1N4-OS-19a-3

効率の良い横断検索を実現するためのメタデータ収集・集積システム

A meta-data gathering system to facilitate an efficient cross search

山本泰智
Yasunori Yamamoto

山口敦子
Atsuko Yamaguchi

情報・システム研究機構 ライフサイエンス統合データベースセンター
Database Center for Life Science, Research Organization of Information and Systems

We describe a way of gathering metadata of datasets scattered around the globe. Our aim is to realize an environment where any life science data are inter-linked and accessible on the Web of Data. We gather several metadata by issuing SPARQL queries and collecting HTTP response header information. The obtained data are publicly available. In addition, to ensure credibility of the data and mutual understanding with data providers, we release our methods to collect the data.

1. はじめに

生命科学分野では様々な生物種のたんぱく質に関する包括的な情報を含む UniProt[UniProt Consortium 2014]をはじめとして, The European Bioinformatics Institute (EBI)や The National Center for Biotechnology Information (NCBI), ライフサイエンス統合データベースセンター(DBCLS)など, 近年様々な機関がたんぱく質や遺伝子, 化合物などに関するデータベースを Resource Description Framework (RDF) を用いて提供している. これらはまた, Linked Data としてのアクセスを可能にしており, ウェブオブデータの世界が実現しつつある. たとえば, UniProt や Ensembl[Yates 2016], MeSH, PubChem[Kim 2015], Life Science Dictionary (LSD)[金子 2005]などが参照解決可能(dereferenceable)な URI を含む RDF データベースを構築している. このため, 特定の URI に HTTP を用いてアクセスすれば, 当該 URI に関する情報を, HTML あるいは Turtle 形式などの特定のシリアライズされた RDF データで取得できる.

さらに, 得られたデータセット中のリンクを頼りにすることで, 探索的に関連する他のデータベース中のデータも取得できる. これは現在広く普及している WWW においてリンクを辿ることと等しいが, WWW における問題と同様に, 対象 URI にリンクしている他のデータベースを網羅的に発見するには手間がかかる. 特に語彙を提供する DBpedia や MeSH, LSD などのデータベースでは, 他のデータベースから参照される, すなわちリンクされる程度が高くなり, ある語彙を軸とした関連データを収集するなどの目的を達成しようとすると多くの時間と記憶装置が必要となる. Linked Data としての RDF データを記述するためのメタデータの語彙として Vocabulary of Interlinked Datasets (VoID)[1]が提案されているが, 残念ながら多くのデータベースについて, 上記の目的を達成するために必要なリンク情報が含まれていないことが事前の予備調査で示唆されている.

そこで, 我々は, ある URI で示される概念について, それを収めるすべてのデータベースにおける関連トリプルを網羅的に効率よく収集できる環境を構築する計画を立てている. 本計画を進めるに当たり, 現状を適切に把握し、データ提供者やデータ利用者の間で共有することを目的として, 様々な SPARQL エンドポイントに関する統計情報などを収集している. 具体的には死活情報, データの更新情報, Service Description (SD)[2]や VoID の有無, Linked Data の 4 原則[Berners-Lee 2006]における各原則への遵守状況などが含まれる. 以下, 具体的な収集事項とこれまでに得られている結果を報告する.

連絡先:山本泰智, ライフサイエンス統合データベースセンター, 〒277-0871 千葉県柏市若柴 178-4-4, 04-7135-5508, yy@dbcls.rois.ac.jp

2. 収集方法

2.1 収集対象エンドポイント

現在収集対象としているエンドポイントは生命科学に関するデータを提供している, 筆者の知る範囲で収集, 選択したもの 93 件である.

2.2 収集事項

種々の SPARQL クエリやクエリ発行時に得られる HTTP ヘッダ情報から抽出した情報をデータベース化しており, 以下に具体的な収集事項を示す.

(1) 死活情報

各エンドポイントの信頼性を図るための重要な指標として, 当該エンドポイントの死活情報を取得することが必要と考える. なお, 参照解決可能な URI に対する HTTP リクエストによる情報取得においては SPARQL エンドポイントへのアクセスは不要である. しかし, 実際的な運用形態として内部的に SPARQL クエリを発行している場合が多いと考えられることから, このような場合においてもエンドポイントの死活情報を得ることは重要であると判断した.

(2) メタデータの提供状況

SPARQL1.1 の仕様に含まれる SD には, エンドポイントの URL に対する HTTP GET クエリへの応答として, 当該エンドポイントに関するメタデータを提供する方法が定められている. また, 複数のデータセット間のリンク情報を提供するメタデータとして VoID が規定されており, 具体的な提供方法も記述されている. そのため, それらの方法を用いてメタデータが取得できるか否かを確認している.

(3) データセットの更新情報

本事項については, MeSH や LSD のような辞書データの場合, 必ずしも更新が頻繁に行われないことが当該データの信頼性を下げるものではないが, 一つの指標として利用者が対象データの特性を判断した上で利用する際に有益であると判断した.

[1] https://www.w3.org/TR/void/

[2] https://www.w3.org/TR/sparql11-service-description/

具体的には，SD や VoID を用いて更新情報が書かれている場合はそれを取得し，これらから得られない場合は SPARQL クエリを発行し，トリプル数の変化を追うなどの手法を用いて推定している．

(4) Linked Data の 4 原則への遵守情報

RDF データとしてアクセス可能なデータベースについてはその提供方法として，SPARQL API によるものとバルクダウンロードによるもののほかに，Linked Data の原則に従う HTTP GET によるものがある．この方法では SPARQL クエリインターフェースを公開する方法に比べると，取得データに対して利用者が設定可能な条件が制限される一方で，サーバーへの負担が少なく URI さえわかればそれに関連するデータが効率的に得られるという利点がある．このため，我々は Linked Data の原則に対する遵守状況も調査している．利用者はこの情報を手がかりに，対象データセットへのアクセス方法を選択できる．個々の原則に対する調査方法は以下の通りである．

1. モノコトには URI を用いて名前を付ける

本原則については，RDF の仕様として主語は URI かブランクノードのみを要求しているため，現状，主語に一つでも URI があれば，すなわち，主語が全てブランクノードのみの場合を除き，遵守していると判断している．

2. URI のスキームには http を使い，当該モノコトに問い合わせできるようにする

本原則については主語に当たる URI が http から始まるか否かを調査している．

3. URI への問合せに対しては，RDF や SPARQL などの標準を用いて有益な情報を提供する

本原則については，いくつかの主語の URI に HTTP GET を要求して某かのデータが得られるか否かを調査している．

4. より多くの関連情報を得られるよう他の URI へのリンクを含める

本原則については，SPARQL クエリを発行して述語に owl:sameAs[1]もしくは rdfs:seeAlso[2]が含まれているか否かを調査している．

本件に関し，RDF データの公開方法に関するベストプラクティスとして知られている方法[3]では，特定の URI への HTTP GET 要求に対して，サーバーは，クライアント側の指定したデータ形式に応じたデータの提供方法が推奨されている．具体的には HTTP の仕様で規定されているヘッダフィールドの Accept フィールドを用いたコンテントネゴシエーションに対応し，HTML および Turtle 形式を要求した際に得られる HTTP ヘッダ情報に含まれるコンテントタイプを調査している．このため，必ずしも Linked Data に記述されている要件を満たしているとは限らないこともある．

(5) 応答速度

応答速度は対象エンドポイントの処理能力を測る一つの指標となることから，これを計測している．

(6) CORS 対応状況

SPARQL エンドポイントに対する問い合わせを行う Web アプリケーションは JavaScript を利用して実現されている事例が多いと考えられるが，その際に問題となるのが異なるドメインへの HTTP アクセスである．クロスオリジン HTTP リクエストと呼ばれ，セキュリティ上の理由から多くのウェブブラウザはこれを認めていないが，サーバー側が明示的にこれを許可できる仕組みが整えられており，Cross-Origin Resource Sharing (CORS)[4]と呼ばれている．利用者が予め対象のエンドポイントが CORS に対応しているか否かを簡単に知ることができるよう，エンドポイントから返される HTTP ヘッダ情報に基づき本情報を取得している．

3. 結果と考察

現在得られている情報を以下に記述する．データは 2016 年 3 月 28 日から 29 日にかけて取得した．アクセスを試みたエンドポイントは 93 件で，そのうち，データが得られたエンドポイントは 45 件であった．残りの 48 件については Not Found などの理由で適切に HTTP でのアクセスができなかった．さらに，SD あるいは VoID データを提供しているエンドポイントは 37 件あり，そして Linked Data の 4 原則に対応しているものはわずかに 2 件であった．調査結果は http://d.umaka.dbcls.jp/ から公開しており，適宜情報の更新と新たな調査対象の追加を行う予定である．

問題点としては，Linked Data としては公開していても，PubChem のように SPARQL エンドポイントが公開されていない場合もあり，このような場合については関連情報を取得できていない．今後は，すでに取得されているデータから適宜 URI を抽出し，エンドポイントが公開されていないデータについても死活情報などのデータを収集する予定である．

現在，メタデータの取得できないエンドポイントに対して更新情報の取得のために SPARQL クエリを用いているが，一部の実装では本来得られるべきデータが適切に得られない問題点がある．各実装について，すべての問題点を把握することは困難であるが，このようなバグに関する情報は一般的に知られていないことから，利用者に分かりやすく問題点を提供するなどの対策をとる計画である．

Linked Data の原則への対応状況については，特に原則 3 や 4 については技術的な条件が明確ではないため，我々独自の解釈が含まれている．本件に関しては[Berners-Lee 2006]に書かれている通り，データの相互リンクの機会が失われると，ウェブが提供している付加価値である想定外の情報の再利用が制限される，という考えに基づき，より高い有益性を持つウェブオブデータの世界が広がることを期待し，一つの指標として有用であると判断している．ただ，実際に得られているデータを調査すると，原則 3 を満たすために実施しているアクセスに失敗している例が多数あることから，データを取得する際の技術的な手法について検討が必要と考えている．

また，様々な指標や結果については，各エンドポイントを立ち上げ，運営している主体，ひいてはすべての利用者への信頼性を確保するため，その根拠をわかりやすく提供し，相互理解が円滑に進むことを目指している．

4. 結論

我々は生命科学分野の Linked Data に網羅的かつ効率的にアクセスできる環境構築のため，公開されている国内外の SPARQL エンドポイントを調査した．現状ではデータセットへの効率的なアクセスを行うために有益なメタデータを提供している事例が少数であることが判明した．我々は取得したデータだけでなく，取得方法についても詳細に公開することで透明性を確保し，データ提供者との相互理解を目指している．この結果としてより高い有益性をもつウェブオブデータの世界が広がることを期待するものである．

[1] http://www.w3.org/2002/07/owl#sameAs
[2] http://www.w3.org/2000/01/rdf-schema#seeAlso
[3] https://www.w3.org/TR/swbp-vocab-pub/
[4] https://www.w3.org/TR/cors/

謝辞

本研究は独立行政法人科学技術振興機構(JST), バイオサイエンスデータベースセンター (NBDC) の助成による.

参考文献

[UniProt Consortium 2014] UniProt Consortium : UniProt: a hub for protein information, Nucleic Acids Res. 43(Database issue):D204-12, Oxford University Press, 2015.

[Yates 2016] Yates A et al. : Ensembl 2016, Nucleic Acids Res. 44(D1):D710-6, Oxford University Press, 2016.

[Kim 2015] Kim S et al. : PubChem Substance and Compound databases, Nucleic Acids Res. 44(D1):D1202-13, Oxford University Press, 2016.

[金子 2005] 金子周司, 鵜川義弘, 大武博, 河本健, 竹内浩昭, 竹腰正隆, 藤田信之: ライフサイエンス辞書から生命科学オントロジーへ, 情報知識学会誌, Vol. 15, No. 4, pp. 1–10, 2005.

[Berners-Lee 2006] Berners-Lee T : Linked Data, https://www.w3.org/DesignIssues/LinkedData.html, 2006.