1N3-OS-39b-3in1

より良い生命科学データ利用環境の構築を目指して

Towards constructing a better environment to use life science data

山本泰智
Yasunori Yamamoto

山口敦子
Atsuko Yamaguchi

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

We have developed a set of services to increase usability of RDF datasets. These services include searching for endpoints that provides data you want, visualizing the structure of a given dataset, monitoring each endpoint from various aspects, and providing forums where data providers and data users can share knowledge with each other.

1. はじめに

生命科学分野ではこれまで多くのデータベースが様々な観点から構築され，研究に利用されてきた．研究の進展に伴い，ゲノムデータを収める GenBank [Benson 2013]のように，これまでのデータベースが充実される一方で，新たな研究領域の誕生とともに新規のデータベースが構築されている．これらは現在インターネットを経由してアクセス可能であるものだけでも 1685 に上る[Rigden 2016]．このような状況において，研究者は複数のデータベースを検索して自身の研究に利用する場合も多いが，データベース毎に所望のデータを取得するための方法や，得られたデータの形式がまちまちであることが多いことが問題である．

そこで近年では The European Bioinformatics Institute（EBI）や National Bioscience Database Center（NBDC）などのデータベース提供機関が Resource Description Framework（RDF）を採用し始めており，その結果として SPARQL エンドポイントが立ち上がり，エンドポイントの所在が分かれば標準化されたアクセス方法でデータが取得できるようになってきた．しかしながら，依然として残る問題として次の三点がある．第一に所望のデータが提供されているエンドポイントの所在が分からない．第二にエンドポイントの所在が得られたとしても，当該エンドポイントを通じて得られるデータの構造が簡単には分からない．そして第三に当該エンドポイントから得られるデータが定期的に更新されている，あるいはダウンタイムが十分短いなどの維持管理や運用体制が分からない，ということである．

これらの問題に対処するために我々はエンドポイントの検索サービス Umaka Search，対象データセットの構造可視化サービス Umaka Viewer，そしてエンドポイントの評価サービス Umaka-Yummy Data（http://yummydata.org/）を構築している．まず，Umaka Search であるが，これは，URI あるいはリテラルの全てあるいはその一部を入力として受け取り，対応する RDF データが提供されているエンドポイントの所在を出力するものである．続いて Umaka Viewer は，対象データセットのオントロジーと，当該データセットのメタデータ，すなわち，利用されている述語と当該述語により結ばれているクラス間関係などを入力として受け取り，それをウェブブラウザー上に可視化するものである．なお，ここで必要となるメタデータの生成ツールも開発している．そして，Umaka-Yummy Data は，定期的に死活確認やトリプル数の増減などの情報をエンドポイント毎に取得して独自の指標を用いて結果を公表するものである．さらに，データ提供者およびデータ利用者双方の相互理解がより良いデータ利用環境の醸成には欠かせないと考え，エンドポイント毎にフォーラムと呼ばれる，対象エンドポイントに関する情報共有が行える場を提供することとした．

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 東京大学柏の葉キャンパス駅前サテライト 6 階，04-7135-5508，yy@dbcls.rois.ac.jp

2. エンドポイント検索サービス

本サービス（Umaka Search）は各エンドポイントが収める RDF データセットのうち，主語および目的語の URI あるいはリテラルを検索対象とし，対応するエンドポイントの所在を出力するサービスである．本サービスでは検索語にマッチする RDF トリプルは得られないが，エンドポイントの所在が分かれば，適宜 SPARQL クエリを発行することでそれが得られるため，この仕様とした．なお，検索対象を RDF トリプルとしなくても，索引付けを行うデータサイズが膨大であり，たとえば，タンパク質に関するデータベースである UniProt [The UniProt Consortium 2017]の RDF データセットについては，索引付けを行う対象となるデータだけでも 17.8 億レコードに及ぶ．

このため，検索には接尾辞配列を用いた高速全文検索システムの Sedue を用いている．UniProt のデータを例にとると，主メモリで 512G バイト超，インデックス構築用の SSD を 2.4T バイト必要としている．なお，本サービスは現在開発途上であり，検索可能なエンドポイントの数は限られている．

3. データ構造可視化サービス

本サービス（Umaka Viewer）は，上述の通り，所望のデータが取得できると思われる SPARQL エンドポイントの所在が分かり，SPARQL クエリを発行できる状態であっても，検索対象のデータセットの構造が不明のままではクエリを構築することが困難であるという問題に対応するために開発している．

SPARQL エンドポイントの設置主体によりデータ構造が静的な図を用いて解説されている例があるが，図の表現方法や説明の詳しさなどがまちまちであるほか，RDF データは概念の階層化がなされたオントロジーを用いていることが多く，静的な図だけでは分かりにくいという問題点がある．そこで，Umaka Viewer では対話的な図を提供することとし，初期状態では最も抽象的なデータ構造のみが表示され，適宜マウスを用いて焦点を変えていくことでより詳しいデータ構造を知ることができるようなインターフェースを開発した（図 1）．

Umaka Viewer を用いたデータ構造の可視化サービスを提供するには予め対象 RDF データセットのオントロジーと，当該デ

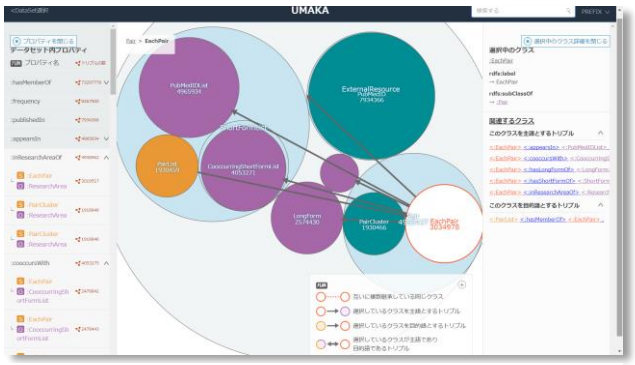

図 1 可視化サービス Umaka Viewer の一画面

ータセットのデータ構造を表現するデータをそれぞれ RDF で用意する必要がある. オントロジーは必須ではないが, 利用者がデータ構造を把握しやすいオントロジーを提供することが望ましい. また, データ構造を表現するデータは, SPARQL Builder Metadata [Yamaguchi2016]により規定されている語彙(SBM オントロジー)を用いて生成されていることを前提とする. 当該語彙は対象データセットのクラス間関係や利用されている述語の種類と数など, 静的な面と量的な面の双方から様々な統計情報を表現するもので, Service Description(SD)[1]および VoID[2]を拡張して作られている. なお, SPARQL エンドポイントを指定するか, あるいは手元にある RDF データセットを与えることで, SBM オントロジーに従うデータを出力するツールも別途配布している[3].

Umaka Viewer は大きく分けて二種類の利用者を想定している. 第一に, 自身の構築したデータセットの構造を可視化しようとする利用者であり, 第二に, これから利用しようとするデータセットの構造を知りたい利用者である. 前者は, オントロジーの構築や, SBM オントロジーに従うデータの準備を行い, それらを Umaka Viewer の入力として与える. それに対応した, 対話的な操作が可能な図にアクセスするための URL が結果として得られるので, それを適宜公開し, 後者が利用することを想定している. 可視化に必要なデータを生成するツールは Python で書かれており, PyPI を利用して取得可能である[4].

4. エンドポイント評価サービス

本サービス(Umaka-Yummy Data)は, 各エンドポイントに対して, 死活確認やトリプル数など, 様々な角度から定期的に情報を収集し, 独自の得点方法を用いて数値化した結果(Umaka Score)をホームページにて公開するものである(図 2). さらに, より利用価値の高い RDF データが流通する環境を醸成するためには, データ提供者と利用者の双方が情報共有しやすい場を構築することが大切であるとの見地に立ち, 各エンドポイントの数値情報を提供するページから当該エンドポイントについて議論することができるフォーラムへのリンクを用意している.

現在, 毎日一回各エンドポイントから下記の情報を収集している.

1. 過去 30 日間の稼働日数
2. 最終更新日
3. メタデータ(SD および VoID)の提供の有無
4. 広く使われているオントロジーの利用の有無

[1] https://www.w3.org/TR/sparql11-service-description/
[2] https://www.w3.org/TR/void/
[3] https://github.com/sparqlbuilder/metadata
[4] https://pypi.python.org/pypi/umakaviewer/0.0.1

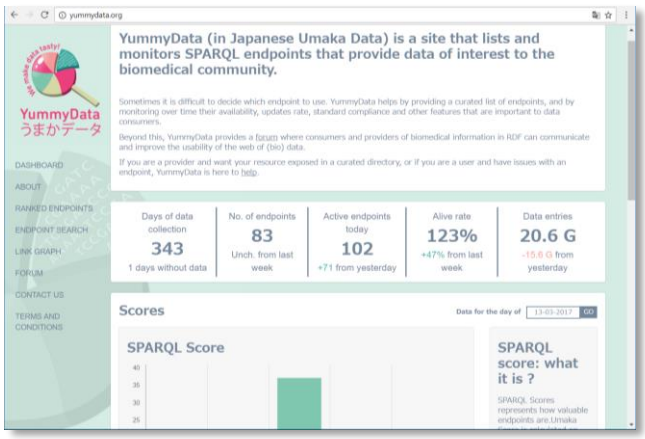

図 2 エンドポイント評価サービス Umaka-Yummy Data の一画面

5. クラス名や型指定の有無
6. 他のデータセットへのリンクの有無
7. コンテントネゴシエーション利用の有無
8. 参照解決な URI であるか否か
9. 適切な長さの URI であるか否か
10. クエリ処理時間

これらの情報は逐一蓄積され, 利用者が指定した日の取得結果を容易に確認できるほか, 独自の計算式により数値化されていることから, その経時変化を把握しやすいように, 過去 30 日分の情報については折れ線グラフで値の変化を追うことができる.

各エンドポイントの Umaka Score の算出方法は, 次の通りである. まず, 各項目を大きく 6 種類に分け, それぞれに 100 点ずつ振り分け, その合計値を最終値とする. この 6 種類と上記の項目との対応を示す.

- 稼働率 1
- 新鮮度 2
- 運用度 3
- 活用性 4,5,6,7
- 適正度 8,9
- 処理速度 10

なお, 現時点では新鮮度について適切に情報が得られないため, 一律に全てのエンドポイントに対して 50 としている.

また, フォーラムについては現在, github の Issues を利用しており, 各エンドポイントに対してそれぞれ Issue を立てている. このため, github のアカウントを作ればだれでも議論に参加できる.

5. 考察

Umaka Search については, 現在, 索引付けを行うデータの縮小方法について検討している. 例えば数値のみのリテラルや, 単なるハッシュ値を URI 化しているものなど, 索引付けが不要と思われるリテラルを検討し, 適宜適用していく予定である.

Umaka Viewer については, 現在 3 種類の RDF データセットを対象として実証実験を進めており, 今後は, より多くのデータセットを対象にする. また, 現在は日本語のインターフェースのみを提供しているが, 英語も追加する予定である.

Umaka-Yummy Data については情報収集対象となるエンドポイントを適宜追加したり取り除いたりしていく. また, 採点方法について現在は含まれていない CORS 対応の有無など指標の洗い出しを行うとともに, 得られた結果を見ながら計算方法の妥当性を検証する.

6. 結論

世界各地で構築された多様なデータベースが存在する生命科学分野において，それらを目的に応じて横断的に検索しやすい環境を構築するために RDF が用いられつつある．このような状況において，我々はそれらデータベースの構築者と利用者の双方に有益となるようなサービスを構築している．具体的には次の通りである．

- 所望のデータを探す，
- 所望のデータベースの構造を把握する，
- 所望のデータの信頼性や利用しやすさを調べる，
- 特定のデータベースについて質問したり提案したりする

今後は，Umaka Search や Umaka Viewer の対象を拡充する，Umaka-Yummy Data の採点方法を改良するなど，有用性を高める技術開発を引き続き進めていく．

参考文献

[Benson 2013] Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W.: GenBank, Nucleic Acids Res., 41(D1): D36-D42, 2013.

[Rigden 2016] Rigden DJ, Fernández-Suárez XM, Galperin MY.: The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection, Nucleic Acids Res., 44(D1):D1-6, 2016.

[The UniProt Consortium 2017] The UniProt Consortium: UniProt: the universal protein knowledgebase, Nucleic Acids Res., 45 (D1): D158-D169, 2017.

[Yamaguchi 2016] Yamaguchi A., Kozaki K., Lenz K., Yamamoto Y., Masuya H., Kobayashi N.: Semantic Data Acquisition by Traversing Class–Class Relationships Over Linked Open Data, LNCS 10055, pp 136-151, 2016.