2

生命科学分野における LOD の構築と利用
― DBCLS における活動事例―

応用

山本泰智（ライフサイエンス統合データベースセンター／ DBCLS）

生命科学のデータベース

● データの多様性と急増

生命科学は生物を生物たらしめている仕組みを解き明かすことを目指す非常に長い歴史を持つ学問であり，少なくとも文献として確認されている最古の研究成果は紀元前のギリシアにまでさかのぼるとされる．紀元前 500 年頃に Alcmaeon が目の神経を記載していたことから，これは生命科学の研究分野のうち，解剖学にあたることになる．その後，生命科学の研究は下火になる時代を経ながらも細菌から植物，動物に至るまで地球上に生きる膨大な数と種類の生物を対象にしてさまざまな研究者がさまざまな切り口で実験や観察を行い，得られた知見を自然言語で記述してきた．現在（2016 年），生命科学に関する文献情報データベースとして世界で広く利用されている PubMed/MEDLINE には，1809 年以降に出版された約 2,600 万件分が納められている．物理学や化学などのほかの自然科学とは異なり，いわゆる基本法則を用いた知見の記述が困難であるため，自然言語による記述が非常に重要であり，また，そこからさまざまな視点でデータベースが構築されている．たとえば，生物を構成する分子であるたんぱく質に関するデータベースとして UniProt があり，個々のたんぱく質を生成するために必要な設計情報を担う遺伝子に関するデータベースとして Gene があるという具合だ．そのほかにも疾患に関するデータベースや生物種に関するデータベースなど，大小あわせると，論文として報告されているものだけで 1,500 を超えるデータベースがインターネットを介して誰でも自由に無償でアクセスできる形で公開されている．

なお，生命科学分野のデータベースには，実験で得られたデータをそのまま納めるものも含まれる．たとえば，各生物におけるすべての核酸配列情報を納めるゲノム配列データベースとして GenBank がこれにあたる．ゲノム情報について論文を発表するにあたり出版社が研究者にデータベースへの登録を求めることや，近年の実験機器の高度化に伴い，より多くの配列情報がより安価に得られる（図 -1）ことから，これらのデータが急増している．生物のゲノム配列を読み取るシーケンサから得られる配列情報を納めるデータベースとして Sequence Read Archive（SRA）があるが，その容量は 2009 年以降，指数関数的に増加している．

したがって，生命科学分野におけるデータベースには実験機器から取得されたままで生物学的な解釈が十分になされていないデータも大量に含まれており，最新の研究成果を受けて人手により意味付けを行う作業であるアノテーションが広く行われている．ただ，生成されるデータすべてに人手によるア

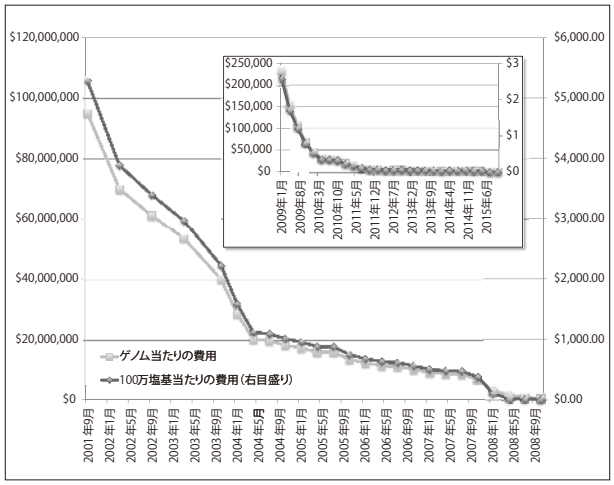

図 -1　ゲノム配列を読むために必要な費用の変遷
（https://www.genome.gov/sequencingcosts/ より）

ノテーションを付けられないほど急増しているので，計算機による自動アノテーションも行われているが，信頼性という点では人手によるものに劣る．

以上のような背景から，生命科学分野では，研究の進展に伴い，研究対象領域の細分化と専門化が進み，分野横断的な知識の整理が困難になりつつある．この問題に対処するための 1 つの方法として，オントロジーの編纂が研究コミュニティの中で盛んに行われている．

● オントロジーの大量出現

生命科学分野においては特に医学に関する研究において用語の整理を目的としたオントロジーに関する論文が 1995 年頃から発表され，その後分子生物学におけるオントロジーの必要性が 1997 年に提起された．そして，利用者数や対象生物種数などの観点から最も成功を収めた Gene Ontology（GO）が 2000 年に Michael Ashburner らにより発表される．これは，それぞれ異なる研究コミュニティでさまざまな生物種のゲノム配列が解析され，そこに含まれている遺伝情報を見つけたときに，共通の語彙を用いてそれらを記述できるようにしたものである．生物横断的に共通の特徴，たとえば，代謝に関する機能などを生物横断的に共通した語彙で記述できるため，各生物固有の特徴と，生物横断的な特徴を明確に区別して記述したり比較したりできるなどの利点が得られる．

GO の成功により，その後多くの生命科学分野のオントロジーが編纂され，その結果としてオントロジー間の相互運用性に深刻な問題を生じる事態を招いた．そこで生命科学をより広く横断的に俯瞰できるような資源や環境の構築を目指して Open Biomedical Ontologies（OBO）Foundry が 2007 年に Barry Smith らにより設立された．誰でも OBO にオントロジーを登録でき，またレビューを受けられる．現在 OBO には 180 のオントロジーが登録されており，すべて自由に取得でき，Creative Commons Attribution（CC BY）などのオープンライセンスで提供しているものも多い．

これらの活動を通して統制語を用いたデータベースのアノテーションが非常に盛んに行われており，高度に複雑な生物の仕組みを少しでも理解するために行われる研究には欠かせない資源となっている．そしてすべてのアノテーションには根拠が求められることから論文情報との紐付けも大切である．

LOD の公開とデータベースの RDF 化

● データベースへのアクセス方法の問題

さて，生命科学分野におけるデータベースの発達とアノテーションについて説明してきたが，続いてこれらのデータベースに対するアクセス方法の問題について説明する．これまでに多くのデータベースが MySQL や PostgreSQL などに代表される関係データベースシステムに格納され，データベースごとに Web サイトを通じたアクセス手段が提供されている．所望のデータを取得するために必要なアクセス方法はそれぞれの Web サイト設置者が用意した規則に従う必要があり，また多くの場合は人が Web ブラウザに対して検索語を入力して結果を HTML で記述された文書として取得することを想定している．

これらのデータベース間においては，互いに関連するデータを納めていることも多く，対応するデータの識別子を参照し合い，Web オブドキュメントとして密にリンクしている．このため，必要な情報が複数のデータベースに散在していても HTML のリンク情報を辿ることで効率良く取得できる．しかし，必要な情報が大量にあり，たとえば，ある特定の研究課題に関連する論文と対応する遺伝子およびタンパク質の一覧を効率良く取得したい，というような場合には手作業では限界があるため，機械的な作業が必要になる．しかし，上述の通り，データベースごとに取得方法が異なることが多く，それぞれに特化したアクセス用のプログラムを用意しなくてはならなくなり，利用者の頭を悩ませてきた．

● リンクト・オープン・データの構築

そこで，統一的なアクセス方法で必要なデータを取得できるようにするためにリンクト・オープン・データ（LOD）を利用したデータベースの統合化が進められている．前述の PubMed/MEDLINE の検索サービスで「Linked Open Data」を検索してみると，執筆時点（2016 年 4 月 20 日）で 37 件ヒットした．最初に使われたのは 2010 年であり [1)]，LOD Cloud サイトが開設された 2007 年から 3 年で論文が発表されていることになる．初出論文での研究目的は，関連データベースの LOD 化により，漢方薬で利用されている植物の成分のなかから抗鬱剤として利用可能なものを探すことだった．LOD 化する過程で，関連する概念の意味や関係がより明確になることが LOD を採用する理由としている．

翌年の 2011 年には，創薬に関係するさまざまなデータベースの RDF（Resource Description Framework）化と LOD 化が行われたプロジェクト，Linked Open Drug Data（LODD）が発表される．このプロジェクトでは関連する 11 のデータセットとそこからリンクする 13 のデータセットを主に利用して創薬への LOD 利用の可能性を検証している．また，論文が発表されてから 6 年が経過しているものの，依然として多くのデータセットに現在でもアクセスできることや，LOD と謳いながら実際にはリンクトデータの四原則に必ずしも従わないデータセットが多く存在する中，LODD プロジェクトによるデータセットは現在アクセスできるものについて確認できる分についてはすべて従っていることも特筆すべき事項である．

● データベースの RDF 化

生命科学分野において LOD を利用する取り組みが早くからなされていた背景としては前述のように複数のデータベースを統合的に利用するニーズがあるにもかかわらず，技術的に使いやすい共通語彙の利用基盤が欠如しているなどの理由で相互運用性が困難となる事態が生じていたことが大きい．その一方で，それが実現しやすい環境が整えられていたことも大きいと考えられる．まず，さまざまな公共データベースの第三者による RDF 化を行うプロジェクト Bio2RDF が 2008 年に発表され，多くの RDF 化されたデータベースが公開された． 続いて，タンパク質のデータベース UniProt の RDF によるデータ提供が 2009 年に開始されている． さらに，先述の通り，生命科学分野におけるさまざまな概念が GO を始めとして 2000 年代初頭からオントロジーとして構築されていたが，それらを OWL 形式で提供するポータルサイト，BioPortal が 2008 年から運用されている．なお，前出の LODD は関連データを RDF を用いて表現する RDF 化と，含まれる URI を参照解決可能にするなどの LOD 構築を同時に行う事例であるが，一方で，LOD 構築に伴うサーバ設定などの作業負担から，データベースを RDF 化するだけの活動もある．

以上の背景から生命科学分野における各種データベースの RDF 化やそれを利用した研究も活発になり，たとえば，生命科学分野における LOD の応用に繋がる基盤整備として遺伝子やタンパク質の配列情報を表現するための FALDO という語彙が提案され，UniProt や Ensembl など主要なデータベース運用主体が採用し始めている．そして 2014 年には欧州におけるバイオインフォマティクスの研究拠点である European Bioinformatics Institute（EBI）が，自身の運用する主要なデータベースを RDF で提供し始めたほか，GenBank や PubMed/MEDLINE を運用する National Center for Biotechnology Information（NCBI）においても MeSH（後述）や PubChem が RDF でも提供され始めた．さらに，生命科学分野におけるさまざまなデータベース間で同じ対象を表現する異なる識別子の関係を LOD で提供する Identifiers.org が運用を開始している．

国内ではライフサイエンス統合データベースセンター（DBCLS）とバイオサイエンスデータベースセンター（NBDC）が共同でさまざまなデータベースの RDF 化を推進するプロジェクトを，関連データベースを開発もしくは維持管理する研究機関とともに 2012 年から進めている．また DBCLS ではそれ

に先立ち，国内外の研究開発者が集まり議論しながら適宜開発を行うバイオハッカソンや SPARQLthon を通じて RDF 化に資する技術開発を 2010 年から進めている．具体的な開発事例は後述するが，これらの活動の成果の 1 つとして生命科学分野のデータベース開発者が自身のデータの RDF 化を行う際に参照していただくことを目的としたガイドラインを発表している．また，すでに RDF データを公開しているデータベース群に関する情報をまとめた RDF ポータルサイトを立ち上げている（図 -2）．

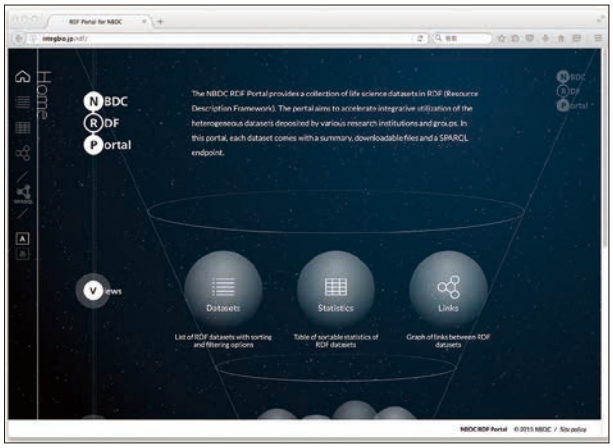

図 -2　RDF ポータルサイト

論文情報のリンクトデータ化に関連する言語資源

DBCLS を中心として構築し公開してきた RDF データベースおよび LOD として論文情報に関連するものがあり，筆者がその作業に従事していることからこれらについて関連資源とともに詳細を紹介したい．

● Medical Subject Headings（MeSH）

MeSH は PubMed/MEDLINE を始めとする米国 National Library of Medicine（NLM）で提供されているデータベースの索引付けや検索効率の向上を主な目的として NLM により開発，維持されている，概念階層構造を持つ統制語彙であり，1960 年から提供されている．毎年内容が更新され，最新版では 27,883 の概念を含む．多くの米国内の国立図書館が検索用に構築している書誌情報などのメタデータを RDF データで構成されたリンクトデータとして公式に提供していることを受け，2014 年に RDF 化した MeSH の初期バージョンを公開した．それまでにすでに Bio2RDF など第三者により RDF 化されたものが複数提供されていたことも，開発主体である NLM が正式に公開を始めた理由である．

MeSH RDF では，http://id.nlm.nih.gov/mesh/ を接頭辞とし，たとえば糖尿病を示す D003920 といった従来から採用している MeSH Unique ID をそれに続ける形で URI を生成している．リンクトデータとしてこれらの URI に HTTP でアクセスすると，MeSH 原典に書かれている定義や，糖尿病の関連語として胃不全麻痺など，MeSH 関連語などへのリンク情報が得られる．ただ，現時点では MeSH 以外の資源へのリンクは含まれていない．

● PubMed/MEDLINE

生命科学分野において最も広く利用されている文献検索サービスおよび文献情報データベースであり，誰でも無償で利用できる．執筆時点（2016 年 4 月 20 日）での検索対象文献数は 25,968,015 件であり，最古の文献は 1809 年に発表された頸動脈瘤に関する報告である．検索対象となる文献は米国内外併せて合計 5,600 を超える学術誌で発表されたものであり，書かれている言語は 40 程度になる．なお，実際に納められている書誌情報は英語で書かれている．近年の生命科学の発展に伴い追加される書誌情報が急増しており，2015 年の 1 年間だけで 100 万件超が追加されている．これは 1 日あたりにすると実に 2,882 件という計算になる．

歴史的には 1971 年に NLM の旗艦データベースとして MEDLINE が構築され，それに対するインターネットを介した無償検索サービスとして PubMed が 1997 年から提供され現在に至っている．インターネットを介して誰でも自由に無償で MEDLINE の検索が可能となる環境の出現は米国内の医療の質の向上に資する非常に大きな事象として当時の Gore 副大統領が大々的にこれを発表した．

各書誌情報には人手により文献中に書かれている内容を端的に表す MeSH タームが複数付けられており，これが PubMed/MEDLINE の存在価値を高めている．人手による作業のため，発表されてから MeSH タームが付けられるまで平均で 3 週間程度かかり，それまでは出版社が提供した書誌情報のみが納められている．

現在 PubMed/MEDLINE は XML 形式での提供はなされているが，公式な RDF データおよび URI 化された関連語彙の提供はなされていない．このため，DBCLS では MeSH など既存語彙と，既存語彙では表現できないデータについては独自の語彙を用いてこれを RDF 化している．

● ライフサイエンス辞書

ライフサイエンス辞書（LSD）はライフサイエンス辞書プロジェクトが編纂している生命科学分野で利用される専門用語の日英対訳辞書およびそれらの用法や実際の論文中での利用統計情報を含むデータベースであり，1993 年に初版が公開され，適宜更新されている．執筆時点では 2016 年 3 月版が最新で，英和 113,986 語，和英 128,073 語，用例 5,031 文，音声 16,144 語が納められている．また，MeSH 2016 に準拠したシソーラス，標準病名などの辞書 MEDIS v3.10 に準拠した病名および国際疾病分類 ICD-10 コードも含まれている．

本辞書はさまざまな形で利用可能で，たとえば，WebLSD として無償で自由に利用可能な検索用の Web サイトが公開されている．また，Windows や Mac OSX 向けのかな漢字変換用辞書という形でも無償で提供されている．そのほか関連アプリケーションや書籍の販売もなされている．

DBCLS では本言語資源がリンクトデータとして提供されることで言語横断的な生命科学の知識の整理に繋がると考え，ライフサイエンス辞書プロジェクトから許諾を受けて RDF 化を行い，公開している．現在公開されているリンクトデータは 2015 年 3 月版に基づくもので，CC 表示改変禁止 3.0 非移植ライセンスで提供している．URI の構造は，MeSH RDF と同様にすでに LSD 内で利用されている用語の ID を，接頭辞 http://purl.jp/bio/10/lsd/term/ に繋げる形で実現している．

言語資源を利用して構築されている LOD の事例紹介

前章で紹介した言語資源などを利用して DBCLS にて構築し提供している LOD を 2 点紹介する．

● 新着論文レビューナビゲーション

DBCLS では Nature や Science，Cell などのトップジャーナルと呼ばれる学術誌に掲載された論文のうち，日本人が著者に含まれていると判断される場合に，その論文の日本語によるレビューを当該著者に執筆していただき，誰でも自由に閲覧・利用できるよう Web 上にていち早く無料で公開する「新着論文レビュー」を提供している．対象読者を専門分野の異なる生命科学研究者としており，生命科学専門の編集者の視点から一文一文を吟味してさまざまな修正を行っている．また，すべての記事と図表が CC 表示 2.1 日本ライセンスで公開されているため，再利用性が高いことも特徴である．

本サービスは執筆時点で 900 を超える記事が閲覧可能で，編集者により付けられている記事あたり 4 つのキーワードに基づく絞り込みや検索ボックスからの全文検索が可能である．さらに，興味のある記事を探すための適切な検索語が思い浮かばないことや，現在提供されているフラットなキーワードリストに加えて，より体系化された概念構造を利用した絞り込みを可能にするために，MeSH シソーラスの階層を用いたナビゲーションサービスも提供している．本サービスを実現するにあたり，すべての記事や執筆者などのメタデータを RDF で表現し，記事中に出現する LSD の見出し語と LSD の各 URI とを関連付けている．LSD は上述の通り，MeSH シソーラスとの対応を含むほか，日英対訳も含まれていることから，NLM の提供している MeSH RDF の URI とのリンクも実現している．この結果として，

たとえば，「消化器系 > 膵臓 > ランゲルハンス島 > インスリン分泌細胞」という流れで概念を絞り込み，インスリン分泌細胞に関する記述のある記事を効率的に検索できる．

本サービスで利用している URI は参照解決可能としており，各記事と章，文それぞれに個別の URI を与えた上で，それらの関係を dcterms 語彙により hasPart や isPartOf を用いて記述している．これにより，LSD の見出し語に対応する URI をもとに，それが出現する記事のセクションや文を効率的に検索できる．これらのデータも CC 表示 2.1 日本ライセンスで公開しているほか，SPARQL エンドポイントも公開している．

● Colil

ある論文について，それを引用する論文（引用論文）リストだけでなく，各引用論文の本文中において実際に対象論文を引用している記述（引用文脈）も検索できるサービスとして Colil が DBCLS から提供されている．Colil では引用される論文（被引用論文）について，引用論文の重なり度の高いほかの論文，すなわち共引用度の高い論文のリストを提示する機能も持つ．検索対象の引用論文は，PMC Open Access（OA）サブセットに含まれるもので，被引用論文については PubMed ID がつけられている論文である．最新のデータセットでは引用論文，被引用論文がそれぞれ 1,312,546 件，8,144,216 件である．

Colil のデータセットはすべて RDF で記述されており，Web サイトで提供されている検索サービスも内部で SPARQL クエリが発行されている．発行されているクエリを確認することもでき，SPARQL エンドポイントで試すことができる．

Colil では bibo や dcterms などの既存関連語彙と独自に用意した語彙を用いてデータを表現しており，相互運用性を高めているほか，CC 表示 2.1 日本ライセンスの下で公開している．また，Colil において各論文は PubMed ID を基にした独自の URI で表現し，参照解決可能としている．そのほか，PMC OA サブセットについては，すでにほかの組織により RDF 化が行われており，これに対するリンクを含めることで，引用論文の内容に関する情報も取得できる．

今後の展望

前章まで生命科学分野におけるデータの RDF 化に関する状況や DBCLS における LOD 関連サービスの提供について述べてきた．

上述の通り，生命の機能を理解するために，これまで得られている膨大な知見を効果的に参照できる環境が必須であり，セマンティック Web 技術はそのための基盤技術の 1 つとして有望である．依然として不明な点が多い生命現象について，過去には顕微鏡の出現が，近年では塩基配列を大量かつ高速に読み取るシーケンサの出現が新たな知見を得るブレークスルーをもたらしており，今後も新たな実験技術が開発されるのに伴い今では考えられないような知見が得られるだろう．これに伴い，新たな概念が現れたり，それまでの概念体系が修正されたりする．これは既存オントロジーの構造が変わることであり，常にデータベース間の相互運用性と過去の知見との関連づけは問題になり続けると思われる．セマンティック Web 技術の有用性は，このような作業が常に行われることを想定して研究活動の効率化に資するような環境を提供し続けられるときに認められると思われる．

参考文献

1）Samwald, M., et al. : Integrating Findings of Traditional Medicine with Modern Pharmaceutical Research : The Potential Role of Linked Open Data, Chin Med (2010).

（2016 年 4 月 25 日受付）

山本泰智 ■ yy@dbcls.rois.ac.jp

生命科学分野における知識の整理を効率的に行えるよう，データの Reduce，Reuse，Recycle が進む環境構築を行う研究者．