BioLOD.org: 生命科学系 LOD データベースと既存アプリケーションによる可視化

BioLOD.org links open data of biological databases and supports visualization by existing applications

西方公郎 [1], 石井学 [1], 吉田有子 [1], 小林紀郎 [1], 高橋聡 [1], 望月芳樹 [1], 松嶋明宏 [1], 田中芳幸 [1], David GIFFORD[1], 土井考爾 [1], 原田えりみ [1], 蒔田由布子 [1], 豊田哲郎 [1]

Koro NISHIKATA[1], Manabu ISHII[1], Yuko YOSHIDA[1], Norio KOBAYASHI[1], Satoshi TAKAHASHI[1], Yoshiki MOCHIZUKI[1], Akihiro MATSUSHIMA[1], Yoshiyuki TANAKA[1], David GIFFORD[1], Koji DOI[1], Erimi HARADA[1], Yuko MAKITA[1], and Tetsuro TOYODA[1]

[1] 理化学研究所 生命情報基盤研究部門

[1]Bioinformatics And Systems Engineering division (BASE), RIKEN

Abstract: The vast amount of various life sciences data at RIKEN and other institutes including genome, transcriptome, proteome, metabolome, and phenome data are ontologically integrated into a common system. The challenge is to facilitate data retrieval, integration and collaboration. BioLOD.org - the Biological Linked Open Data database (http://biolod.org) - provides over 6,800 downloadable OWL/RDF graph files of mutually linked public biological data organized as a semantic web using standardized formats of the World Wide Web Consortium Linking Open Data (W3C LOD) project. BioLOD.org mines numerous semantic links from original databases and re-classifies them into graph files based on ontology classifications. Relationships between the files are mutually and clearly referenced so it is easy to find other files associated by semantic links included in detailed data instances. BioLOD.org intensively surveyed both forward and reverse semantic link relationships from 36 databases for humans and mice, 33 databases for plants and 16 databases related to protein experiments and structures. BioLOD summarizes this information as archive files available for download in various useful formats. The BioLOD.org database uniquely provides Linked Open Data annotated contextually with biological vocabulary and supports visualization services to browse LOD data through SciNetS.org, repository services to deposit users' LOD through LinkData.org and SPARQL endpoint service for BioLOD data is through BioSPARQL.org.

1 はじめに

近年、理化学研究所(理研)では、理研内外の世界の研究機関より提供されている、哺乳類・植物・タンパク質を含む 192 個の公共データベースを公開している。これらのデータベース群には 820 万件以上のデータレコードが登録されており、SciNetS.org (Scientists' Networking System) という共通の基盤で統合されている [1]。このシステムには 28 万件のトリプルが登録されており、ここから生成される 4.5TB の関連ファイルに世界中の研究者がアクセスできる。最近、我々は Semantic-JSON という Application Programming Interface (API)を開発し、公開した [2]。ユーザは SciNetS のセマンティック Web データにプログラム的にアクセスでき、SPARQL クエリを容易に作成できる。

この 7 月には BioLOD.org (http://biolod.org) を公開した。これは、理研内外のバイオ系の Linked Open Data のダウンロードハブとなるデータベースである。BioLOD データは、W3C LOD プロジェクト[1]に基づく多様な形式でダウンロード可能である。BioLOD は、2011 年 11 月時点で 756 クラス、824 万インスタンスの構造化されたデータセットを提供している。これらの生命科学関連のセマンティック Web データは、オントロジー分類に基づいて生物種横断的に統合されている。

[1] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

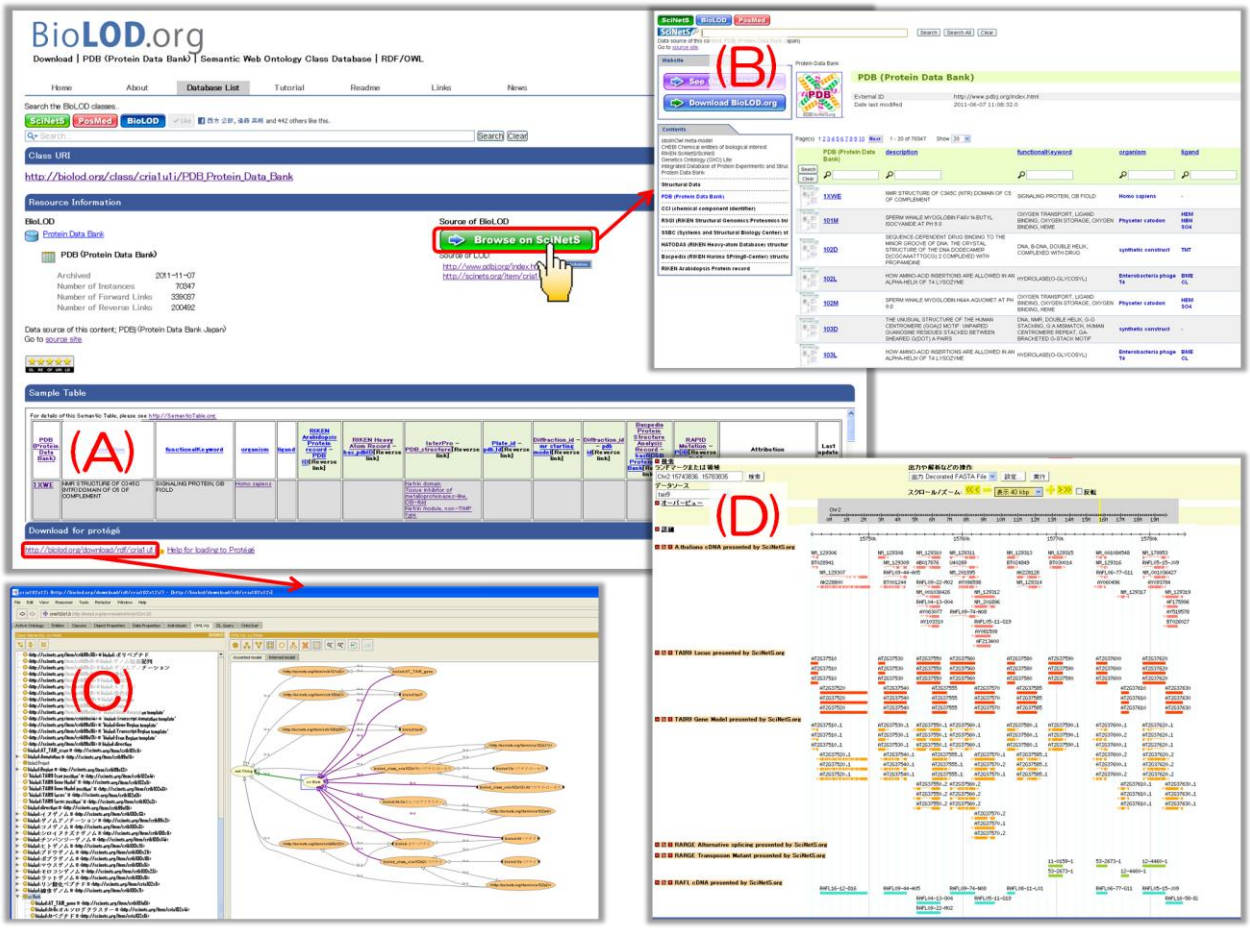

図 1: BioLOD データ (A) の SciNetS (B)、Protégé (C)、GBrowse (D) による可視化

2 BioLOD.org のサービスと利点

BioLOD では、バイオインフォマティクス研究者が、大規模なデータ交換と通常のデータ検索を行うために、以下に掲げるサービスと利点を提供する。

1) BioLOD では、RDF, RDFa, OWL, OBO, Turtle, N-Triples, TSV を含む様々な形式で LOD データをダウンロードできる。

2) BioLOD データにはそれ自身の attribution 情報が含まれているので、オリジナルデータのリソースを正しく参照できる。

3) BioLOD のデータはほぼ 1 ヶ月ごとに更新され、利用度を伝えるための統計情報が閲覧できる。そのため、最新かつ利用価値の高いデータにアクセスできる。

4) BioLOD のメタデータは Creative Commons ライセンスの基に提供されているので、法的に適切な形でのメタデータの配布を促進する。

5) BioLOD サイトでは、複数の LOD 間の関係は、セマンティックリンク関係とオントロジー階層構造によって表現されるので、複数の LOD をまるごと取得することが出来る。

6) BioLOD データの閲覧サービスは、SciNetS.org で提供される。

7) BioLOD レポジトリへのデータ投稿サイトは、LinkData.org で提供される。

8) SPARQL 機能は、BioSPARQL.org プロジェクトによって、サポートされる。

3 バイオ研究への応用

以下では、BioLOD データの生命科学研究への応用例を紹介する。

3.1 BioLOD のテーブルデータによるセマンティック Web データの人間可読性

BioLOD ではセマンティック Web データを、バイオ研究者がよく利用するテーブル形式で提供している。そのため、既存のバイオデータ解析プログラムを BioLOD データに適用する事が容易にできる。BioLOD の RDFa は、人間可読性の高い HTML テー

ブルの形式で、我々はこれを SemanticTable (http://semantictable.org) と呼んでいる。機械可読なメタ情報は RDFa の仕様 [3]に基づいて埋め込まれているので、RDFa Distiller のような変換ツールを用いて、RDF グラフを抽出する事ができる。BioLOD の TSV 形式では、データアイテムが BioLOD URI で識別できるので、RDF の概念に詳しくない人でも、カラムの交換やテーブルのマージの様なテキスト処理を行うことができる。したがって、テキストエディタやスクリプト言語を使うなど比較的容易な方法で、セマンティック Web データを高度に処理する事ができる。

3.2 既存アプリケーションによる BioLOD のセマンティック Web データの可視化

BioLOD のメタデータはセマンティック Web データベースシステム SciNetS (http://scinets.org) で閲覧できる。また、Protégé [4] や OBO edit [5] のような既存の可視化ツールで読み込むことができる (**図 1**)。特に、ゲノム情報に関連したセマンティック Web データは、Generic feature format Version 3 (GFF3) 形式で提供している。これは、ゲノムリソースを簡潔に記録したいというバイオインフォマティクス・コミュニティの要望から生まれた特殊な TSV 形式である。従って、Ensembl や Mouse Genome Informatics (MGI), The Arabidopsis Information Resource (TAIR) の様なゲノム・アノテーションを、Generic Genome Browser (GBrowse) [6]で可視化できる。

4 BioLOD.org のデータ検索

セマンティック Web のアーキテクチャは、Tim-Berners Lee によって提唱された、言語の階層構造によって説明される。標準的なセマンティック Web のアーキテクチャは、RDF, RDF Schema (RDFS), OWL, SPARQL によって構成される。我々の基盤を **図 2** に示す。データ交換および登録は RDF を利用して LinkData.org によってサポートされる。SciNetS.org は RDFS に基づいたクラスとプロパティの階層構造を創ることで、意味論的に統合された生命科学データベース群を構築する。BioLOD.org は、RDFS の拡張版である OWL を用いて、オントロジー階層に基づいてバイオデータを統合し、LOD をファイルとして提供する。OWL は記述論理に基づいているので、セマンティック Web に推論能力をもたらす。

BioLOD のファイルは、ファイルのコンテンツにマッチする任意のキーワードで検索できる。検索結果は、ファイルのダウンロード数や閲覧数に加え、他の LOD ファイルからの被リンク数、Fischer の正確検定の p-値によってランク付けされる [7]。BioLOD のユーザは、SciNetS や PosMed も同時に検索する事が出来る。PosMed とは、推論型のテキスト検索エンジンで、代謝化合物、変異体、疾患、研究者、ドキュメントセットやデータベースを検索できる [8]。現在 SciNetS では、プロパティとキーワードを指定しての検索が可能である。

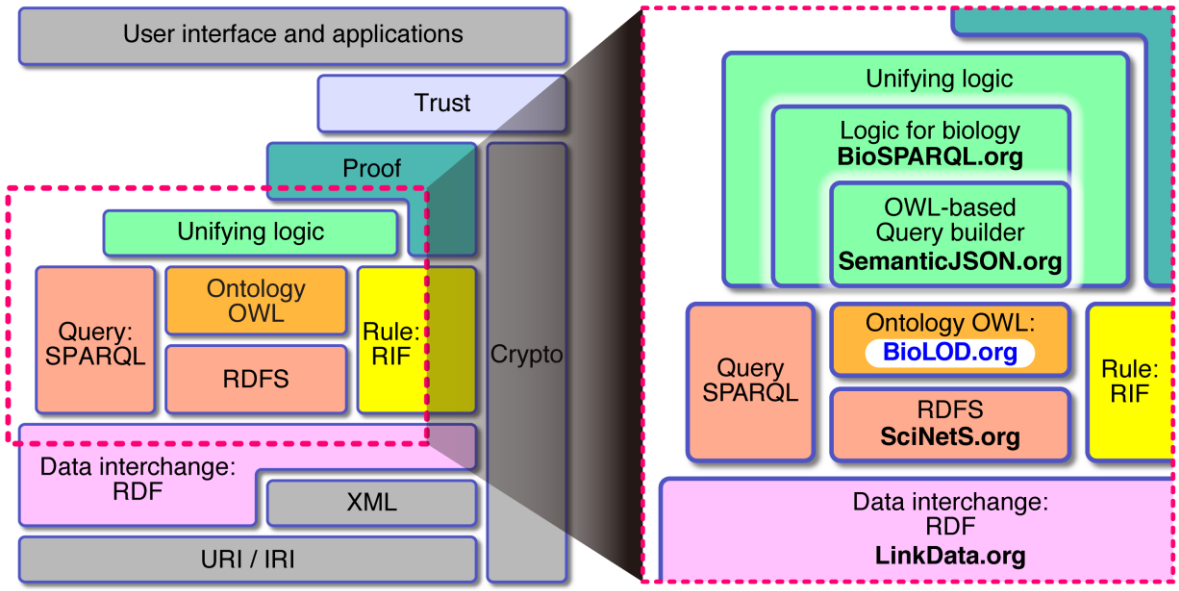

図 2: セマンティック Web のレイヤーケーキ(左)と我々の生命情報基盤の対応関係(右)

5　結論および今後の展望

BioLOD.org は、データ登録、ダウンロード、データの可視化、データ検索、そしてバイオ研究の論理の統一を容易にするため、オントロジーに基づいて統合されたデータベースハブとして機能する (**図 2**)。生命科学系のセマンティック Web データの先行事例として UniProt RDF[2]や BioPax [9]が挙げられる。しかし、バイオ系を含む多種多様な大規模データを、OWL を用いて共通の形式で Linked Open Data として公開したのは BioLOD.org が初の試みである。

セマンティック Web のレイヤーケーキの上位のレイヤーにおいては、Semantic-JSON.org がデータにアクセスする為の論理を提供して、OWL ベースのクエリを生成する。また、BioSPARQL.org がデータ検索の為の論理を提供して、RDF グラフを得るための SPARQL クエリを生成する(**図 2**)。DBpedia [10]や Bio2RDF [11]のような典型的な LOD データベースは SPARQL エンドポイントを提供しているが、通常のバイオ研究者が SPARQL クエリ文を作成して検索を実行するのは困難を伴う。我々は現在 BioSPARQL という SPAQRL エンドポイントインターフェースを開発中である。これにより、予め RDF グラフ構造を知らなくてもユーザは SPARQL クエリを容易に構築・実行し、目的の RDF グラフを取得できるようになるであろう。

謝辞

BioLOD.org プロジェクトは、科学技術振興機構(JST)のバイオサイエンスデータベースセンター(NBDC)によるサポートを受けて実施しています。

参考文献

[1] Masuya H, Makita Y, Kobayashi N, et al.: The RIKEN integrated database of mammals, Nucleic Acids Res., Vol. 39, Suppl.1, D861-870, (2011)

[2] Kobayashi N, Ishii M, Takahashi S, et al.: Semantic-JSON: a lightweight web service interface for Semantic Web contents integrating multiple life science databases, Nucleic Acids Res., Vol. 39, Suppl. 2, W533-540, (2011)

[3] Adida B, Birbeck M, McCarron S, and Pemberton S: RDFa in XHTML: Syntax and Processing, W3C Recommendation 14 October 2008.

[4] Rubin DL, Noy NF, and Musen MA: Protege: A Tool for Managing and Using Terminology in Radiology Applications, J Digit Imaging, Suppl. 1, pp. 34-46, (2007)

[5] Day-Richter J, Harris MA, Haendel M, et al.: OBO-Edit--an ontology editor for biologists, Bioinformatics, Vol. 23, No. 16, pp. 2198-2200, (2007)

[6] Donlin MJ: Using the Generic Genome Browser (GBrowse), Curr Protoc Bioinformatics, Chapter 9, Unit 9.9, (2009)

[7] Kobayashi N, and Toyoda T: Statistical search on the Semantic Web, Bioinformatics, Vol. 24, No. 7, pp. 1002-1010, (2008)

[8] Yoshida Y, Makita Y, Heida N, et al.: PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning, Nucleic Acids Res., Vol. 37, Suppl. 2, pp. W147-152, (2009)

[9] Demir E, Cary MP, Paley S, et al.: The BioPAX community standard for pathway data sharing, Nat Biotechnol. Vol. 28, No. 9, pp. 935-942, (2010)

[10] Auer S, Bizer C, Kobilarov G, et al.: DBpedia: a nucleus for a web of open data, In Proceedings of the 6th international semantic web and 2nd Asian conference on Asian semantic web conference, Vol. 4825, pp. 722-735, (2007)

[11] Belleau F, Nolin MA, Tourigny N, et al. : Bio2RDF: towards a mashup to build bioinformatics knowledge systems, Journal of Biomedical Informatics, Vol. 41, No. 5, pp. 706–716, (2008)

[2] http://dev.isb-sib.ch/projects/uniprot-rdf/