

川島 秀一, 千葉 啓和, 五斗 進, 細田 正恵, 池田 秀也, 片山 俊明, 三橋 信孝, 守屋 勇樹, 山本 泰智

Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems

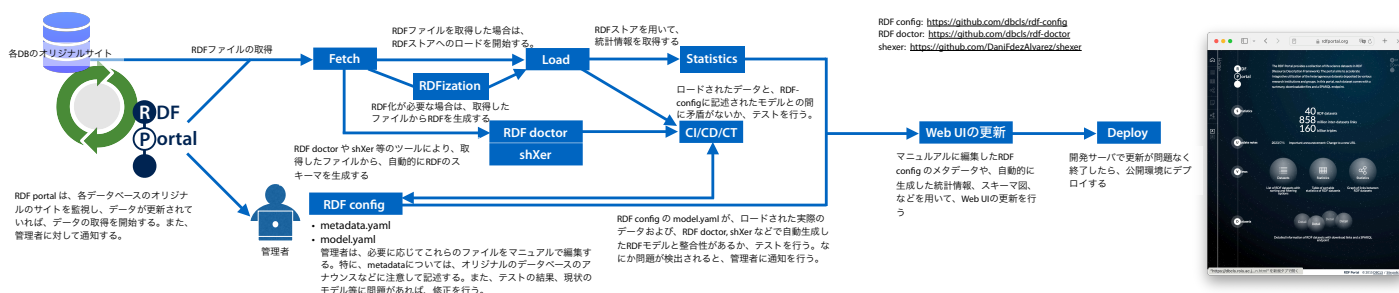
RDFポータルの開発状況とその進展について紹介する。RDFポータルは、生命科学におけるRDFデータへのアクセスを容易にすることを目的として、2015年よりサービスが開始された。また本ポータルは、単なるRDFデータの収集にとどまらず、データの相互運用性を向上させるためにガイドラインとレビュープロセスを導入し、高品質なRDFデータ基盤を構築してきた。これにより、DB統合化推進プログラムを含む国内RDF開発者の協力のもと、包括的なRDFデータ・リソースの提供が可能となっている。

現在、RDF-configを用いて記述されたRDFデータセットのメタデータを活用し、RDFポータルのデータ更新プロセスの自動化を進めている。RDF-configの仕様準拠したメタデータを整備することで、最新のRDFデータと、各種メタデータ、データスキーマ図、サンプルクエリなどが同期され、データ更新プロセスのさらなる自動化と、利用者の利便性の向上が期待される。

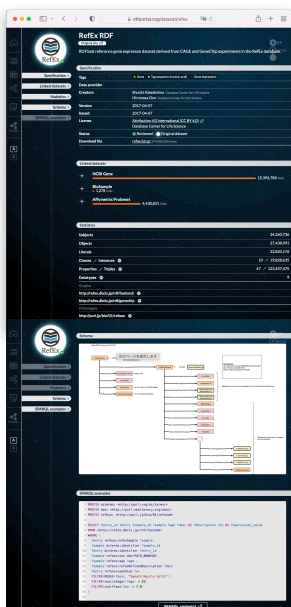
今後のデータ更新プロセス

現在、RDFポータルのデータ更新システムの自動化を進めている。現状では、以下の2点が特に問題となっている。1) 更新プロセスの大半がマニュアル作業に依存しているため、更新作業に時間がかかる。2) RDFデータのモデルが変更された場合、スキーマ図に表現されたモデルと、実際にロードされたデータとの間に齟齬が生じる。つまり、RDF-configに記述されたモデルと、実際のRDFとの間に齟齬があることになり、これは、自動的なデータベースの統合利用を実現するうえで障害となる。

1) については、rdf-portalの更新を自動化するプログラム (rdfportal-config) を開発することで解決しようとしている。また、2) については、RDF-configの整備と、RDF-doctor, shXerなどの、RDFデータの検証や自動的にモデルを生成するツールを利用して、更新時にテストを行う、RDFモデルの変更やエラーの検出することで対応することを計画している。



Web interface

Metadata
section

RDF schema

SPARQL
examples &
link to endpoint

Example of a dataset page in RDF portal (RefEx database)

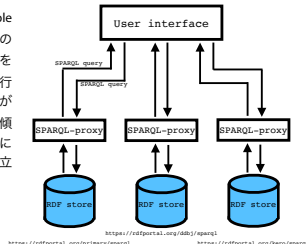
Current datasets

2015年に運用が開始されたNBDC RDFポータルは、ライフサイエンス分野におけるRDFデータのアクセシビリティの課題に対処することを目的としてきた。単にRDFデータを集めるだけではなく、ガイドラインとレビュープロセスを導入することでデータの相互運用性を高め、高品質なRDFデータ基盤の確立を実現してきた。これらは、科学技術振興機構の「ライフサイエンスデータベース統合推進事業」が支援する「データベース統合化コーディネートプログラム」に参加する国内のRDF開発者との連携により実現したものである。また、RDF登録の範囲を拡大しており、従来のようなサブミッションされたデータに加え、UniProtやPubChemなどのパブリックなRDFデータや、DBCLSが開発したRDFデータも収録対象になっている。これにより、RDFの特性を最大限に活かしたアプリケーション開発が促進されることが期待される。2023年度より、本サービスの管理が、NBDCからDBCLSに移管され、これに伴い、サービス名称が「NBDC RDF Portal」から「RDF Portal」に変更された。URLもhttps://integbio.jp/rdfからhttps://rdfportal.orgに変更され、エンドポイントのURLも変更されている。

Dataset	Major topic	Description	#triples	Dataset	Major topic	Description
UniProt	Protein	Comprehensive resource for protein sequence and annotation data	54,381,877,065	QCD3 Ortholog database RDF	Gene	13,652,175
DBJ	Gene	Semantic Representation of DBJ Annotated Sequence Records	36,574,632,446	Dataset of WURCS-RDF	Glycan	6,213,789
molSOL/RDF	Protein structure	Translation of PDB/POBMA data into RDF	16,203,398,274	KNhSack RDF	Compound	3,776,538
PubChem	Compound	Open chemistry database	15,079,698,568	GlyTouCan	Glycan	1,748,948
QSPERO RDF	Genes	Multi-omics data sets including RNA-seq, ChIP-seq, BS-seq and TSS-seq	11,186,913,244	HomoGene RDF	Gene	1,593,100
Tagged RDF	Metadatas	Public biotechnology databases	7,770,752,116	HCNC	Gene	1,314,653
Open TIS-GETA	Transcriptomics	Links between database identifiers	6,800,384,600	Metadatas of JCM resources	Bioresource	1,106,776
PubMed	Literature	Biomedical literature from MEDLINE, life science journals, and online books	2,531,937,601	Dataset to NBRC bioresource	Bioresource	868,818
MBGD RDF	Gene	Microbial Genomes Database for Comparative Analysis	1,609,018,143	NLM Catalog	Literature	203,480
ClinVar	Variation	Relationships between medically important variants and phenotypes	1,247,658,822	Integbio Database Catalog/RDF	Metadatas	117,526
PubMed Central	Literature	Automatic annotations of biomedical concepts in PubMed and PMC articles	1,107,708,438	Dataset of NARD Genbank	Gene	132,210
BMRB/RDF	Protein structure	Translation of NMR-STAR data into RDF	643,511,337	PAConto	Disease	81,785
Linked KEGG RDF	Transcriptomics	Small chemical compounds for drug screening on computer docking studies	577,082,774	SBSD	Images	40,300
NBDC Nihari/RDF	Compound	Japan Chemical Substance Dictionary (Nihari)	344,246,952	GDConto	Disease	39,440
Lingfai/RDF	Compound	Small chemical compounds for drug screening on computer docking studies	283,808,849	GlycoFitpe	Glycan	30,150
POST database RDF	Proteins	Re-analyzed proteins dataset in the POST project	259,817,763	NARDO	Disease	28,425
Muslin	Disease	Information related to human medical genetics	126,654,418	PhyloR, BCL, S, D, I definition	Polymer	5,448
Raffi RDF	Transcriptomics	Gene expression dataset derived from CAGE and GenChip	123,447,475	Nuclear Acid Drug Database	Drug	348
Quanto	Metadatas	Sequencing quality of public high-throughput sequencing data	107,782,639	In progress	In progress	In progress
ChGNET RDF	Disease	A collection of genes and variants associated to human diseases	88,376,594	CHEMBL	Bioassay	NA
FAMBASE GPCR	Protein structure	Predicted protein structures of GPCR	21,297,798	Ensembl	Gene	NA
PubCafFinder RDF	Disease	Disease-Phenotype associations and Gene-Phenotype associations	16,367,386	BioSample	Bioresource	NA

SPARQL endpoints

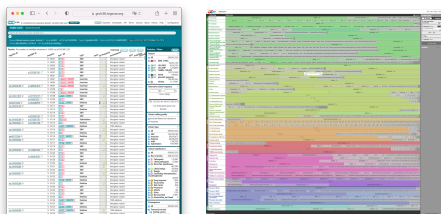
RDFポータルでは、Virtuoso Triple Storeを使用している。これまでの経験では、トリプル数が200億を超えると、必要な統計処理が実行できないなどの問題が発生する傾向がある。そのため、次のように複数のVirtuosoインスタンスを立ち上げている。



	Endpoint URLs
Primary	https://rdfportal.org/primary/sparql
DBJ	https://rdfportal.org/ncbi/sparql
DBKERO	https://rdfportal.org/ker/sparql
PDB	https://rdfportal.org/pdb/sparql
SIB	https://rdfportal.org/sib/sparql
EBI	https://rdfportal.org/ebi/sparql
NCBI	https://rdfportal.org/ncbi/sparql
PubChem	https://rdfportal.org/pubchem/sparql

As a Foundation for other DB services

RDFポータルは、データポータルとしてだけでなく、TogoDX、TogoVar、TogoIDなどDBCLSのサービスや、NITENBRCのDBRP Stanzaのような外部のサービスまで、さまざまなサービスの基盤となるデータソースとしての役割も果たしている。



RDF data submission

RDFポータルでは、RDFデータの投稿を受け付けています。投稿可能なデータは、ライフサイエンス分野のRDF化された知識ベースに限定されます。実験機器が生成した生データのみを表すRDFデータは投稿できません。提出の際には、RDFファイルに加えて、RDF-config(https://github.com/dbcls/rdf-config)の仕様に基づく一連のYAML形式の構成ファイルも提供する必要があります。RDFポータルに提出されたRDFデータは、RDFポータルチームによる審査プロセスを経ます。審査では、RDFデータがDBCLSの「データベースのRDF化ガイドライン」(https://github.com/dbcls/rdfizing-db-guidelines)に記載された要件を満たしているかどうか評価されます。審査に同意される場合は、https://dbcls.rois.ac.jp/contact.htmlまでご連絡ください。

