

ライフサイエンス統合データベースセンター (Database Center for Life Science)

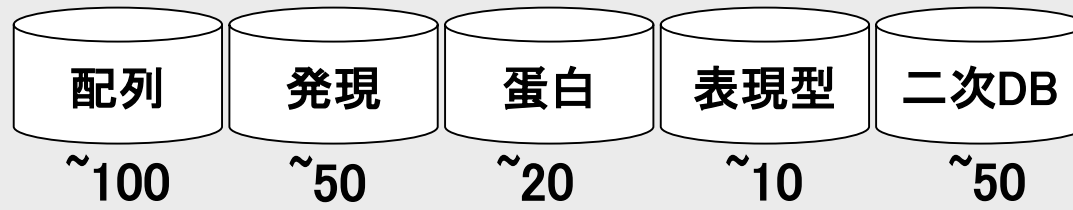
ライフサイエンス統合DBプロジェクト開始時点の 我が国におけるライフサイエンスDBの問題点

- DBが散在していて所在情報や利用法が分からない
 - 我が国で1,000、世界で10,000ものDB
 - 似たようなものがいくつもありどれを使ってよいか分からない
- DBやDBのエントリに信頼性の高い注釈がついていない
 - DB構築、維持を行える人材不足、DB構築への評価の低さ
- 大型プロジェクトの成果公開、共有が不十分
 - 我が国のライフ関係予算は年間3,000億円以上
 - 公開されているものもプロジェクトが終了すると更新ストップ
- ばらばらに構築・管理されていて検索・解析・応用が困難
 - 現在の統合化は分子レベルで行われていて医療などへの応用困難
 - 日本語化されていないので研究動向や分野の状況の把握困難
- 我が国にはDB戦略、中核センターがない

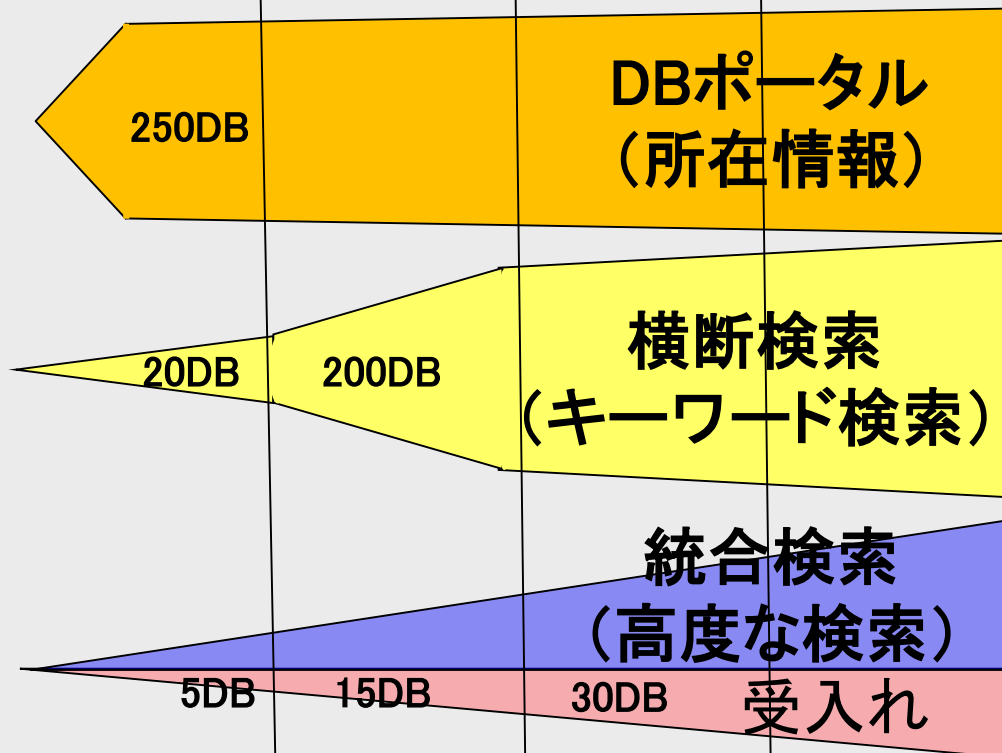
⇒DBCLSの設置(H19)、統合DBプロジェクトの開始

平成18年度調査結果 >>>>> 平成28年度現在

国内データベース(約250について)



~ H19年度 H20年度 H21年度 H22年度



H23以降はJST/NBDCとの
一体的運用/共同研究

Integbio

データベースカタログ

1543件(うち国内DB 1105件)

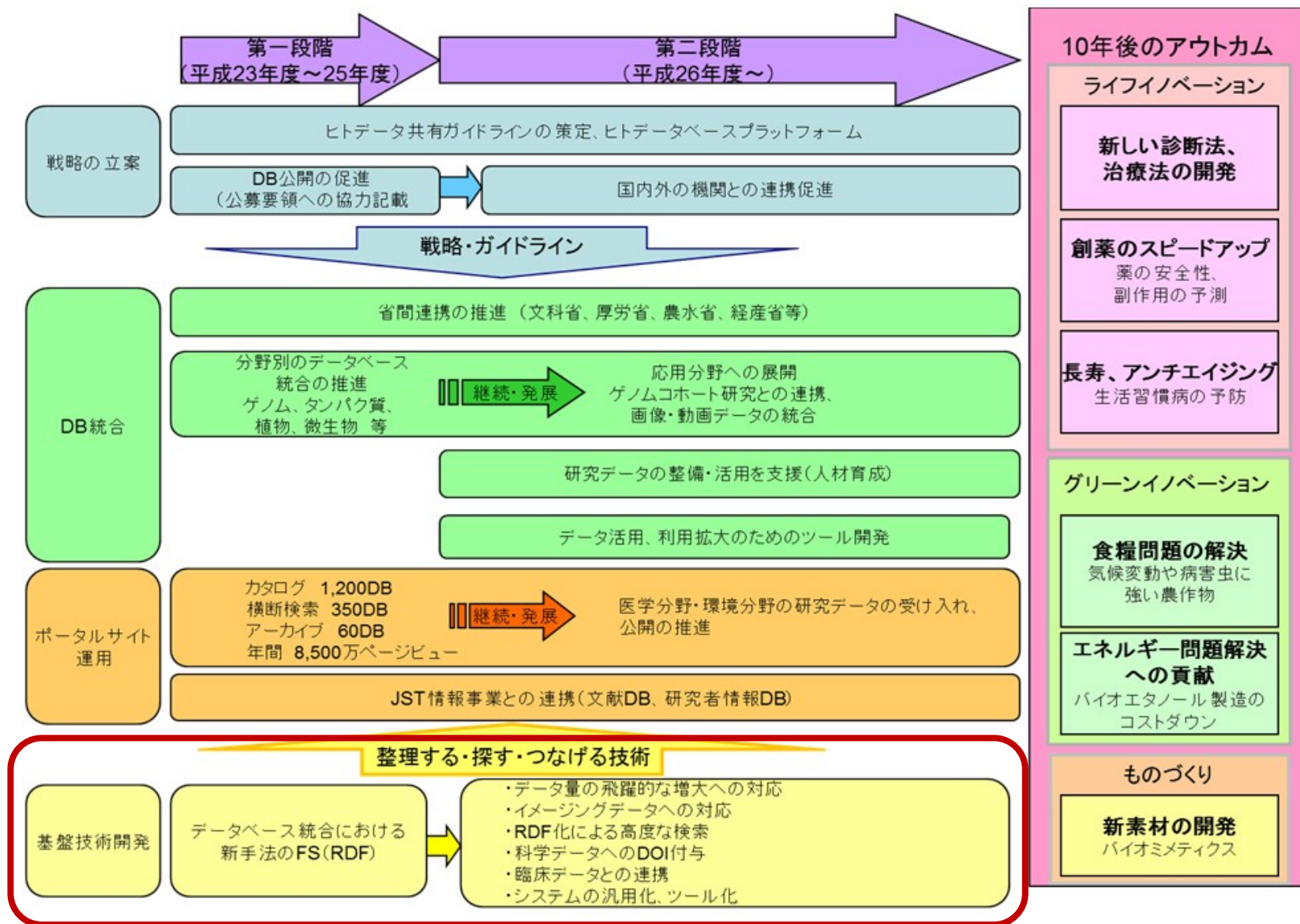
生命科学DB横断検索

586DBが対象

生命科学系DBアーカイブ

113DBがダウンロード可能

ライフサイエンスデータベース統合推進事業のロードマップ



次世代生命科学データベースの実現に向けて

- ・ 次世代生命科学データベース＝**データ駆動型サイエンスを実現**するデータベース
- ・ データ駆動型サイエンスにおいては、新規データ生成も必要ながら、膨大に蓄積されたデータを効率的・効果的に再利用する必要がある→データインフラの整備
- ・ そのためには、データのセマンティクス(**データの意味**)を扱うことが不可欠
- ・ また、データ処理の大幅な**省力化**も必要



RDFの採用

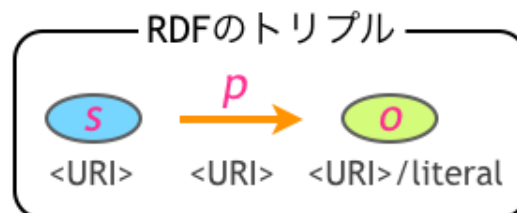
= セマンティック・ウェブ
= Linked Open Data

RDFのデータ構造

- 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)

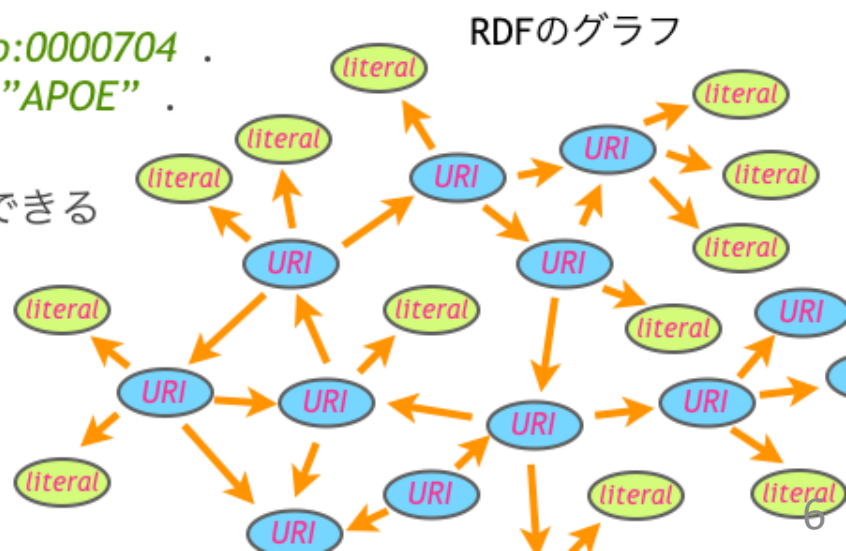


- つまり RDF は非常にシンプルなデータ形式

- `<http://genome.db/gene1> rdf:type so:0000704 .`
 `<http://genome.db/gene1> rdfs:label "APOE" .`

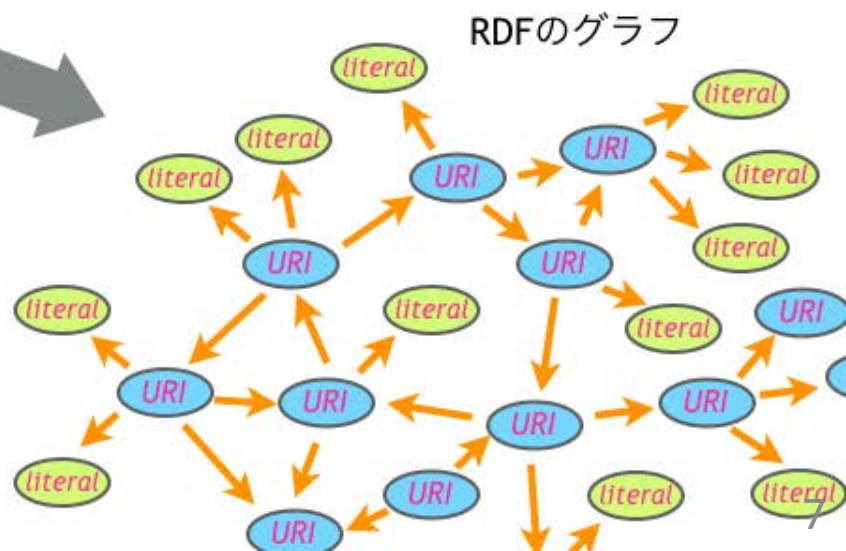
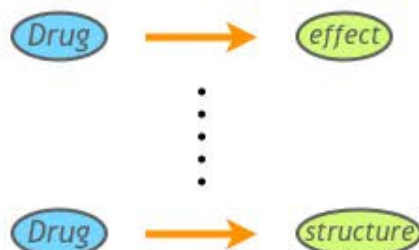
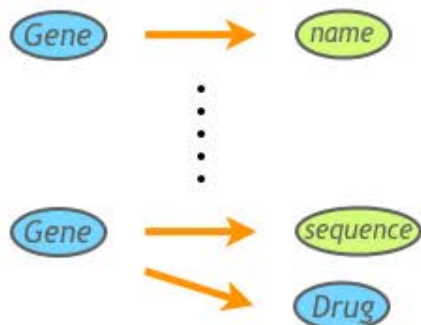
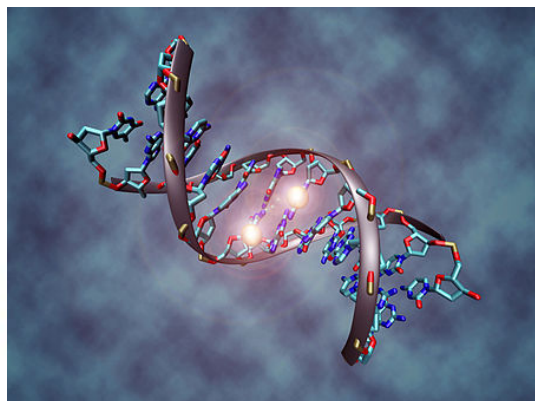
- そして RDF では多様なデータを容易に統合できる

- 同じモノは ID = URI も同じ
- 重ねあわせていくとグラフになる



RDFによるデータ統合のイメージ

例として遺伝子と薬のデータを RDF 化して統合した（混ぜこんだ）とします



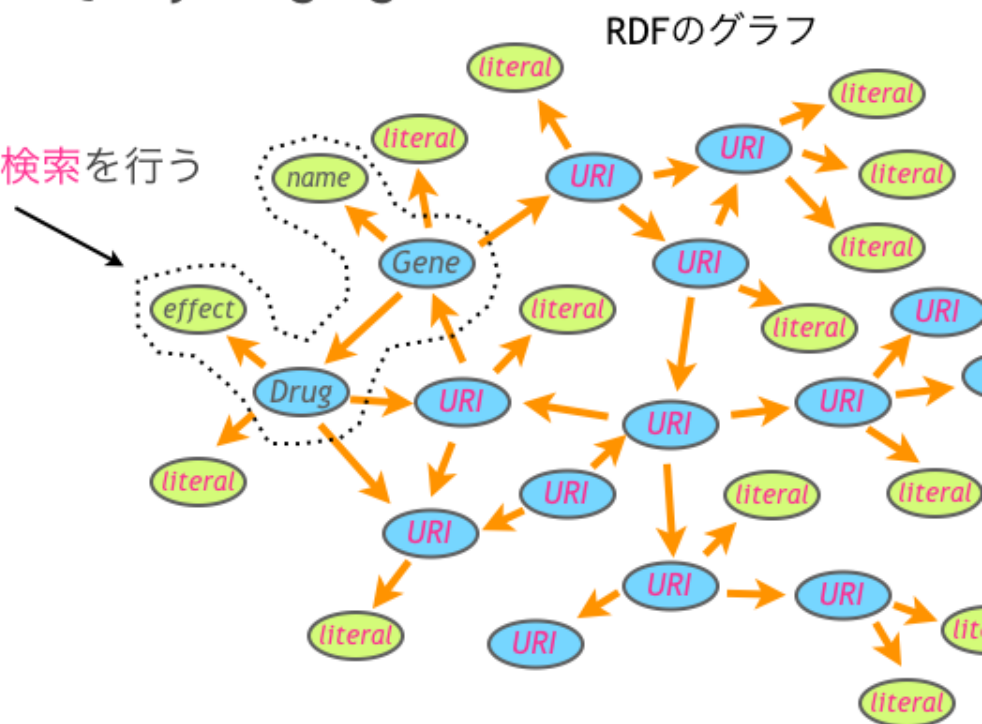
RDFデータのSPARQLによる検索

- SPARQL: SPARQL Protocol and RDF Query Language

- RDF を検索するための標準言語

- パターンマッチにより部分グラフ検索を行う

```
SELECT ?gene ?name ?drug ?effect
WHERE {
  ?gene rdf:type      so:0000704 .
  ?gene rdfs:label    ?name .
  ?gene :dose         ?drug .
  ?drug :has_effect   ?effect .
}
```



→ 統合するデータベースが同じIDを共有して有機的につながっていることが重要

自然文によるRDFデータベースの検索

LODQA@QALD-BioMed

START TARGETS DOC GRAPH

QALD-BioMed

Step 1. Enter a query and parse it.

What genes are associated with alzheimer disease?

Parse

Step 2. Check the graph and edit it if necessary.

New Node + to be connected as



Step 3. Find the terms for each part of the graph.

F	Text	Term
☒	genes	☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/genes
☐	alzheimer disease	☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74 ☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74 ☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74 ☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74 ☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74 ☒ http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74
	associated with	

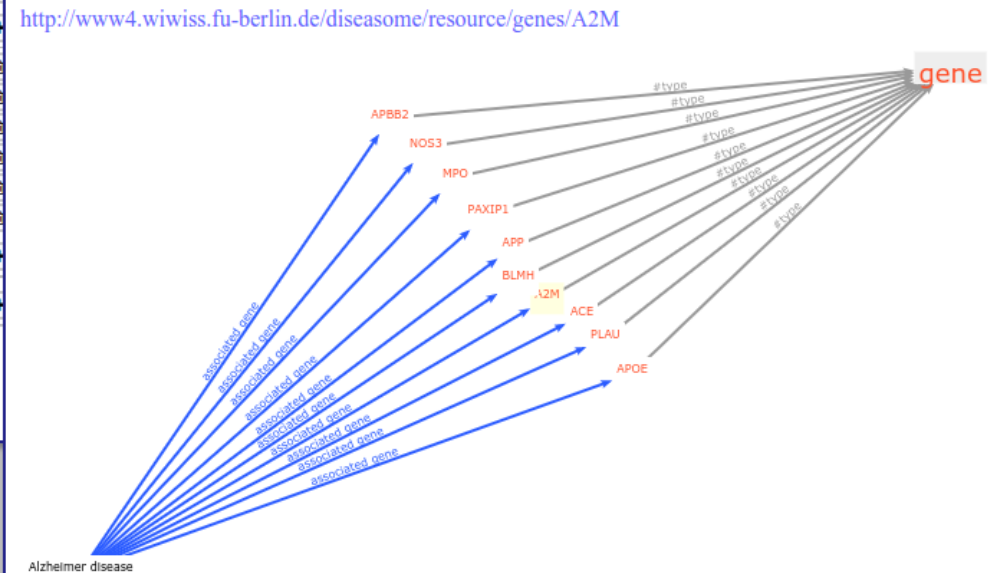
Ready for Search

Begin Search ☐ verbose

Results		
	head	text
t1	genes	http://www4.wiiss.fu-berlin.de/diseasome/resource/diseasome/genes
t2	alzheimer disease	http://www4.wiiss.fu-berlin.de/diseasome/resource/diseases/74

Sparql 2 Show

sparql	answer
<pre>SELECT ?it1 ?st1 ?p01 WHERE { ?it1 ?st1 <http://www4.wiiss.fu-berlin.de/diseasome/resource/diseasome/genes> . <http://www4.wiiss.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</pre>	A2M ACE APBB2 APOE APP BLMH MPO NOS3 PAXIP1 PLAU



■ RDF化ガイドライン／RDFポータル構築

- ・ SPARQLthon（および国際版・国内版BioHackathon）を継続して開催し、統合化推進プログラムのRDF化のサポートを行う。そこでの知見を基に**RDF化のガイドライン**を作成した
- ・ それらガイドラインに沿って作られたRDFを一覧できるポータルサイトを構築し**RDFデータの利用促進**および**国際的認知度を高める** → <http://integbio.jp/rdf/>

現在、協力を表明頂いているグループ一覧

	グループ	RDFコンテンツの種類
統合化推進	微生物統合	微生物のゲノム、オーソログ、環境、菌株情報等
	疾患ヒトゲノム統合	転写開始点
	植物ゲノム統合	植物オーソログ
	PDBJ	タンパク立体構造
	糖鎖統合	糖鎖構造、糖蛋白等
	生命と環境のフェノーム統合	理研バイオバンクの細胞情報
その他	Open TG-GATEs	遺伝子発現（毒性）
	RefEx	遺伝子発現（リファレンス）
	DBCLS SRA	NGSメタデータ
	FAMSBASE	予測タンパク立体構造

国際的な潮流

平成13 セマンティックウェブの提唱



Tim Berners-Lee

平成19 DBCLS設置

平成20



平成22 第3回BioHackathon開催

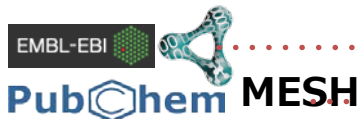
平成23



平成25



平成26



平成27

RDF portal

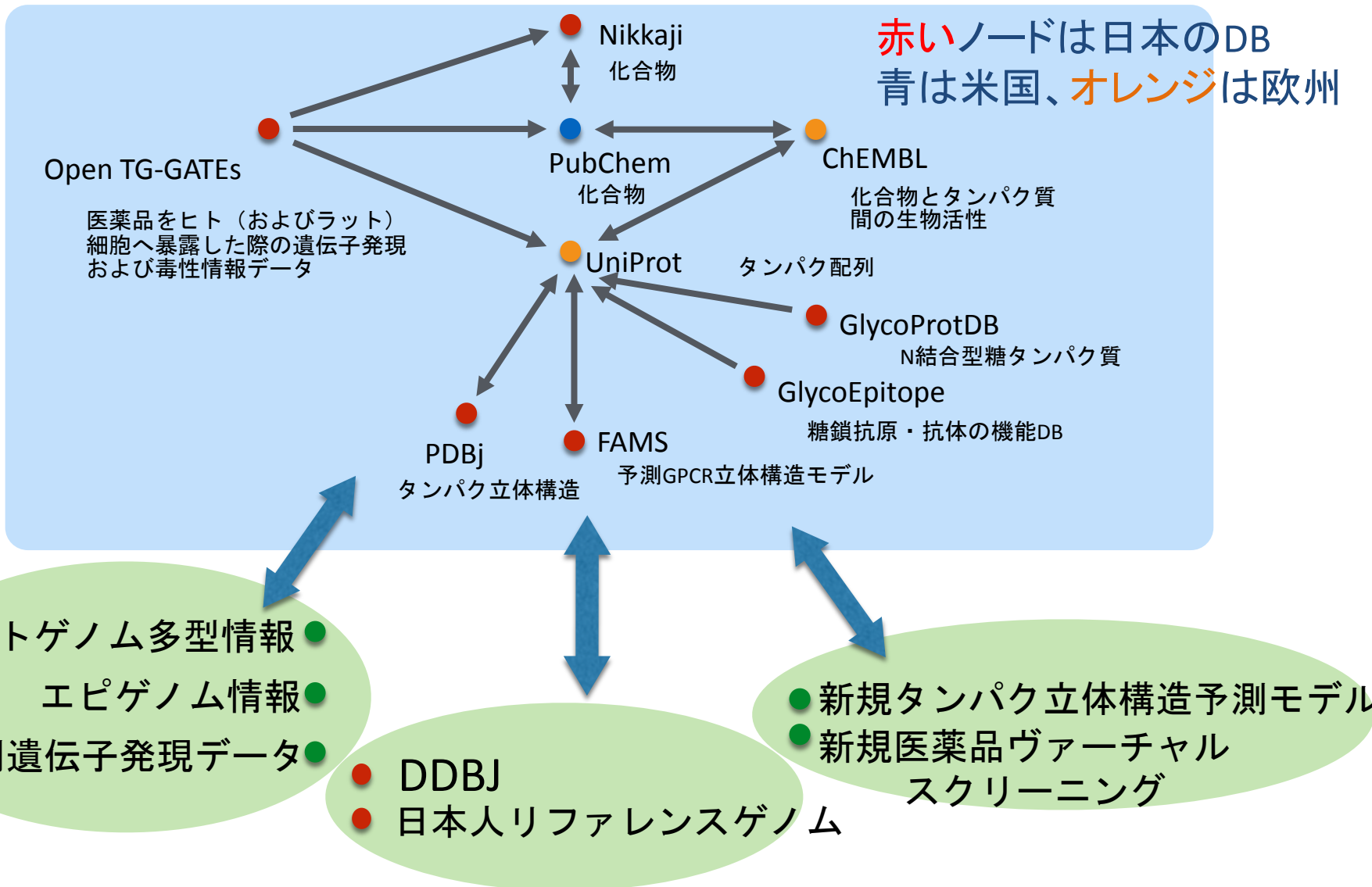


国際開発者会議BioHackathon



例：データ統合による個別化医療のサポート

-分野や国境を越えたDB統合-



スタッフ



GOTO Susumu
五斗 進
教授



OKUBO Kousaku
大久保 公策
教授 (兼務)



KAWANO Shin
河野 信
特任准教授



KIM Jin-Dong
金 進東
特任准教授



BONO Hidemasa
坊農 秀雅
特任准教授



MINOWA Mari
箕輪 真理
特任准教授



YAMAGUCHI Atsuko
山口 敦子
特任准教授



YAMAMOTO Yasunori
山本 泰智
特任准教授



ONO Hiromasa
小野 浩雅
特任助教



KATAYAMA Toshiaki
片山 俊明
特任助教



KAWASHIMA Shuichi
川島 秀一
特任助教



NAITO Yuki
内藤 雄樹
特任助教



NAKAZATO Takeru
仲里 猛留
特任助教



IIDA Keisuke
飯田 啓介
特任研究員



OHTA Tazro
大田 達郎
特任研究員



MORIYA Yuki
守屋 勇樹
特任研究員



WANG Yue
王 悦
特任技術専門員



OKAMOTO Shinobu
岡本 忍
特任准教授



KAWAMOTO Shoko
川本 祥子
特任准教授



FUJIWARA Toyofumi
藤原 豊史



WATANABE Tadayoshi
渡辺 忠剛

客員部門・研究協力者



NAKAMURA Haruki
中村 春木
客員教授



KINJO Akira
金城 玲
客員准教授



OGASAWARA Osamu
小笠原 理
業務参加者

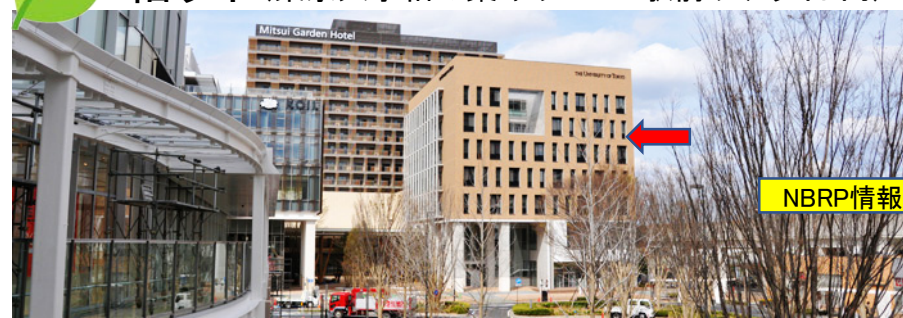


IWASAKI Wataru
岩崎 渉
業務協力者

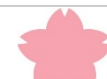


Nicola Colic
外来研究員

 **柏ラボ** (東京大学柏の葉キャンパス駅前サテライト内)



NBRP情報センター

 **三島ラボ** (国立遺伝学研究所内)



DDBJ

スパコン棟