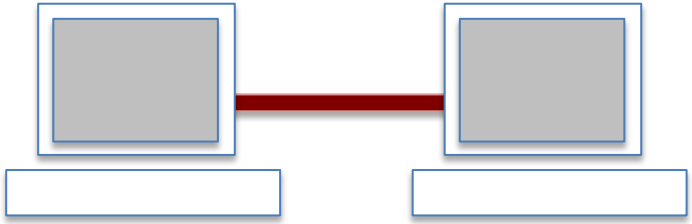
DBCLSにおけるデータベースRDF化への取組みと今後の展望

情報・システム研究機構
ライフサイエンス統合データベースセンター
山口　敦子

Web of Data (1)

インターネットの出現



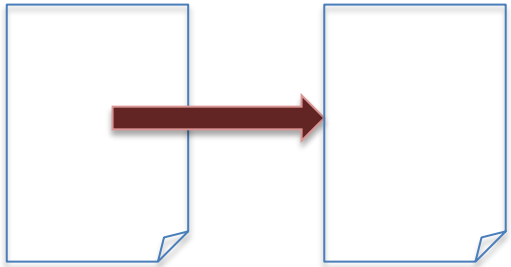
ケーブルを意識せずに
コンピュータをつなぐ世界の出現

各コンピュータに蓄積される文書を得るためには,
文書の在処を何らかの手段で知り,
アクセスするしかなかった

ftp
telnet
gopher
archie
…

Web of Data (2)

WWWの出現



文書がどこにあるかを意識せずに
利用できる世界(Web of Documents)の出現

各文書に含まれるデータを得るためには,
文書ごとに解析してデータを
抽出していくしかなかった.

HTMLパーザ
自然言語処理
…

Web of Data (3)

ライフサイエンスDBにおいて,
現状: DB数増大＆大規模化
課題: データを合わせることで新しい知識が生まれる
　そのためにはデータ共有と統合的利用の仕組みが必須

データの置き場や形式を意識せずに
データをつないで利用できる世界(Web of Data)

↓

RDF 技術の利用

RDF

データ記述・交換のための標準的枠組み(W3C仕様)

の三つ組で

もの(ex. 遺伝子, タンパク質, 生物種, 化合物…)の関係を表現

→シンプルな形式

主語，述語は IRI(ex. URL)
目的語は IRI かリテラル(具体的な値)

→ものに対してグローバルにユニークなID付け

三つ組の集合＝データ

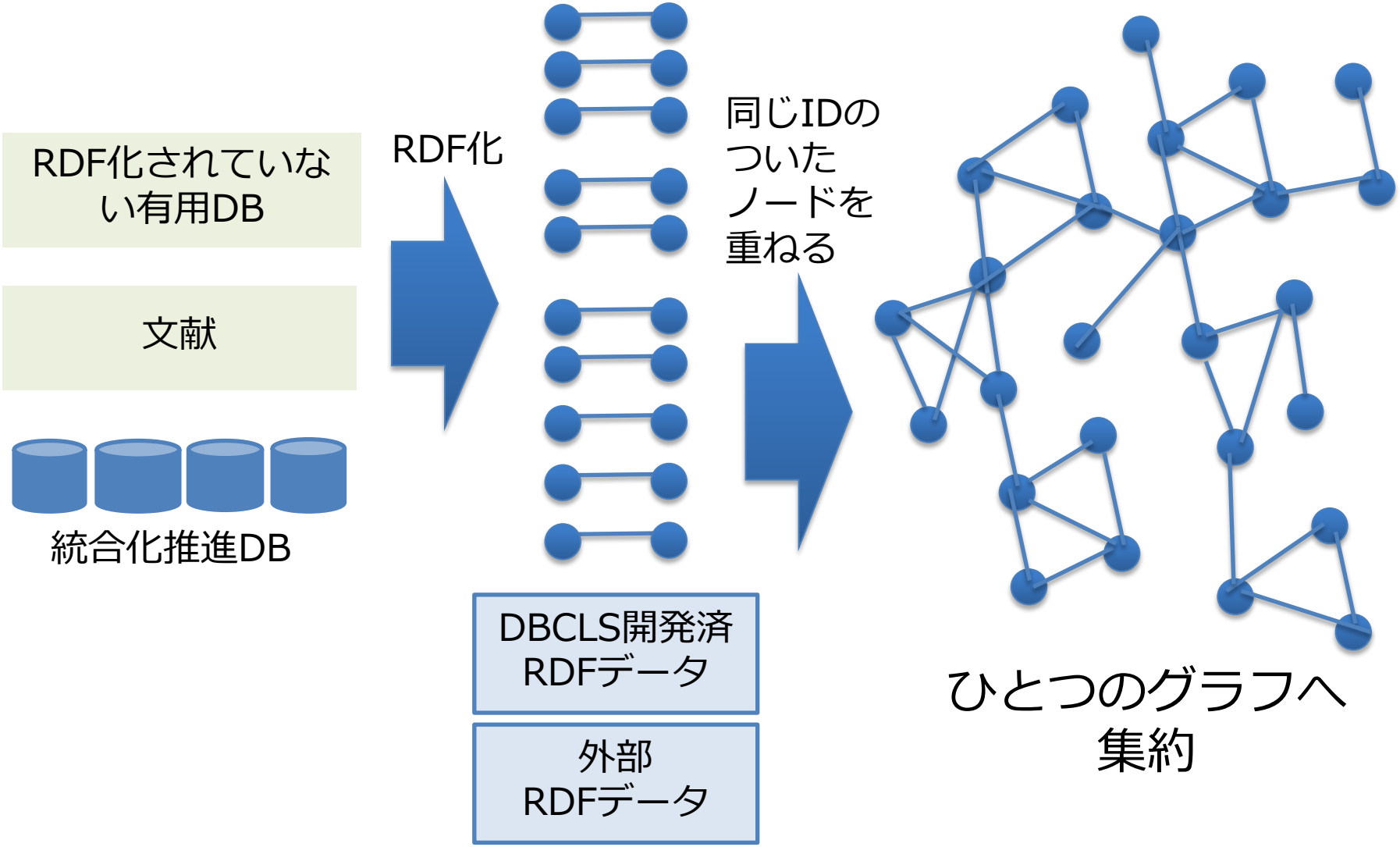
→ひとつひとつのものにURLなどのグローバルなIDをつけ，
グラフとしてつなぐことで，
ひとつのグローバルなデータ空間(Linked Data)を実現

世界的にRDF化が加速

UniProt (2008-)
Bio2RDF (2008-)
BioHackathon (2010以降RDF化とその周辺がテーマ)
NBDC 基盤技術開発プログラム(2011-2013)
wwPDB (2011-)
EBI RDF platform (2013-)
- BioModels
- BioSamples
- ChEMBL
- Expression Atlas
- Reactome

PubChem (2014-)
Ensembl (TBA)

RDF化の取り組み

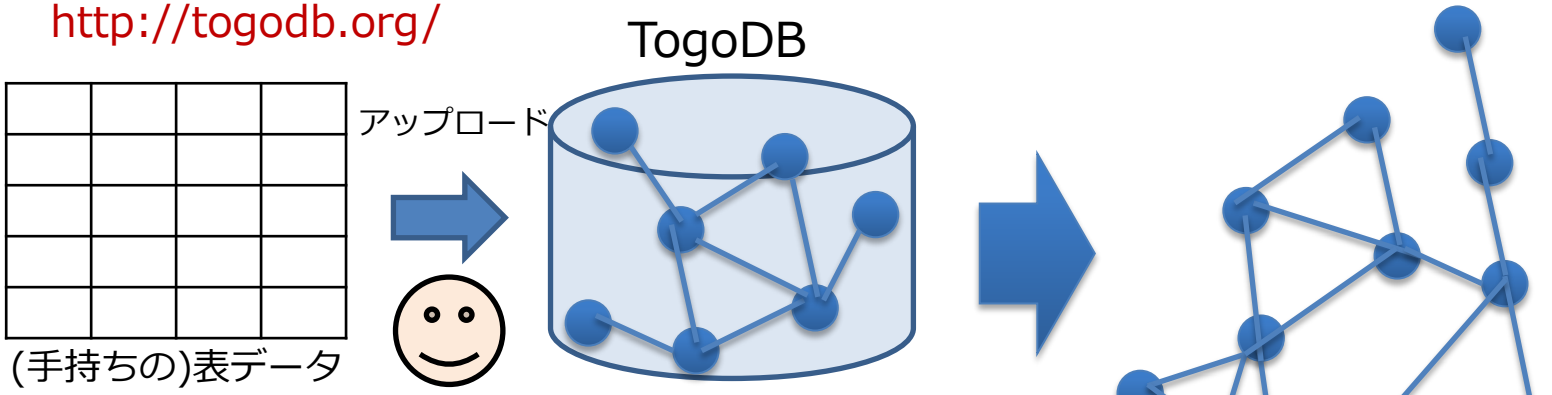

RDF化されていない有用DB
文献
統合化推進DB
RDF化
DBCLS開発済
RDFデータ
外部
RDFデータ
同じIDのついたノードを重ねる
ひとつのグラフへ集約

RDF化の取り組み

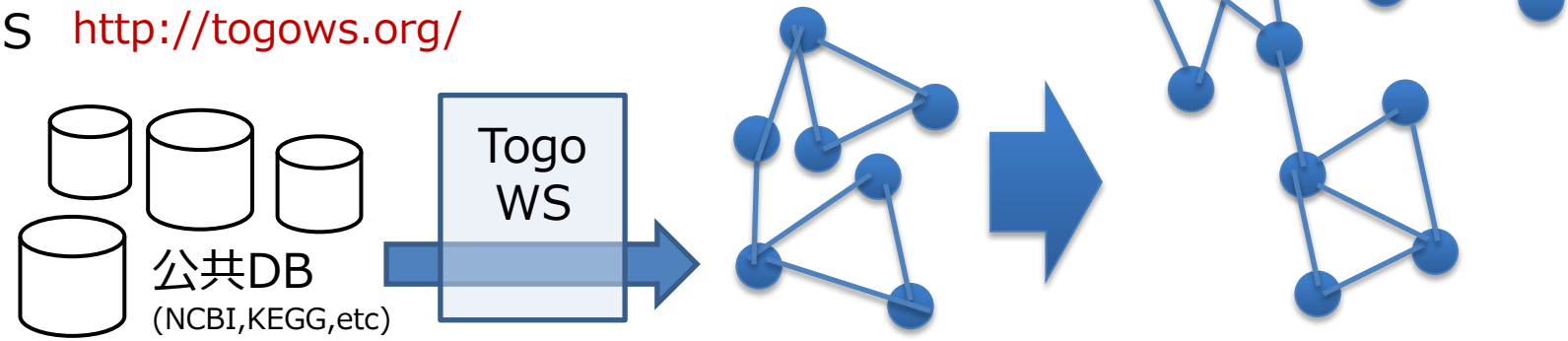
RDF化されていない有用DB

RDF化ツールの開発，データのキュレーション

TogoDB http://togodb.org/



TogoWS http://togows.org/



ポスター05 RDFによるデータ統合と相互運用性のための技術開発

一つのグラフへ

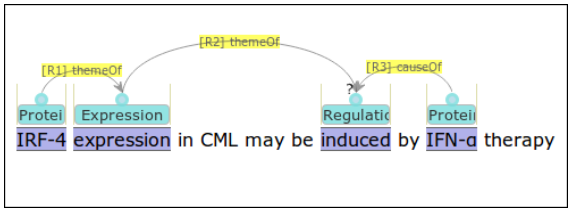
RDF化の取り組み

文献

文献内の知識をRDF化するツールの開発

PubAnnotation http://www.pubannotation.org/

PubMedの文献に埋もれた知識のRDF化

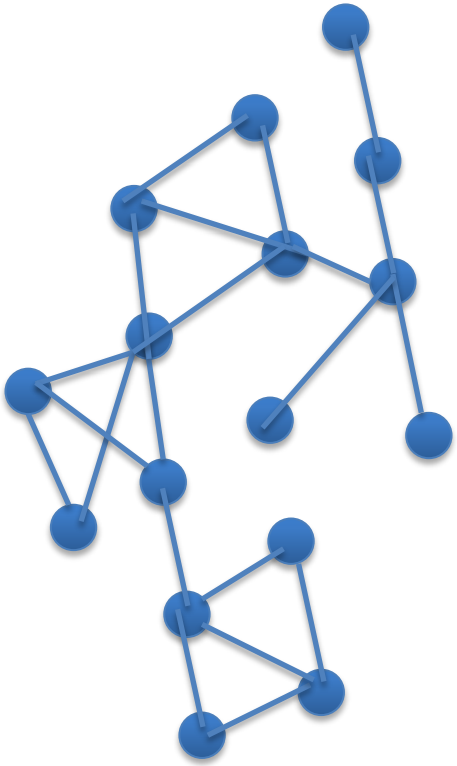


新着論文レビューのナビゲーションシステム開発

http://navi.first.lifesciencedb.jp/stanza/top

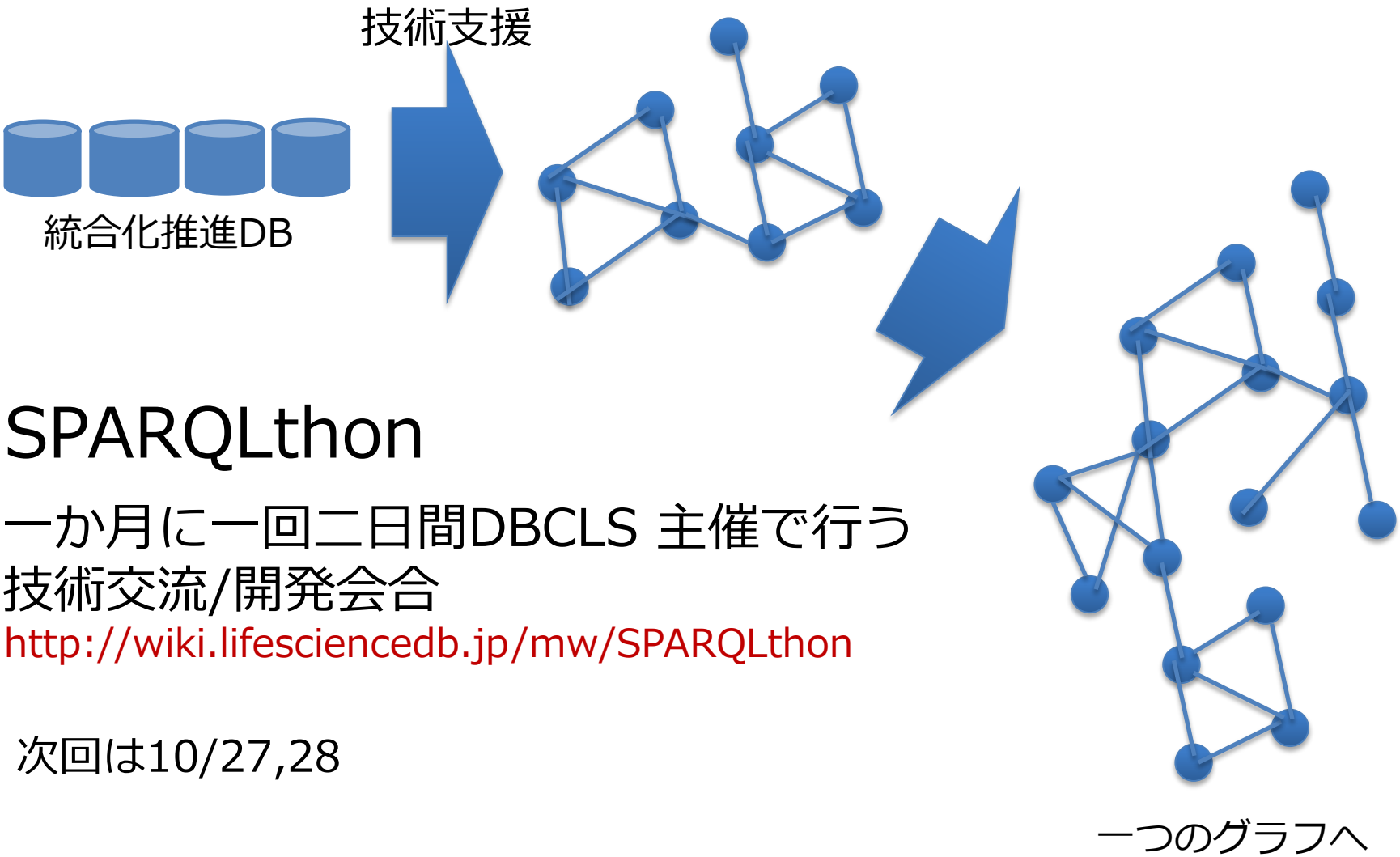
DBCLSで作成している日本語文献のRDF化

ポスター07 日本語コンテンツに対する
セマンティックウェブ技術の適用



一つのグラフへ

RDF化の取り組み



SPARQLthon

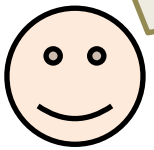
一か月に一回二日間DBCLS 主催で行う
技術交流/開発会合
http://wiki.lifesciencedb.jp/mw/SPARQLthon

次回は10/27,28

めざす世界

データの置き場や形式を意識せずにデータをつないで利用できる世界

日本人のBRAFのSNPs の統計情報とそれに関連する病気を知りたい.



RDF検索言語(SPARQL)へ変換

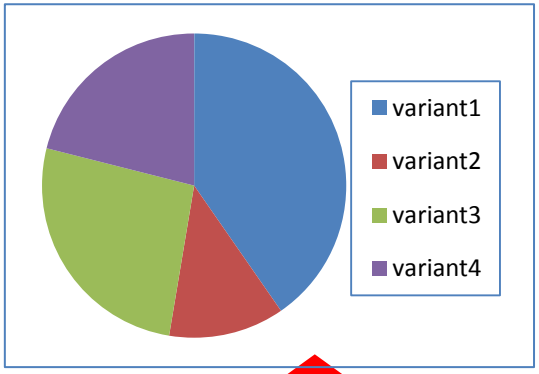
```
SELECT ?person ?variant ?disease
WHERE {
 ?gene has_relation ?disease
 ?gene type:label “BRAF”
 ?person has ?gene .
 ?person nationarity “Japanese“ .
}
```

検索

結果

person	variant	disease
person1	variant1	colorectal cancer
person2	variant3	lung cancer
person3	variant2	cardiofaciocutaneous syndrome 1
person4	variant1	colorectal cancer
person5	variant3	lung cancer

可視化

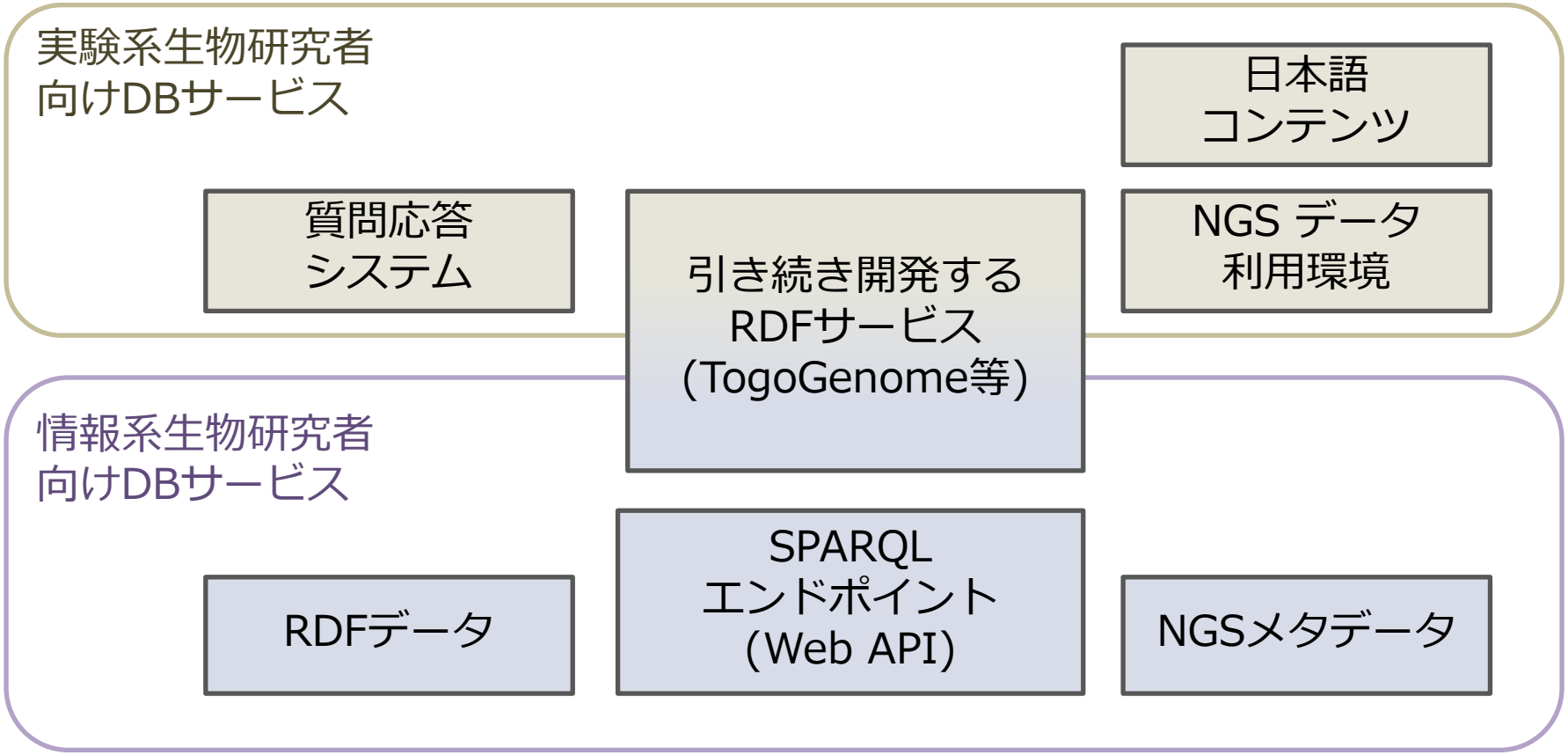


variant	disease
variant1	colorectal cancer
variant2	cardiofaciocutaneous syndrome 1
variant3	lung cancer
variant4	Noonan syndrome 7

H28年度達成目標

DBのRDF化による分野を超えた統合化の実現

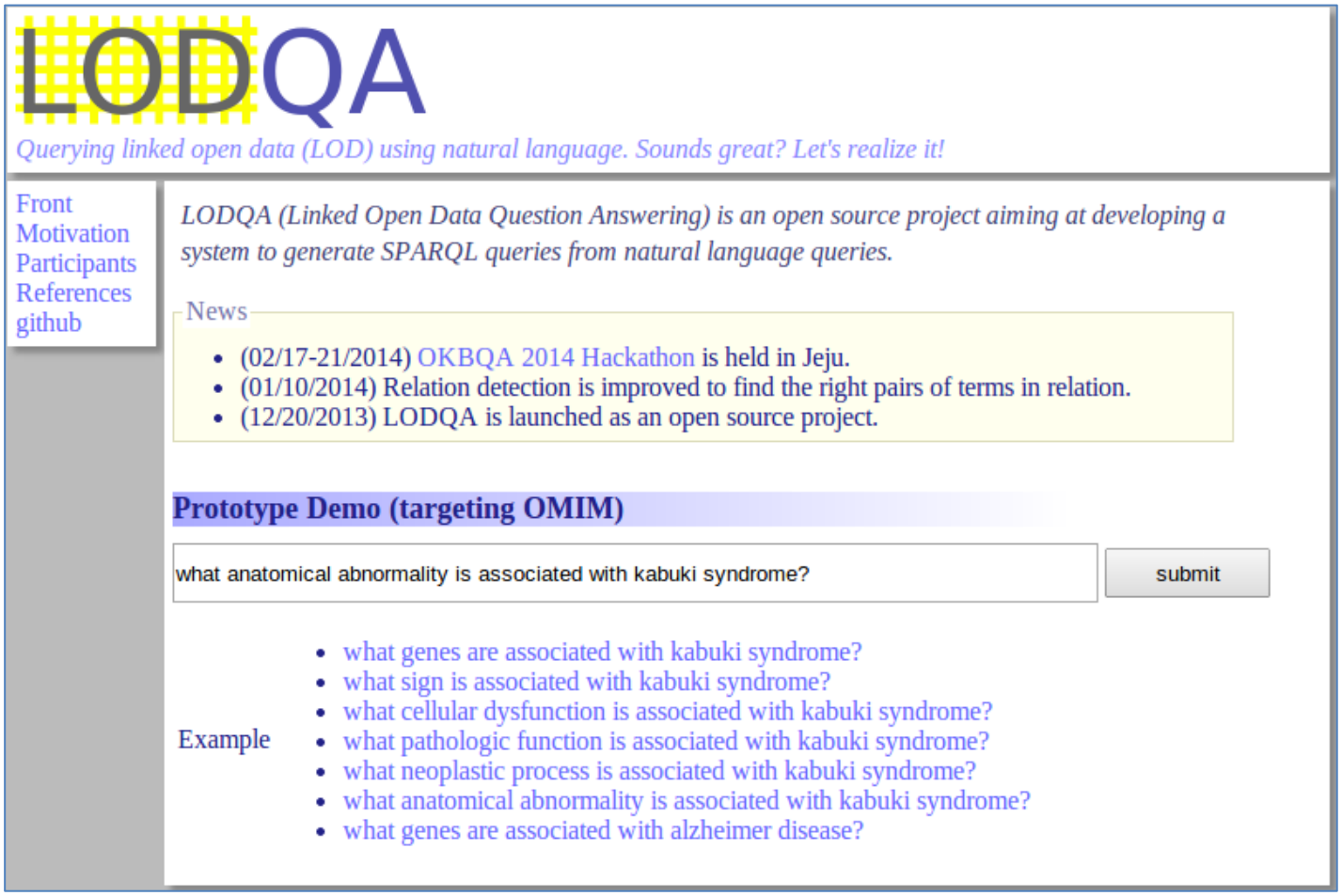
RDFデータを揃えるのみならず,
実験系生物研究者のニーズに沿ったサービスを開発



利用可能な技術/サービス (検索言語変換): LODQA

http://lodqa.org/

質問応答システム



利用可能な技術/サービス (検索言語変換): LODQA

http://lodqa.org/

入力:“what anatomical abnormality is associated with kabuki syndrome?”

SPARQL

検索言語(SPARQL)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?t1 ?l1
WHERE {
  GRAPH <http://bioportal.bioontology.org/ontologies/OMIM> {
    {
    ?t1 <http://bioportal.bioontology.org/ontologies/umls/tui> "T019"^^xsd:string .
    ?t2 ?d1 <http://purl.bioontology.org/ontology/OMIM/147920> .
```

Results

Send query to SPARQL endpoint

http://purl.bioontology.org/ontology/OMIM	
http://purl.bioontology.org/ontology/OMIM	
http://purl.bioontology.org/ontology/OMIM/124500	DEAFNESS, CONGENITAL, WITH KERATOPACHYDERMIA AND CONSTRICTIONS OF FINGERS AND TOES
http://purl.bioontology.org/ontology/OMIM/158900	FACIOSCAPULOHUMERAL MUSCULAR DYSTROPHY 1A
http://purl.bioontology.org/ontology/OMIM/114300	GORDON SYNDROME
http://purl.bioontology.org/ontology/OMIM/277590	WEAVER SYNDROME
http://purl.bioontology.org/ontology/OMIM/226980	EPIPHYSEAL DYSPLASIA, MULTIPLE, WITH EARLY-ONSET DIABETES MELLITUS
http://purl.bioontology.org/ontology/OMIM/138800	GOITER, MULTINODULAR 1, WITH OR WITHOUT SERTOLI-LEYDIG CELL TUMORS
http://purl.bioontology.org/ontology/OMIM/610651	XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP B
http://purl.bioontology.org/ontology/OMIM/265800	PYCNODYSOSTOSIS
http://purl.bioontology.org/ontology/OMIM/255310	MYOPATHY, CONGENITAL, WITH FIBER-TYPE DISPROPORTION
http://purl.bioontology.org/ontology/OMIM/154500	TREACHER COLLINS SYNDROME 1

検索結果

利用可能な技術/サービス(検索): TogoGenome

http://togogenome.org/

RDFで集積したゲノム情報を検索するシステム

遺伝子機能，生物種，表現型，環境等
の条件の組み合わせによる
ゲノム情報の絞り込み検索

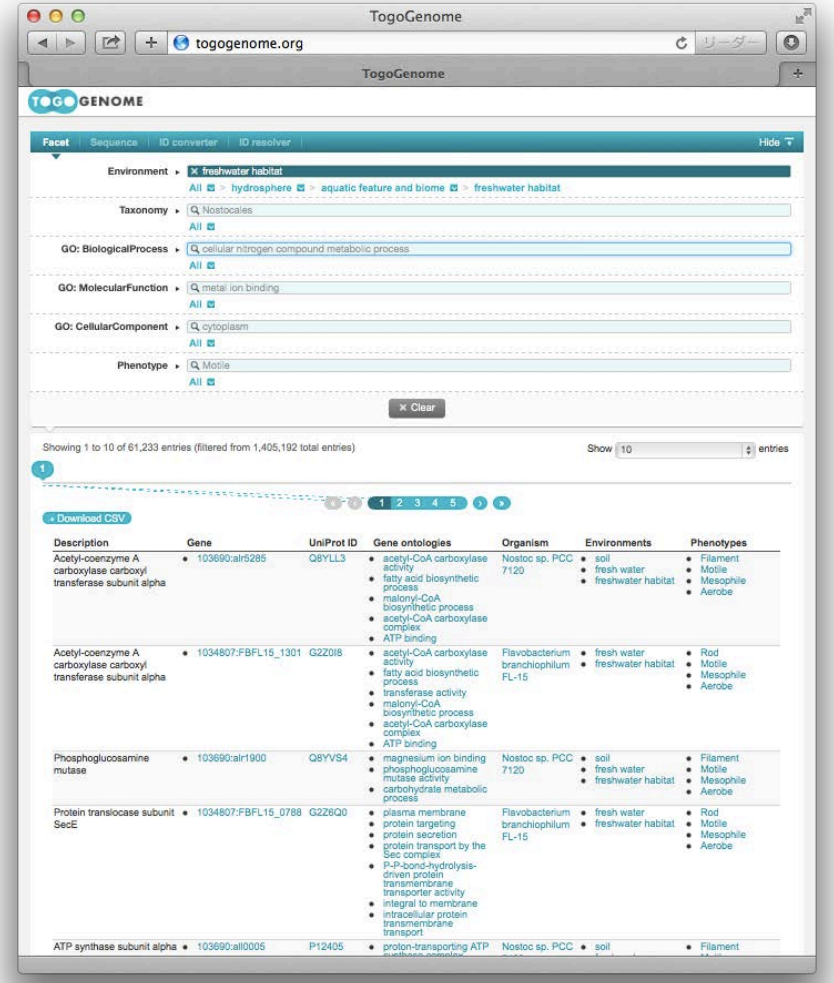
絞り込み条件軸
(オントロジー)

- 環境
- GO細胞局在
- GO分子機能
- GO生命プロセス
- 微生物表現型
- 生物分類

SPARQL
検索

検索結果を可視化した
一覧性の高い
レポート表示

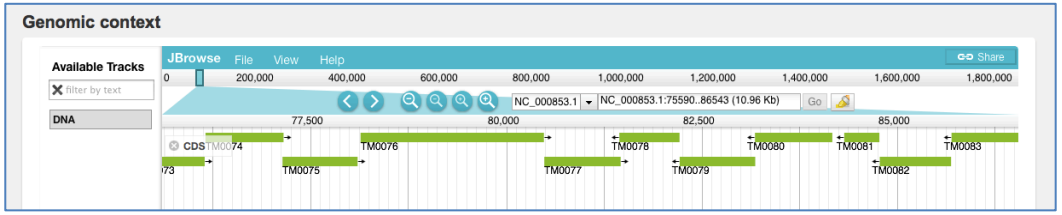
- 遺伝子情報
- 生物種情報
- 環境情報



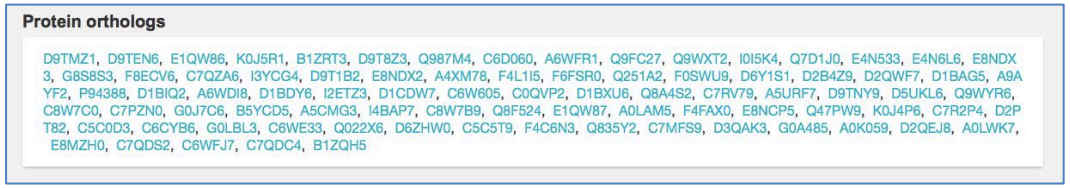
利用可能な技術/サービス(可視化): TogoStanza

SPARQL検索した結果を可視化するためのシステム

TogoGenome, MicrobeDB.jp, CyanoBase等での表示に利用されている



Genomic context スタンザ



Protein orthologs スタンザ

Nucleotide sequence スタンザ

現在，TogoGenome上では約40種のスタンザを提供

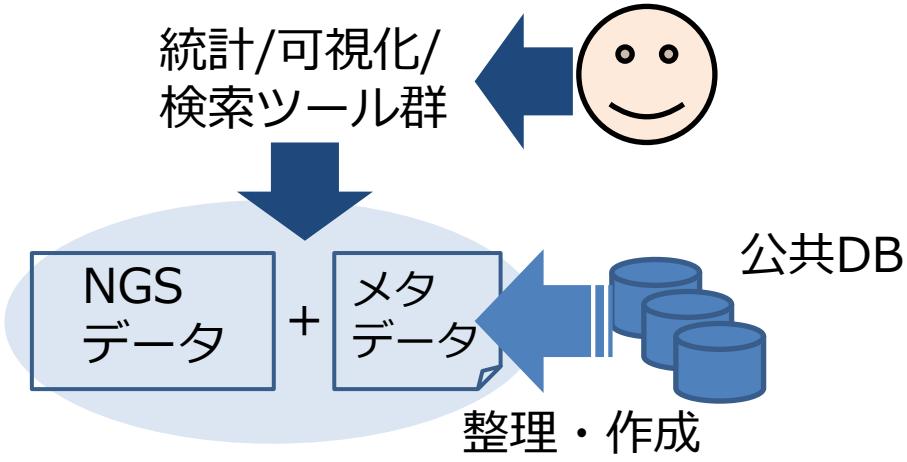
ポスター06 TogoGenome / TogoStanza の新規機能開発

実験系生物研究者ニーズに沿ったサービス開発

NGS データ
利用環境

NGS実験データを簡単に再利用／検索できる環境の構築

・メタデータの整理
・DDBJ との連携
・検索・解析ツールの開発



ポスター27 NGSデータの利用を促進する統合環境の構築とサービスの提供
28 公共NGSデータの活用を促進する検索システムの構築

利用可能な技術/サービス

NGS データ
利用環境

遺伝子発現データのリファレンス RefEx　http://refex.dbcls.jp/

RefEx

異なる実験手法の遺伝子発現データを統合し，
参照データとして整理

検索技術 GGRNA, GGGenome

統合遺伝子検索 GGRNA　超絶高速ゲノム配列検索 GGGenome

http://GGRNA.dbcls.jp/
http://GGGenome.dbcls.jp/

ポスター29 遺伝子発現リファレンスデータセット RefEx
30 統合遺伝子検索GGRNAと高速塩基配列検索GGGenome：
塩基配列データベースをすばやく検索するウェブサーバ

実験系生物研究者ニーズに沿ったサービス開発

日本語
コンテンツ

日本語コンテンツの拡充

・ライフサイエンス分野のDBやツールの動画教材
TogoTV の整備 http://togotv.dbcls.jp/

・ライフサイエンスを解説する日本語テキストの整備

http://first.lifesciencedb.jp/

http://leading.lifesciencedb.jp/

ポスター03 新しい日本語Webコンテンツ、
「新着論文レビュー」と「領域融合レビュー」
07 日本語コンテンツに対するセマンティックウェブ技術の適用

まとめ

有用な統合データベースの実現には統合化推進プログラムのデータを含め，網羅的に幅広い種類のデータをRDF化することが望ましい(Web of Data).

DBCLSはその実現へ向けて
・DBのRDF化の取り組み
・RDFでアプリケーション開発
・実験系生物研究者ニーズに沿ったサービス開発
これまでの技術の蓄積を生かして行う