3C2 -OS -13b -1

大規模 Linked Data のオントロジーによる統合化と SPARQL の高度化

Ontology-based integration of huge linked data and upgrading SPARQL

豊田 哲郎[*1]
Tetsuro Toyoda

小林 紀郎[*1]
Norio Kobayashi

[*1] 理化学研究所 生命情報基盤研究部門

Bioinformatics And Systems Engineering (BASE) Division, RIKEN

In the life-science field, huge amounts of various raw data sets are generated through experimental technical innovations such as next-generation sequencers. We categorized such data sets including 826 million data records into 764 classes on the basis of OWL ontologies and published these as downloadable files for each class on our BioLOD.org web site. Since conventional SPARQL query techniques cannot be applied against such huge data sets, we also developed the BioSPARQL.org web site that allows those who do not know the structure of the linked data in the web to generate and execute a SPARQL query by supporting them with an intelligent ontology based query and data management system using a graphical user interface that downloads a "snapshot" locally for local processing of only the needed necessary subsets of data of the huge datasets.

1. はじめに

生命科学においては、測定技術の進歩に伴って大量でかつ多様なデータが産出される。生命の理解に迫るにはこのようなデータは広く研究現場で活用されることが不可欠であり、このためにはデータにオントロジー等の標準語彙を用いて意味論を与え、バイオ研究者に理解しやすく、かつ機械可読性の高いデータを実現し広く公開されることが望まれる。さらに、生命科学のデータは、遺伝子発現と表現型の組のように複数のデータセットを組み合わせて解析されることが多いため、複数のオントロジークラスで意味づけされたデータのつながりを効率よく取得することが求められる。既存のクエリ技術の問題点は、ユーザがあらかじめデータのつながりを調べ解析し、適切なクエリを構築しなければならないことである。このことが多くのバイオ研究者の利用を難しくしているため、プログラムによる知的なクエリ構築支援技術の開発が目下の課題となっている。

我々は以上の課題を解決し実用的な大規模生命科学データの高度な知的処理を実現するため、セマンティックウエブ技術に基づいたデータ公開基盤を構築してきた。具体的には、SciNetS (Scientists' Networking System; http://scinets.org) [Masuya2011] と呼ぶデータ編纂、閲覧のための基盤を構築し、世界の研究機関より提供されている、827 万件のデータレコードを含む 205 個の様々な生物種のオミックデータベースをセマンティックウエブの Linked Data 形式に変換、統合し閲覧可能にした。次に BioLOD（http://biolod.org）と呼ぶデータダウンロードサイトを構築し、10 万件に及ぶオントロジークラス毎に上記データを分類、編纂してアーカイブ化し、ダウンロード可能にした。さらには、BioLOD が提供するデータを検索するための、最適かつ容易な SPARQL クエリ構築とそのクエリ実行をユーザに支援する BioSPARQL(http://biosparql.org)と呼ぶサイトを構築した。

連絡先：豊田 哲郎，理化学研究所 生命情報基盤研究部門，〒230-0045 神奈川県横浜市鶴見区末広町 1-7-22，toyoda@base.riken.jp

本稿では、上記の各ツールについて論じ、セマンティックウエブ技術に基づいた高度な生命情報の知的処理について概観する。

2. オントロジーに基づいた生命情報の編纂

2.1 SciNetS による Raw データの収集と Linked Data の研究者による共同編纂

SciNetS は、文書、画像、動画などの Raw データを登録し、そのメタデータを研究者同士の共同編纂により記述することで、Linked Data 作成と共有をクラウドの中で行うための生命情報基盤である。SciNetS では、仮想ラボと呼ばれる、セキュリティで仮想的に区切られた共同研究スペースを提供する。仮想ラボでは、公開データだけでなく非公開のデータもアクセス権を考慮したメタデータの編纂が可能である。すでに SciNetS には 400 を超える仮想ラボが存在し、このうち 205 の仮想ラボが公開されている。

2.2 オントロジーに基づいて分類されるダウンロード可能なデータファイルの生成

SciNetS に登録されている公開 Linked Data を世界標準形式でダウンロードできるようにデータを加工し、誰でもダウンロードできるサイト BioLOD を構築した。SciNetS は仮想ラボと呼ぶ複数のクラスから構成されるデータベースを単位としたデータセットが構成されるのに対し、BioLOD では知能知識処理に好適なオントロジークラスを単位としたデータセットが構成される。具体的には、遺伝子オントロジーGO (Gene Ontology) や表現型・性質オントロジー PATO (Phenotypic Quality Ontology)など世界標準のオントロジー、理化学研究所が開発した生物学上の統一的な知識構造 YAMATO-GXO (Yet Another More Advanced Top-level Ontology-Genetics Ontology)や SciNetS 上で研究者が新たに構築したオントロジーを用い、Web Ontology Language (OWL) クラスにデータを分類し、OWL クラス毎にデータアーカイブファイルが作成される。データアーカイブファイルは、セマンティックウエブ標準形式である RDF/OWL、RDFa、OBO、Turtle、N-Triples で作成されており、ユーザはこれら形式のファイルを

ダウンロードすることができる。2012 年 4 月現在、BioLOD.org からは 765 クラス、827 万データレコードを含むデータがダウンロード可能である。

3. SPARQL クエリ構築支援機能の実現

SPARQL [Prud'hommeaux2008]は RDF 形式で保存されるデータセットに対して検索問い合わせを行う世界標準言語である。SPARQL はグラフ構造をなす RDF データに対してグラフパターンをクエリとして記述し、グラフパターンマッチングを行うことでデータの検索を行う。

生命情報の解析にあたっては、オミックデータ、すなわち遺伝子(ゲノム)、遺伝子発現(トランスクリプトーム)、タンパク質(プロテオーム)、代謝物質(メタボローム)、表現型(フェノタイプ) 等の様々な網羅的分子情報を複合的、多面的に解析する必要があり、必然的にクエリ対象となるオントロジークラスが多くなる。SPARQL クエリを記述するためには、ユーザはクエリ対象となる RDF データセットを、データリポジトリを巡回して特定しかつそのデータ構造を調べなければならないが、大規模で多くのオントロジークラスを含む RDF データセットに対してこのような作業を行うことは一般には難しい。この課題に対し、BioSPARQL は、我々が実装したサービスで、機械がオントロジーで体系化されたデータを知的に解析することで、ユーザはデータ構造を知らなくとも対話形式でクエリ生成を容易にすることを目指す。

3.1 生命科学研究を推進する BioSPARQL の機能

BioSPARQL は、BioLOD.org からダウンロード可能な RDF データセットに対して検索を行う SPARQL クエリを生成する。しかし BioSPARQL は BioLOD のウエブサービスとして SPARQL エンドポイントを提供するものではなく、ユーザのローカルサーバ又は PC に BioLOD から必要な RDF データファイルをダウンロードし、その PC 上で SPARQL を実行するよう設計されている。

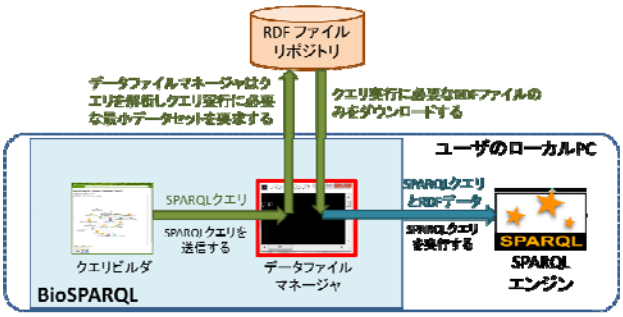

図1. BioSPARQL のアーキテクチャ

図 1 は BioSPARQL のアーキテクチャを示したもので、以下に掲げる生命科学研究において望まれる機能を実現する。

1. BioLOD のオントロジーを用いて大規模データを小さなデータセットの集合に分類したデータを利用することで、公開されているデータセットの中から最小限必要なものを選び出し、効率よくクエリを実行することが可能である。
2. ローカル環境において、適切なバージョン管理の元で自動的にクエリ実行に必要なデータをダウンロードして「スナップショット」を作成し、ユーザの非公開データと共に公開データとを組み合わせた SPARQL クエリを実行することが可能である。バージョン管理を行うことで、データの更新が行われてもユーザは一定クエリ結果を得ることができ、クエリ結果の科学的証拠に用いることができる。
3. SPARQL クエリの実行には大規模な計算リソースを要することが知られているが、大規模なクエリ検索に対してもタイムアウトが起こりにくくなるようローカル PC の計算リソースを管理できる。

3.2 BioSPARQL の実装

BioSPARQL の核となるソフトウエアコンポーネントはクエリビルダとデータファイルマネージャである。

クエリビルダは、ユーザが興味のあるオントロジークラスを始点としセマンティックリンクで関連付けられるオントロジークラスを選び出すために、オントロジークラスネットワークを描画し操作するための直感的なグラフィカルユーザインターフェイス(GUI)を提供するウエブサービスである。ここで、セマンティックリンクとは、オントロジークラス間に直接張られているリンクの他に、2 つのオントロジークラスそれぞれに含まれるインスタンス間のリンクも含まれる。さらに上記始点オントロジークラスと選ばれた別のオントロジークラスとを結ぶリンクデータを検索するための SPARQL クエリを、当該 RDF/OWL データセットのデータスキーマを論理的に解析することにより自動的に生成する。

データファイルマネージャは、ローカル PC 上で動作するアプリケーションソフトウェアである。データファイルマネージャはクエリビルダで生成された SPARQL クエリを解析し、適切なバージョン管理のもとでクエリ実行に必要な RDF データセットを BioLOD から自動的にダウンロードして RDF データファイルのローカルコピーを作成する。さらにユーザが管理する SPARQL エンドポイントにアクセスして該クエリを実行する。この時、ユーザはクエリを編集することで、ローカル PC 上の非公開データをも組み合わせたクエリを作成し、実行することができる。

3.3 解を持たない BioSPARQL クエリの除去

BioSPARQL はオントロジークラスやデータクラスなどのクラス間の関係をデータスキーマとしてユーザに提示し、該スキーマを満足するデータのつながりを検索する SPARQL クエリを生成する。最初のステップとなる SPARQL クエリの生成は任意の 2 クラス間の関係を RDFS のデータスキーマや、両クラスそれぞれのインスタンス同士をつなぐセマンティックリンクが存在する等の方法を用いて調べ、このような 2 クラス間の関係を直列につなげていくことで実現される。ただし、クラス A、B、C が存在し、A と B、B と C の間にそれぞれインスタンス同士をつなぐ関係があったとしても、A から B を経由し C に至る連続したインスタンスのつながりは必ずしも存在せず、このようなクエリは解を持たない。そこで BioSPARQL では、クラス間関係をつなぐスキーマを構築する際にこのようなインスタンスの連続したつながりを調べ、解を持たないクエリを生成しないようにしている。我々が構築したデータにおいては、連続する 3 つのセマンティックリンクを介して関係付けられる 2 クラス間の関係は 188,553 件存在するが、このうち解を持つものは高々43,962 件(全体の 23.3%)であった。このことから、残り 76.7%の解をもたないクエリをあらかじめ除外できる本手法の有効性が示された。

4. 結論

大規模で多様な生命科学データに対して高度な知識処理を実現するために、これらデータにオントロジーに基づいた意味論をメタデータとして与え、765 クラス 827 万件に及ぶセマンティックウエブデータを構築し、BioLOD サイトからダウンロード可能にした。さらに構築したデータに対し、SPARQL 検索のためのクエリをデータのスキーマを知らずとも適切に生成するサービス

BioSPARQL を構築した。その際、複雑なバイオデータのつながり検索において、約 8 割の解をもたない無駄なクエリをあらかじめ除外することで、検索の効率化を実現した。

参考文献

[Masuya2011] Masuya H., Makita Y., Kobayashi N., et al.: TheRIKEN integrated database of mammals, Nucleic Acides Res., Vol. 39, Suppl. 1, D861-870, 2011.

[Prud'hommeaux2008] Prud'hommeaux E., Seaborne A.: SPARQL Query Language for RDF, W3C Recommendation 15 January 2008, 2008.