バイオインフォマティクスを使い尽くす秘訣教えます！【第6回】

RDFによるデータベース統合化技術

河野　信

テクノロジーの発展に伴って，日々大量のデータが産出されるようになった．また，昨今のオープンサイエンスの流れで，論文の根拠となったデータは公共データベースに登録することが求められ，これまた大量の公共データが利用可能となっている．しかし，我々はこれらのデータを十分に活用できているだろうか？

目的のデータがどこにあるかわからない，使い方がわからないといった問題の解決方法は，本連載の第1回[1]ならびに第3回[2]で紹介した．本稿ではデータ自体を統合化してつなげ，それを活用する技術について紹介したい．

文章のウェブとデータのウェブ

なぜ我々はデータの海の中を自由に泳ぎ回れないのか？我々はWorld Wide Web（WWW）でつながった文章の海であれば，ある程度自由に泳ぎ回ることができる．それは，文章のウェブがHyper Text Markup Language（HTML）という共通の形式で記述され，また文章と文章の間がハイパーリンクでつながっているからである．一方でデータの場合，テキストファイルであったり，表計算ソフトファイルであったり，PDFファイルであったり，データベース（DB）であったり，さまざまな形式で記述され，またその中身もそれぞれ独自の表現が使われているため，そもそもつながりようもなく，我々はそれぞれの池の中でしか泳ぐことができない．これらの分断された池をまたいで自由に泳ぎ回るためには，文章のウェブと同じようにデータのウェブも共通の形式で記述し，それらの間にリンクを張ってやる必要がある．この際，文章は人が書くので自分自身で他の文章とのリンクを指定できるが，データの場合は測定機器や解析ソフトなどから出てくるものも多く，そもそも大量であるため，なるべくコンピュータに自動的にリンクを生成してもらう工夫が必要である．

このような研究データの統合的な利用を可能とするために，バイオサイエンスデータベースセンター（NBDC）ならびにライフサイエンス統合データベースセンター（DBCLS）では現在，Resource Description Framework（RDF）と呼ばれる技術を採用し，データのウェブ化の実現を目指している．しかしながら，実際のところRDFの技術は，より便利な検索を実現するために裏側で使われている基盤的な技術であり（みなさんも自身のホームページを作成する際に直接HTMLは書かないであろう），みなさんの目に触れる機会は少ないかもしれないが，最先端のデータ統合技術について知っていただければ幸いである．

RDFとは？

RDFはデータを記述・相互交換するためのもので，文字通りリソース（resource）を記述する（description）ための枠組み（framework）である．たとえば，遺伝子や論文，人など，あらゆるものがリソースになりうる．これらリソースに対して，リソース間の関係や，リソースの持つデータを紐付けていく．これを実現するためには，前述した通り，1）データの形式が揃っていること，2）データを表現するための用語が揃っていること，これに加えて，3）リンク先を特定するためにリソースが一意に特定できること，が必要である．

データ形式　RDFはインターネットの標準化団体World Wide Web Consortium（W3C）によって規格化されている世界標準である．RDFは非常にシンプルな構造をしており，すべてのデータは主語・述語・目的語からなる三つ組で表現される（図1）．これをトリプルと呼ぶ．主語と述語はURI（後述）で記述され，目的語が他のリソースの場合はURIで，値（データ）の場合は数字や文字列で記述される．

データを表現するための用語　データを統合的に扱うためには，それぞれのデータを説明するための統一された用語が必要である．RDFではオントロジーと呼ばれるものを利用する．オントロジーは，ただ単に統一した用語を定義するだけでなく，用語間の関係についても定義する．本稿では詳しく触れないが，オントロジーを利用することで，コンピュータによる推論が可能になる．たとえば，Gene Ontology（GO）ではcellulase activity（GO:0008810）はhydrolase activity（GO:0016787）の

著者紹介　情報・システム研究機構データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター（特任准教授）　E-mail: kawano@dbcls.rois.ac.jp

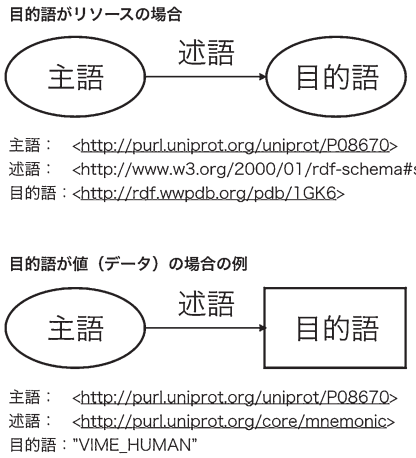

図1．RDFトリプルの例．すべてのリソースは主語・述語・目的語の三つ組で表現される．

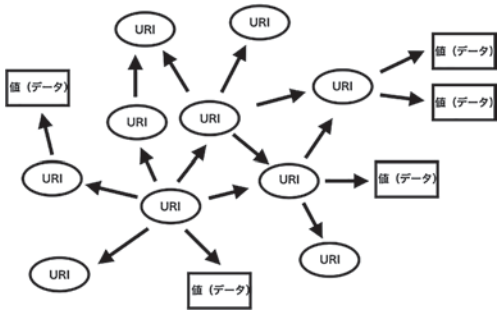

図2．さまざまな種類のデータがURIを介して相互に接続されてLODを形成する．

下位概念として定義されているので，とある遺伝子産物の機能としてcellulase activityとしか記述されていなくても，コンピュータはhydrolase activityの一種であることをオントロジーから自動的に認識することができる．

URI　RDFでは，リソースを一意に表すためのIdentifier（ID）として，Uniform Resource Identifier（URI）と呼ばれるものを利用する．特にインターネット上では，Uniform Resource Locator（URL）が利用される．たとえば，UniProtではヒトのVimentinは<http://purl.uniprot.org/uniprot/P08670>と記述される．このようにURI（URL）を使うことで，インターネット上でリソースを一意に特定することができ（誰がどこからアクセスしても上記アドレスはVimentinのページである），異なるDB中であっても同じURIで記述されていればコンピュータにも同じリソースであることが認識可能である．逆に，多義語（たとえば，生き物の“mouse”とコンピュータデバイスの“mouse”）の場合も，異なるURIを使えばコンピュータにも区別可能になる．

LOD　RDFのトリプルではすべてのリソースをURIで表現するため，共通のURIを介してネットワークが形成される．同じURIが使われていれば，個々のDBを越えてデータが自動的にリンクする．このようなURIがリンクしてつながったデータをLinked Open Data（LOD）と呼ぶ（図2）．

SPARQL　SPARQL（SPARQL Protocol and RDF Query Language）は，RDFデータを検索するための言語である．LODから目的のデータを検索するために，RDFのパターンをSPARQLで指定する．パターンを記述して検索するため，通常のキーワード検索と比べて非常に柔軟な検索が可能となるが，記述が非常に複雑になるという問題があり，これを使いこなすためには相当の習熟が必要となる．

RDFデータセット　このようなRDFを利用した統合化の試みは日本独自のものではなく，海外の主要なDBも徐々にRDFへの対応が進んでいる．特に欧州のEuropean Bioinformatics Institute（EBI）ではEBI RDF Platform[3,4]として，ゲノムDBのEnsembl[5]やタンパク質DBのUniProt[6]を含む7個のDBが，米国National Center for Biotechnology Information（NCBI）からは，化合物DBのPubChem[7]や本連載第2回[8]で紹介された文献用の統制語彙を集めたMeSH[9]などのDBが，すでにRDF化されて公開されている．また，日米欧3極体制で構築されているタンパク質の立体構造DBであるPDB[10]もすでにRDF化されている．

RDFデータの作り方

自身の持つデータをRDF化してLODと接続すれば，これまでにRDF化されているリソースと自動的にリンクすることになる．“つながる”RDFを作るためには上述の通り適切なURIとオントロジーを選択する必要がある．このようなRDFを設計するうえで考慮すべき事柄は「データベースのRDF化ガイドライン[11]」としてまとめられているので，自身のデータをRDF化する際の参考になるであろう．また，DBCLSでは大量データをRDFに変換するためのツールも用意している．

TogoDB　TogoDB[12]では，表計算ソフトなどで作成されたタブ区切りやコンマ区切りのデータをアップロードして，表をまるごとRDF化する（図3）．各行にURIを割り当て（主語），カラムを述語，セルの値を目的語としてRDFを自動的に作成する．値のURI化，オ

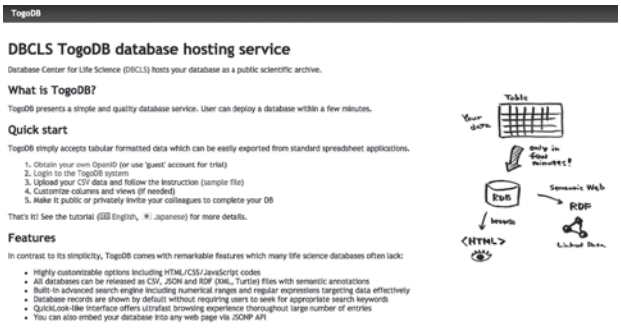

図3．TogoDBでは表形式のデータをアップロードすることでRDFを作成する．

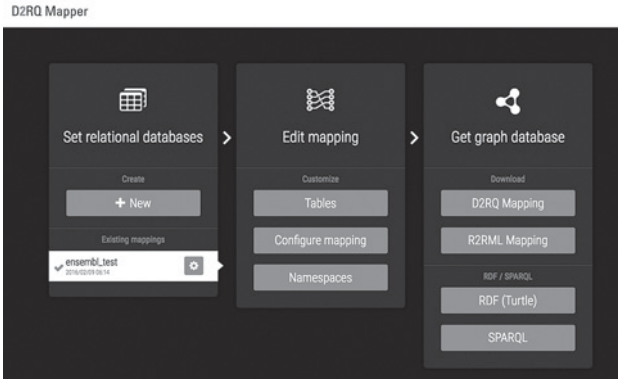

図4．D2RQ Mapperでは，リレーショナルDBからRDFを作成する．

ントロジーの指定も可能であるため，即座にLODにつながるRDFを作成することができる．

D2RQ Mapper　D2RQ Mapper[13]はデータがMySQLなどのリレーショナルデータベースに格納されている場合に利用可能なRDF化ツールである（図4）．TogoDBの場合と同様に，どの値やカラムがどのURIやオントロジーに対応するか指定（mapping）することでRDFを作成する．

SPARQLthon　自身のデータをRDF化するにあたって，上記のようなガイドラインや支援ツールはあるものの，やはり未経験者にとってRDF化の障壁は高い．そのような方たちをサポートするため，DBCLSではSPARQLthon[14]というデータをRDF化するためのイベントを毎月開催している．興味のある方はぜひ一度ご参加いただければと思う．

RDF Portal

これまで述べてきたように，異なるDBに収録されているデータを共通のIDやオントロジーを使って記述す

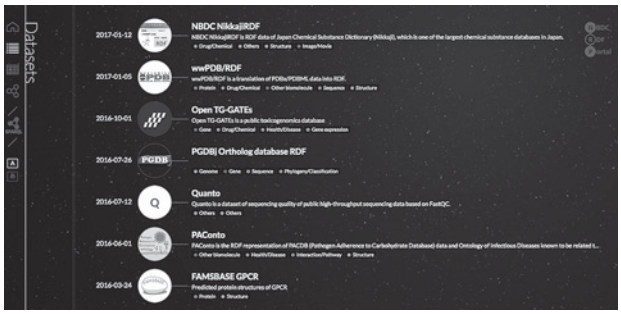

図5．RDF portalに収録されているRDF化されたデータベースを一覧できる．

ることで，それぞれのDBの中でしか通用しなかったデータが，インターネットを通じてつながるようになる．本来，RDFデータは必ずしも物理的に同じ場所にある必要はないが，どういうRDFデータがあるか，というRDF版のDBカタログがあると便利であろうという考えから，NBDCではRDF Portal[15]を提供している（図5）．ここでは，国内で公開されているRDFデータセットの一覧ページのほか，統計情報や他のデータセットとのリンクに関する情報がまとめられている．また，RDF形式のデータファイルをダウンロードできるほか，ここからSPARQL検索も可能となっている．もし，ご自身でRDFデータをお持ちの場合，掲載が可能なのでぜひご一報いただきたい．

RDFを活用したサービス

RDFによってLOD化されたデータは，相互につながっているものの，それゆえにリンクが複雑に絡み合っており，また検索用言語であるSPARQLも難解なものとなっていて，そのままの状態では使いづらい．そこで，より容易にRDF化・LOD化されたデータを活用してもらうことを目的として，DBCLSではいくつかのサービスを開発して提供している．

TogoStanza　Stanzaとは意味を成す一つのまとまりを指す．TogoStanza[16]は，RDF化されたデータの中から，意味のある単位でSPARQL検索を実行し，結果を可視化したものである（図6）．それぞれのStanzaはウェブページに埋め込むことが可能で，Stanzaを組み合わせて自由にページをデザインすることができる．

TogoGenome　TogoGenome[17]は，ゲノム・遺伝子・タンパク質・生物種・表現型・生育環境などの多面的な情報をRDF化し，DB化したものである．性質の異なるデータを共通のID体系（URI），共通のオントロジーを

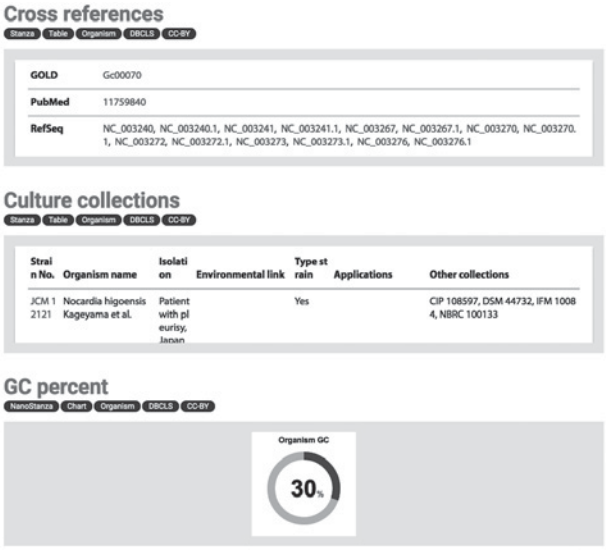

図6．TogoStanzaはSPARQL検索の結果を可視化する．一つのBoxが一つのStanzaになっており，Stanzaを自由に組み合わせて表示画面を作ることができる．

用いて共通の形式（RDF）で記述しているため，データ横断的にさまざまな切り口からの検索が可能になっている．各種条件でフィルタリングした結果から得られたリンクをクリックすると，それぞれのリソースに対するさまざまなアノテーション情報がTogoStanzaを組み合わせて表示されるようになっている（図7）．

TogoTable　TogoTable[18,19]は，なんらかのDBのID（たとえば，UniProt ID，PDB ID，PubMed IDなど）をカラムに含む表形式のデータをアップロードして，そのIDからたどれるアノテーションを指定することで，すべての行にわたって関連する情報を一括で取得できるウェブツールである（図8）．LODに対して検索を行っているため，たとえばPDB IDからUniProtに収録されているアノテーションデータを直接取得するなど，DBをまたいだ情報取得が可能である．

SPARQL Builder/LODQA　前にも触れたように，RDFを検索するためにはSPARQL言語を取得する必要があり，特にWet研究者にとっては敷居の高いものになっている．より多くの方にRDF化されたデータを使ってもらい，LODの恩恵に預かってもらうため，これまでに紹介してきたツール以外にも，SPARQL言語を意識することなくRDFデータを利用可能にするツールの開発を進めている．たとえばSPARQLbuilder[20]ではグラフィカルに項目を選択することで，それらを検索するためのSPARQLを自動生成する（図9）．またLODQA[21]

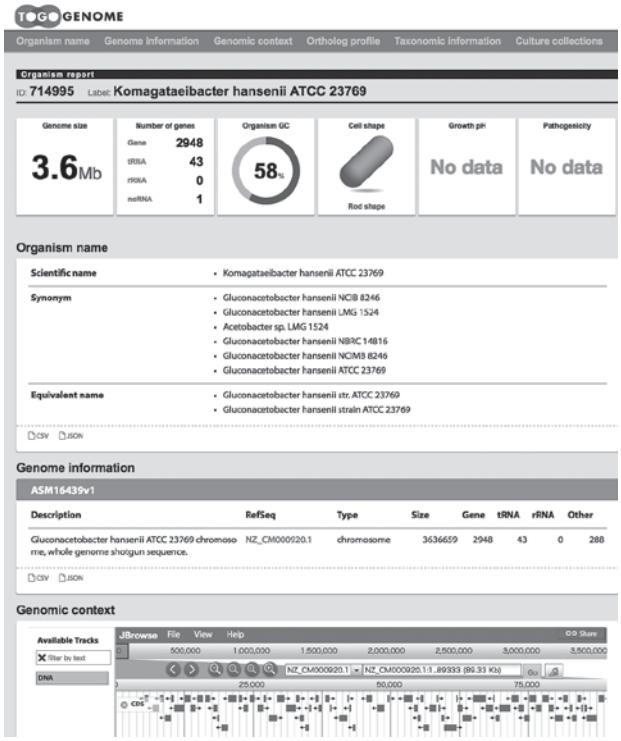

図7．TogoGenomeで生物種情報を表示した画面．白いBoxがそれぞれTogoStanzaで作られており，それらを呼び出して画面を構成している．

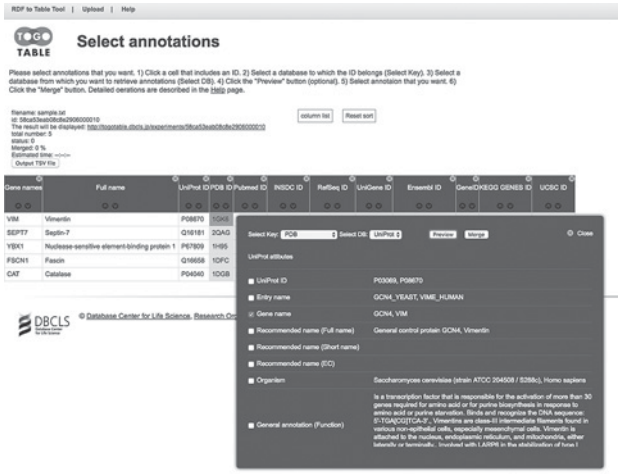

図8．TogoTableでPDB IDからUniProtに収録されている情報を取得している．

では，“Which diseases are associated with SAR1B?”などのように自然文で質問すると，自動的に構文解析を行ってSPARQLを生成し，検索が行えるサービスを提供している（図10）．

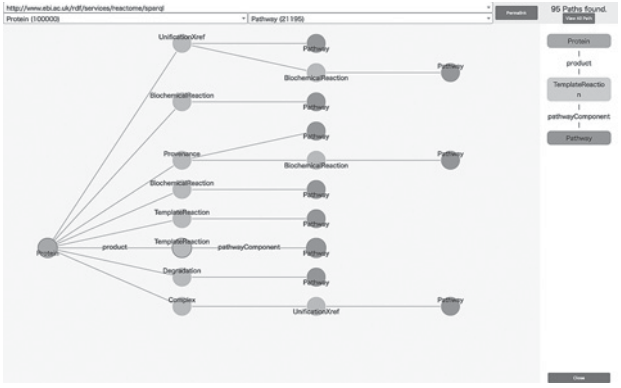

図9. SPARQL builderでは，画面上で始点と終点を指定することで，クエリを作成できる．

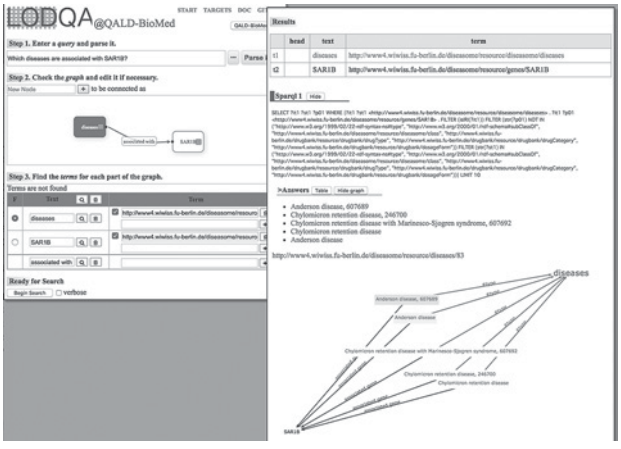

図10. LODQAで“Which diseases are associated with SAR1B?”を検索した結果．右側のSparql 1に記述されているのが実際に検索に使われたSPARQL（自動生成）で，Answersが質問の答えである．

おわりに

本稿では，さまざまなデータを活用するためにNBDC/DBCLSで進めているRDFによるDB統合化技術について紹介した．RDFはまだまだ開発途上の技術であり，使いにくい部分もあるだろうが，本稿でも触れたように世界のトレンドであるので，これから大きく発展していくものと思われる．今後に期待していただきたい．

文　献

1) 坊農秀雅：生物工学，**94**, 572 (2016).
2) 飯田啓介，小野浩雅：生物工学，**95**, 40 (2017).
3) RDF Platform: https://www.ebi.ac.uk/rdf/ (2017/04/13)
4) Jupp, S. *et al.*: *Bioinformatics*, **30**, 1338 (2014).
5) Ensembl RDF: https://www.ebi.ac.uk/rdf/services/ensembl/ (2017/04/13)
6) The UniProt Consortium: *Nucleic Acids Res.*, **42**, D191 (2014).
7) Fu, G. *et al.*: *J. Chemonform.*, **7**, 34 (2015).
8) 山本泰智：生物工学，**94**, 722 (2016).
9) Bushman, B. *et al.*: *J. Libr. Metadata*, **15**, 157 (2015).
10) Kinjo, A. R. *et al.*: *Nucleic Acids Res.*, **40**, D453 (2014).
11) データベースのRDF化ガイドライン：http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline (2017/04/13)
12) TogoDB: http://togodb.org/ (2017/04/13)
13) D2RQ Mapper: http://d2rq.dbcls.jp/ (2017/04/13)
14) SPARQLthon: http://wiki.lifesciencedb.jp/mw/SPARQLthon (2017/04/13)
15) NBDC RDF Portal: https://integbio.jp/rdf/ (2017/04/13)
16) TogoStanza: http://togostanza.org/ (2017/04/13)
17) TogoGenome: http://togogenome.org/ (2017/04/13)
18) TogoTable: http://togotable.dbcls.jp/ (2017/04/13)
19) Kawano, S. *et al.*: *Nucleic Acids Res.*, **42**, W442 (2014).
20) SPARQL Builder: http://sparqlbuilder.org/ (2017/04/13)
21) LODQA: http://lodqa.org/ (2017/04/13)