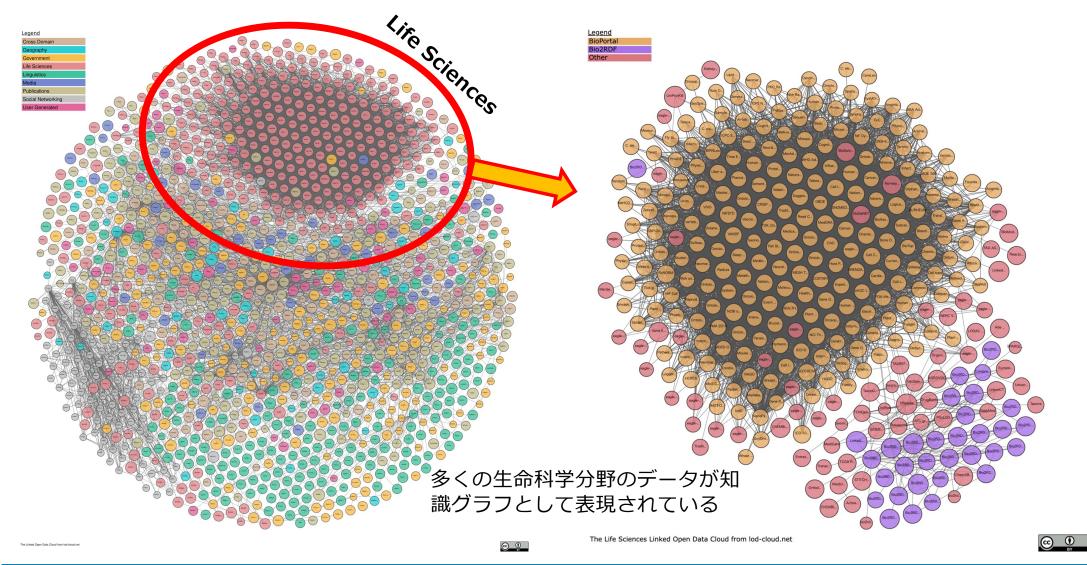
6

より繋がりやすいRDFデータの構築に向けて

〇山本 泰智1、藤澤 貴智2

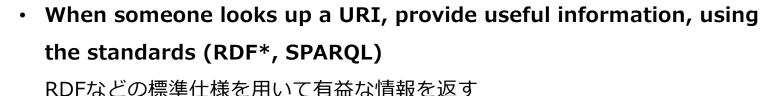
1. 情報・システム研究機構 ライフサイエンス統合データベースセンター (DBCLS)、2. 国立遺伝学研究所





Linked Data

- Use URIs as names for thingsすべてのモノやコトにURIを用いて名前をつける
- Use HTTP URIs so that people can look up those names
 httpで始まるURIでネット経由のアクセスを可能に



Include links to other URIs. so that they can discover more things
 関連外部データのURIを含めて情報を芋づる式に

--- Tim Berners-Lee

http://www.w3.org/DesignIssues/LinkedData.html





LODの特徴



- だれでもWeb上でデータを公開できる
- 語彙を自由に選んでデータを表現できる
- だれでもデータに情報(注釈)を付加できる
- データの選択や加工の自由度が高い
- お互いにデータを公開し、共有することで、これらの利便性が高まる



- Web上のデータは玉石混交
- 語彙がばらばらで同じ概念同士でも繋がらない
- 注釈の信頼性や一貫性は期待できない
- どこに必要なデータがあるのか分かりにくい
- お互いにデータを公開し、共有するのは難しい



コンピューターを用いて

散在するデジタルデータを効率良く見つけ

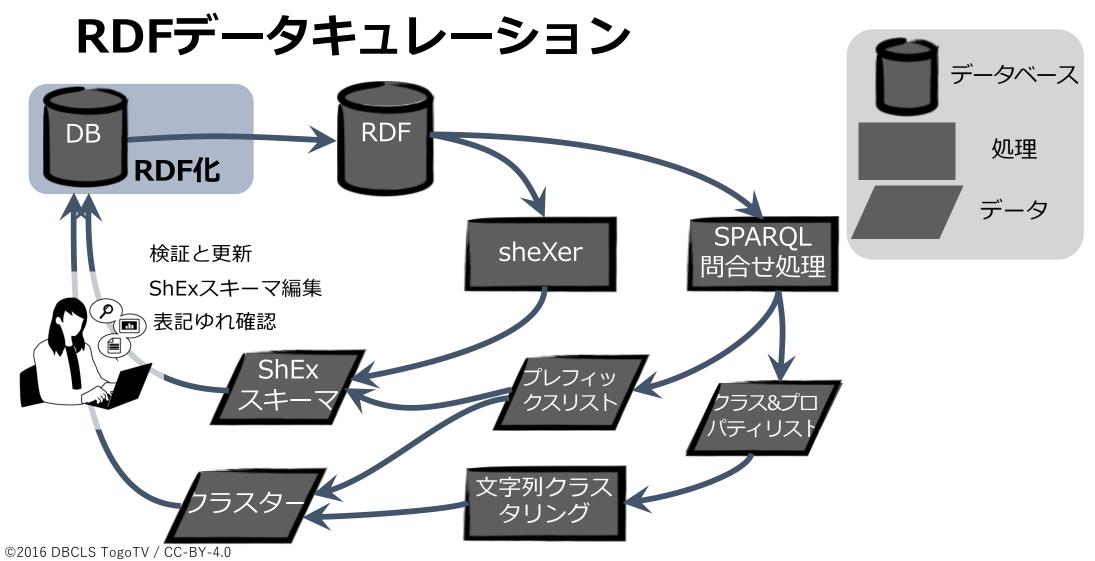
効率よくデータを統合したい

果たして本当に繋がっているのか? RDFデータ作成者の意図通りに作られているのか?



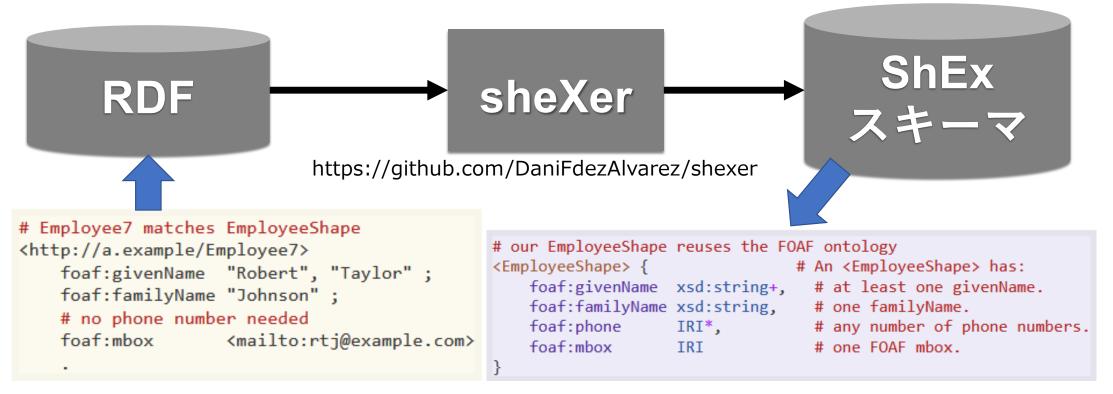
実際に見つかった例

- http://www.w3.org/2000/01/rdf-schema#Label
- http://www.w3.org/2000/01/rdfschema#label
- http://purl.jp/bio/12/glyco/glycan#component
- http://purl.jp/bio/12/glyco/glycan/component
- http://purl.jp/bio/12/glyco/glycan#Saccharide
- http://purl.jp/bio/12/glyco/glycan/saccharide
- http://glycoinfo.org/dbid/imex/IM-14106
- http://rdf.glycoinfo.org/dbid/imex/IM-14106



sheXer





https://shex.io/



プレフィックス

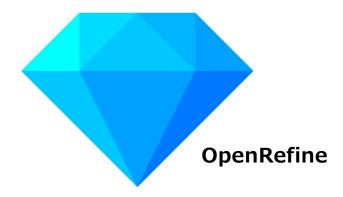
URIの最後のスラッシュもしくは#以降を除去した部分

```
SELECT distinct ?pref_subj
WHERE {
    ?subj ?prop ?obj .
    bind(replace(str(?subj),"[/#][^/#]+$","") as ?pref_subj)
}
```

クラスやプロパティ

SELECT distinct ?class SELECT distinct ?prop

WHERE { [] a ?class . }
WHERE { [] ?prop ?obj . }

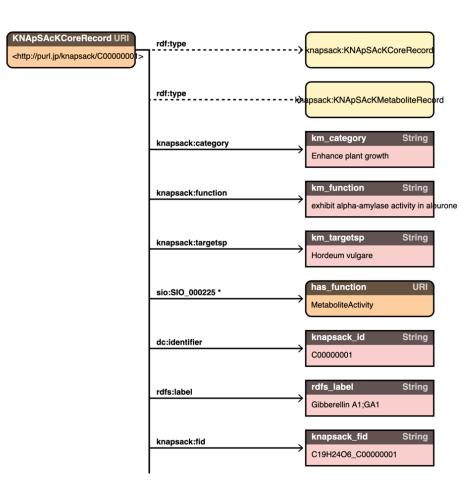


プレフィックスリストやクラス名やプロパティ名のリストを、 OpenRefineを用いてクラスタリング、表記ゆれをチェック



ユースケース:天然化合物データベースのRDF化

- KNApSAcK Familyは、生物種を軸に整理された天然化合物データベースであり、食・健康・環境情報など多岐にわたる主に文献情報に基づいた生物機能情報を扱っている。
- 化合物、生物種、文献情報とデータ統合を 実現し、植物や微生物データベースとの統 合を目的として、我々はKNApSAcK Family をRDF化を実施した。
- KNApSAck RDFに対して本研究で提案する RDFデータのキュレーションワークフロー を適用した。





```
:KNApSAcKCoreRecord
  rdf:type [<http://purl.jp/knapsack/resource#KNApSAckCoreRecord>] ;
                                                                                  # 100.0 %
                                                                # 100.0 %
  rdfs:seeAlso IRI ;
  dc:hasPart IRI +;
                                                                # 100.0 %
           # 100.0 % obj: @:CDXfile. Cardinality: {1}
           # 100.0 % obj: @:SDfile. Cardinality: {1}
           # 100.0 % obj: @:CHEMINF 000058. Cardinality: {1}
  <http://purl.jp/knapsack/resource#fid> xsd:string ;
                                                                   # 100.0 %
  <http://semanticscience.org/resource/CHEMINF 000200> IRI +;
                                                                           # 100.0 %
           # 100.0 % obj: @:CHEMINF 000059. Cardinality: {1}
           # 100.0 % obj: @:CHEMINF 000018. Cardinality: {1}
           # 100.0 % obj: @:Start substance. Cardinality: {1}
           # 100.0 % obj: 0:CHEMINF 000113. Cardinality: {1}
  foaf:homepage IRI ;
                                                                # 100.0 %
  <http://semanticscience.org/resource/SIO 000008> IRI +;
                                                                       # 100.0 %
           # 100.0 % obj: @:CHEMINF 000043. Cardinality: {1}
           # 100.0 % obj: @:CHEMINF 000334. Cardinality: {1}
           # 100.0 % obj: @:CHEMINF 000042. Cardinality: {1}
  <http://semanticscience.org/resource/SIO 000255> IRI +;
                                                                       # 100.0 %
           # 72.04371622438232 % obj: IRI. Cardinality: {1}
           # 12.88350972910444 % obj: IRI. Cardinality: {2}
           # 100.0 % obj: @:KNApSAcKCoreAnnotation. Cardinality: +
           # 9.445950378977537 % obj: @:KNApSAcKMetaboliteActivityAnnotation. Cardinality: +
  dc:references IRI +:
                                                                # 100.0 %
           # 65.57951622092922 % obj: IRI. Cardinality: {1}
           # 17.755831419741362 % obj: IRI. Cardinality: {2}
           # 6.762893005749408 % obj: IRI. Cardinality: {3}
           # 100.0 % obj: @:KNApSAcKReference. Cardinality: +
           # 100.0 % obj: @:Article. Cardinality: +
  rdfs:label xsd:string +;
                                                                # 100.0 %
           # 94.51475336245446 % obj: xsd:string. Cardinality: {1}
           # 5.471434244375766 % obj: xsd:string. Cardinality: {2}
  <http://purl.org/dc/elements/1.1/identifier> xsd:string +;
                                                                          # 100.0 %
           # 94.51475336245446 % obj: xsd:string. Cardinality: {1}
           # 5.485246637545538 % obj: xsd:string. Cardinality: {2}
  rdf:type [<http://purl.jp/knapsack/resource#KNApSAcKMetaboliteActivityRecord>] ?;
           # 5.485246637545538 % obj: <a href="http://purl.jp/knapsack/resource#KNApSAckMetaboliteActivityRecord">http://purl.jp/knapsack/resource#KNApSAckMetaboliteActivityRecord</a>. Cardinality: {1}
  <http://purl.jp/knapsack/resource#function> xsd:string *;
           # 5.485246637545538 % obj: xsd:string. Cardinality: +
  <http://semanticscience.org/resource/SIO 000225> @:KnapsackMetaboliteActivity *;
            # 5.485246637545538 % obj: @:KnapsackMetaboliteActivity. Cardinality: +
  <http://purl.jp/knapsack/resource#category> xsd:string *;
           # 5.485246637545538 % obj: xsd:string. Cardinality: +
  <http://purl.jp/knapsack/resource#targetsp> xsd:string *
           # 5.485246637545538 % obj: xsd:string. Cardinality: +
  skos:closeMatch IRI *
           # 89.67350955644953 % obj: IRI. Cardinality: +
           # 81.6450560265198 % obj: IRI. Cardinality: {1}
           # 7.669331307515669 % obj: IRI. Cardinality: {2}
```

スキーマの基本構造の 確認

Cardinalityの確認

独立して生成した RDFのリソース間関係 の確認

開発中に追加したトリ プルの確認

今後の予定

・RDFデータキュレーション支援システム構築

検討項目

- UI
- 配布方法(Webサービス / デスクトップ)

終

ご意見等おまちしています。 yy@dbcls.rois.ac.jp

