

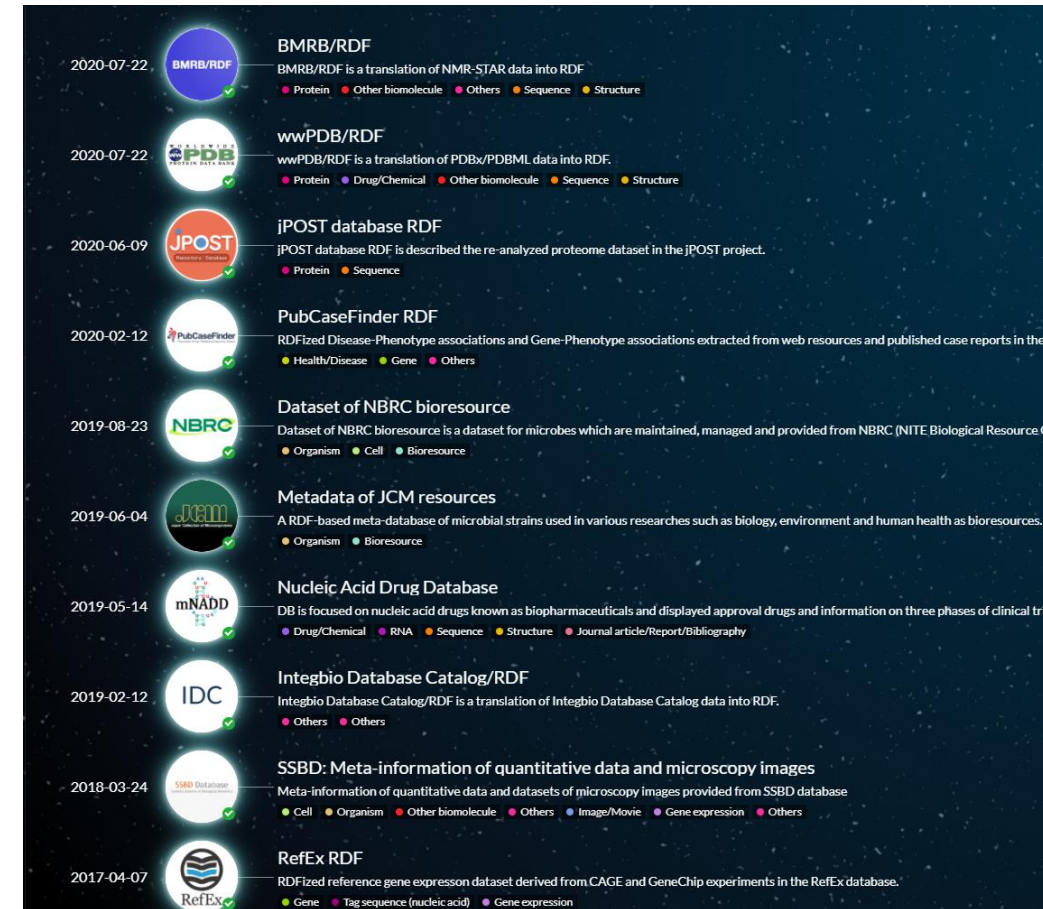
○八塚 茂¹、片山 俊明²、川島 秀一²、畠中 秀樹²

1. 国立研究開発法人科学技術振興機構バイオサイエンスデータベースセンター(NBDC)、2. 大学共同利用法人情報・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)

■ システムおよびRDFデータセットの現状

2020年9月初旬現在で、27のRDFデータセットが公開されており、総データサイズは1000億トリプルである。2015年11月の運用開始時から約17倍のデータサイズとなっている。

SPARQLエンドポイントとしては、Virtuoso を利用している。すべてのRDFを同じVirtuosoインスタンスにロードしてある方が検索の利便性は高いが、一方でデータサイズが巨大になると運用上に支障をきたすことがあり、現状では、DDBJ、KERO、UniProtなどの大きなデータセットについて個別のインスタンスをたてることで対処している。個々のURIは、<http://integbio.jp/rdf/{データセット名}/sparql> のように設定している。



■ ガイドラインの設定とレビューの実施

NBDC RDFポータルでは、登録依頼のあったRDFデータセットに対して独自のガイドラインに基づきレビューを実施している。これにより、RDFデータの質とデータ間の相互運用性を高めることができた。

skwsm Correct URI for purf.jp 344e7c8 on 10 May 22 commits

RDF-portal-guidelines-en.md	Add about HCO in the RDF portal guidelines	2 years ago
README.md	Edit the README.md	2 years ago
dbcls-rdfizing-db-guidelines-ja.md	Correct URI for purf.jp	4 months ago

README.md

DBCLS guidelines for RDFizing databases

The DBCLS guidelines for RDFizing databases are a collection of useful practices for developers who expose life science databases as RDF. DBCLS have been compiled the guidelines based on experience and knowledge gathered from several hakathon events organized by DBCLS and NBDC such as the DBCLS/NBDC BioHackathons, the SPARQLthons and the RDF summits.

■ SPARQLエンドポイントのフロントエンドにSPARQL-proxyを利用

SPARQLエンドポイントのフロントエンドとして、DBCLSで開発しているSPARQL-proxyを利用している。これにより、クエリ実行のジョブスケジューリング、キャッシュによる応答性能向上、実行結果の複数フォーマットによるダウンロードを実現することができた。

SPARQL Proxy

```

1 # Retrieve a full list of Ids with Japanese and English labels.
2
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 PREFIX dct: <http://purl.org/dc/terms/>
5 PREFIX owl: <http://www.w3.org/2002/07/owl#>
6
7 SELECT STR(?id) AS ?ID STR(?nameJ) AS ?Label_Ja STR(?nameE) AS ?Label_En
8 FROM <http://nanbyodata.jp/ontology/nando>
9 WHERE
10 {
11   ?s a owl:Class ;
12     dct:identifier ?id
13   OPTIONAL {
14     ?s rdfs:label ?nameJ
15     FILTER (lang(?nameJ) = "ja")
16   }
17 }

```

Run Query (Ctrl-Enter)

[2348 bindings. -- 1.2 sec.]

ID	Label_Ja	Label_En
"0000001"	"難病"	"intractable disease"
"0000002"		"obsolete class"
"1000001"	"指定難病"	"designated intractable disease"
"1100001"	"神経・筋疾患"	"Neuromuscular disease"
"1100002"	"代謝系疾患"	"Metabolic disease"
"1100003"	"皮膚・結合組織疾患"	"Skin and connective tissue disease"

Download
JSON
CSV
TSV

ダウンロード
フォーマット
を選択

■ 登録対象データセットの拡張

2019年度からは、登録対象について大きく方針転換し、これまでは登録依頼のあったRDFに限っていたものを、UniProtやPubChemといった一般に公開されているRDF（この場合、SPARQLエンドポイントのみ提供）や、NBDC等が独自に開発したRDFにまで対象を広げることにした。

➤ 既公開

- UniProt
- PubChem

➤ 今後公開予定

- BioSamples
- Ensembl
- Rhea

など

データの登録依頼や
その他のお問い合わせは、
rdf-portal@integbio.jp
までお気軽に

The screenshot shows the UniProt website interface. At the top, it says "UniProt" and "Original site". Below that, a description states: "The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. For the detailed information,".

The "Specification" section includes the following details:

Tags	Protein Sequence
Data provider	European Bioinformatics Institute, Swiss Institute of Bioinformatics, and Protein Information Resource
Creators	UniProt Consortium
Version	2019-12-19
Issued	2019-12-19
License	Creative Commons Attribution 4.0 International
Status	Unreviewed 3rd party dataset
Download file	0 bytes

The "Linked datasets" section shows two entries:

- Reactome: 282,912 links
- NCBI Gene: 15,648 links

■ インフラの増強（予定）

データ登録対象の拡張に伴い、2021年度からはインフラを増強して、安定したSPARQL検索の提供と、よりタイムリーなデータ追加・更新を目指す。

