第3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

2018/11/1

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

DBCLS
Database Center for Life Science

■ 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。

According to a 2016 survey, data scientists across a wide array of fields said they spend most of their work time (about 80 percent) doing what they least like to do: collecting existing datasets and organizing data.

NIH STRATEGIC PLAN FOR DATA SCIENCE
https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

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

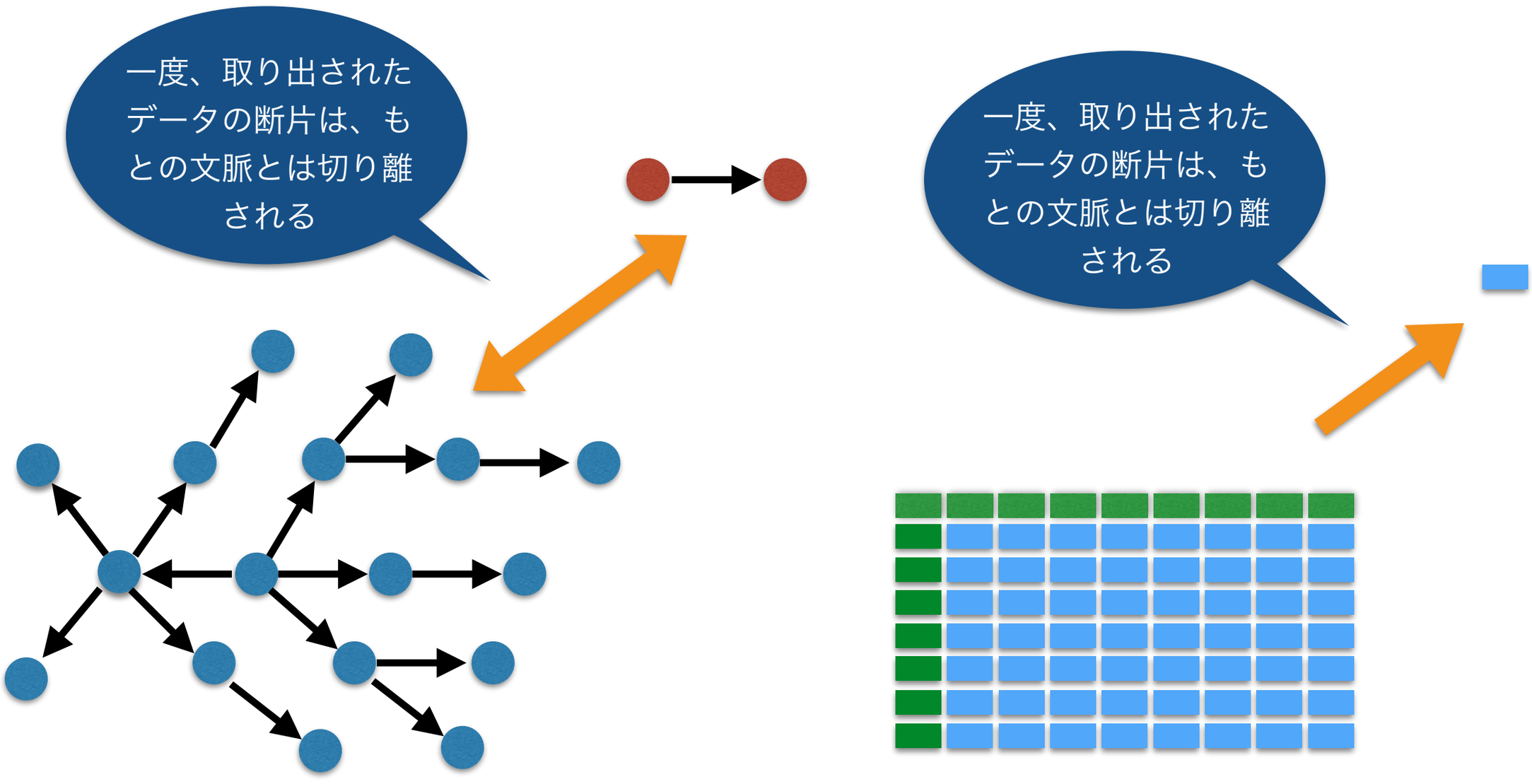
☑ RDFでは新規のデータもそのまま追加できる

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

■ 1つのRDF（トリプル）は、URIの組であり、それだけで自立した情報（もとの文脈から切り離されない）として扱える。



DBCLS基盤技術開発

RDFの利点：多分野に渡る相互運用性の向上

様々な分野からRDFでデータ公開されることで、想定していなかった統合も低コストで可能になる

医薬品申請での例

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、2015年６月にRDFバージョンの標準仕様公開

https://www.cdisc.org/standards/transport/rdf

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

DBCLS基盤技術開発

RDFの利点：多分野に渡る相互運用性の向上

様々な分野からRDFでデータ公開されることで、想定していなかった統合も低コストで可能になる

臨床ゲノム情報での例

国立研究開発法人 日本医療研究開発機構
Japan Agency for Medical Research and Development

AMEDの「臨床ゲノム情報統合データベース整備事業」に、京都大学のグループによる「ゲノム医療を促進する臨床ゲノム情報知識基盤の構築」が採択された。そこで、医療分野における学術文献や公共DBを集約した知識ベースを**RDFにより構築**し、それに対して機械学習技術を応用することで、臨床解釈の推定や、その根拠となるエビデンスおよび治療薬候補などを出力するシステムの構築を行う計画している。

日本人ゲノム配列情報および疾患情報がRDFで構築される可能性がある

DBCLS基盤技術開発

RDFの利点：多分野に渡る相互運用性の向上

様々な分野からRDFでデータ公開されることで、想定していなかった統合も低コストで可能になる

農業分野での例

Food and Agriculture Organization of the United Nations

AGROVOC は、FAO（国際連合食糧農業機関）とCEC（欧州共同体委員会）により、1980年代より開発されている、農林水産、食糧安全保障等に関するシソーラス。日本語を含む、多言語（現在23言語）に対応している。

http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

現在では、オントロジーとして**RDF形式で公開**されており、多くの農業関連オントロジーからリンクされている。農林水産技術会議による　JAT（日本農業シソーラス）も、日本固有の農林水産業・食品およびその関連分野の用語を、AGROVOCに追加する形で構築されている。

http://www.agropedia.affrc.go.jp/dictionary/

農林水産分野における知識の記述も、RDF形式のオントロジーが利用されている。

RDFの長所・短所

■現状、RDFは、データベース開発者向けの技術

■利用者にとってのメリットが大きい（再利用したいデータの記述に向いている）

■生データの記述には向いていない

■自分だけで使うデータ、一度だけ使うデータを記述するには向いていない

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

2017 DDBJ RDF

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。
- 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、GitHub上で構築・公開している。
- 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。

https://github.com/dbcls/rdfizing-db-guidelines

データベースのRDF化ガイドライン

■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

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

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

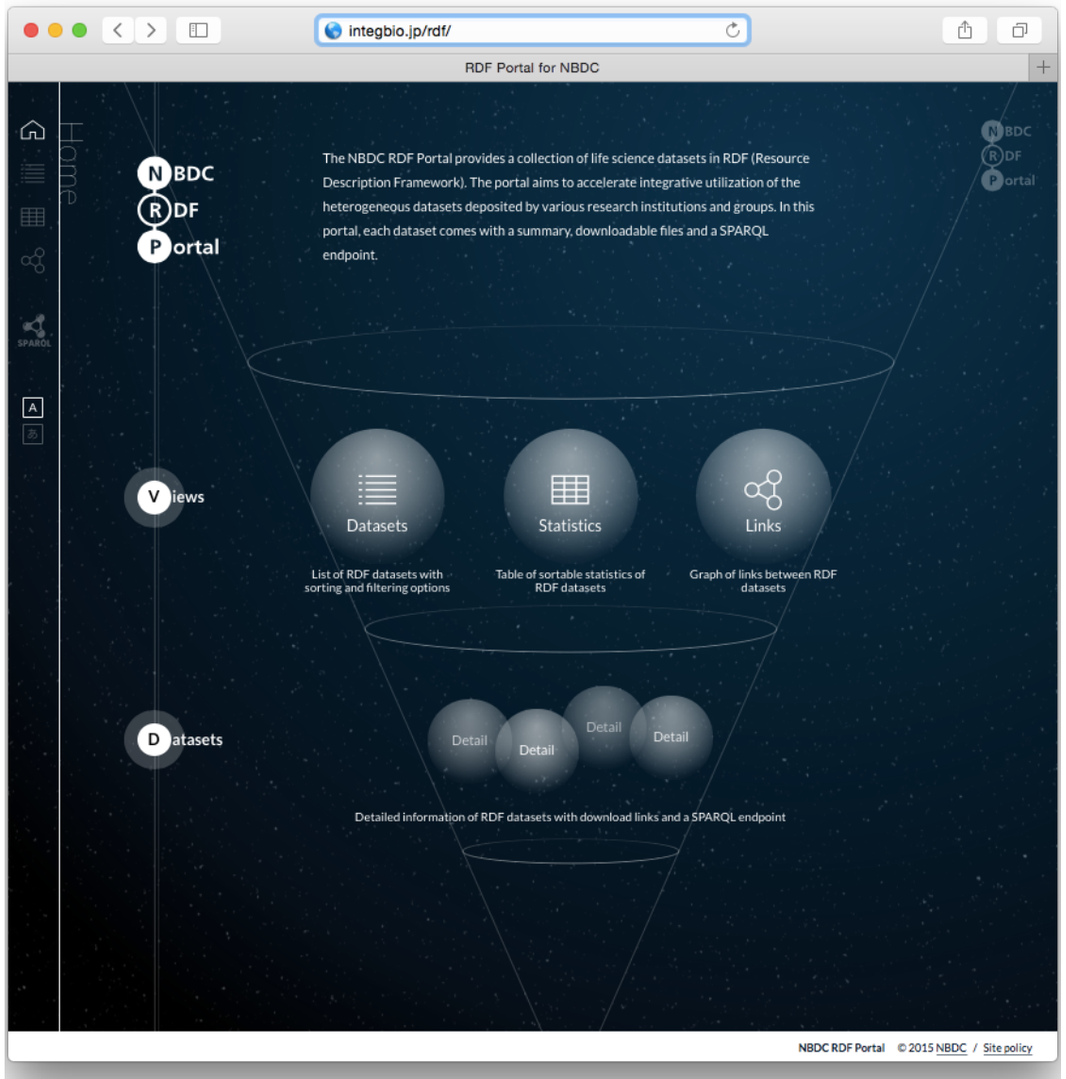
DBCLS
Database Center for Life Science

（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

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/

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。