ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
平成26年度キックオフミーティング

NBDC/DBCLS 共同研究計画

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

2014/6/2 @ NBDC (JST東京本部別館)

平成26年度からの NBDC 統合化推進プログラム募集要項

RDFによるすべてのDBの統合 ← 支援します

平成２６年度

ライフサイエンスデータベース統合推進事業
統合化推進プログラム

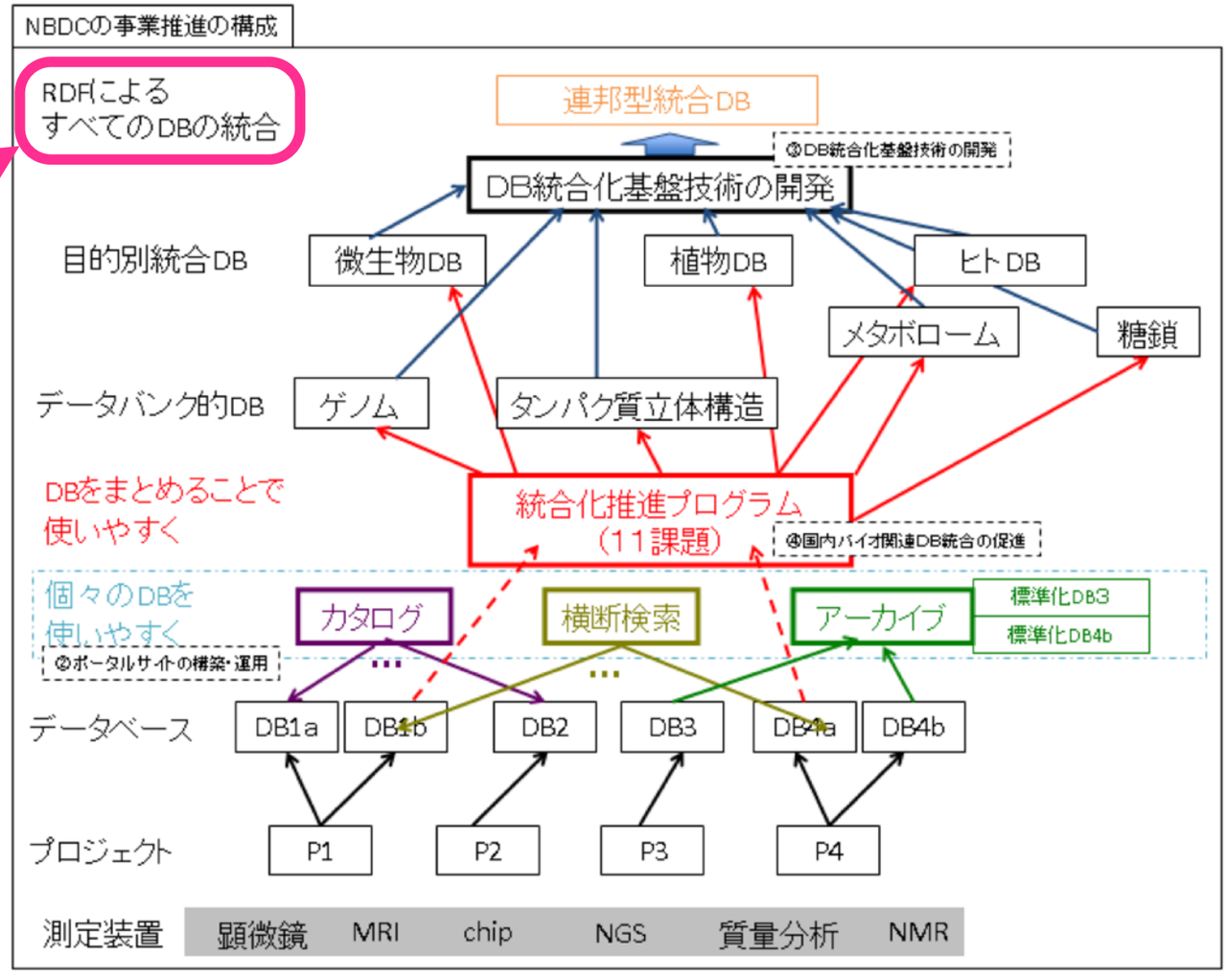
研究開発提案募集のご案内
［募集要項］

独立行政法人科学技術振興機構（JST）
バイオサイエンスデータベースセンター（NBDC）

平成２５年１２月

独立行政法人
科学技術振興機構
Japan Science and Technology Agency

国内外に散在しているライフサイエンス分野のデータやデータベースについて、それらの共有を強力に促進し、公共財として誰でもが自由に活用できるようにするとともに、生物種や個々の目的やプロジェクトを超えて幅広い統合化を実現することにより、データがより多くの分野の研究者、開発者、技術者に簡便に利活用できるようにして、データの価値を最大化することを目指すものです。



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

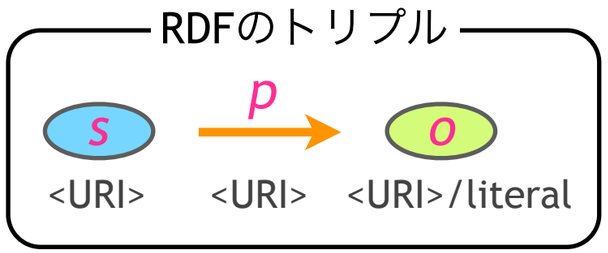
何のために？

→ 広大なデータのウェブからコンピュータで適切に知識発見を支援

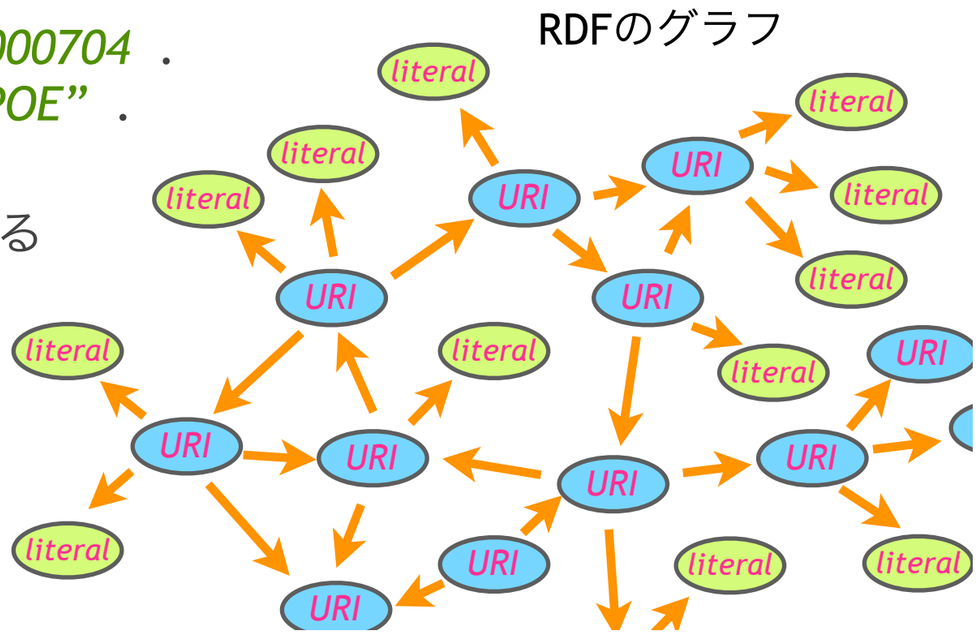
- 曖昧な記述をやめよう！曖昧なフォーマットをやめよう！

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** - モノや概念の ID (URI)
    - **述語** - オントロジーで定義された属性 (URI)
    - **目的語** - 別のモノや概念のID (URI) または 値 (literal)



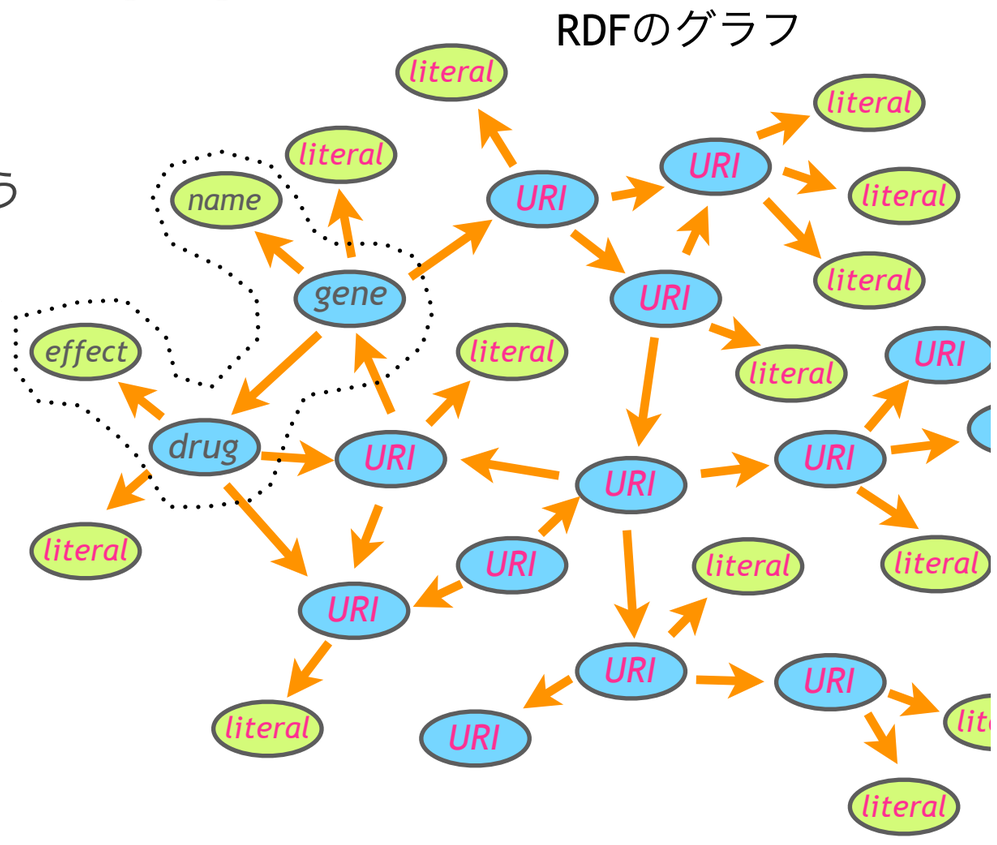
  - つまり RDF は非常にシンプルなデータ形式
    - <http://genome.db/gene1> rdf:type so:0000704 .
      <http://genome.db/gene1> rdfs:label "APOE" .
  - そして RDF では多様なデータを容易に統合できる
    - 同じモノは ID = URI も同じ
    - 重ねあわせていくとグラフになる



セマンティック・ウェブ: 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 RDF で有機的につながっていることが重要

セマンティック・ウェブ利用に必要となる技術

- **オントロジー (OWL, RDF, RDFS, DC, SKOS, OBO, ...)**
  - 既存オントロジーの活用、正確な統制語彙のデザイン、クラス設計、ドメインとレンジ
- **W3C による RDF の仕様 (RDF/XML, Turtle, TriG, ...)**
  - どのようにデータを RDF に変換するか、暗黙のセマンティクスを明示化
- **W3C による SPARQL 1.1 の仕様**
  - SPARQL エンドポイントをどのように検索し、グラフのデータをどう更新するか
- **フリーや商用の各種トリプルストア**
  - RDF データベースの性能、管理、セキュリティ、スケーラビリティ
- **ウェブアプリケーション (REST API, HTML, CSS, JavaScript, JSON, ...)**
  - ウェブインターフェイスによる検索〜可視化のデザイン
- **分散データベースや巨大データとの連結 (NGS, 画像, ...)**
  - データベース間の連携、公開・非公開データの連結、ビッグデータ対応

→ ベストプラクティスの蓄積やガイドラインが必要

平成25年度までの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

DBインフラ

- TogoDB
- TogoWS
- TogoTable
- OReFiL

ゲノム情報

- TogoGenome
- TogoStanza
- GGGenome
- GGRNA
- SRA

遺伝子発現

- RefEx

フェノーム

- Bodyparts 3D

オントロジー

- LODQA
- OntoFinder
- OntoFactory

テキストマイニング

- PubAnnotation
- TogoDoc
- inMeXes
- Allie
- Colil

国際連携・標準化

- BioHackathon
- SPARQLthon
- DBメタデータ

コミュニティ支援

- OpenID
- LSQA
- Galaxy

日本語コンテンツ・教材

- 新着論文レビュー
- 領域融合レビュー
- 統合TV

これらのサービスやコンテンツ資産を活用
連携的な開発で一段と幅広い統合化を実現

平成26年度からの NBDC/DBCLS 共同研究プラン

NBDC/DBCLS共同研究の４本柱

---

1. **RDF統合化のための基盤技術開発**
   - 高度な RDF 技術の実用化
2. **統合化支援**
   - 統合化推進プログラム等の RDF 化をサポート
3. **エンドユーザ向けデータベース利用技術開発等**
   - 大規模 RDF データを活用したアプリケーション
4. **既存・新規サービス/データベースの運用と拡張**
   - 公開サービスやデータベースの持続的な運用

NBDC/DBCLS 共同研究のアウトライン

1. RDF統合化のための基盤技術開発

高度なRDF技術の実用化
分散環境・セキュリティ

2. 統合化支援

DB構築ユーザ向け

統合化推進Pの支援
補完的有用DBのRDF化

国内の技術交流支援
国際連携による標準化

標準化と国際・国内連携

W3C-HCLS など DB メタデータ
FALDO などの共通オントロジー
Identifiers.org などの URI

BioHackathon / SPARQLthon

DDBJ/EBI/NCBI/UniProt 等 RDF 連携
共通ソフトウェア・技術開発

3. エンドユーザ向けデータベース利用技術開発等

DB利用ユーザ向け

大規模データの活用
QAシステムを目標
日本語コンテンツ拡充

個人ゲノムや非公開データのアクセス制御

4. 既存・新規サービス/データベースの運用と拡張

インフラとしてウェブサービスやトリプルストアを持続的に運用 (NBDC/DDBJ連携)

NBDC/DBCLS 共同研究のアウトライン

1. RDF統合化のための基盤技術開発

高度なRDF技術の実用化
分散環境・セキュリティ

シーケンシングなど測定技術の発展

ヒト

個人ゲノム情報・コホート研究
↓
アクセス制限のあるデータも多い
↓
分散環境の構築とセキュリティ確保
↓
RDFで公共データと個人データを連携

動物・植物・微生物

幅広い生物種のゲノム情報を網羅的に集約
↓
大量データの解釈に参照DBの整備が不可欠
↓
表現型～環境まで多様な情報の標準化
↓
RDFで多種多様なデータを統合し相互利用

NBDC/DBCLS 共同研究のアウトライン

---

1. RDF統合化のための基盤技術開発

高度なRDF技術の実用化
分散環境・セキュリティ

- **RDF による DB 統合化に求められる技術**
  - **分散環境** - 各研究機関や医療機関などで産出される分散データ利用と大規模データ処理
  - **セキュリティ** - 個人ゲノム情報などアクセス制限のあるデータとの連携
  - **相互運用性** - 分散データ運用のためのエンドポイント連携や RDF モデルの共通化

- **これらの課題に対する RDF データ運用 (トリプルストア) の技術は発展途上**
  - データ量の増大に耐えうる分散検索の安定性や高速化・効率化についての技術開発
  - オープンなデータが前提のセマンティック・ウェブにおける情報保護の技術開発

NBDC/DBCLS 共同研究のアウトライン

2. 統合化支援

DB構築ユーザ向け

統合化推進Pの支援
補完的有用DBのRDF化

国内の技術交流支援
国際連携による標準化

～昨年度

国際版 BioHackathon などでの標準化と技術開発

↓

TogoGenome/MicrobeDB.jp/CyanoBase/MBGD 間での連携

↓

TogoStanza・オントロジー・ノウハウなどの共有

今年度～

NBDC/DBCLS と全統合化推進プログラムの実務者間で連携強化

↓

国内版バイオハッカソンや SPARQLthon での技術交流をすでに開始

↓

（統合化推進プログラム = TPP と呼ぶことに）

↓

RDF 化の支援や TPP 間での RDF 相互利用

NBDC/DBCLS 共同研究のアウトライン

2. 統合化支援

DB構築ユーザ向け

統合化推進Pの支援
補完的有用DBのRDF化

国内の技術交流支援
国際連携による標準化

- **統合化推進プログラム (TPP) の RDF 化支援**
  - **技術共有** - RDF 化のガイドライン整備
  - **技術開発** - RDF 化支援ツールや共通アプリケーションの開発
  - **データ整備** - TPP 連携に必要な補完的データの RDF 化
  - **技術交流** - 国内版バイオハッカソン、SPARQLthon の開催
  - **国際連携** - 国際版 BioHackathon、RDF summit の開催
- **これまで以上に TPP 間の連携を強化**
  - 毎月2日間開催 SPARQLthon の1日目を TPP 主体に再編成
- **必要な国際標準化を推進**
  - 5 月に RDF サミットを開催しゲノム情報の RDF 化方針で合意
  - ヒトゲノム情報を中心に INSDC <-> Ensembl の RDF を共通化
  - 11 月の国際版 BioHackathon では変異・多型情報なども

NBDC/DBCLS 共同研究のアウトライン

大規模データ

NGS など実験やサンプルのメタデータを DDBJ と連携して整備

↓

正常な遺伝子発現のリファレンス RefEx

＋

RNAi, CRISPR などの実験による発現変動解析支援

質問応答システム

SPARQL を隠蔽しつつ RDF から適切な情報を検索取得したい

↓

BioPortal のオントロジー活用、文献からの知識抽出

↓

自然言語による質問応答、TogoStanza によるレポート生成

3. エンドユーザ向けデータベース利用技術開発等

DB利用ユーザ向け

大規模データの活用

QAシステムを目標

日本語コンテンツ拡充

個人ゲノムや非公開データのアクセス制御

NBDC/DBCLS 共同研究のアウトライン

- **大規模データ - 利活用促進のためのシステム開発**
  - **NGSデータ** - メタデータを活用したデータ検索性の向上
  - **高速検索技術** - 効率的な発現解析の実験支援サービス

- **質問応答システム - RDF DBとオントロジーを活用**
  - **ウルフラムα型** - 比較的定形の質問に知識レポートを返す
  - **IBMワトソン型** - 自然言語による問合わせに対話的に答える

- **日本語コンテンツ整備**
  - **統合TV** - DBやソフトウェアの利用に関する教材の整備
  - **テキスト** - 新着論文レビュー、融合領域レビュー

- **基盤技術開発に基づく安全なデータ管理**
  - エンドユーザが利用するアプリケーションのアクセス制御

3. エンドユーザ向けデータベース利用技術開発等

DB利用ユーザ向け

大規模データの活用
QAシステムを目標
日本語コンテンツ拡充

個人ゲノムや非公開データのアクセス制御

NBDC/DBCLS 共同研究のアウトライン

- **大規模な RDF データベースの分散的運用**
  - **負荷分散** - 組織内外のサーバ間での負荷分散と冗長性の高いサービス構成
  - **可用性** - ダウンタイムの少ない運用ノウハウの構築
  - **安定性** - 公共 RDF データの分散検索などにおける安定性や整合性などでの連携
- **分散 DB のための統合的な検索インターフェイス**
  - **統合検索** - 分散 DB 検索をシームレスかつ統合的に利用するためのアプリケーション開発
- **既存および新規サービスの安定的な運用とユーザ対応**

4. 既存・新規サービス/データベースの運用と拡張

インフラとしてウェブサービスやトリプルストアを持続的に運用 (NBDC/DDBJ連携)

平成26年度　　今後の主な

統合化推進プログラム関連のミーティング予定

- **RDF summit**
  - **5/17-20** (開催済み)：INSDC, Ensembl でのゲノム情報 RDF 標準化

- **SPARQLthon**
  - **6/18-19**：第21回@DBCLS柏の葉、**7/15-16**：第22回、以後も毎月開催

- **統合の日**
  - **10/5**：例年通りトーゴーの日にシンポジウムの開催

- **BioHackathon 2014**
  - **11/9-14**：東北大学メディカル・メガバンクと共催（仙台・松島）

- **国内版バイオハッカソン 14.14**
  - 日程未定

→ biohackathon-jp@googlegroups.com のメーリングリストで告知しています