

採択課題名：データベース統合に関わる基盤技術開発

データベース統合の実現に向けて 基盤技術開発 (Ⅰ)

(Ⅱ)は明日10:10～

片山 俊明 <ktym@dbcls.jp>

<http://jp.linkedin.com/in/toshiakikatayama>

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



2013/10/4 @ トーゴーの日シンポジウム 2013 ～ここまできたライフサイエンスデータベースの統合～

基盤技術開発プログラム

データベース統合に関わる基盤技術開発

研究開発課題

1. データベースの RDF による統合化
2. 解析プラットフォームによる統合利用環境の整備
3. インターネットを活用した高度検索技術の開発
4. RDF 化に資するオントロジー、辞書、コーパス整備、標準化技術開発
5. 大規模データの利用技術開発
6. 情報統合化・知識発見のためのキューレーション支援
7. 統合データベースに関わるコンテンツの作成、整備

生命科学分野におけるデータベース統合化のための基盤技術開発



ライフサイエンス統合データベースセンター

研究代表者：小原雄治

解析プラットフォームによる統合利用環境の整備



産総研・生命情報工学研究センター

研究代表者：浅井潔

データ統合と新規分野データ活用のための基盤技術開発



京都大学・化学研究所

研究代表者：五斗進

データベース統合に関わる基盤技術開発（Ⅰ）

研究開発課題

1. データベースの **RDF** による統合化
2. **解析プラットフォーム**による統合利用環境の整備
3. インターネットを活用した**高度検索技術**の開発
4. RDF 化に資するオントロジー、辞書、コーパス整備、**標準化技術開発**
5. **大規模データ**の利用技術開発
6. 情報統合化・知識発見のためのキューレション支援
7. 統合データベースに関わるコンテンツの作成、整備



そもそも RDF とは何なのか？
RDF によるデータベース統合とは？
なぜ RDF なのか？



セマンティック・ウェブ = ウェブ 3.0

RDF は次世代のインターネットといわれる**セマンティック・ウェブ**の標準データ形式

では、セマンティック・ウェブとは何か？

- すべてのモノに ID を付けましょう
 - **URI**: ID には世界中でユニークである URL を使用
- モノを説明するときは共通のコトバを使いましょう
 - **Ontology**: 意味を説明する標準語彙としてオントロジーを利用
- モノとモノの関係を記述して世界中の情報を繋げましょう
 - **Linked Data**: 上記 URI と Ontology を用いて作った RDF を公開

何のために？

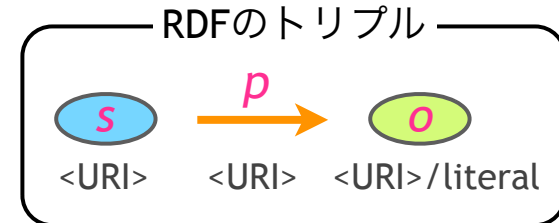
- コンピュータで知識発見をするために、曖昧な記述をやめよう！
(効率的な検索技術なしでは、もはや生きていけませんよね?)

セマンティック・ウェブ: RDF とは

- 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)

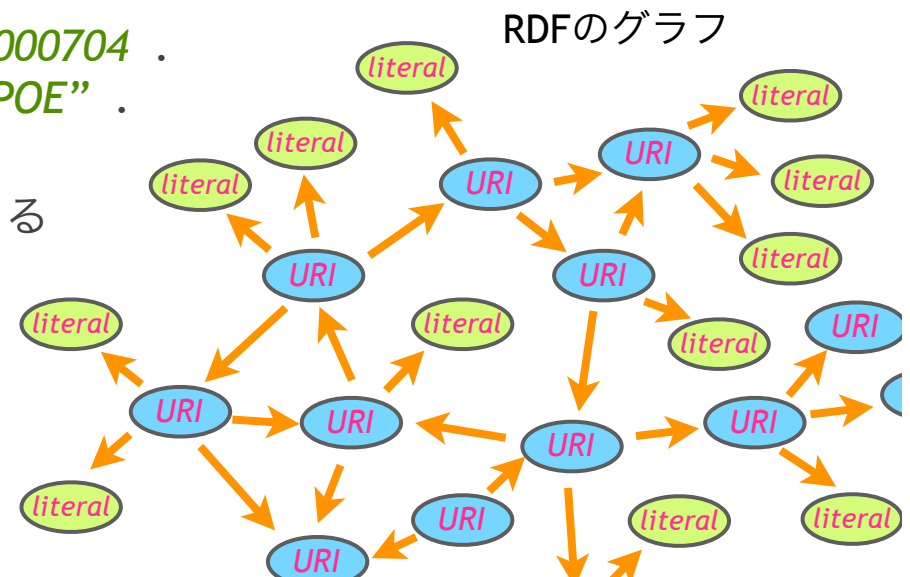


- つまり RDF は非常にシンプルなデータ形式

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

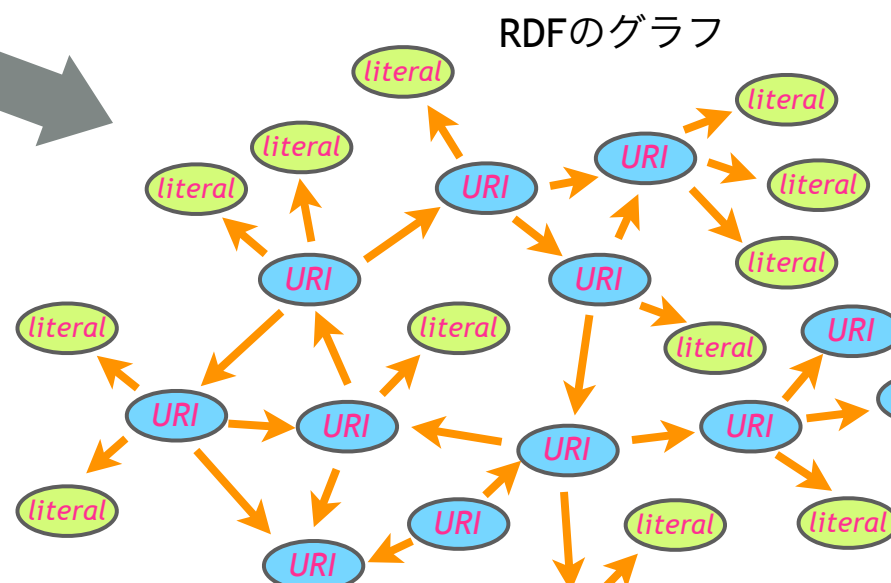
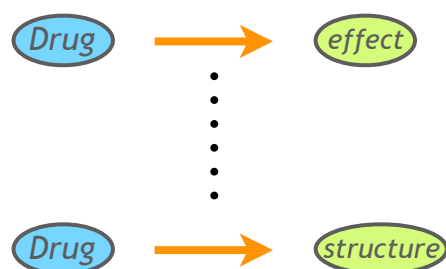
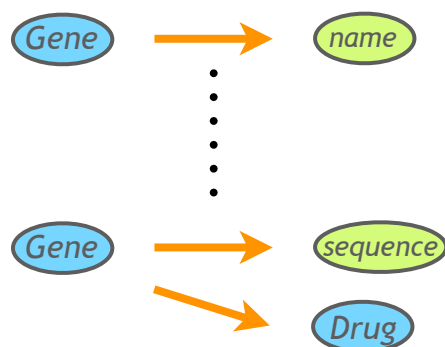
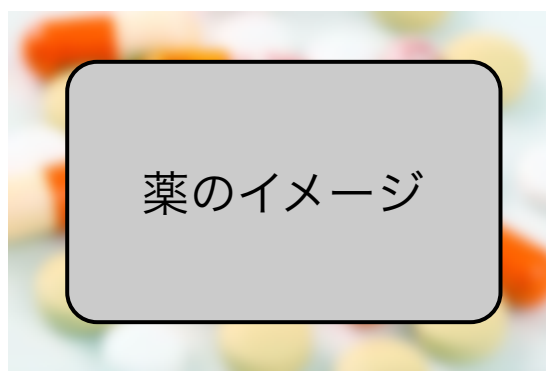
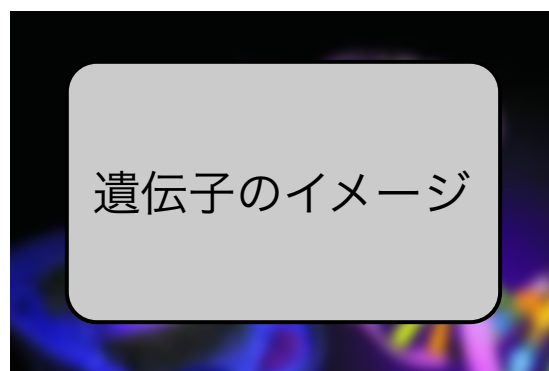
- そして RDF では多様なデータを容易に統合できる

- 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



セマンティック・ウェブ: RDF (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



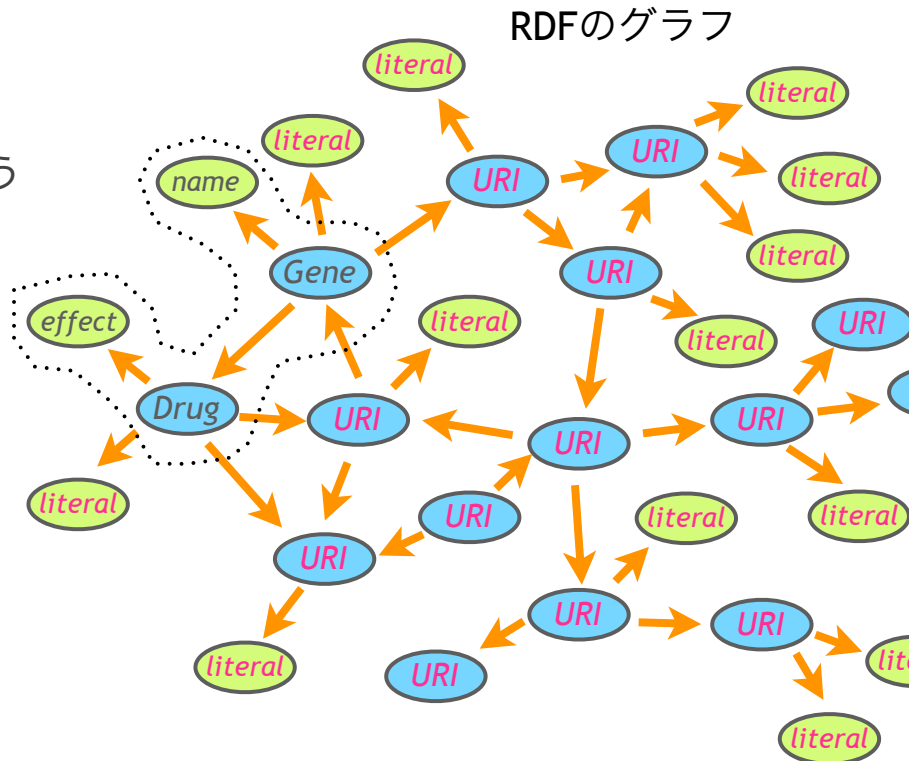
セマンティック・ウェブ: SPARQL 検索

- 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

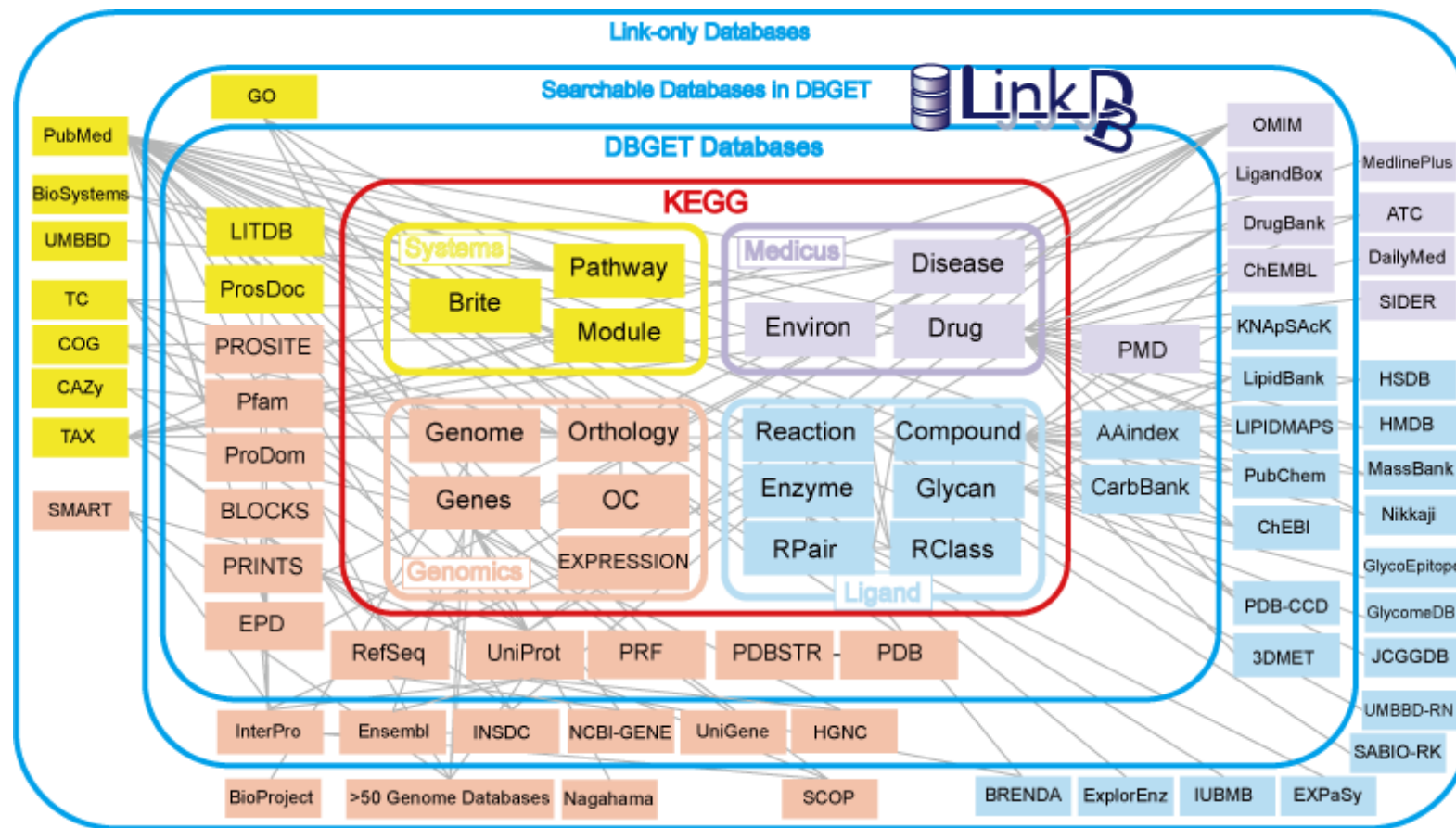
```
SELECT ?gene ?name ?drug ?effect
WHERE {
  ?gene rdf:type      so:0000704 .
  ?gene rdfs:label    ?name .
  ?gene :dose         ?drug .
  ?drug :has_effect   ?effect .
}
```



→ データベース間が有機的につながっていることが重要

京都大学：データベース間のリンク情報のRDF化

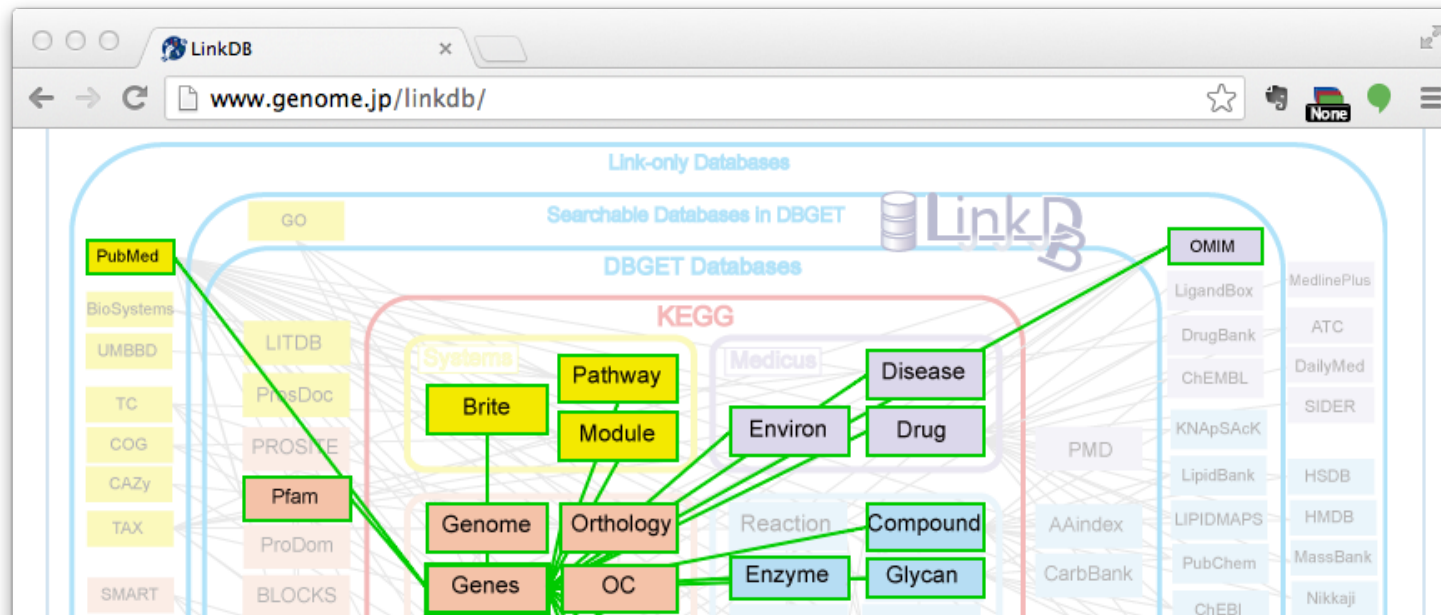
- LinkDB: 京都大学で開発されているリンク情報のデータベース



<http://www.genome.jp/linkdb/>

京都大学：データベース間のリンク情報のRDF化

- KEGG を含む国内外の様々な DB 間のリンク情報を RDF でも取得可能に



# KEGGのヒト遺伝子ID	リンクのタイプ	リンク先のパスウェイID
<http://www.kegg.jp/entry/hsa:2>	<http://www.genome.jp/linkdb/original>	<http://www.kegg.jp/entry/hsa04610> .
<http://www.kegg.jp/entry/hsa:9>	<http://www.genome.jp/linkdb/original>	<http://www.kegg.jp/entry/hsa00232> .
<http://www.kegg.jp/entry/hsa:9>	<http://www.genome.jp/linkdb/original>	<http://www.kegg.jp/entry/hsa00983> .

例：全てのヒト遺伝子からパスウェイへのリンク情報をRDF(N-triple)形式でダウンロード

Click database names to download link information between two databases

from hsa to in format

CBRC：解析ワークフローの入出力をRDF化

- 各種ツールをRDFデータを元利用できる解析プラットフォームを構築



<http://togo.cbrc.jp/>



入力：タンパク質や化合物などのRDFデータ



産業技術総合研究所
生命情報工学研究センター

ドッキングシミュレーションなどの計算

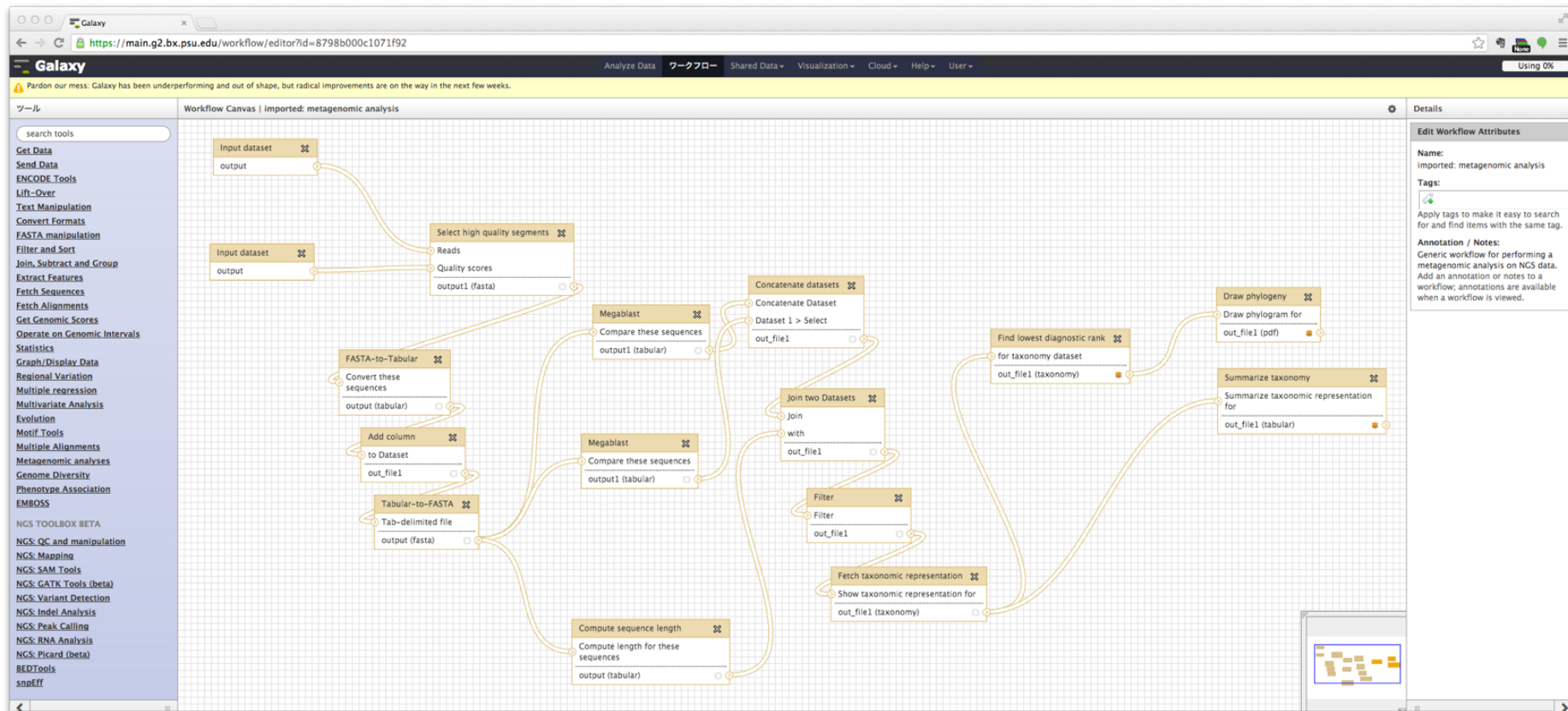


出力：解析結果のRDFデータ

SADI フレームワークを使用し
データに合ったサービスを選択

DBCCLS+京大+CBRC連携でのワークフロー構築例

- **Galaxy** - ウェブ上で解析ワークフローを構築・実行するフリーソフトウェア
 - DBCLS でも独自に拡張したものを提供中 <http://galaxy.dbcls.jp/>
 - 例：タンパク質と化合物のドッキングシミュレーションを実行してみる



DBCLS+京大+CBRC連携でのワークフロー構築例

- **DBCLS** Galaxy からインフルエンザの **ノイラミニダーゼ** の立体構造を取得
- このタンパク質と結合が知られている化合物として **シアル酸** を指定

The image displays two screenshots of the Galaxy web interface, illustrating a workflow for docking a ligand into a protein structure.

Left Screenshot: The interface shows the 'Galaxy' header with navigation tabs: 'データ解析', 'ワークフロー', '共有データ', '情報可視化', and 'ヘルプ'. The 'ワークフロー' tab is active. The 'ツール' (Tools) sidebar on the left lists various tools under categories like 'Get Data', 'DBCLS GGENOME', 'CGGenome', 'SADI SUPPORT TOOLS', 'Preparation for AutoDock', 'DBCLS TOGOWS TOOLS', 'TogoWS Search', 'DBCLS TEXT MINING TOOLS', 'OREFIL Search', 'Gendoo Table', and 'DBSearch Main'. The main panel shows a form for 'Structure ID' (2BAT) and 'Chain ID' (ex) A. Below these are buttons for 'Select chain', 'Ligand' (require SMILES format. ex) CCC, 'Find similar compounds', and 'Submit'.

Right Screenshot: This screenshot shows the same interface but with a 3D molecular model of the protein structure (2BAT) displayed in the center. The protein is shown as a yellow ribbon structure. The 'Chain ID' dropdown menu is set to 'A'. The 'Ligand' field is empty. The 'Submit' button is visible at the bottom of the form.

DBCCLS+京大+CBRC連携でのワークフロー構築例

- 京都大学の simcomp サービスでシアル酸の類似化合物を検索
- ノイラミニダーゼと類似化合物をRDFデータとしてCBRCに送信

Galaxy

galaxy.dbcls.jp/root?workflow_id=c34129969476e859

データ解析 ワークフロー 共有データ 情報可視化 ヘルプ User

Using 0%

ツール

search tools

Get Data

- Upload File from your computer
- DBCLS GGGENOME
- GGGenome Ultrafast sequence search
- GGGenome with History Ultrafast sequence search
- SADI SUPPORT TOOLS
- Preparation for AutoDock generate AutoDock Input File
- DBCLS TOGOWS TOOLS
- TogoWS Search search with a direct-input query
- TogoWS Search with History search with a query from history pane
- TogoWS Get Value of a Given Field get the value of a given field for a selected ID
- DBCLS TEXT MINING TOOLS
- OReFiL Search OReFiL search with a direct-input query
- OReFiL Search with History OReFiL search with a query from a history pane
- GenDoo Gene-to-PMIDs-to-MeSH terms relation table.
- GenDoo Table Gene/MeSH relation table on PMID
- DBCLS DBSEARCH TOOLS
- DBSearch Main 横断検索ツール

Structure ID 2BAT

Chain ID A

Select chain

Ligand CC(=O)N[C@@H]1[C@@H](O)C[C@H]2[C@@H](CO)O[C@H]2O

Find similar compounds

Compound ID	Name	Drug ID	Score	SMILES
☐ C00270	N-Acetylneuraminate	-	1.00	<chem>CC(=O)NC1C(C(C(OC1C(C(CO)O)O)(C(=O)O)O)O</chem>
☐ C19910	N-Acetyl-beta-neuraminate	-	1.00	<chem>CC(=O)NC1C(C(C(OC1C(C(CO)O)O)(C(=O)O)O)O</chem>
☐ C06241	N-Acetylneuraminate 9-phosphate	-	0.83	<chem>CC(=O)NC1C(C(C(OC1C(C(COP(=O)(O)O)O)O)O</chem>
☒ C02713	N-Acetylmuramate	-	0.81	<chem>CC(C(=O)O)OC1C(C(OC(C1O)CO)O)NC(=O)C</chem>
☐ C20418	N,N'-Diacetyllegionamine	-	0.75	<chem>CC(C1C1C(C(CO1)(C(=O)O)O)O)NC(=O)C)NC(=O)C)O</chem>
☐ C20082	Pseudaminic acid	-	0.75	<chem>CC(C1C1C(C(CO1)(C(=O)O)O)O)NC(=O)C)NC(=O)C)O</chem>
☒ C08095	Zanamivir	D00902	0.73	<chem>CC(=O)NC1C(C=C(C(OC1C(C(CO)O)O)C(=O)O)N=C(N)N</chem>
☐ C03525	O-Acetylneuraminic acid	-	0.73	<chem>CC(=O)OCC(C1C1C(C(CO1)(C(=O)O)O)O)N)O</chem>

Submit

ノイラミニダーゼの立体構造を指定

シアル酸の構造をSMILESで指定

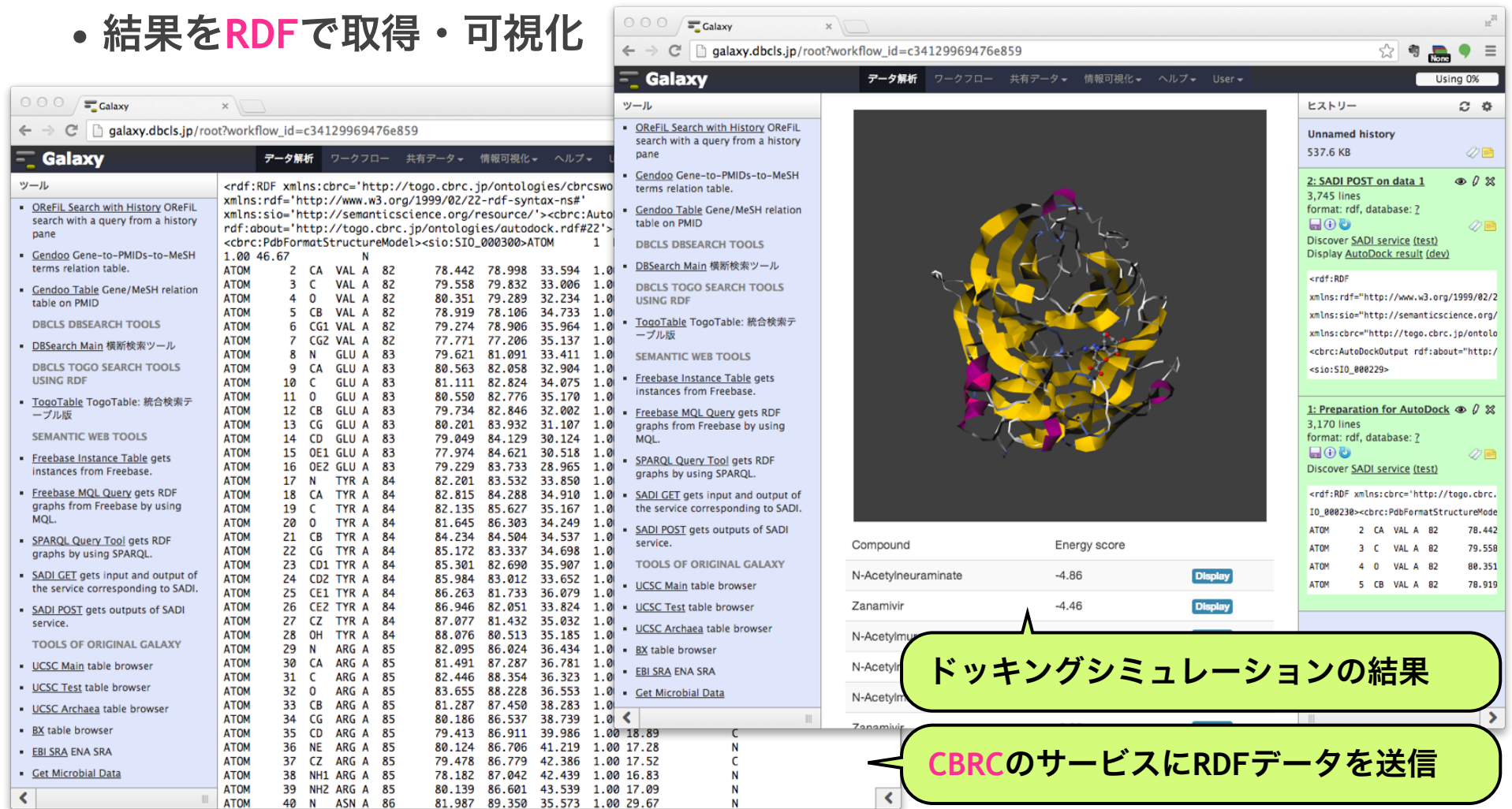
京都大学の類似化合物検索を実行

類似化合物の計算結果一覧

CBRCのサービスにRDFデータを送信

DBCLS+京大+CBRC連携でのワークフロー構築例

- **CBRC** の AutoDock サービスでドッキングシミュレーションを実行
- 結果を**RDF**で取得・可視化



The screenshot displays the Galaxy web interface for a workflow titled "galaxy.dbcls.jp/root?workflow_id=c34129969476e859". The interface is divided into several panels:

- Tool List (Left):** A list of tools available in the Galaxy environment, including "ORFSearch with History", "Gendoo Gene-to-PMIDs-to-MeSH terms relation table", "DBCLS DBSEARCH TOOLS", "DBSearch Main", "TogoTable", "Freebase Instance Table", "Freebase MQL Query", "SPARQL Query Tool", "SADI GET", "SADI POST", "UCSC Main table browser", "UCSC Test table browser", "UCSC Archaea table browser", "BX table browser", "EBI SRA ENA SRA", and "Get Microbial Data".
- Workflow View (Center):** A 3D molecular model of a protein structure, likely the target for docking. The model is shown in a yellow ribbon representation with a purple surface.
- History Panel (Right):** A list of workflow steps. The first step is "2: SADI POST on data 1", which is highlighted. The second step is "1: Preparation for AutoDock".
- Output Data (Bottom Right):** A table showing the results of the docking simulation. The table has two columns: "Compound" and "Energy score". The compounds listed are "N-Acetylneuraminic acid" and "Zanamivir". The energy scores are -4.86 and -4.46, respectively. Below the table, there is a section for "Tools of Original Galaxy" with links to "UCSC Main table browser", "UCSC Test table browser", "UCSC Archaea table browser", "BX table browser", "EBI SRA ENA SRA", and "Get Microbial Data".

Two callout boxes highlight key features:

- ドッキングシミュレーションの結果** (Result of docking simulation)
- CBRCのサービスにRDFデータを送信** (Send RDF data to CBRC service)

DBCLS+京大+CBRC連携でのワークフロー構築例

- サービスの連携や統合には標準化や相互運用性が必須



DBCLS Galaxy の RDF/SADI 対応



CBRC サービスの RDF/SADI 対応



京都大学サービスの REST API 対応



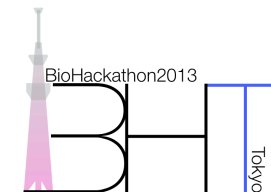
それぞれが標準仕様への対応を進めていた



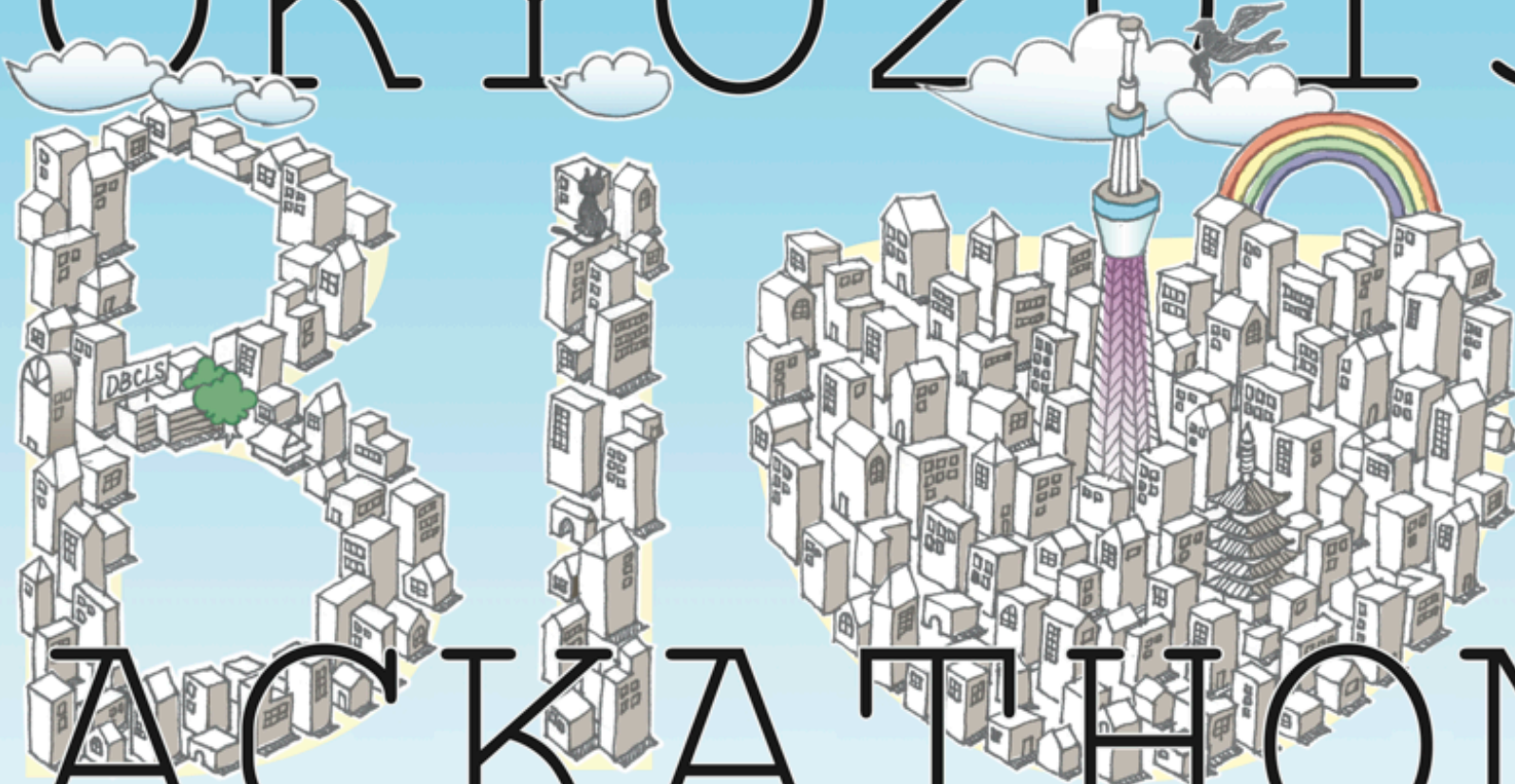
短期間でワークフロー連携できた

- さらに多くのデータやサービスと連携するためには、国際的な技術開発とオントロジーなどの標準化のための継続的な取り組みが必要

- 国際開発者会議 BioHackathon の開催へ



TOKYO 2013



HACKATHON

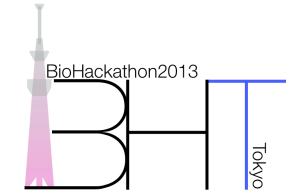
2013.biohackathon.org

Semantic interoperability and standardization of bioinformatics data and Web services

Twitter #biohack13

Mailing list <http://groups.google.com/group/biohackathon>

BioHackathon - 国際開発者会議



- 国内外の実務レベルのDB開発者が参集、1週間にわたる合宿会議を開催
 - 生命科学におけるウェブサービスやセマンティック・ウェブの標準化と技術開発
 - いま直面している技術的な課題をその場で解決

BioHackathon

= bio + hack + marathon
= bioinformatics coding camp
= emergence of
new ideas,
new technologies,
and a community

Happy hacking!



<http://2013.biohackathon.org>

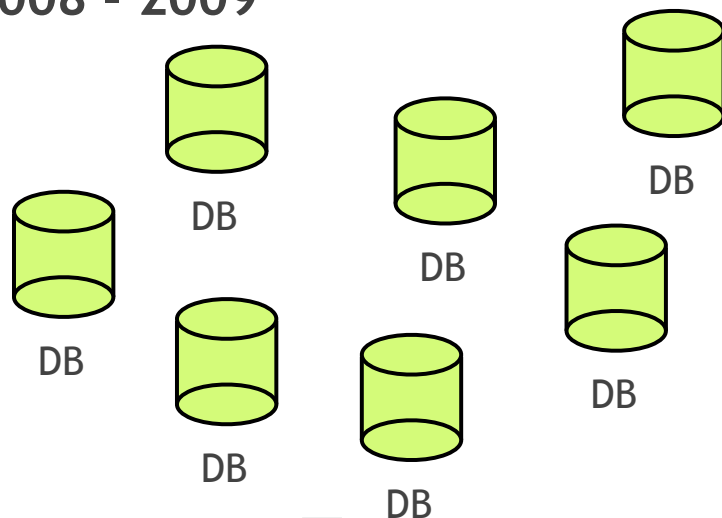
BioHackathon 2013 - 第6回 NBDC/DBCLS BioHackathon

- BioHackathon 2008 in Tokyo
 - <http://hackathon.dbcls.jp>
 - **Web Services** の標準化と相互運用性
- BioHackathon 2009 in Okinawa
 - <http://hackathon2.dbcls.jp>
 - **Web Services** による解析ワークフロー構築
- BioHackathon 2010 in Tokyo
 - <http://hackathon3.dbcls.jp>
 - **Semantic Web** による生命科学の知識統合
- BioHackathon 2011 in Kyoto
 - <http://2011.biohackathon.org>
 - 生命科学の **Linked Data** を活用するための技術開発
- BioHackathon 2012 in Toyama
 - <http://2012.biohackathon.org>
 - **Semantic Web** を利用した生命科学のアプリケーション開発



Web Services から Semantic Web へ

2008 - 2009

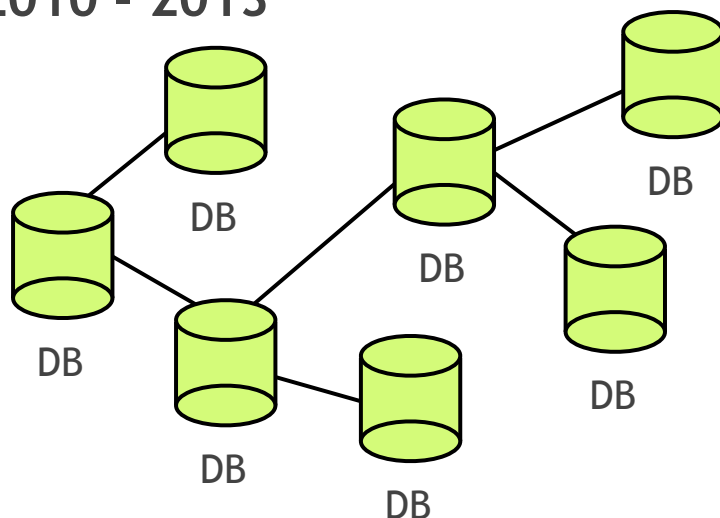


分散したDBをどのように統合・相互利用するか？

Web Services!

ハッカソン参加者

2010 - 2013

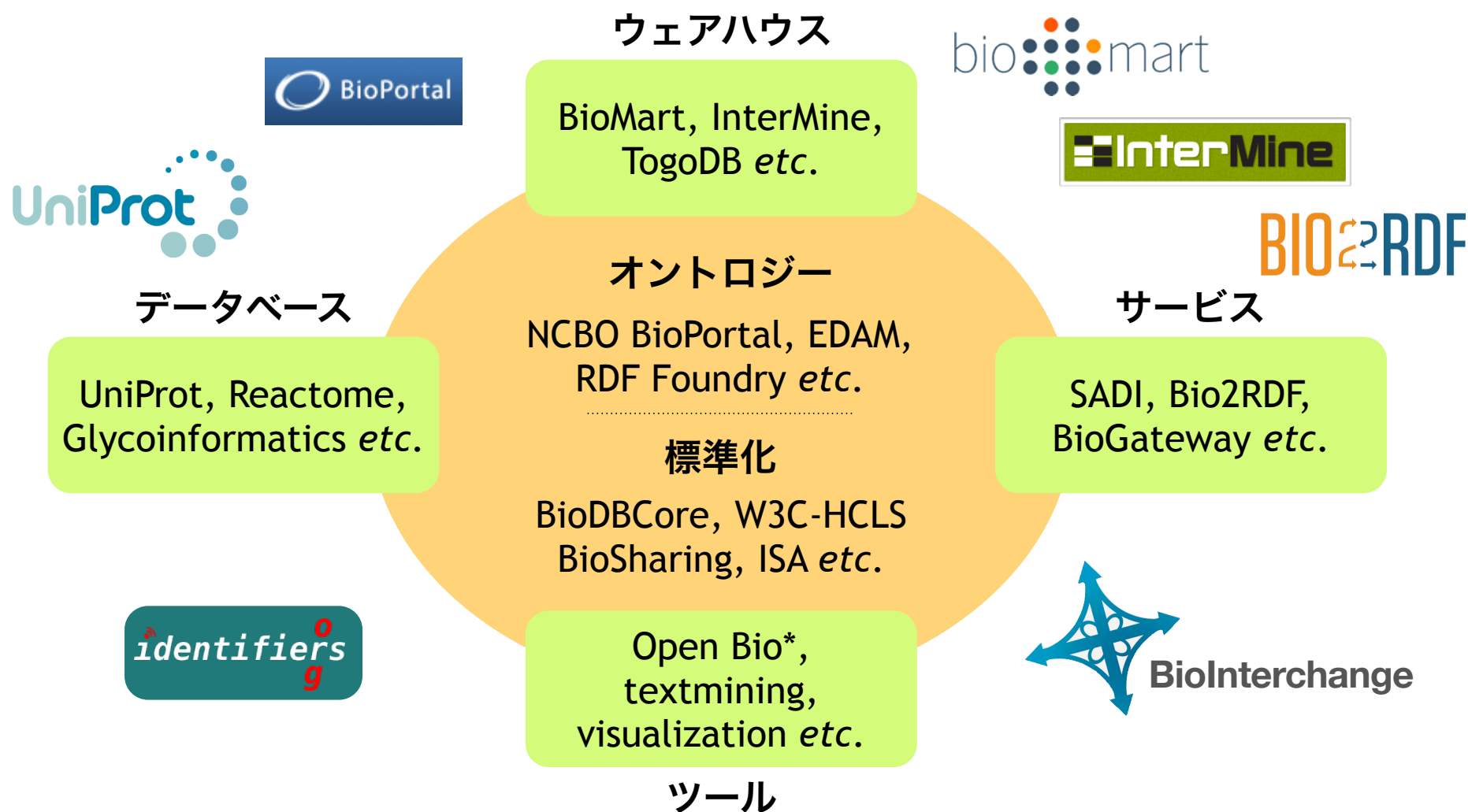
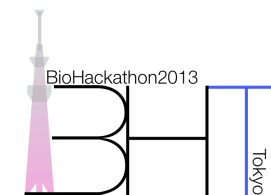


DBのコンテンツを有機的に活用するには？

Semantic Web!

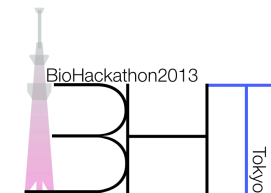
ハッカソン参加者

BioHackathon の参加グループ概要

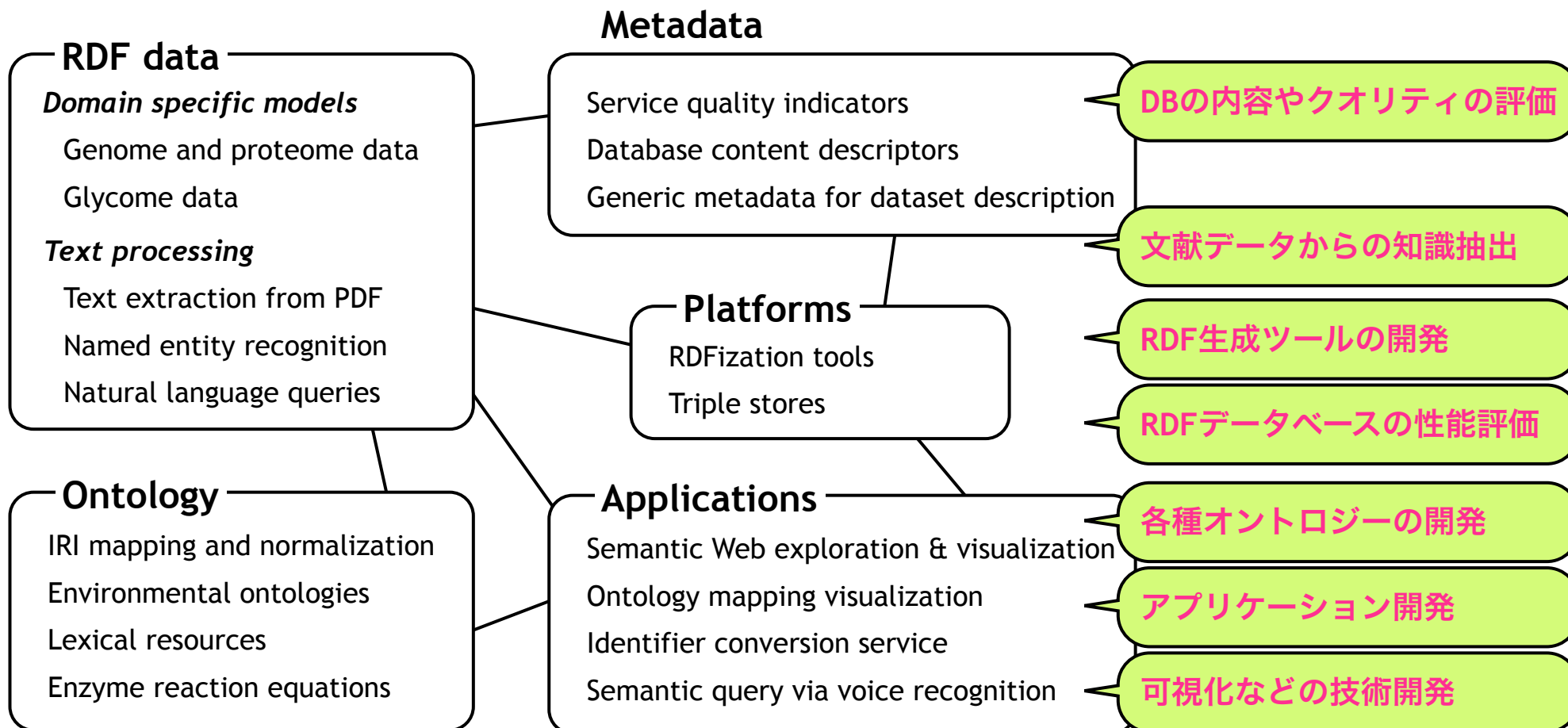


Semantic interoperability and standardization of bioinformatics data and Web services

BioHackathon 最近の成果



- 各分野のデータベースでセマンティック・ウェブ対応が加速
 - **Identifiers.org** や **BioPortal** など共通の **URI/オントロジー** 利用の促進
 - UniProt/DDBJ 共通の **配列位置情報オントロジー FALDO**



国内版 BioHackathon と SPARQLthon

- 国内版 BioHackathon - 国際版で得られた知見を国内の開発者にも還元
 - BH10.10, BH11.11, BH12.12 と3年間継続、BH13.13は次の1月に予定
 - 日本語で国内のリソースについて議論し情報共有できるメリット
 - SPARQLthon - ライフサイエンスデータベース統合推進事業の推進
 - 2012年10月から今日のトーゴーの日に向けて月1回で計12回の開催
 - RDFデータベースの構築とSPARQL検索に関わる諸技術はまだ発展途上
- 
- 統合に必要なRDFデータの生成
 - RDFで利用するオントロジーの開発
 - RDFを格納するトリプルストアの運用
 - より効率的なSPARQLの書き方
 - RDFとSPARQLを活用したアプリケーションの開発
 - スパークリングな懇親

研究課題の基盤技術とアプリケーション

- TogoDB - ユーザのデータを受け入れて DB 化、RDF 化
 - 1, 3, 6, 7. RDF による統合, 高度検索技術, 情報統合化, コンテンツ整備



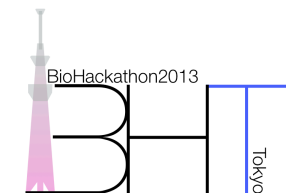
- TogoWS - ネット上の公共 DB を検索、データ取得、RDF 化
 - 3. インターネットを活用した高度検索技術の開発



- TogoGenome - 基盤となるゲノム情報を RDF で集積、DB 化
 - 1, 5. データベースの RDF による統合化, 大規模データの利用技術開発



- BioHackathon - これらに必要な技術開発と国際標準
 - 4. RDF 化に資するオントロジー、辞書、コーパス整備、標準化技術開発



- CBRC, 京大連携 - 解析プラットフォーム, 新規分野データ活用



<http://togodb.org>

TogoDB - あなたのデータをDB化・RDF化



- TogoDB - ユーザのデータを受け入れて DB 化、RDF 化

- 知識やデータは論文の中に埋まっているが表やサプリメントデータは再利用されにくい
- TogoDB では表形式のデータをアップロードするだけで高機能 DB を簡単に公開できます

- TogoDB の特徴

- 手軽に DB を作って公開したい？ → 表形式のデータを簡単に DB 化
- 共同研究者や公開時期を調整したい？ → アクセス権や公開・非公開を設定
- 見た目や作り込みにもこだわり？ → HTML/CSS/JS を柔軟にカスタマイズ
- 中身の分からない DB が多すぎる？ → 最初から中が見えているデータベース
- 検索機能が貧弱でうまく絞り込めない？ → 正規表現を含む豊富な検索オプション
- 大量のデータを見るのに一苦労？ → 矢印キーで効率的な結果のブラウジング
- プログラムから処理したい？ → Open Search 対応 REST API と JSON 出力
- **オントロジー**やメタデータを追加して **RDF** を自動生成、**SPRAQL** 検索にも対応

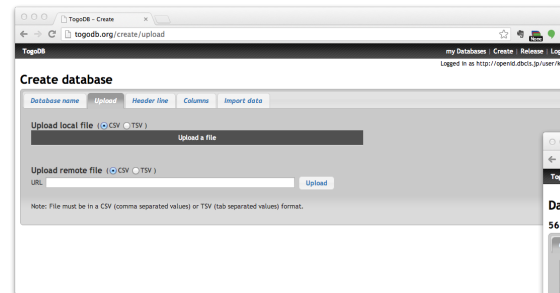
TogoDB - 表形式のデータから高機能DBを構築

TogoDB

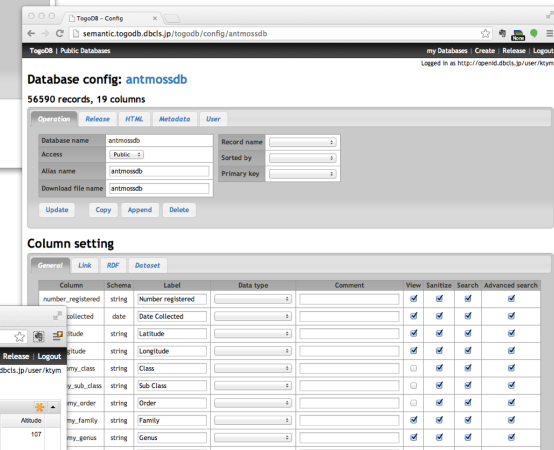
Create your DB in 5min



upload



configure



DBごとに汎用の
オントロジー管理

RDF生成
SPARQL検索

<http://togodb.org/sparql/yourdb>

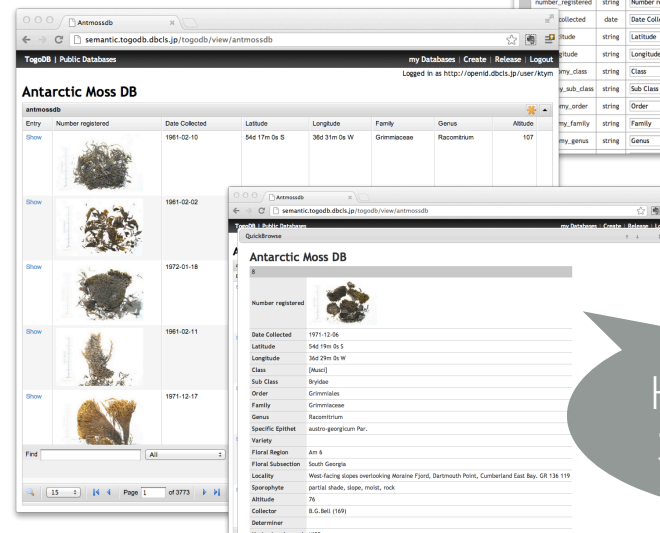
OpenSearch



分散REST検索 → Atom

<http://togodb.org/search/yourdb/query>

endpoint



deploy

HTML, CSS, JS
カスタマイズ

<http://togodb.org/entry/yourdb/123>

<http://togodb.org/db/yourdb?column=/regex/>

TogoDB - カスタマイズ・検索・高速ブラウズ

Antmossdb

togodb.org/db/antmossdb

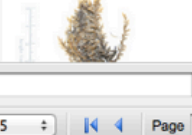
TogoDB | Public Databases

Login

Not logged in

Antarctic Moss DB

antmossdb

Entry	Number registered	Date Collected	Latitude	Longitude	Family	Genus	Altitude
Show		1961-02-10	54d 17m 0s S	36d 31m 0s W	Grimmiaceae	Racomitrium	107
Show		1961-02-02	54d 4m 0s S	37d 9m 0s W	Grimmiaceae	Racomitrium	30
Show		1972-01-18	54d 30m 0s S	36d 9m 0s W	Grimmiaceae	Racomitrium	152
Show		1961-02-11	54d 16m 0s S	36d 31m 0s W	Grimmiaceae	Racomitrium	107

Find All Download:

15

TogoDB - カスタマイズ・検索・高速ブラウズ

The screenshot shows the TogoDB web interface for the Antarctic Moss DB. The browser address bar displays `togodb.org/entry/antmossdb/1`. A green callout bubble points to the URL, stating: "1 エントリごとにURLが振られる (RDFの主語)".

The main content area features a Google Map of Antarctica with a red pin indicating the location. A green callout bubble points to the map area, stating: "表示用の HTML/CSS/JavaScript をカスタマイズ".

Below the map, there is a table with the following data:

Number registered	
1	
Date Collected	1961-02-10
Latitude	54d 17m 0s S
Longitude	36d 31m 0s W
Class	[Musci]
Sub Class	Bryidae

TogoDB - カスタマイズ・検索・高速ブラウズ

Antmossdb

antmossdb:1 - TogoDB

TogoDB | Public

Antarctic Mos

Google Map: Sh

antmossdb

Entry Number registered

Show

Google

1

Number register

Find

15

Date Collected

Latitude

Longitude

Class

Sub Class

Advanced search

Field	Type	Search Format
Number registered	string	
Date Collected	date	from to
Latitude	string	
Longitude	string	
Class	string	
Sub Class	string	
Order	string	
Family	string	
Genus	string	
Specific Epithet	string	
Variety	string	
Floral Region	string	
Floral Subsection	string	
Locality	string	
Sporophyte	string	
Altitude	integer	from to
Collector	string	
Determiner	string	
Herbarium housed	string	

正規表現を含むコラム毎の高度な検索機能

Latitude

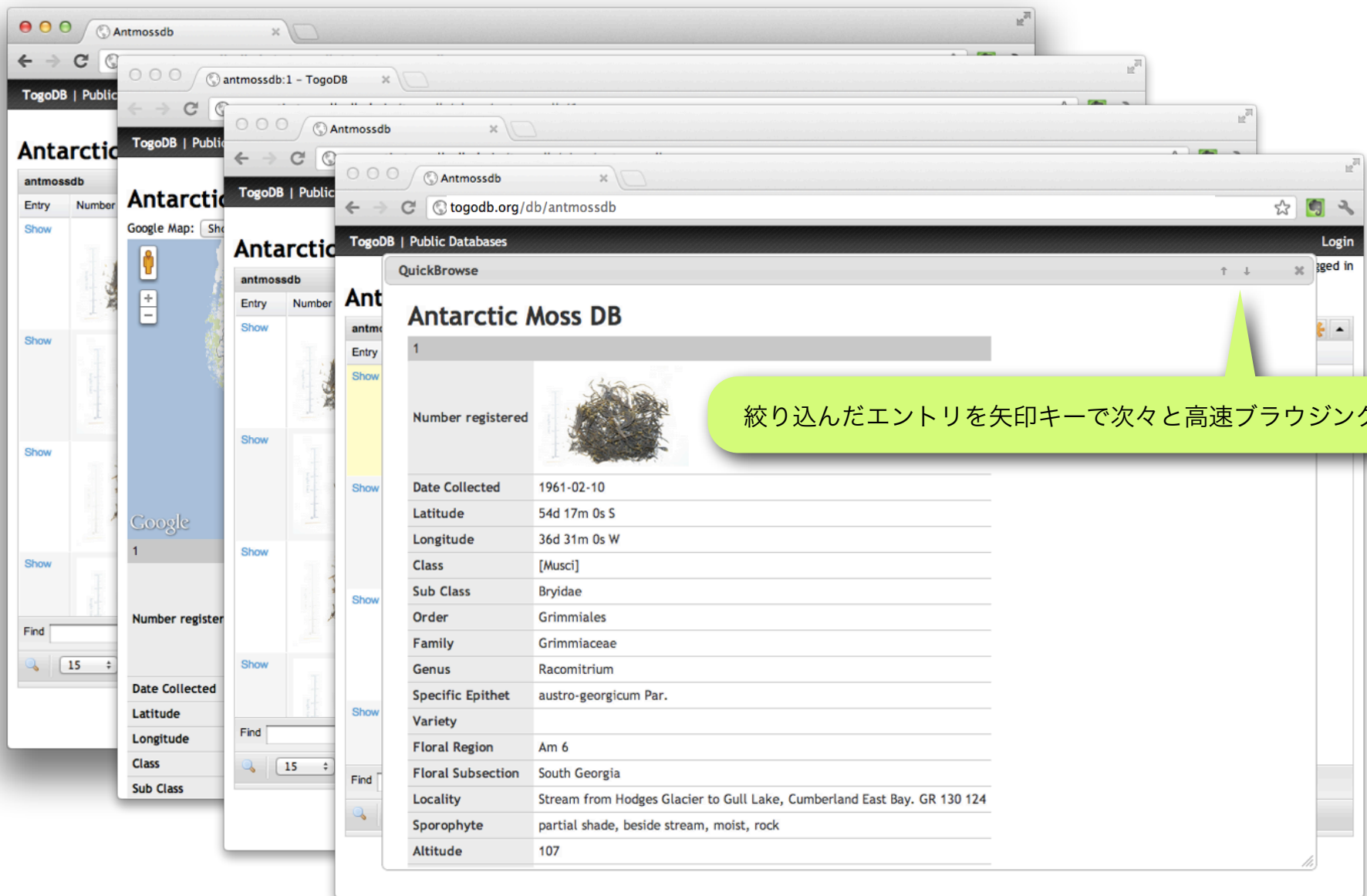
107

30

152

107

TogoDB - カスタマイズ・検索・高速ブラウズ



The screenshot displays the TogoDB web interface for the Antarctic Moss DB. The interface is designed for quick browsing and searching. A green callout bubble points to the 'QuickBrowse' feature, which allows users to navigate through entries using arrow keys.

Antarctic Moss DB

Entry	Number
1	15

QuickBrowse

Antarctic Moss DB

1

Number registered

Date Collected: 1961-02-10

Latitude: 54d 17m 0s S

Longitude: 36d 31m 0s W

Class: [Musci]

Sub Class: Bryidae

Order: Grimmiales

Family: Grimmiaceae

Genus: Racomitrium

Specific Epithet: austro-georgicum Par.

Variety:

Floral Region: Am 6

Floral Subsection: South Georgia

Locality: Stream from Hodges Glacier to Gull Lake, Cumberland East Bay. GR 130 124

Sporophyte: partial shade, beside stream, moist, rock

Altitude: 107

絞り込んだエントリを矢印キーで次々と高速ブラウジング

<http://togows.org>

TogoWS - 公共DBの検索・取得・RDF化



- TogoWS - ネット上の公共 DB を検索、データ取得、RDF 化
 - NCBI, EBI, DDBJ, UCSC, KEGG, PDBj などの DB を REST API で検索・取得
 - BioPerl, BioRuby などの機能を内蔵し、エントリをパース、JSON/RDF への変換
- これまでの課題
 - データベースごとに異なるアクセス方法
 - 結果が HTML ページでは大量処理が難しい
 - 得られた結果をさらにプログラムを書いてパースする必要がある
- TogoWS での解決
 - 全てのデータベースに同じ URL 形式 (REST API) でアクセス可能
 - 取得したデータを XML, JSON, RDF, FASTA など様々なフォーマットに自動変換
 - パーザをサーバに内蔵し、プログラミングなしにエントリの一部を取得可能

TogoWS - 公共DBの検索・取得・RDF化

• 検索

- `http://togows.org/search/db/query[/offset,limit][.format]`
- NCBI, EBI, KEGG, PDB などの主要 DB を **query** に指定したキーワードで検索

• 取得

- `http://togows.org/entry/db/id[/field][.format]`
- 主要な配列DB, KEGG, PDB, dbSNP, PubMed などのエントリ取得
- エントリ中の一部を **field** で指定して切り出せる

• RDF化

- `http://togows.org/entry/db/id[/field][.format]`
- 取得したエントリを **format** で指定して XML, JSON, RDF, FASTA などに自動変換

New!

TogoWS - UCSCからヒトゲノム等の検索・取得

• 塩基配列の取得

- [http://togows.org/api/ucsc/db/position\[.format\]](http://togows.org/api/ucsc/db/position[.format])
- e.g. <http://togows.org/api/ucsc/hg19/chr1:1,000,000-1,234,567.fasta>

```
>chr1:1,000,000-1,234,567
TGGGCACAGCCTCACCCAGGAAAGCAGCTGGGGGTCCACTGGGCTCAGGGAAGACCCCT
GCCAGGGAGACCCAGGCGCCTGAATGGCCACGGGAAGGAAAACCTACCAGCCCTCCGT
GTGTCCTCCTGGCACATGGCGACCTCCATGACCCGACGAGGGTGCGGGGCCCGGGCAGG
:
```

• アノテーションの取得

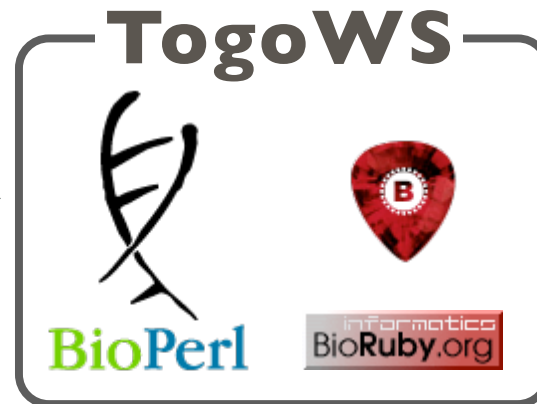
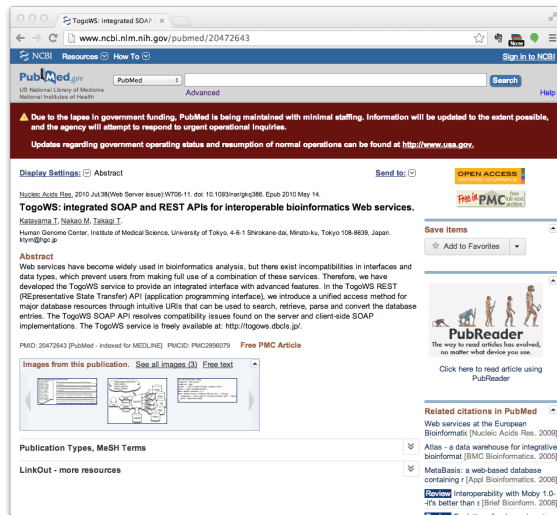
- [http://togows.org/api/ucsc/db/table/\[column=\]query\[.format\]](http://togows.org/api/ucsc/db/table/[column=]query[.format])
- e.g. <http://togows.org/api/ucsc/hg19/refGene/name2=UVSSA.json>
- e.g. <http://togows.org/api/ucsc/hg19/snp137/chrom=chr22;refUCSC=A>

```
[
{
  "bin": 595,
  "name": "NM_020894",
  "name2": "UVSSA",
  "chrom": "chr4",
  "strand": "+",
  "txStart": 1341103,
  "txEnd": 1381837,
  "cdsStart": 1341879,
  "cdsEnd": 1379749,
  "exonCount": 14,
  "exonStarts": "1341103,1341877,1343311,1345502,1346817,1348522,1348904,1360107,1369151,1369821,1373834,1374667,1377553,1379",
  "exonEnds": "1341548,1341977,1343642,1345623,1347201,1348635,1349033,1360219,1369296,1369956,1374018,1374776,1377728,138183",
  "score": 0,
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	exc
1	bin	chrom	chromStart	chromEnd	name	score	strand	refNCBI	refUCSC	observed	molType	class	valid	avHet	avHetSE	func	locType	weight	exc
2	707	chr22	16050251	16050252	rs3016036	0	+	A	A	A/T	genomic	single	by-cluster	0	0	unknown	exact	3	Mult
3	707	chr22	16050251	16050252	rs199856444	0	+	A	A	A/T	genomic	single	unknown	0	0	unknown	exact	1	
4	707	chr22	16050374	16050375	rs2844882	0	+	A	A	A/G	genomic	single	by-cluster.by-2hit-2allele	0	0	unknown	exact	3	Mult
5	707	chr22	16050739	16050740	rs111307625	0	+	A	A	-/A	genomic	deletion	unknown	0.5	0	unknown	exact	3	Mult
6	707	chr22	16051208	16051209	rs7292503	0	+	A	A	A/G	genomic	single	by-cluster.by-2hit-2allele.by-hap	0.5	0	unknown	exact	1	
7	707	chr22	16051391	16051392	rs77125914	0	-	A	A	C/T	genomic	single	unknown	0.5	0	unknown	exact	3	Mult
8	707	chr22	16051452	16051453	rs62224611	0	+	A	A	A/C	genomic	single	unknown	0	0	unknown	exact	3	Mult
9	707	chr22	16051452	16051453	rs143503259	0	+	A	A	A/C	genomic	single	by-1000genomes	0	0	unknown	exact	1	
10	707	chr22	16051496	16051497	rs11089134	0	+	A	A	A/G	genomic	single	by-cluster.by-2hit-2allele	0	0	unknown	exact	3	Mult
11	707	chr22	16051496	16051497	rs141578542	0	+	A	A	A/G	genomic	single	unknown	0	0	unknown	exact	1	

TogoWS - RDFへの変換例

- エントリをウェブサービスで取得、TogoWS サーバ内でパースして動的に変換
- <http://togows.org/entry/pubmed/20472643.ttl>



```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix dcterms: <http://purl.org/dc/terms/>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix prism: <http://prismstandard.org/purl/2003/06/prism/>
@prefix medline: <http://togows.org/ontology/medline/>
```

```
<http://pubmed.org/20472643> medline:pmid "20472643" ;
rdfs:label "pmid:20472643" ;
dc:title "pmid:20472643" ;
dc:identifier "http://pubmed.org/20472643" ;
medline:own "NLM" ;
medline:stat "MEDLINE" ;
medline:da "2010-06-25" ;
medline:dcom "2010-09-27" ;
medline:lr "2013-05-29" ;
medline:is "1362-4962 (Electrophoresis)" ;
medline:is "0305-1048 (Linkage)" ;
medline:vi "38" ;
prism:volume "38" ;
medline:dp "2010 Jul" ;
prism:publicationDate "2010-07-15" ;
medline:ti "TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services." ;
dc:title "TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services." ;
medline:pg "W706-11" ;
prism:startingPage "W706" ;
medline:ab "Web services have become widely used in bioinformatics analysis, but there exist incompatibilities in interfaces and data types, which prevent users from making full use of a combination of these services. Therefore, we have developed the TogoWS service to provide an integrated interface with advanced features. In the TogoWS REST (Representational State Transfer) API (application programming interface), we introduce a unified access method for major database resources through intuitive URIs that can be used to search, retrieve, parse and convert the database entries. The TogoWS SOAP API resolves compatibility issues found on the server and client-side SOAP implementations. The TogoWS service is freely available at: http://togows.dbcls.jp/." ;
medline:ad "Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8538, Japan." ;
medline:fau "Katayama, Toshihiro" ;
dcterms:creator "Katayama, Toshihiro" ;
medline:fau "Nakao, Mitsuteru" ;
dcterms:creator "Nakao, Mitsuteru" ;
medline:fau "Takagi, Toshihiro" ;
dcterms:creator "Takagi, Toshihiro" ;
medline:fau "Katayama, Toshihiro" ;
dcterms:creator "Katayama, Toshihiro" ;
```

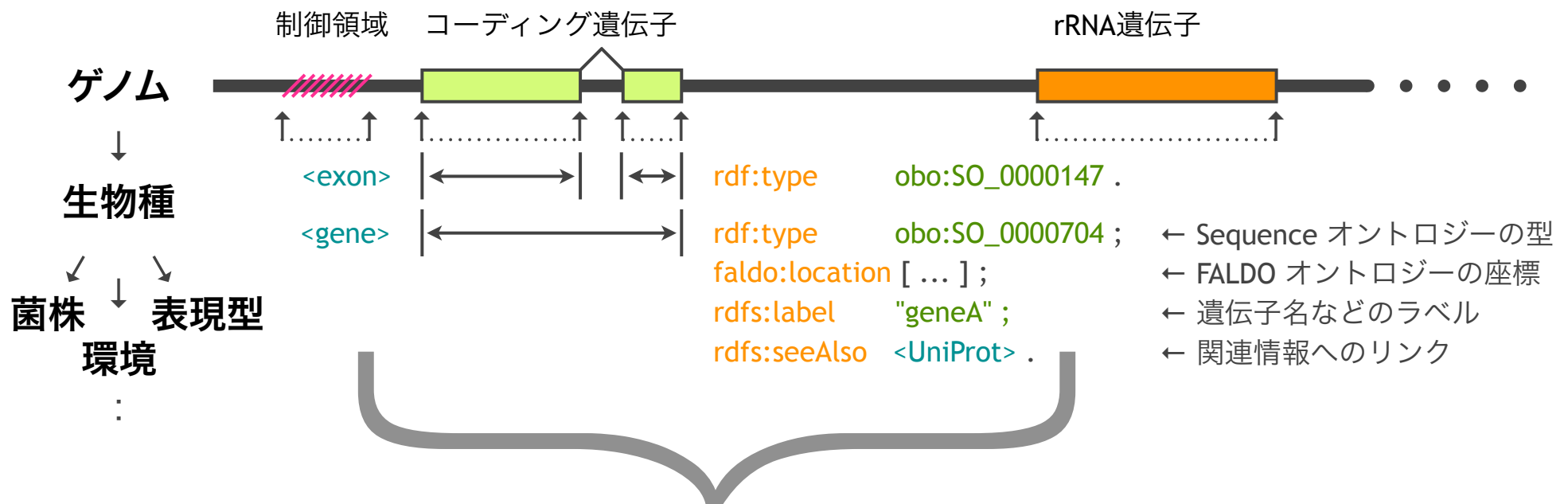


<http://togogenome.org>

TogoGenome - RDFによるゲノムのDB



- 生命科学の白地図であるゲノムにモノをマッピングし、IDと型と座標をつける
- ゲノムを中心にあらゆる情報をRDFで足しこんでゆき、関連情報にリンクする



全てのデータをRDFに統一して蓄積し、SPARQLで検索する

- セマンティック・ウェブを活かしたアプリケーションを開発

TogoGenome - RDFによるゲノムのDB



• データとオントロジーの整備

- RefSeq のゲノム配列データ
- UniProt のタンパク質アノテーション
- INSDC と Taxonomy のオントロジー
- FALDO 配列座標オントロジー
- GOLD の生物種サンプル情報
- MEO, MPO, GMO, MCCV 環境・微生物表現型・培地・菌株ストックのオントロジー
- BioPortal/OBO の各種オントロジー
- :

足りないオントロジーやデータは自前で整備

• セマンティック・ウェブでのアプリケーション開発

- データを様々な切り口で絞り込めるファセット検索

TogoGenome - RDFファセット検索



TogoGenomeでは全ゲノム情報に対し複数のオントロジーをファセットとして用いた絞り込み検索が可能で、検索結果から環境・生物種・遺伝子などのレポートを表示

活用例は明日のプレゼンで

Showing 1 to 25 of 868 entries (filtered from 1,292,481 total entries)

Description	Gene	UniProt ID	Gene ontologies	Organism	Environments
ATP synthase gamma chain	• ali0004	P12408	• ATP synthesis coupled proton transport • ATP binding	Nostoc sp. (strain PCC 7120 / UTEX 2576)	• fresh water • freshwater habitat • soil

1,292,481 total

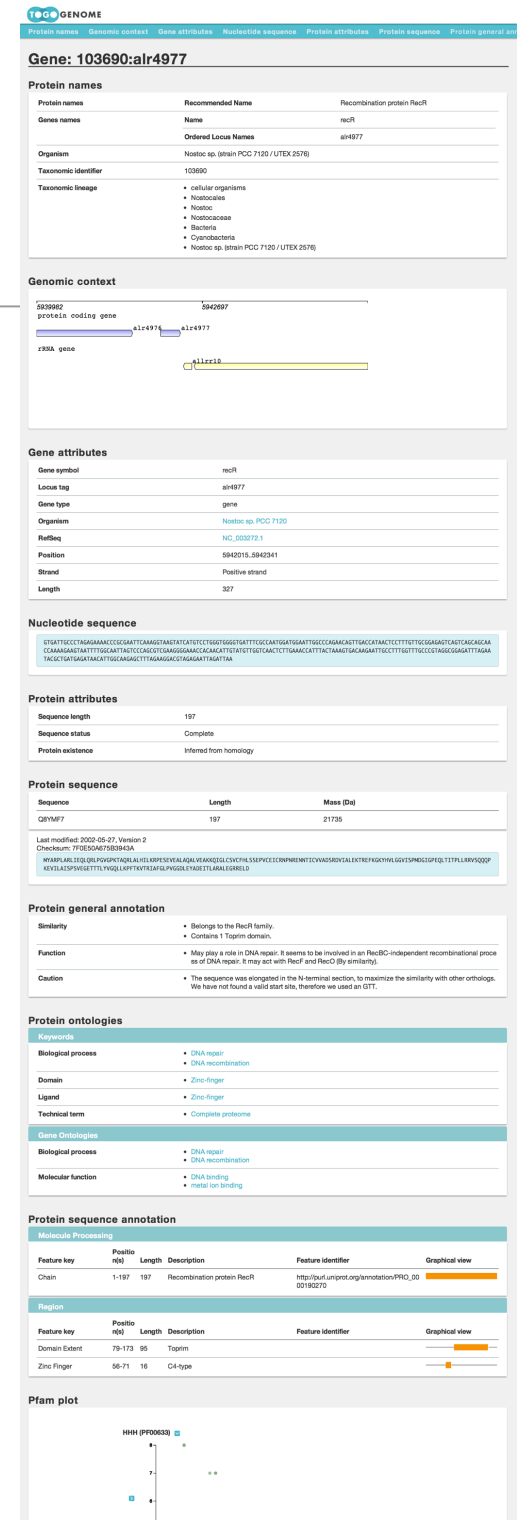
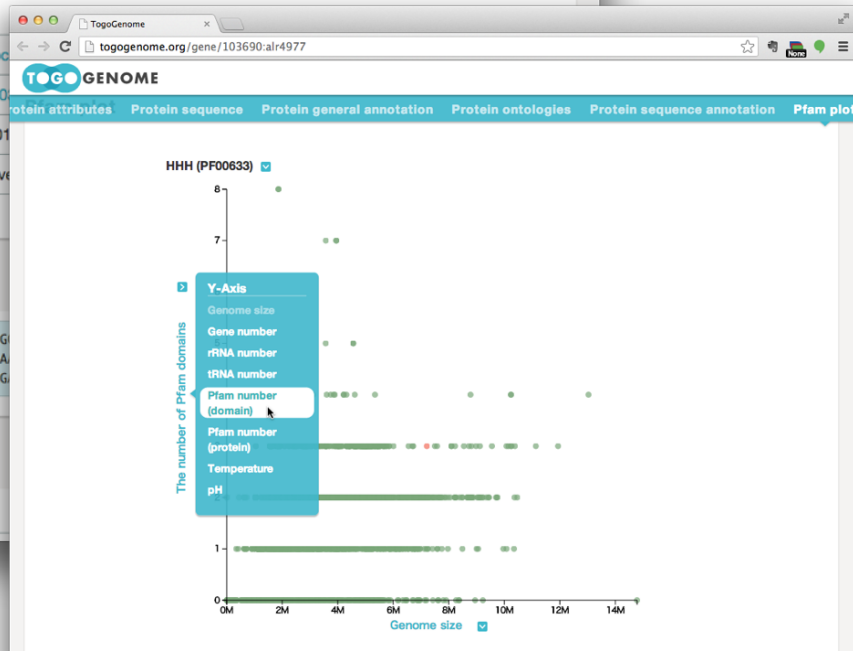
Show 25 entries

29

All columns Search:

UniProt ID	Gene ontologies	Organism	Environments
88YMF7	• DNA repair • DNA recombination • DNA binding • metal ion binding	Nostoc sp. (strain PCC 7120 / UTEX 2576)	• fresh water • freshwater habitat • soil
88YME4	• DNA dealkylation involved in DNA repair • methylated-DNA-[protein]-cysteine S-methyltransferase activity • cytoplasm	Nostoc sp. (strain PCC 7120 / UTEX 2576)	• fresh water • freshwater habitat • soil
88YMD4	• glutamine metabolic process • 'de novo' CTP biosynthetic process • CTP synthase activity • ATP binding	Nostoc sp. (strain PCC 7120 / UTEX 2576)	• fresh water • freshwater habitat • soil
88YME6	• ATP synthase coupled electron transport • NADH dehydrogenase (ubiquinone) activity • quinone binding • integral to membrane	Nostoc sp. (strain PCC 7120 / UTEX 2576)	• fresh water • freshwater habitat • soil

再利用可能なカタチで提供 (MicrobeDB.jp等と共同開発)



TogoStanza - ウェブページへの埋め込み



```
<!DOCTYPE html>
```

```
<head>
```

```
<title>TogoStanza example</title>
```

```
<script src="jquery.min.js" type="text/javascript"></script>
```

```
<script src="http://togogenome.org/stanza/assets/stanza.js" type="text/javascript"></script>
```

```
</head>
```

```
<body>
```

```
<div data-stanza="http://togogenome.org/stanza/protein_names"
```

```
data-stanza-tax-id="103690" data-stanza-gene-id="alr4977"></div>
```

```
</body>
```

基本的には1行HTMLを書くだけでスタンザの埋め込み完了！

The screenshot shows the TogoGenome website interface. The browser address bar displays `togogenome.org/gene/103690:alr4977`. The page title is "Gene: 103690:alr4977". The main content area is titled "Protein names" and contains a table with the following data:

Protein names	Recommended Name	Recombination protein RecR
Genes names	Name	recR
	Ordered Locus Names	alr4977
Organism	Nostoc sp. (strain PCC 7120 / UTEX 2576)	
Taxonomic identifier	103690	
Taxonomic lineage	Bacteria; Cyanobacteria; Nostocaceae; Nostoc sp. (strain PCC 7120 / UTEX 2576)	

✓スタンザのURLに `/help` をつけるとパラメータのヘルプが表示される

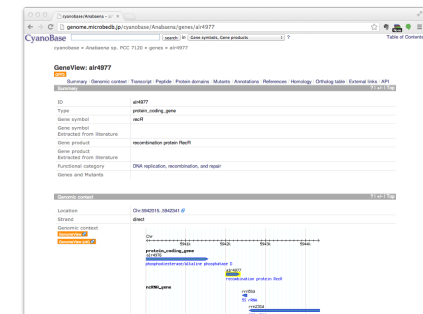
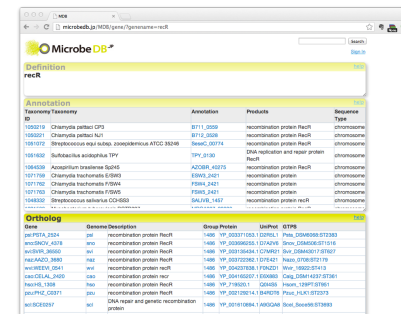
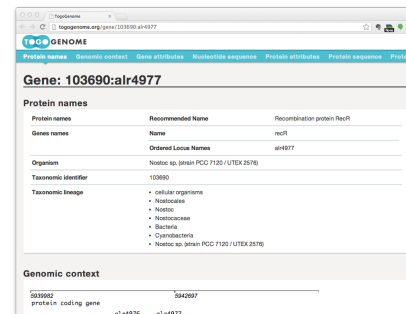
✓スタンザのURLを `stanza/` で終わるとスタンザの一覧が表示される

TogoStanza - スタンザサーバの開発と利用



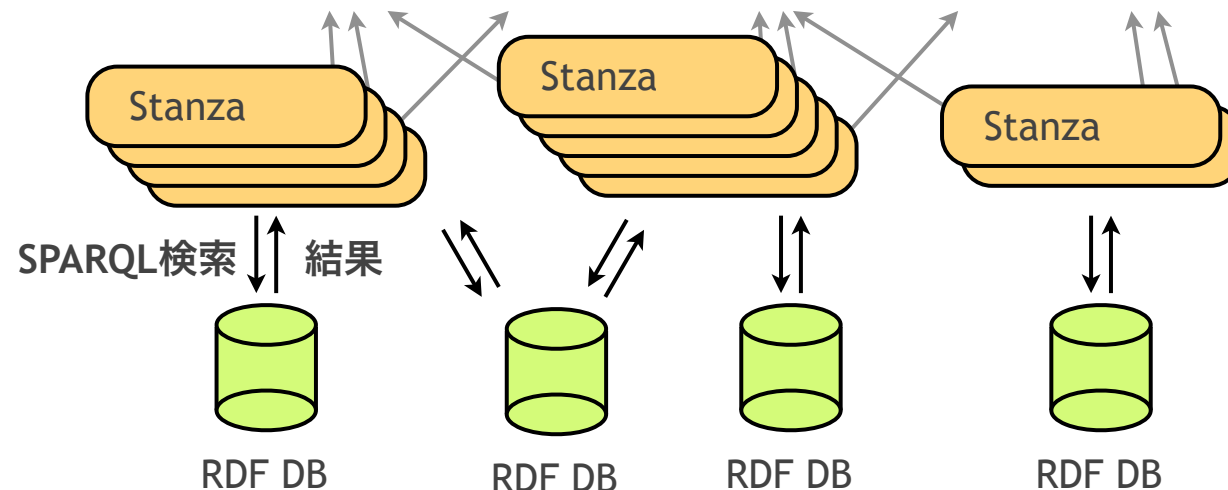
スタンザのユーザ

各地のサーバから自由に
スタンザを組み合わせて
自分のDBで利用



スタンザの開発者

スタンザサーバを提供
SPARQL検索の結果を
HTMLのIFRAMEで可視化



→ DBCLSではスタンザを容易に作成しサーバを公開するためのフレームワークも開発・提供

TogoGenome - RDF利用のメリット

- **多種多様なデータをユニークな ID (URI) とゲノム座標にもとづいて集積可能**
 - Linked Open Data (LOD) では地理情報を Google map などに集約した成功例が多数
 - TogoGenome では、ゲノムの白地図としての座標系とリファレンスデータセットを提供
- **ファセット検索でデータを絞り込み目的のデータセットを取得**
 - キーワードや ID では困難なオントロジーを活用した多様な切り口での検索を可能に
- **世界標準となるオントロジーやデータの開発と提供**
 - これまで暗黙知だった作成者やデータ形式に依存する「意味」を明示的に分かりやすく
 - 塩基配列DBでも INSDC (米NCBI, 欧EBI, 日DDBJ) の RDF 化を DDBJ/DBCLS 発で推進

TogoGenome - 今後の課題



• セマンティック・スタティスティクス

- データが統合されてはじめて可能となる横断的な統計解析

• 統合すべきデータの拡充と標準化

- ヒトゲノムを含む真核生物のサポートとドラフトゲノムへの対応
- UniProt, NCBO などとの国際連携による RDF データの標準化

• キュレーション

- Nanopublication による研究者個々人のアノテーション集約

• データ更新

- バージョンコントロール (hg19→hg20問題など)

