The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

質の高い生命科学 Linked Data 利用基盤の構築に向けて

Constructing a better Linked Data infrastructure in Life Sciences based on our experience

山本泰智 Yasunori Yamamoto
山口敦子 Atsuko Yamaguchi

情報・システム研究機構
データサイエンス共同利用基盤施設
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
Joint Support-Center for Data Science Research
Research Organization of Information and Systems

We provide Umaka-Yummy Data system to facilitate mutual understandings between Linked Data providers and consumers. Our aim is to build a better Linked Data user community in the life sciences domain. Our system monitors and evaluates each Linked Data provider in terms of six aspects by issuing a series of SPARQL queries and the other HTTP requests. Through our three-year experience of operating the system, we learned that a mutual understanding between the Linked Data providers and us is important to provide reliable monitoring results from Umaka-Yummy Data system.

1. 生命科学分野におけるデータの急増への対応

生命科学分野においては実験機器の発展や研究者の増加に伴い，得られる研究成果が大幅に増加している．これに伴い，関連情報を収めるデータベースも増加しており，有用なデータベースを紹介する論文誌において紹介されているものだけでも 1500 を超えている[Rigden 2019]．このような状況において，研究者は自身の研究に関連するデータを適宜効率よく見つけることが困難な状況が発生している．関連するデータを取得するために複数のデータベースを横断的に検索する必要があることも多く，この問題に対する解決策の一つとして Resource Description Framework（RDF）を活用したデータベースが公開されつつある．RDF を採用し，生命科学分野で重要な概念である遺伝子やたんぱく質について，個々の URI を割り当てることにより，複数のデータベースから効率よく必要な情報を取得しやすくなる．また，多くの公開されている RDF データベースは当該 URI にアクセスすることで関連情報が得られるように参照解決可能としている．従って，生命科学分野における Linked Data が充実してきたといえる．また，これらのデータにアクセスできるように SPARQL エンドポイントが公開されていることも多いため，利用者は自身の開発したプログラムを利用して機械的にデータを取得することも容易になってきている．

その一方で，公開されている SPARQL エンドポイントは様々なデータ提供者が維持管理しているため，同じデータが異なる SPARQL エンドポイントから得られることもあり，一方はあまり更新がなされていなかったり，もう一方はアクセスできないことあったりなど，利用者からみて様々な原因により SPARQL エンドポイントが使いにくい状況にある．

そこで，ライフサイエンス統合データベースセンター（DBCLS）では生命科学分野において公開されている RDF データの提供サイトである SPARQL エンドポイントについて，定期的に死活確認や格納されているデータをモニターしてその結果を公開するサービス Umaka-Yummy Data を立ち上げた[Yamamoto 2018]．本サービスの目的は，RDF データを利用したい研究者が信頼できる SPARQL エンドポイントを容易に見つけられるようにすることだけでなく，データ提供者が自身の提供するサービスの状況を確認しやすくすることであり，さらにはデータの利用者と提供者の相互理解を促し，より良い Linked Data の利用基盤を構築することにある．

Umaka-Yummy Data を正式に立ち上げてから既に 3 年が経ち，これまでデータの提供側からの問い合わせを受けながらシステムを改変してきた．本論文ではこれまでに得られた知見を紹介するとともに，質の高い Linked Data を提供するために必要な事項を議論することで，有益な Linked Data がより多く提供される環境の実現に貢献することを目標とする．結論としてデータ提供者側の参加が重要であり，そのために Umaka-Yummy Data が一定の役割を果たしていることが分かった．

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 柏市若柴 178-4-4，04-7135-5508，yy@dbcls.rois.ac.jp

2. Umaka-Yummy Data

Umaka-Yummy Data は毎日各 SPARQL エンドポイントに対して一連の SPARQL クエリを発行してその結果を得たり，VoID などのメタデータの有無を確認したり，CORS 対応の有無を調査したりしてそれらをデータベースに格納している．格納されたデータをもとに，各々の SPARQL エンドポイントに対する評価を 0 から 100 までの整数値で表し，Umaka Score という名前の下で公開している（図 1 参照）．Umaka Score は，6 つの評価軸のそれぞれで得られる値の平均値であり，それらの評価軸は以下の通りである．

1. 利用可能度（Availability）
2. 新鮮度（Freshness）
3. 利便度（Operation）
4. 有用度（Usefulness）
5. Linked Data 原則への対応度（Validity）
6. 処理速度（Performance）

これらの各評価軸の数値は以下のデータから生成している．それぞれの番号が上記の評価軸のそれに対応し，一つの評価軸の数値を得るのに複数のデータを利用している場合には，それぞれアルファベットを付けて区別している．

1. 過去 30 日間の稼働日数
2. 最終更新日

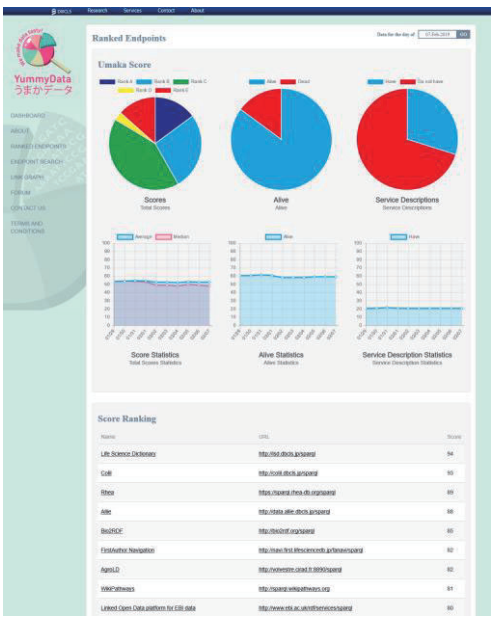

図 1 各エンドポイントのスコア一覧表示画面

3. メタデータ(SD および VoID)の提供の有無
4a. 広く使われているオントロジーの利用の有無
4b. クラス名や型指定の有無
4c. 他のデータセットへのリンクの有無
4d. コンテントネゴシエーション利用の有無
5a. 参照解決な URI であるか否か
5b. 適切な長さの URI であるか否か
6. クエリ処理時間

現在，情報収集する対象としている SPARQL エンドポイントの数は 68 で，長時間アクセスできないエンドポイントは手動で状況を確認したうえで対象から除外している．

3. 得られた意見や改善案

これまでに Umaka-Yummy Data を運用して得られた知見に基づき，大きく分けて次の二点について議論する．すなわち，データ提供者との信頼関係の構築と，データを収集するための SPARQL クエリや HTTP リクエストの見直しである．以下，それぞれ詳細に述べる．

3.1 データ提供者からの信頼を得る

Umaka-Yummy Data は関連する SPARQL エンドポイントに対して毎日一連の SPARQL クエリやその他の HTTP リクエストを発行してデータを取得しているが，特に対象となる SPARQL エンドポイントが格納するデータの容量が大きい場合にはそれなりの負荷をかけることとなる．このため，データ提供者との信頼関係を築くことが，Umaka-Yummy Data システムが安定して有益なデータを提供し続けるために重要である．そのために以下の事項を実践している．

(1) クローラーの出自を明示する

データを取得するためのクローラーは各 SPARQL エンドポイントにアクセスするさいに，HTTP ヘッダの User-Agent の値に Umaka-Crawler という記述と，コンタクトポイントを含めている．これにより，我々のクローラーが先方のサーバーに不適切なアクセスを行っているときには，先方はアクセスログから我々を特定し，対応できる．これは以前強い負荷をかけてしまうクエリを発行したさいに生じた一連の事案を踏まえた結果である．

(2) 得られた結果について詳細に公開する

具体的なクエリと，得られた結果を技術的に詳細に提供しているが，データ提供者側において不明と思われる現象が観測された場合に，その理由が必ずしも自明でない事例がある．このような場合には，データ提供者がシステム運用者である我々に問い合わせをし，当方にて調査する必要がある．そこで，得られた結果の再現性の有無を提供者側が確認しやすい環境を構築し，問い合わせをする前に予め問題点を把握しやすくすることが今後の課題となる．具体的には，クローラーが発行した問い合わせを再現できるようなコマンドの文字列を提示することにより，データ提供者がそれをターミナルにて貼り付けて実行可能とする仕組みを検討している．

(3) 提供者からの質問に対応する

サービスを立ち上げてからこれまでに様々な質問や意見を受けているが，当初の想定に反し，大半が筆者への個人的な問い合わせである．立ち上げ当初から我々は Linked Data の提供者と利用者の間の相互理解を促すことを目標としていたため，様々な形でその場を用意する計画でいた．システムのトップページから問い合わせできるフォームを設置するにとどまらず，システムのソースコード一式を GitHub で公開したり，Twitter アカウントを立ち上げたりした．当初は GitHub の Issues で問い合わせがなされることもあったが，筆者との個人的な関係が構築されると，直接 Skype で問い合わせがなされるようになってきた．これは問い合わせをする側との信頼関係が築けてきた証といえるが，他により容易に質問などを行える場がないか引き続き検討する必要がある．

3.2 クエリの改善

上述の通り，Umaka-Yummy Data ではエンドポイントから取得したデータをもとに，いくつかの観点から評価し，それを分かりやすくするために数値化して Umaka Score として公開している．その中には必ずしも目的の値を取得するために最適とは言えないクエリが含まれていたので，それを見直す作業を進めている．また，評価項目の中には実際にクエリで表現する方法や得られた結果に対する解釈が明確に定義しにくいものも含まれている．そこで，引き続き本来の目的である，より良い Linked Data の利用基盤環境を実現するという見地に照らして改善していく計画である．以下に具体例を示す．

(1) クエリの最適化例

RDF データセットにおいて，クラスを示す URI が含まれている場合には，そのクラス URI に rdfs:label プロパティを用いた名前が付けられていると利用者が当該データセットを理解する助けとなる．このため，Umaka-Yummy Data ではそれを調査するための SPARQL クエリを発行している．

現状では，得られたクラス URI のセットに対して，FILTER と IN を用いた SPARQL クエリを発行しているが，一般的に FILTER を用いるより，VALUES を用いた方が探索範囲を狭められるため，問い合わせ内容は論理的に変わらず検索の高速化が望める．さらに，クラスの数が非常に多い場合は一つのクエリの長さが膨大になり，クエリを発行すらできない可能性もある．

そこで，クラス名を取得することと，それぞれのクラスの名前を取得することを一つのクエリに含めることでこの問題を回避する．また，予めクラス名を取得して保存しておく必要がないので，プ

ログラム全体の効率化に繋がる. 具体的なクエリは以下のようになる.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?class ?label
WHERE {
  GRAPH <named-graph> {
 ?class rdfs:label ?label .
 [] a ?class.
}}
```

(2) クエリ結果の解釈

Umaka-Yummy Data の評価項目として, コンテントネゴシエーションへの対応の有無がある. これは次の二つの理由による. 第一に, Linked Data の原則[1]として, URI へのアクセスに対して, 当該 URI が示す概念に関する有益な情報を返すことが望ましいとされていること, 第二に, それらの情報は人にも機械にも適したものであると理想的であるから, クライアントの要求に基づき, HTML もしくは RDF（Turtle や RDF/XML など）, それぞれ別々の URL にリダイレクトさせることが望ましいとされているためである. その一方で, 特にオントロジーの場合には個々の概念を表す URI には#記号を用いたハッシュ URI を用いることが多い. この場合, コンテントネゴシエーションは行わず, RDF/XML 形式のデータが得られることが多い.

従って, コンテントネゴシエーションに対応していることは評価軸として適切と判断できる一方, オントロジーの場合はそれに非対応であることを理由として評価を低くすることは適当ではない. 従って, アクセスする URI がハッシュ URI であるか否かで対応を変えることを検討している.

3.3 議論の場の設定

各エンドポイントについて GitHub の Issues の機能を利用した議論の場を提供しているが, これまで得られている意見や質問は, Umaka-Yummy Data への質問などと同様に, SNS を利用した筆者への個人的な連絡が多い. また, Umaka-Yummy Data のツイッターアカウントを通した連絡も行われている. その一方で, システムを公開しているサイトにはウェブから意見や質問などを投稿できる場を提供しているが, スパムばかりで実質機能していない.

このことから, Umaka-Yummy Data システム運用者の顔の見えるコンタクトポイントが重要であり, 比較的容易に, そして迅速な対応が期待できるチャネルを用意することで, データ提供者とより良い関係を築けるといえる.

4. 結論

Umaka-Yummy Data は, 各 SPARQL エンドポイントの様々な情報を定期的に取得してその結果を公開している. その目的は, データ提供者とデータ利用者の間の情報交換の場を提供することで両者の理解が深まり, より良い Linked Data 利用環境が構築できると考えているからである. 現状では提供者と利用者の間での情報交換は進んでいない一方で, 提供者と Umaka-Yummy Data の運用者との間の関係は深くなり, データ提供者側における問題点が比較的容易に判明する事例が増えているばかりでなく, Umaka-Yummy Data における評価方法の問題点も見つかりやすい. この状況はデータ提供者と Umaka-Yummy Data 運用者の間での Linked Data 利用環境における問題点を共有できるという利点があるので, 今後も引き続き良好な関係を保つように努めたい. そして生命科学分野における Linked Data コミュニティの発展に貢献できればと考えている.

[1] https://www.w3.org/DesignIssues/LinkedData.html

参考文献

[Rigden 2016] Rigden DJ, Fernández-Suárez XM: The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection., Nucleic Acids Res., 47(D1):D1-D7, 2019.

[Yamamoto 2018] Yamamoto Y, Yamaguchi A, Splendiani A: YummyData: providing high-quality open life science data., Database (Oxford), 2018.