第1回 RDF講習会@JST東京本部別館

DBCLS
Database Center
for Life Science

NBDC, DBCLS における RDF への取り組み

RDF によるデータ統合

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
川島秀一

2016/10/7

生命科学データベースの現状

DBCLS
Database Center
for Life Science

■ データベースの数

- ◆ 世界全体で10,000〜20,000
- ◆ NAR誌のデータベースリストに1,662（2016/6現在）
- ◆ NBDC Integbio DBカタログ（日本のDB）に1,555（2016/6現在）

■ データベースの種類

- ◆ NAR誌のデータベースカタログの分類15カテゴリ40サブカテゴリ

■ ゲノムプロジェクトの数

- ◆ （GOLDデータベース）89,744（2016/6現在）

生命科学データの利活用に関する問題点

■生命科学は研究目的、対象、データの種類、解釈が非常に多様

■無数のデータベースがある

■複数のデータベースを統合して扱う必要がある（データの統合利用）

■どのように得られたデータなのか、どのようにアノテーションを行ったのか、等のメタデータが不十分（信頼性の程度が不明）

■データのフォーマットがデータベース毎に違う

■使われている用語や概念がそれぞれに異なる

■データの文脈依存性、曖昧性、冗長性、複雑性

DBCLSのミッション

■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**するデータベースの研究開発

■**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

セマンティックウェブの採用

セマンティックウェブとは

■ これまでのウェブが、人間が読むことを前提として構築されていたのに対して、**機械が利用する**ことを前提としたウェブ。

■ 様々な技術を利用するが、特に中心となるのが、RDF、OWL、SPARQL。

■ RDF（Resource Description Framework）により情報を記述する。

■ OWL（Web Ontology Language）によりオントロジーを記述する。

■ SPARQL（SPARQL Protocol and RDF query language）によりRDFデータに問い合わせを行う。

生命科学のデータ統合にRDFを利用する利点

DBCLS
Database Center
for Life Science

- フォーマットが共通になる

  ☑ **データ統合に有利**

- データの意味が明確になる

  これまでの大半のデータベースは、計算機には意味が不明確

  従って人の介在なしにデータをつなげることは事実上不可能

  ☑ **データ統合に有利、データ処理の自動化が可能**

- W3C標準

  ☑ **標準規格があるので、特定の企業の意向等に影響されにくい**

- 様々な分野で、また国際的にも、利用されている

  ☑ **統合したデータ活用の可能性が著しく高まる**

  ☑ **国際連携がやりやすい**

RDFを採用する利点　フォーマットの共通化

■ バイオインフォマティシャンのエフォートの多くは、データの取得/整形に割かれている。

データベースエントリーのパーズ、IDの対応付け等

☑ RDFでは不要！

■ RDFはグラフ構造なので、**データスキーマの変更に強い**

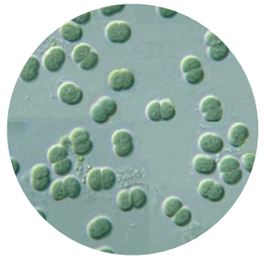
RDBで、スキーマ変更が発生するデータの追加は、高コスト

☑ RDFでは新規のデータもそのまま追加できる

RDFを採用する利点　データの意味の明示化

- URI は、グローバルなID
- グローバルなID＝あるURIを、世界のどこで閲覧しても、同じ情報が得られる。
- 異なる組織が、独自にデータを生成したとしても、グローバルなIDなので、矛盾が起きない。
- 1つのRDFは、URIのリンクであり、それだけで自立した情報（文脈から切り離されることがない）として扱える。

RDFを採用する利点　データの意味の明示化



Synechocystis PCC 6803 が、指す対象が、データベース毎に（さらに時代毎に）変化している例

もともとのタクソノミーIDは、1148

Synechocystis PCC 6803

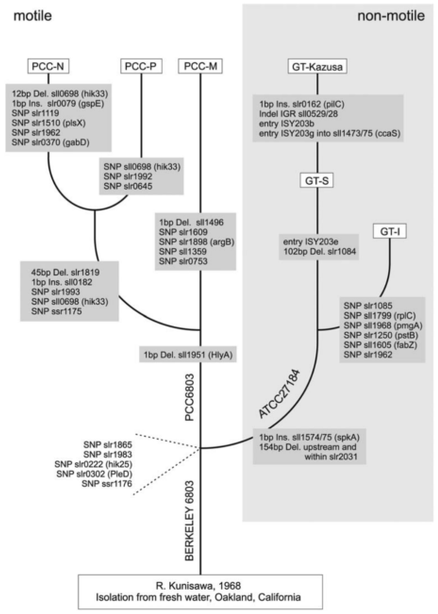
(当初株扱いだったものが種に格上げされ)

→ **UniProt では、1111708**
名前も Synechocystis sp. (strain PCC 6803 / Kazusa)に変更

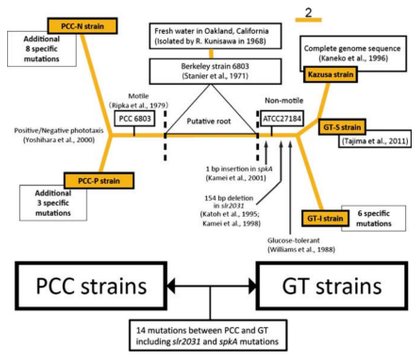
→ **NCBI では、不明**

→ **KEGGでは独自ID、T0004**
（タクソノミーIDは1148のまま）

←現状、ゲノム配列の観点から見るとたぶんこう



Trautmann *et al.* (2012)



Kanesaki *et al.* (2011)

Synechocystis PCC 6803 の場合、最初に単離され、その後パスツールの菌株コレクションに入ってから、世界のラボでどのようにゲノムに変異が蓄積されてきたかについて、論文が報告されており、これらを読んでデータベースと突き合わせれば、人間ならば、状況が把握できるはず。

既存のDBは、各々閉じた世界では整合性があるが、他のDBと統合する事は、実は単純な問題ではないことが多い。

RDFでは、URIの利用やメタデータの重視、共通語彙（オントロジー）を利用等から、こういった問題を回避しやすい。

RDFを採用する利点　データ活用の可能性の向上

同じデータ形式が、多方面で採用されることで、想定していなかった統合も低コストで可能になる

PMDAが、CDISCによる申請の義務化

CDISC: Clinical Data Interchange Standards Consortium

平成26年度よりパイロット実施

1000万人規模の大規模医療情報データベースを目指す

一方、CDISCは、現在RDFバージョンの仕様策定

2015年２月にパブリックレビューの開始

http://www.cdisc.org/standards/dataexchange

⇒ 将来、RDF形式で大規模な医療情報が利用できる可能性がある

RDFを採用する利点　生命科学DBの国際的な歴史

DBCLS
Database Center for Life Science

2001　セマンティックウェブの提唱

2006　UniProt　タンパク質

2008　BIO2RDF

2011　PDBj Protein Data Bank Japan　タンパク立体構造

2013　TOGO GENOME

2014　EMBL-EBI　パスウエイ　バイオアッセイ　遺伝子発現

PubChem　化学化合物

MESH　医学用語

2015　RDF portal

Ensembl　ゲノム

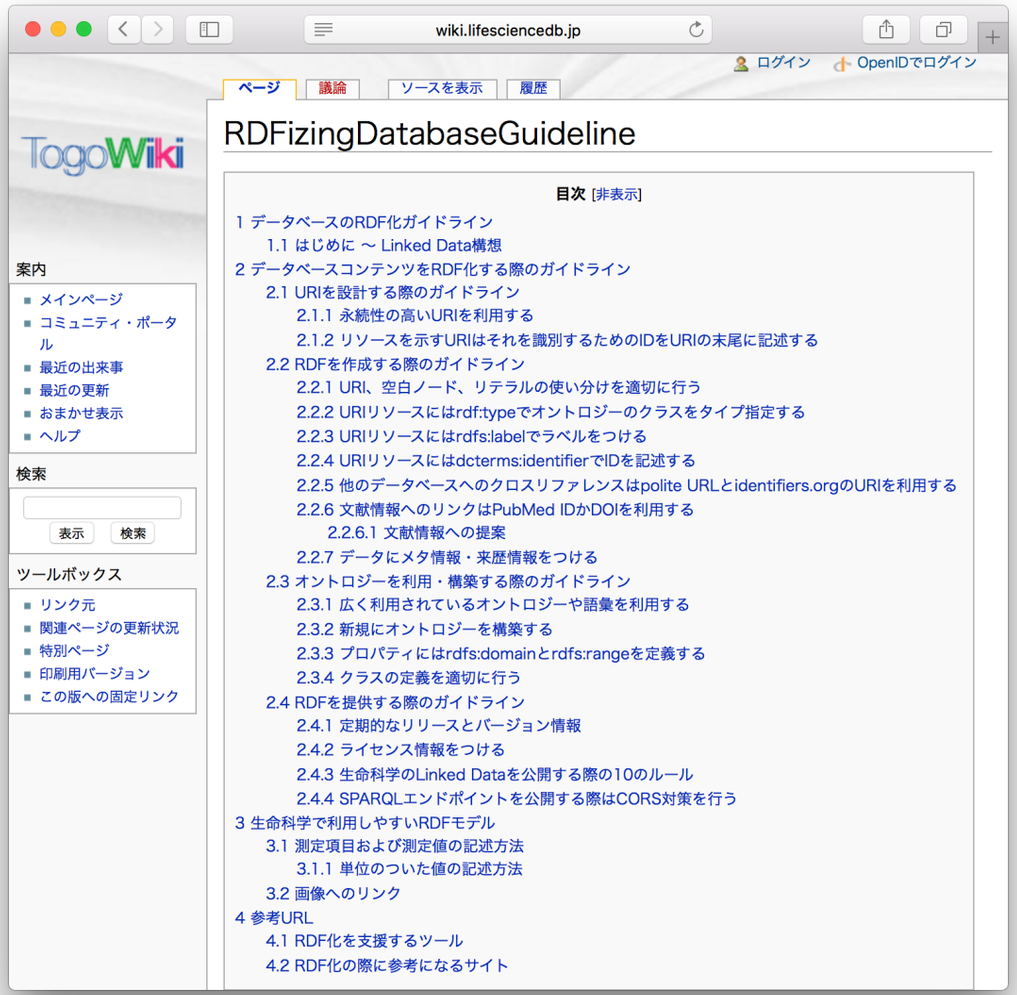
2016　DDBJ Coming soon !

データベースのRDF化ガイドライン

■ DBCLSにおいて、RDF化を行う際に参考にするとよい指針について、特に初心者向けにまとめたもの。

■　定期的に行っているハッカソンイベント等で得られた見地をもとに、**Wiki**上で構築している。

■　特に、頻繁に記述するタイプの情報（例えば、エントリーID、他DBへのクロスリファレンス、文献情報、ライセンス等）について、どのオントロジーを使うのがよいか、提示してある。



http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline

データベースのRDF化ガイドライン

DBCLS
Database Center
for Life Science

DBCLSにおいて、RDF化を行う際に参考にするとよい指針について、
特に初心者向けにまとめたもの。

■RDFの仕様を守る

■URIリソースにrdf:type でオントロジーを指定する

■URIリソースにrdfs:label でラベルをつける

■URIリソースにdcterms:identifierでIDを記述する

■他のデータベースへのクロスリファレンスにidentifiers.orgのURIを仕様する

■データセットへのメタ情報の付与

■文献情報のリンクはPubMed IDかDOIを利用する

■プロパティには、rdfs:domain と refs:rangeをできるだけ定義する

■RDFの仕様を守る

当たり前のように思えますが、初心者がやりがちな間違いとして、つぎのようなものがあります。

- リテラルを主語に使ってしまう
- オントロジーを検索してきて、それがプロパティなのに、目的語として使ってしまう
- 適切なデータ・タイプを使っていない

データベースのRDF化ガイドライン

■他のデータベースへのクロスリファレンスにidentifiers.orgのURIを仕様する

生命科学データベースで、同じリソースを指すURIが複数存在することがあり、
RDFの作成者は、そのうちのどれを利用するか分からないため,
identifiers.orgのURI を使うことで、データが繋がらなくなる問題が避けられる

http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2153
http://pubchem.ncbi.nlm.nih.gov/compound/2153
http://identifiers.org/pubchem.compound/2153

http://www.uniprot.org/uniprot/Q9F9F2
http://purl.uniprot.org/uniprot/Q9F9F2
http://identifiers.org/uniprot/2153

RDFガイドラインに紹介されている語彙・オントロジー

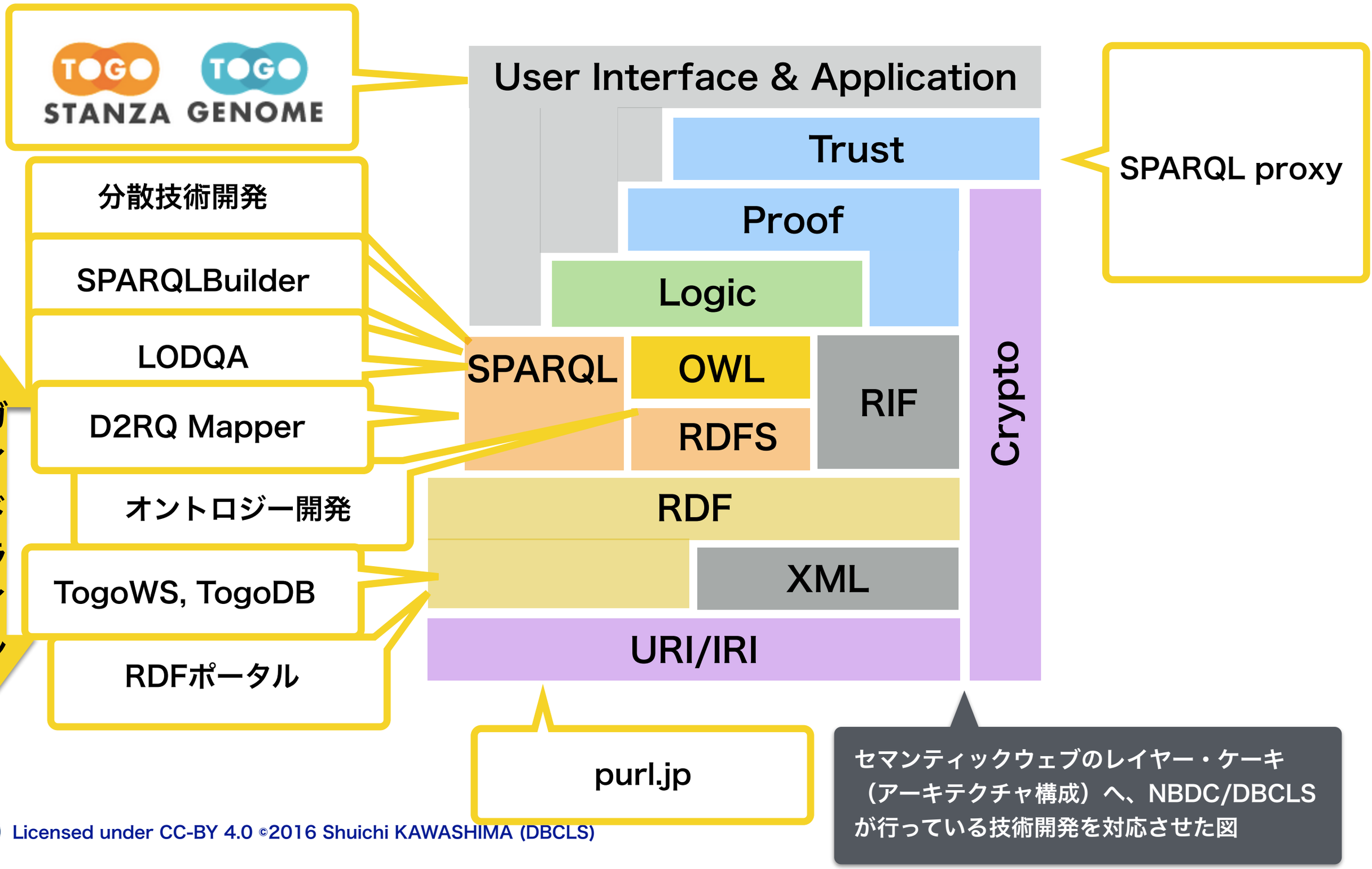
DBCLS Database Center for Life Science

（生命科学ではない）一般的な語彙・オントロジー

オントロジー名		主な内容
RDF	Resource Description Framework	RDFの基本語彙
RDFS	RDF Schema	RDFの構造を記述する語彙
OWL	Web Ontology Language	オントロジーを記述
DC	Dublin Core	基本メタデータ
DC terms	DCMI Metadata Terms	DCの拡張語彙
SKOS	Simple Knowledge Organization System	既存知識間のマッピング
FOAF	Friend of a Friend	人間／組織の関係
VoID	Vocabulary of Interlinked Datasets	データベース間の関係
UO	Ontology of Units of Measurement	単位
QUDT	Quantities, Units, Dimensions, and Types Ontology	単位、次元、量
PROV-O	PROV Ontology	由来情報
PAV	Provenance, authoring and versioning	由来情報、著者情報等
XSD	W3C XML Schema Definition Language	データ型

レイヤーケーキにおける位置

DBCLS Database Center for Life Science



セマンティックウェブのレイヤー・ケーキ（アーキテクチャ構成）へ、NBDC/DBCLSが行っている技術開発を対応させた図

NBDC RDFポータル　開発の理由

どこにいけば、日本で作られたRDFがあるの？一覧が欲しい

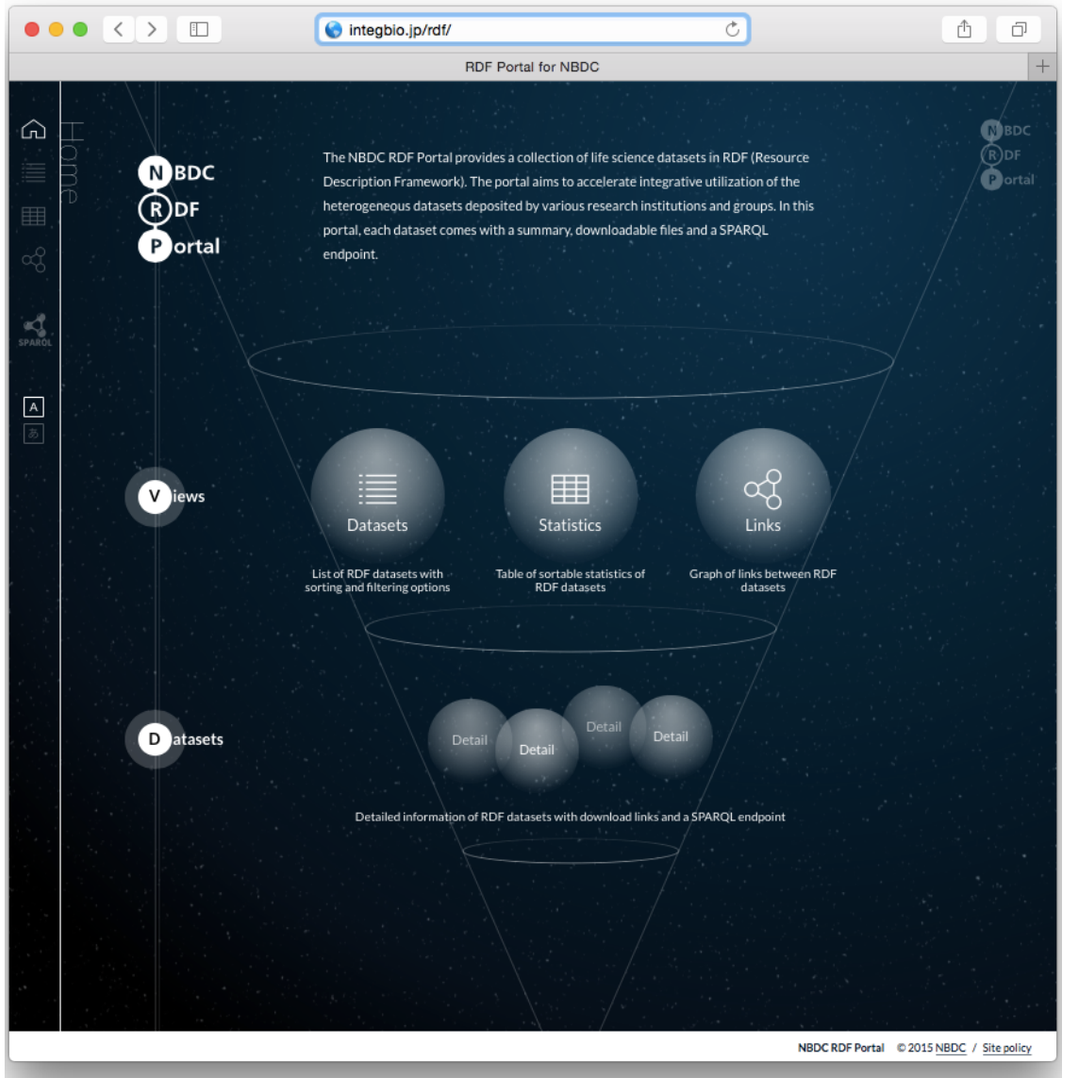
RDFファイルだけあっても、使うのは難しい。もう少し情報があれば。

とりあえず試してみるのに、自前でエンドポイント用意するのは大変。

このような要望に答えるべく、RDFポータルの開発を開始した。

NBDC RDFポータル

- 国内の研究グループが構築したRDFデータを一覧するためのポータルサイト
- 掲載されているRDFデータは、RDF化**ガイドラインに準拠**しているか**事前にレビュー**されている
- 全てのRDFファイルをダウンロードすることが可能
- ライセンス情報、作成者、作成した日付、RDFデータの統計値等のメタデータも閲覧できる
- またSPARQLエンドポイントのサービスを提供しているため、本ポータルサイトへ直接SPARQLで問い合わせを行うことも可能



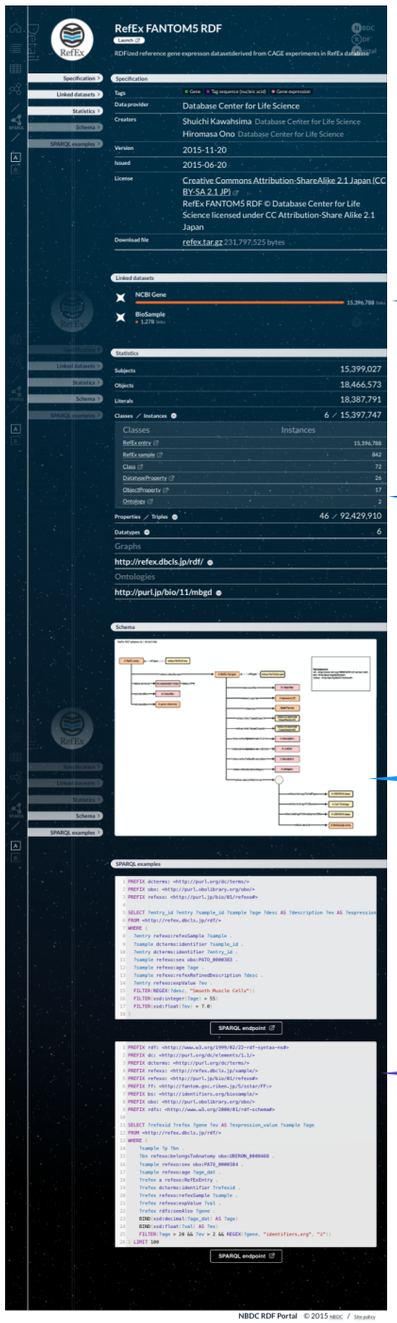
http://integbio.jp/rdf/

NBDC RDFポータル データセットの一覧

DBCLS
Database Center
for Life Science

http://integbio.jp/rdf/

NBDC RDFポータル 各データセットエントリー



データセットエントリーの例

RDFデータセットの各種メタデータ。データ提供者、開発者、バージョン、ライセンス等。RDFファイルのダウンロードもここから。

外部のデータセットへのリンクに関する統計情報。

統計情報。主語、目的語の数、オントロジーのクラス毎にそのインスタンスの数、プロパティ毎のトリプルの数等。統計情報からおおまかな内容の把握ができる。

データセットのスキーマ図。RDFは人間には分かりにくいため、構造を理解する際、助けになる

SPARQLによる問い合わせサンプル。クリックするだけで、実行し結果を閲覧することができる。初めての場合は、サンプルを改造して、オリジナルのSPARQL文を書くことで、容易に問い合わせが行える。

NBDC RDFポータルのデータセット 統計

DBCLS Database Center for Life Science

integbio.jp/rdf/?view=matrix

Statistics

SPARQL

Name	Triples	Links	Classes	Instances	Literals	Subjects	Properties	Objects
Open TG-GATEs	6,803,095,323	583,068	65	1,498,632,260	1,267,408,613	1,498,632,260	38	2,766,567,833
wwPDB/RDF	3,558,962,316	2,988,687	716	248,830,859	5,146,538	249,095,010	4,226	255,461,646
MBGD RDF	1,609,018,143	43,260,205	32	273,443,876	74,289,149	309,702,751	79	472,067,973
Linked ICGC Dataset	577,082,774	57,483	9	51,410,015	20,735,685	51,410,016	66	30,906,909
BMRB/RDF	368,147,209	161,277,662	391	18,844,789	1,627,291	19,082,731	2,604	20,764,246
NBDC NikkajiRDF	160,428,782	0	9	16,354,532	50,484,510	33,481,886	24	85,914,019
Quanto	107,782,575	1,995,973	9	21,955,725	10,369,612	21,955,725	30	31,483,981
RefEx FANTOM5 RDF	92,429,910	15,398,066	6	15,397,747	18,387,791	15,399,027	46	18,466,573
FAMSBASE GPCR	21,297,786	1,328,293	16	5,858,908	488,759	5,858,909	30	6,378,250
Dataset of WURCS-RDF	6,213,789	0	14	817,535	40,435	1,365,653	56	1,138,140
GlyTouCan	1,749,648	0	20	375,657	126,879	375,657	30	502,463
PAConto	81,785	1,133	63	9,296	8,586	9,329	117	16,396
GlycoEpitope	27,796	0	24	5,726	5,453	8,678	35	9,769
SSBD: Meta-information of qua…	18,752	0	18	2,686	1,739	2,686	32	3,712
Metadata of JCM resources	8,896	0	6	1,789	2,574	1,854	25	4,104

NBDC RDF Portal © 2015 NBDC / Site policy

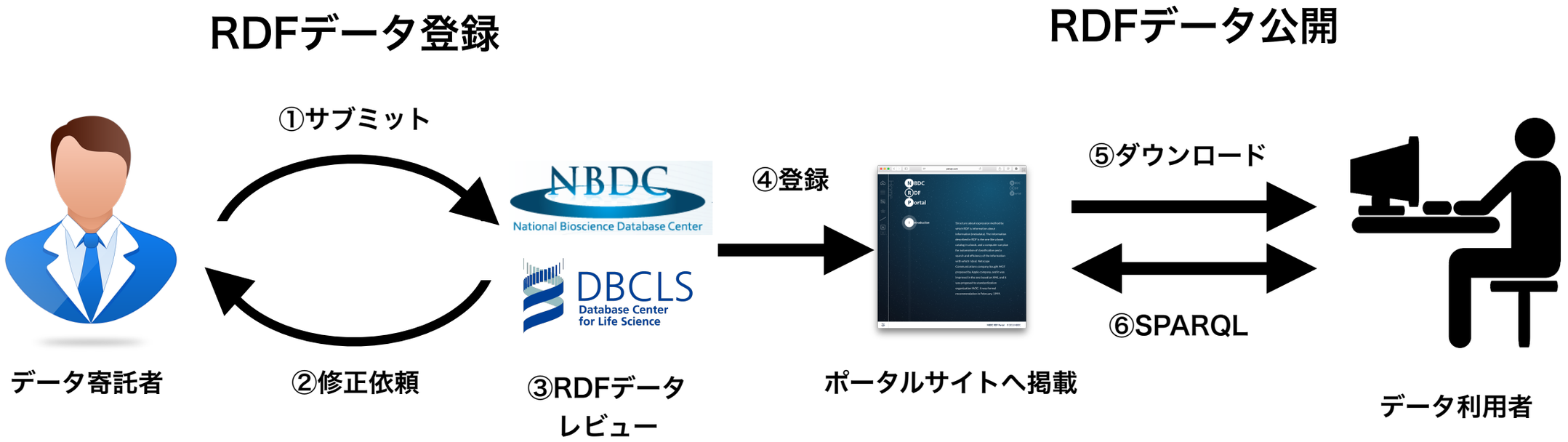
NBDC RDFポータルのデータセット統計

DBCLS Database Center for Life Science

RDFデータセット名	内容	トリプル数
Open TG-GATEs	トキシコゲノミクス	6,803,095,323
wwPDB/RDF	タンパク立体構造メタデータ	3,558,962,316
MBGD RDF	遺伝子オーソログ	1,609,018,143
Linked ICGC Dataset	国際癌ゲノム	577,082,774
MBRB/RDF	タンパク質NMR実験メタデータ	368,147,209
NBDC Nikkaji RDF	化学化合物	160,428,782
Quanto	SRAデータのクオリティ情報	107,782,575
RefEx FANTOM5 RDF	遺伝子発現	92,429,910
FAMSBASE GPCR	GPCR立体構造予測	21,297,786
Dataset of WURCS-RDF	糖鎖構造	6,213,789
GlyTouGan	糖鎖レポジトリ	1,749,648
Metadata ofJCM resources	理研微生物リソース	654,211
PACONTO	糖鎖と病気	81,785
GlycoEpitope	糖エピトープ	27,796
SSBD: Meta-information of quantitative data and microscopy images	生物の定量データと顕微鏡イメージのメタデータ	18,752

2016/10 現在

RDFポータル データ登録／公開／利用の流れ



RDFデータ開発者から①サブミットされたRDFデータデータは、③NBDC/DBCLSによって、ガイドラインに準拠しているかレビューされ、修正点がある場合は寄託者に②修正依頼がなされ、ガイドラインに準拠していることが確認できたRDFデータは、④ポータルサイトへの登録がされる。データ利用者は、RDFポータルサイトから、⑤データをダウンロードして利用するか、⑥SPARQLによる問い合わせによりデータの検索を行う。

データ／データベースの将来像

DBCLS
Database Center for Life Science

	これまでのデータベース	未来のデータベース
データベースの構造	データとデータベースシステムとユーザインターフェースが不可分な構造	データとデータベースシステムとユーザインターフェースは独立した存在
データの単位	特定のまとまり（ファイルや、データベースシステム、ウェブサイト等）に依存	個々のデータが最小単位で自立的に存在する
データの場所	特定の場所	分散
データの利用者	人間	人間＋コンピュータ

まとめ

- セマンティック・ウェブ技術により生命科学データベースの統合をすすめている。
- RDFは、生命科学データを表現する上で、様々な長所がある。
- 世界的にも、データベースのRDF化は注目されている。
- RDFはオントロジーを利用することが前提となっており、そのため高度な利用が可能になると考えられるが、一方で、RDFの構築を難しくしている。
- 国内の生命科学データベースのRDF化を促進するため、定期的なハッカソンイベントの開催や、RDF化のガイドライン等を構築している。
- またRDFデータの利用を促進するために、NBDC RDFポータルを開発した。