

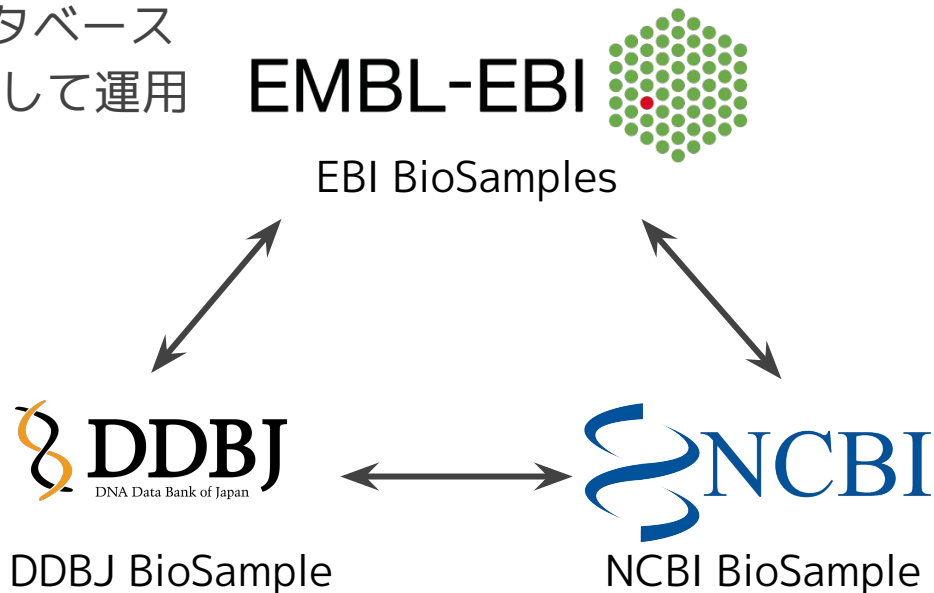
18. BioSamplePlus: BioSampleメタデータの オントロジーマッピングとRDF化

○池田秀也¹、藤澤貴智²、川島秀一¹、大田達郎¹

1. 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

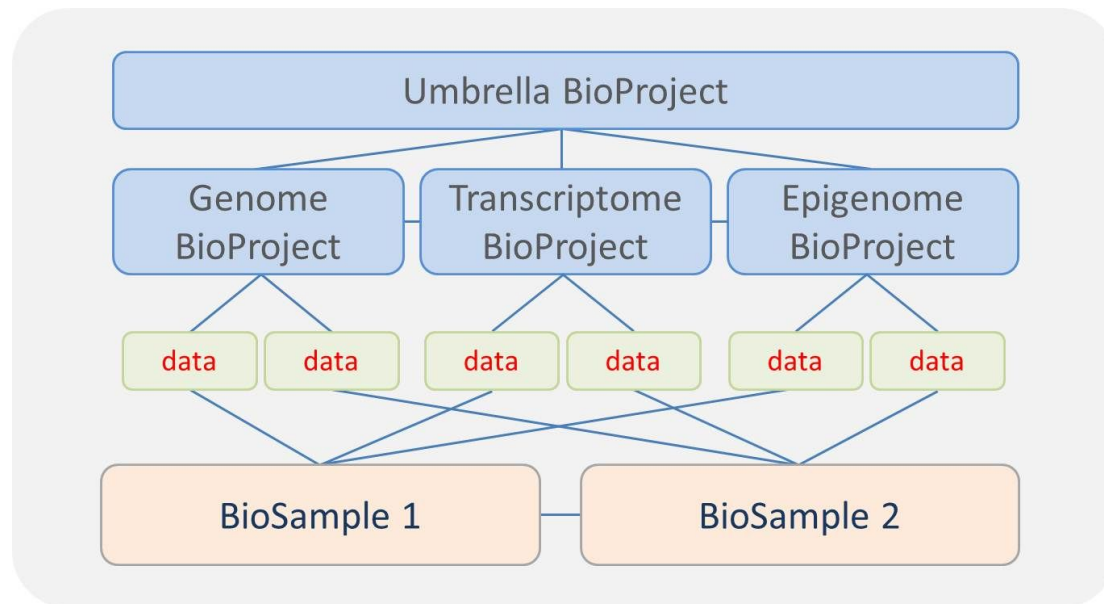
BioSample

- 実験サンプルのメタデータのデータベース
- INSDC 三極で相互にデータを交換して運用



BioSample

- サンプルに対して行われた各種実験のデータへ紐付く
- 同一サンプルが異なる実験に用いられた場合も辿れる



※ [DDBJ BioSample Handbook](#) より

サンプルメタデータの記述

- name: value のペアで属性を記述
- 記法が統一されていないため、同一属性のサンプルを一括して検索するのが難しい

```
tissue: blood  
age: 45  
sex: male
```

```
tissue: Blood  
age: 45 years old  
biological_sex: M
```

```
cell type: HeLa
```

```
cell_line: HeLa
```

```
cell type: adipocyte
```

```
cell type: fat cell
```

従来のサンプル検索

テキストベースのため網羅性に限りがある
(↓ NCBI BioSample の例)

Search results

Items: 1 to 20 of 763

<< First < Prev Page 1 of 39 Next > Last >>

 Filters activated: human. [Clear all](#) to show 3562 items.

- [TEAD1 ChIP-seq Adipocyte D1 Experiment #1](#)
 1. Identifiers: BioSample: SAMN14214211; SRA: SRS6220155; GEO: GSM4340705
Organism: **Homo sapiens**
Accession: SAMN14214211 ID: 14214211
[BioProject](#) [SRA](#) [GEO DataSets](#)

- [TEAD1 ChIP-seq Adipocyte D1 Experiment #2](#)
 2. Identifiers: BioSample: SAMN14214210; SRA: SRS6220156; GEO: GSM4340706
Organism: **Homo sapiens**
Accession: SAMN14214210 ID: 14214210
[BioProject](#) [SRA](#) [GEO DataSets](#)

Find related data

Database:

Find items

Search details

adipocyte[All Fields] AND "Homo sapiens"[Organism]

Search

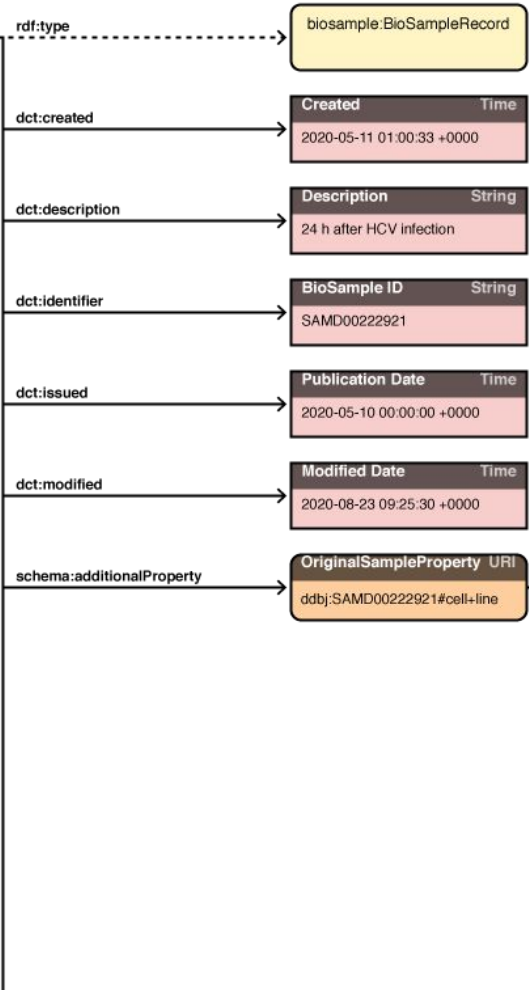
See more...

Recent activity

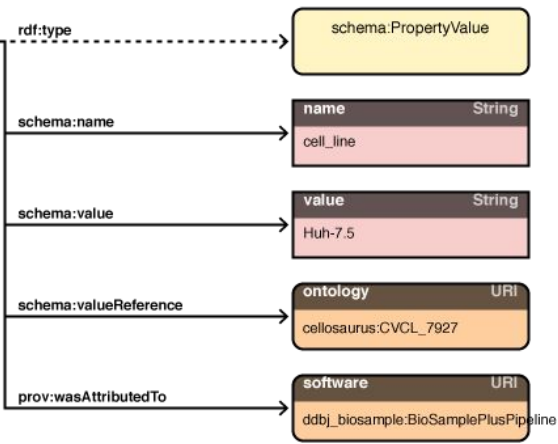
DDBJ BioSample RDF

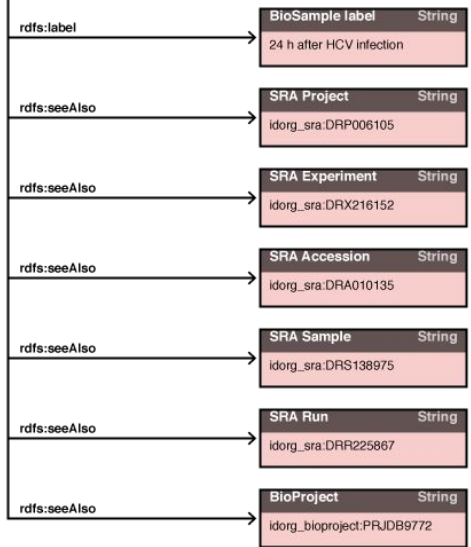
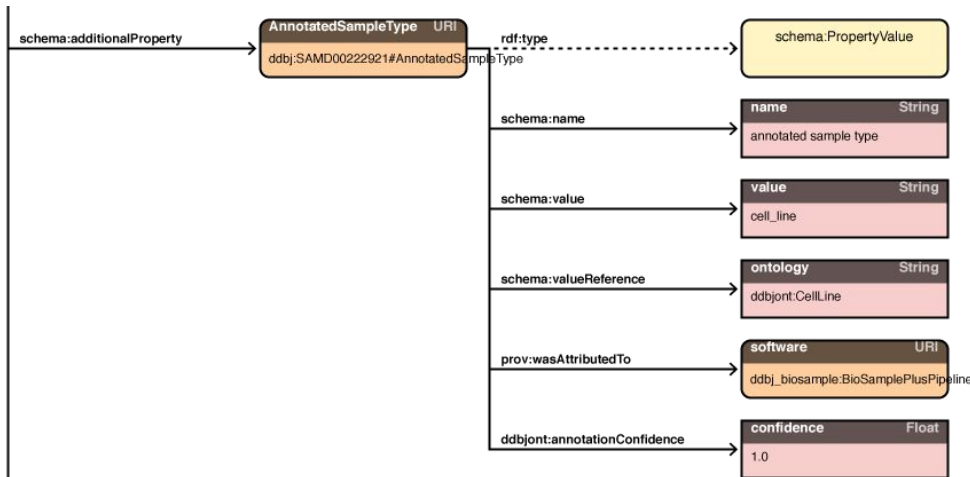
- BioSample メタデータの RDF 化
SPARQL による柔軟な検索
- オントロジーマッピング
統一された語彙による網羅性の高い検索

※現在はヒトサンプルのみをカバー



biosample: <http://ddbj.nig.ac.jp/biosample/>
 samea5619337: <http://ddbj.nig.ac.jp/biosample/SAMEA5619337#>
 ddbj_bioproject: <http://ddbj.nig.ac.jp/ontologies/bioproject/>
 ddbj_biosample: <http://ddbj.nig.ac.jp/ontologies/biosample/>
 ddbj_dra: <http://ddbj.nig.ac.jp/ontologies/dra/>
 idorg_bioproject: <http://identifiers.org/bioproject/>
 idorg_biosample: <http://identifiers.org/biosample/>
 idorg_sra: <http://identifiers.org/insdc.sra/>
 schema: <http://schema.org/>
 dct: <http://purl.org/dc/terms/>
 rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
 rdfs: <http://www.w3.org/2000/01/rdf-schema#>





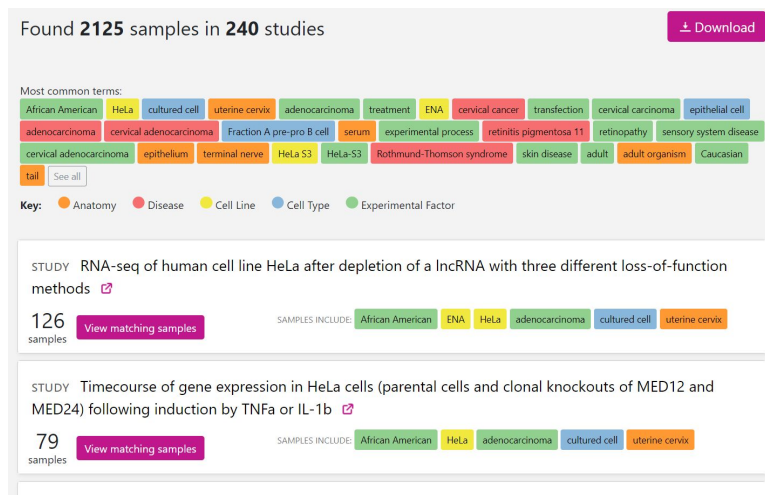
```

biosample: <http://ddbj.nig.ac.jp/biosample/>
samea5619337: <http://ddbj.nig.ac.jp/biosample/SAMEA5619337#>
ddbj_bioproject: <http://ddbj.nig.ac.jp/ontologies/bioproject/>
ddbj_biosample: <http://ddbj.nig.ac.jp/ontologies/biosample/>
ddbj_dra: <http://ddbj.nig.ac.jp/ontologies/dra/>
idorg_bioproject: <http://identifiers.org/bioproject/>
idorg_biosample: <http://identifiers.org/biosample/>
idorg_sra: <http://identifiers.org/insdc.sra/>
schema: <http://schema.org/>
dct: <http://purl.org/dc/terms/>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
  
```



オントロジーマッピングの自動化

MetaSRA(<http://metasra.biostat.wisc.edu/>)



ヒト・マウス RNA-seq サンプルをオントロジーに基づいて検索

UBERON (組織), Cell Ontology (細胞タイプ), EFO (実験変数),

Cellosaurus (細胞株), Disease Ontology (疾患)

自動オントロジーマッピングのパイプラインを提供(<https://github.com/deweylab/MetaSRA-pipeline>)

オントロジーマッピングの自動化

- MetaSRA パイプラインを独自に改良 <https://github.com/sh-ikeda/MetaSRA-pipeline>
- 高速化
5000サンプルのテストで、改良前 3~6 時間、改良後 0.5~1 時間
- 高精度化
 - 長文で記述された value を対象外化
 - 細胞株として判定されるのを許可する属性を追加
 - マッピングに使用しない属性を追加
 - オントロジーインポートのバグ修正
 - etc.

206サンプルのテストセットを用いた評価
(※精度比較の妥当性については要検討)

	改良前	改良後
Precision	0.940	0.968
Recall	0.831	0.871

展望

- 対象生物種の拡大
現在はヒトのみ → 主要なモデル生物をカバーしたい
- 実際に BioSample RDF をつないで使える RDF データを増やす
BioSample ID を軸としたデータ検索が可能な範囲を拡大する
- マッピングの改善
RNA-seq サンプルの発現プロファイルを利用したサンプル属性予測