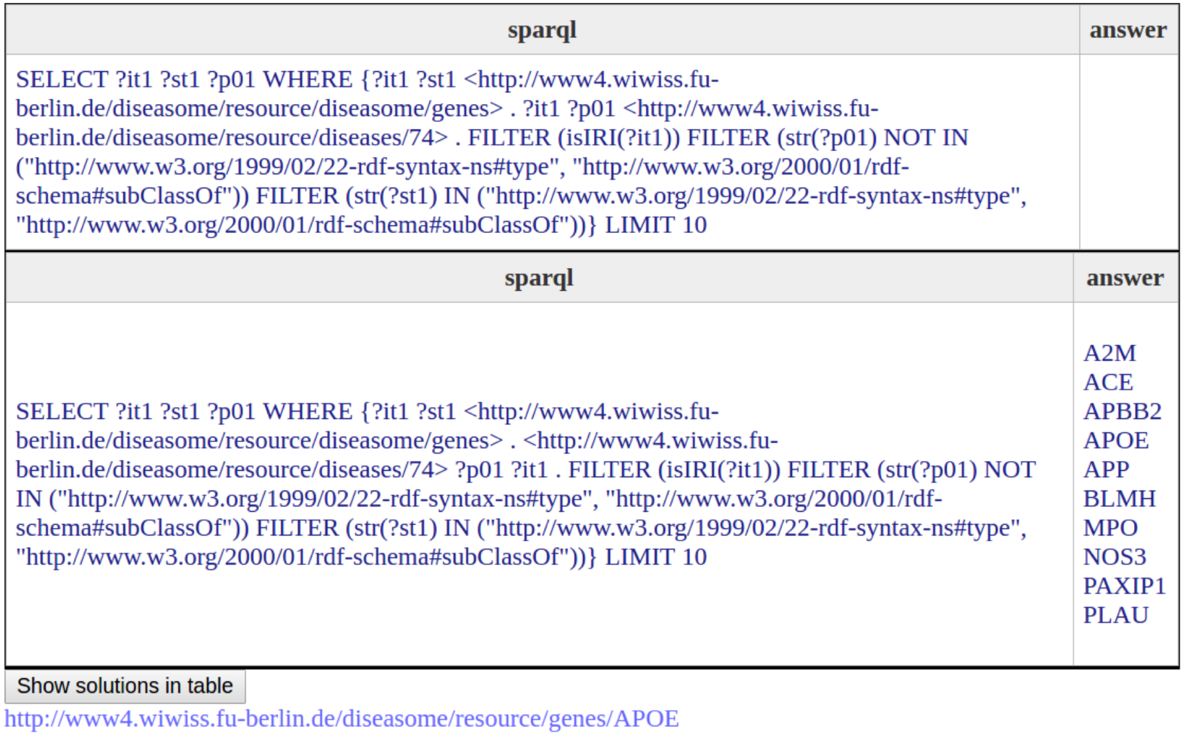
Graph Editor

New Node + to be connected as *chain* ● or *star* ○ .

Term Finder

f	nodes	term
●	genes	☑ http://www4.wiwiss.fu-berlin.de/diseason
○	alzheimer dise	☑ http://www4.wiwiss.fu-berlin.de/diseason

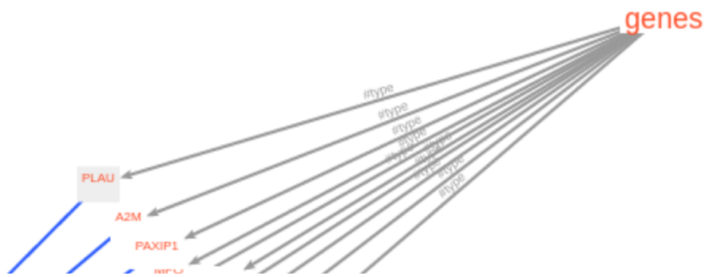
Graph Finder



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

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



http://lodqa.org

Summary

- Database integration via semantic web technology
  - RDF, Linked Open Data
  - RDF Portal and converting tools
- Tools to utilize integrated database
  - http://dbcls.jp/services
- Community for the development and utilization
  - Biohackathon
  - SPARQLthon
  - Lecture series, TogoTV for lecture videos

Acknowledgements

Director

KOHARA, Yuji

KAWANO, Shin

KIM, Jin-Dong

BONO, Hidemasa

MINOWA, Mari

YAMAGUCHI, Atsuko

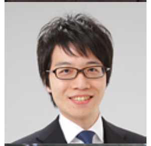
YAMAMOTO, Yasunori

ONO, Hiromasa

KATAYAM, Toshiaki

KAWASHIMA, Shuichi

CHIBA, Hirokazu



NAITO, Yuki

NAKAZATO, Takeru

MORIYA, Yuki

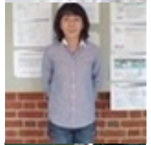
IIDA, Keisuke

OHTA, Tazro

FUJIWARA, Toyofumi

WANG, Yue

OKUBO, Kousaku



KAWAMOTO, Shoko

OKAMOTO, Shinobu

NBDC
National Bioscience Database Center

JST Japan Science and Technology Agency