

○片山 俊明

ライフサイエンス統合データベースセンター(DBCLS)

生命医科学における主要なDBのRDF化が進展してきた
(BioHackathon, SPARQLthon等での国内外の連携も貢献)

- | | | | |
|--|---|---|--|
| <ul style="list-style-type: none"> Nucleotide seq & annot <ul style="list-style-type: none"> INDSDC (DDBJ/DBCLS) Genome <ul style="list-style-type: none"> Ensembl (EBI) RefSeq (TogoGenome) Protein seq & annot <ul style="list-style-type: none"> UniProt (SIB) Protein structure <ul style="list-style-type: none"> PDB (PDBj) BMRB (PDBj) FAMSBASE (Chuo U) Compounds <ul style="list-style-type: none"> PubChem (NCBI) ChEMBL (EBI) Nikkaji (JST) Gene expression <ul style="list-style-type: none"> RefEx (DBCLS) ExpressionAtlas (EBI) Samples <ul style="list-style-type: none"> BioSamples (EBI/DDBJ) JCM (RIKEN) | 












 | <ul style="list-style-type: none"> Biomedical (Med2RDF) <ul style="list-style-type: none"> ICGC, COSMIC, CIViC DGIdb, OpenTG-Gates ClinVar, dbSNP, dbVar ExAC, gnomAD HiNT, INstruct Glycome <ul style="list-style-type: none"> GlyTouCan, GlycoEpitope, WURCS, GGDonto, PAConto Proteome <ul style="list-style-type: none"> jPOST The Human Protein Atlas Pathway <ul style="list-style-type: none"> Reactome (EBI) Others <ul style="list-style-type: none"> MeSH (NCBI) BioModels (EBI) MBGD (NIBB/DBCLS) Quanto (DBCLS) : | 











 |
|--|---|---|--|

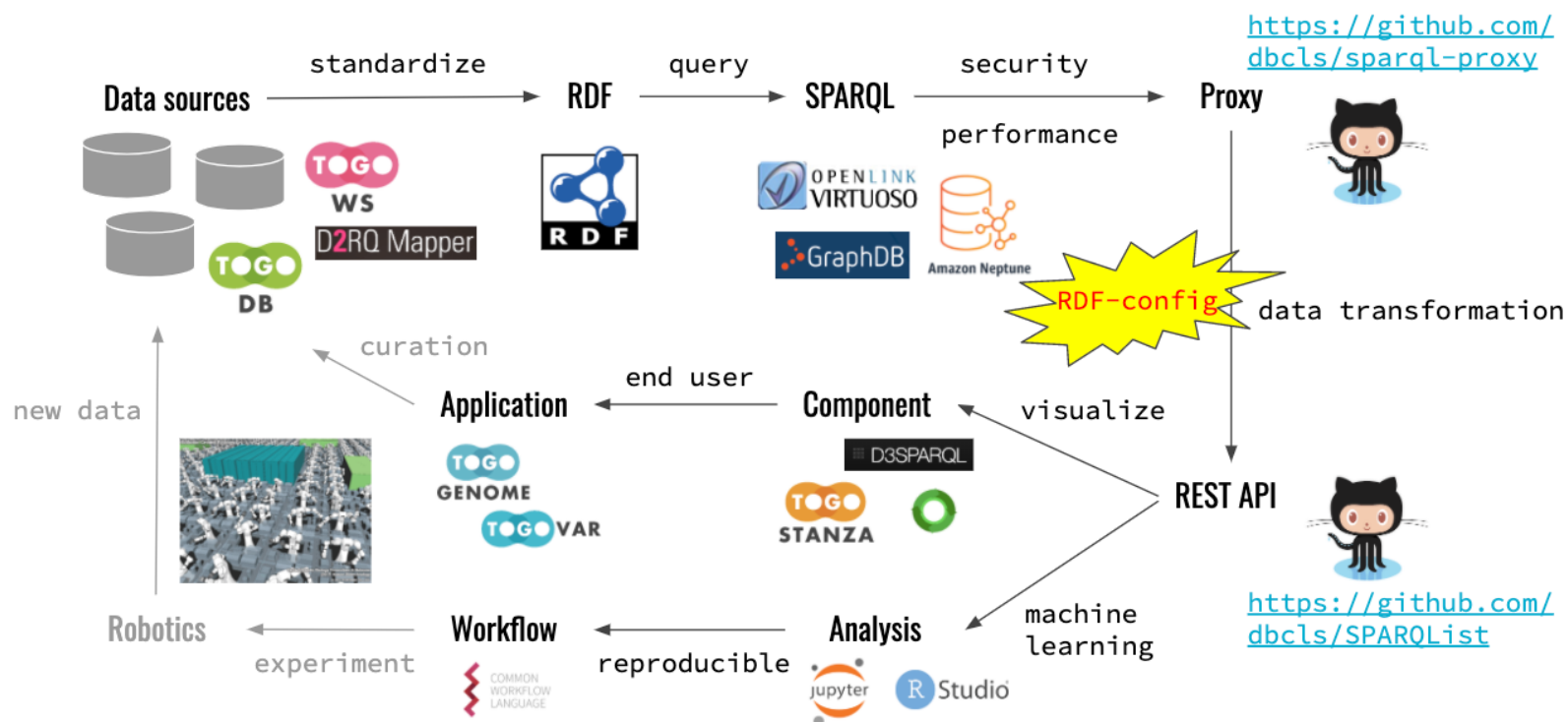
NBDC RDFポータルでRDFデータを蓄積

<https://integbio.jp/rdf/>

Name	Triples	Links	Classes	Instances	Literals	Subjects	Properties	Objects
UniProt	35,998,089,442		0	123	7,814,089,346	805,432,887	7,565,403,474	120
DDBJ	20,067,185,022	389,818,023		109	1,878,437,947	950,890,006	2,913,415,115	144
PubChem	11,484,688,918		0	125,668	2,815,763,079	631,121,824	3,539,254,440	67
DBKERO RDF	11,017,998,412		0	38	929,457,877	1,764,181,364	929,551,492	60
wwwPDB/RDF	10,575,823,782	4,270,350		591	747,934,581	7,762,574	748,285,479	3,624
Open TG-GATES	6,800,384,609	1,181,705		65	1,497,955,718	1,267,395,887	1,497,955,718	38
MBGD RDF	1,609,018,143	43,260,265		32	273,443,876	74,289,149	309,702,751	79
BMRB/RDF	586,142,662	21,689,368		419	32,451,702	1,822,147	32,696,228	2,823
Linked ICGC Dataset	577,082,774	57,483		9	51,410,015	20,735,685	51,410,016	66
NBDC NikkajiRDF	333,968,051		0	38,937	23,738,365	63,322,504	60,432,596	41
jPOST database RDF	239,817,763	1,159,902		86	68,502,884	11,392,480	68,502,952	77
RefEx RDF	123,447,475	19,828,887		10	19,828,635	22,820,176	24,260,736	47
Quanto	107,782,639	1,995,973		9	21,955,729	10,369,656	21,955,729	37
FAMSBASE GPCR	21,297,786	1,328,293		16	5,858,908	488,759	5,858,909	30
PubCaseFinder RDF	18,579,330	775,640		7	2,213,172	19,030	2,222,999	11
PGDBJ Ortholog database RDF	13,652,175	499,798		11	1,963,733	1,858,372	1,963,741	35
Dataset of WURCS-RDF	6,213,789		0	14	817,535	40,435	1,365,653	56
GlyTouCan	1,749,648		0	20	375,657	126,879	375,657	30
Dataset of NBRC bioresource	787,964	14,082		9	147,718	127,928	147,718	27
Metadata of JCM resources	760,271	20,048		10	145,138	155,593	145,138	39
Integbio Database Catalog/RDF	98,913	1,536		9	8,523	15,353	11,595	41
PAConto	81,785	4,362		63	9,296	8,586	9,329	117
SSBD: Meta-information of quant...	40,300		0	18	6,644	3,807	6,644	33
GGDonto	39,439	1,024		23	1,705	4,963	1,782	943
GlycoEpitope	27,796		0	24	5,726	5,453	8,678	35
NANDO	24,952		0	3	2,850	8,485	2,350	17
Nucleic Acid Drug Database	948	16		5	112	162	112	16
Total	99,584,784,788	485,906,755		166,328	16,186,525,971	5,634,400,144	17,774,947,031	8,653

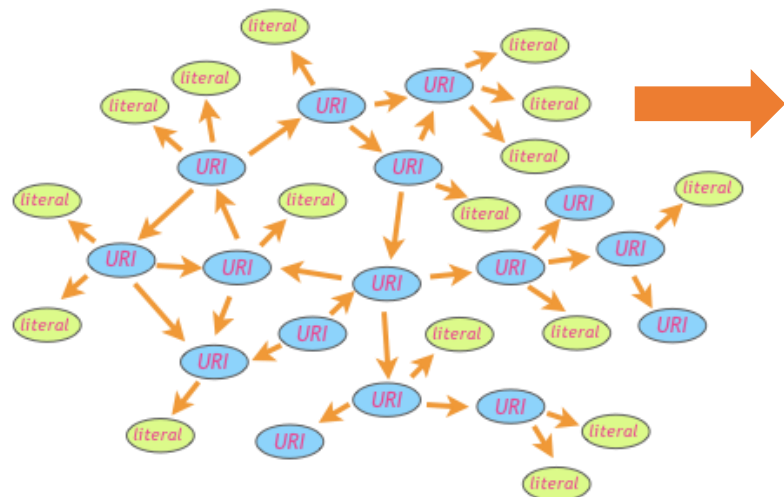
NBDC RDF Portal © 2015 NBDC / Site policy

DBCLSでは、RDFの生成・利活用のためのソフトウェア開発を進めてきたが、RDFはグラフデータであるため、多様なデータを統合できるメリットを持つ一方、複雑なグラフ構造を読み解いて必要な情報を取り出すにはデータモデルの理解が必要



グラフデータにおけるデータモデルを簡潔に表現し、スキーマ図やSPARQL検索クエリを自動生成できれば、統合されたRDFデータの利活用やアプリケーション開発が飛躍的に容易になるのでは？

RDFデータ

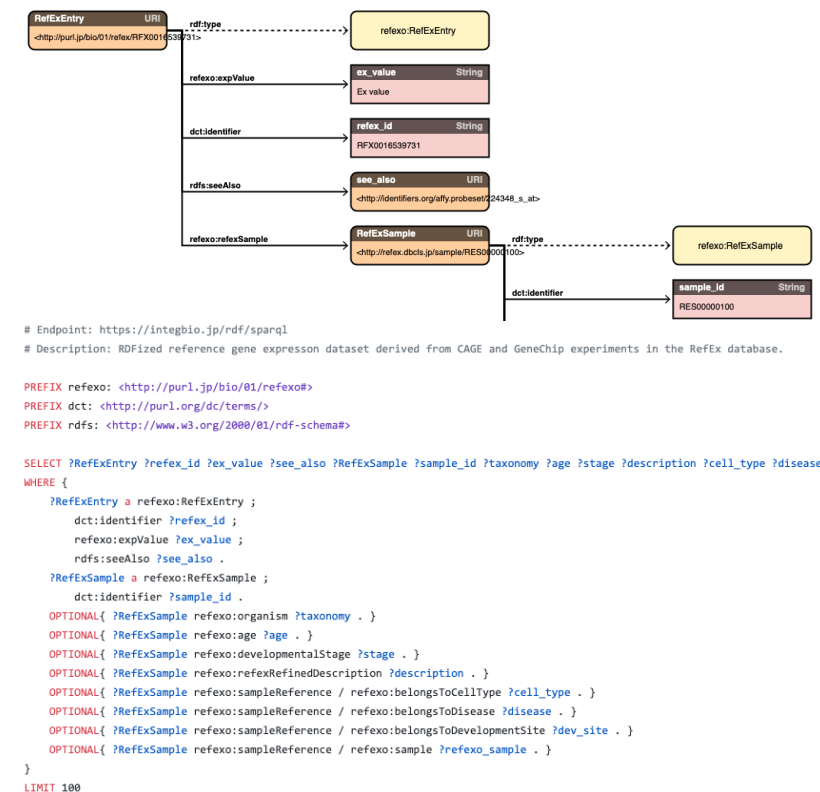


データモデル定義ファイル (YAML)

```
- RefExEntry <http://purl.jp/bio/01/refex/RFX0016539731>:
  - a: refexo:RefExEntry
  - refexo:expValue:
    - ex_value: Ex value
  - dct:identifier:
    - refex_id: RFX0016539731
  - rdfs:seeAlso:
    - see_also: <http://identifiers.org/affy.probeset/224348_s_at>
  - refexo:refExSample:
    - sample: RefExSample

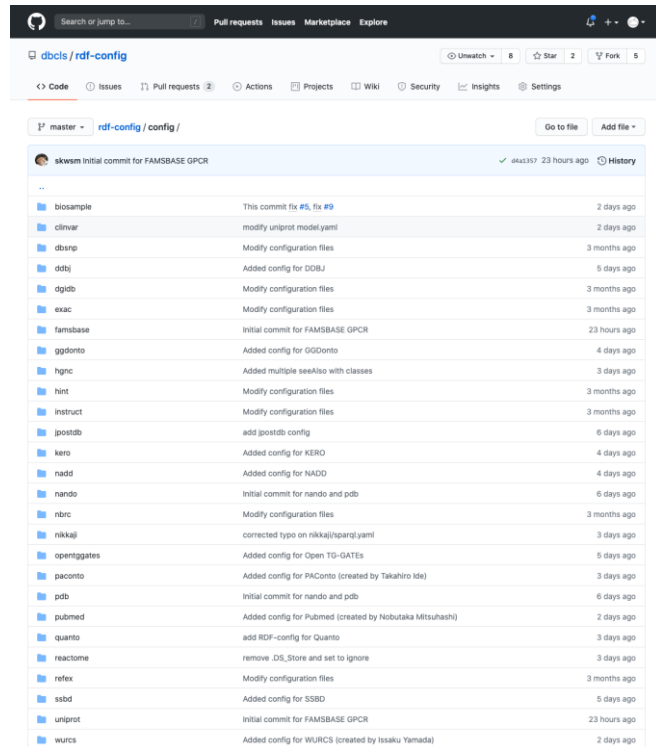
- RefExSample <http://refex.dbcls.jp/sample/RES00000100>:
  - a: refexo:RefExSample
  - dct:identifier:
    - sample_id: RES00000100
  - refexo:organism?:
    - taxonomy: <http://identifiers.org/taxonomy/9606>
  - refexo:age*:
    - age: obo:PATO_0000383
  - refexo:sex?:
    - sex: obo:PATO_0000383
  - refexo:developmentalStage*:
    - stage: adult
```

スキーマ図とSPARQLクエリ (自動生成)



RDFのグラフ構造を、主要な要素とその属性、値を取得する変数名と、値の例をYAMLによる定義ファイルに記述することで単純化

現在、主要なRDFデータのデータモデル定義ファイルの開発が進行中



<https://github.com/dbcls/rdf-config>

統合TV ([1](#), [2](#))と[スライド](#)もご参照下さい

RDF-configのポイント

- 複雑なグラフデータを、骨格となる「**主要な要素**」とその「**属性**」および、その「**値**」、の木構造に単純化してDBごとに把握しやすく（RDFの「**主語**」、「**述語**」、「**目的語**」に相当）
- 「**目的語**」の「**値**」を取り出すために、「**変数名**」をつけることで、SPARQLクエリやウェブアプリケーションの見通しをよくする
- 汎用的なYAML形式を採用することで、定義ファイル自体の検証や、他のアプリケーションでの再利用性を向上
- 手間のかかるスキーマ図やSPARQL検索クエリの自動生成により、RDFデータの提供者および利用者の効率を大幅に改善

今後の予定

- 対応DBの拡充と、複数のDBを統合的に利用する仕組みを標準化する
- MetaStanzaによりSPARQL検索結果を汎用的に可視化する
- ShEx/SHACLによるRDFデータの検証を実現する
- SPARQListによるREST API や、GraspによるGraphQL APIを提供する