1 NBDC RDFポータル

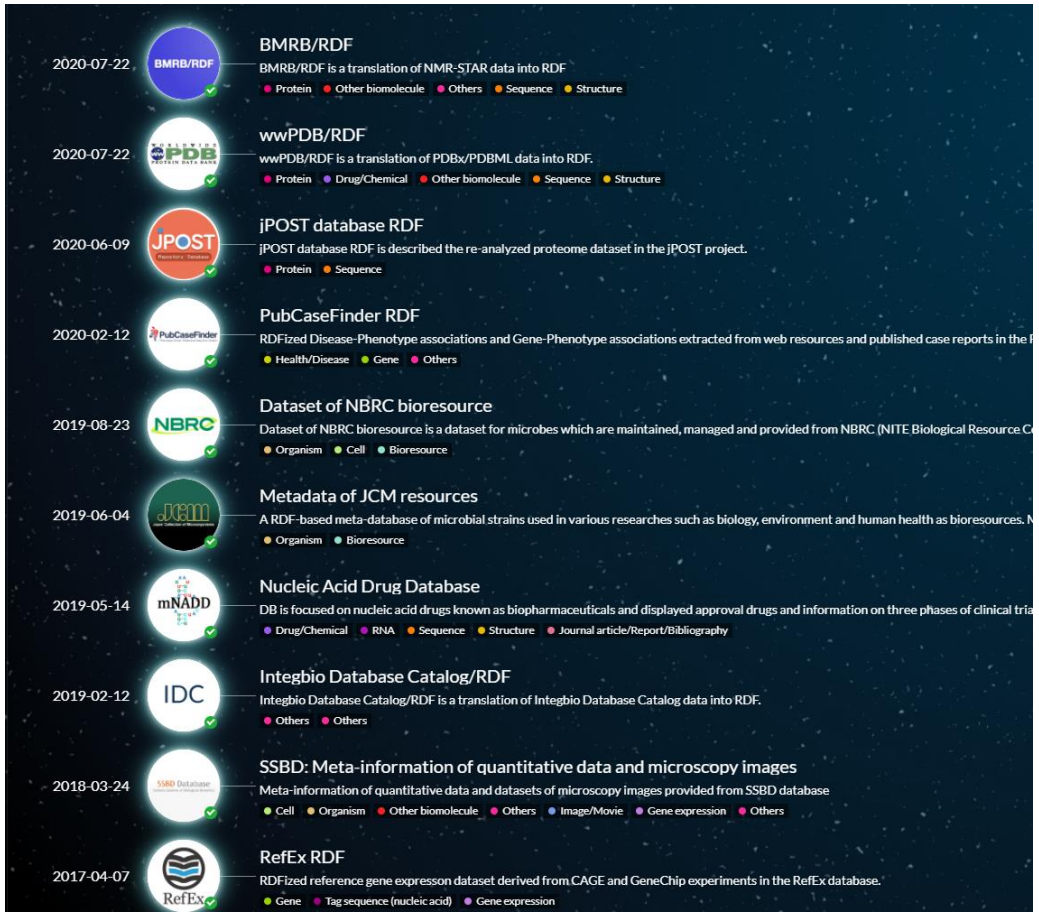
トーゴーの日シンポジウム2020

○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

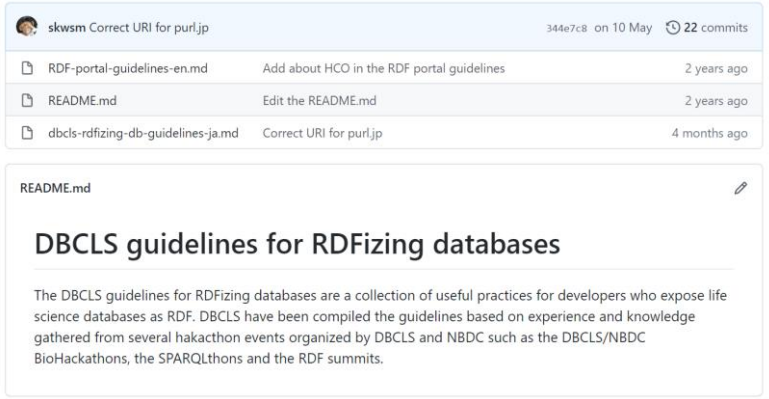
2020年9月初旬現在で、27のRDFデータセットが公開されており、総データサイズは1000億トリプルである。2015年11月の運用開始時からは約17倍のデータサイズとなっている。SPARQLエンドポイントとしては、Virtuoso を利用している。すべてのRDFを同じVirtuosoインスタンスにロードしてある方が検索の利便性は高いが、一方でデータサイズが巨大になると運用上に支障をきたすことがあり、現状では、DDBJ、KERO、UniProtなどの大きなデータセットについて個別のインスタンスをたてることで対処している。個々のURIは、http://integbio.jp/rdf/{データセット名}/sparql　のように設定している。



NBDC RDFポータル

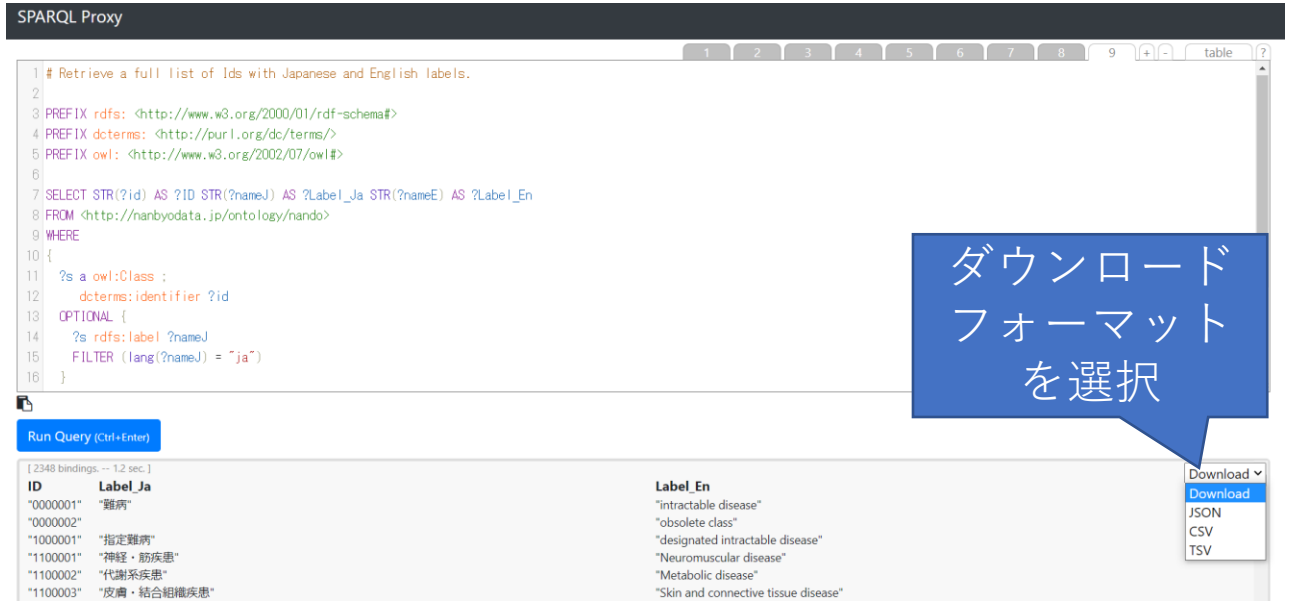
- ガイドラインの設定とレビューの実施

  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く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。



- 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。



NBDC RDFポータル

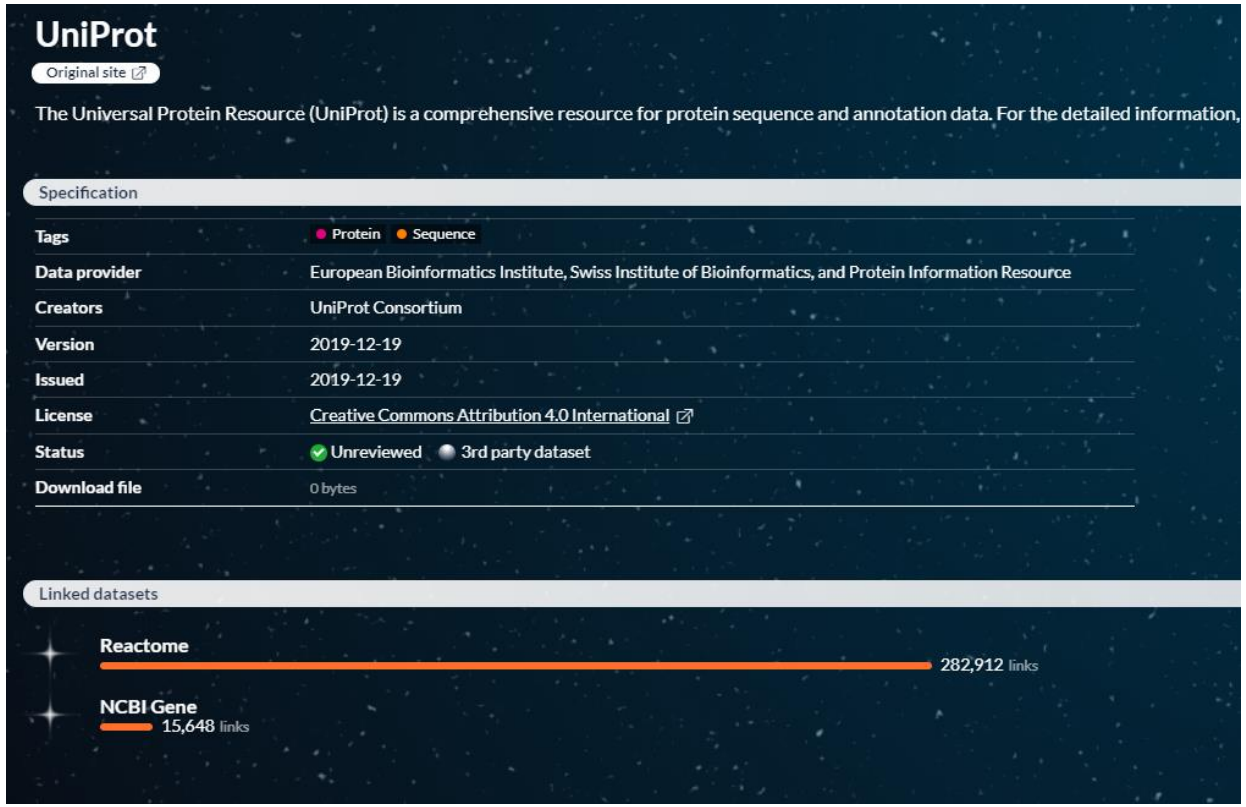
登録対象データセットの拡張

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



NBDC RDFポータル

- 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。