ライフサイエンスの RDF データベースにおける スキーマ定義の現状分析

A survey of schema definitions for life-sciences RDF databases

山口敦子 [1]　櫛田達矢 [2]　山本泰智 [1]　古崎晃司 [3]

Atsuko Yamaguchi [1], Tatsuya Kushida [2], Yasunori Yamamoto [1] and Kouji Kozaki [3]

[1] 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
[1]Database Center for Life Science, ROIS
[2] 科学技術振興機構 バイオサイエンスデータベースセンター
[2] National Bioscience Database Center, JST
[3] 大阪大学 産業科学研究所
[3] The Institute of Scientific and Industrial Research, Osaka University

概要: ライフサイエンスのデータベースは，測定機器の発展に伴い，ますます巨大化し複雑化している．これらのデータベースを統合的に利用可能とするために，セマンティックウェブ技術の利用が広まっており，セマンティックウェブ技術の標準モデルである RDF を用いた RDF データベースが数多く作られてきている．RDF データベースを公開する際には，RDF の標準的な検索言語である SPARQL を用いた検索が可能なウェブ API である，SPARQL エンドポイントが提供されることが多い．SPARQL エンドポイントに効率的に検索をかけるためは，RDF データベースのスキーマを予め取得し利用できることが望ましい．しかし，巨大なデータベースが多いライフサイエンス分野においては，SPARQL エンドポイントから網羅的にスキーマを取り切れないことも多い．一方，スキーマ定義を適切に記述して提供してあれば，網羅的にスキーマを取る必要はなく，スキーマ定義のみを取得すればよい．そこで，ライフサイエンスデータベースにおけるスキーマがどの程度適切に定義され取得可能になっているかを，特にプロパティの主語目的語クラスに着目して調査した．その結果，一部のデータセットについては，スキーマ定義用の語彙の一つである RDF Schema 1.1 の domain/range 定義は利用できる可能性があるものの，OWL2 はデータセットのプロパティのスキーマ定義としてはほとんど利用されていないことがわかった．

1. はじめに

ライフサイエンスでは，日々実験によってデータが生産され，蓄積されている．これらのデータを再利用可能とし，新たな知識を得るための基盤とするために，多くのデータベースが RDF 化され公開されてきている． RDF データベースの検索には RDF の標準的な検索言語である SPARQL が主に用いられる．したがって，RDF データベースを公開する際には，SPARQL エンドポイントとよばれる，SPARQL を用いた検索が可能なウェブ API がしばしば提供される．ライフサイエンス分野でも，SPARQL エンドポイントは次々と新たに提供されている．そのため，ユーザが公開され続ける RDF データベースの全体像を理解するのは困難である．

しかしながら，ユーザが SPARQL エンドポイントを適切に選び，SPARQL 検索を行うためには，どのようなデータがどの SPARQL エンドポイントによって提供されるかを知る必要がある．さらに，その SPARQL エンドポイントで提供される RDF データベースの構造を把握して検索クエリを記述する必要がある．これまで，著者らは，ユーザが SPARQL エンドポイントを選択するための情報を提供するサービス Umaka-Yummy Data [1]や，SPARQL エンドポイントで提供される RDF データベースの構造を利用して SPARQL クエリ記述を補助するサービス SPARQL Builder [2]，RDF データベースのスキーマ構造を表示するサービス Umaka Viewer [3]を開発してきた．

これらのサービスに共通して有用となる RDF データベースの構造の情報として，プロパティの情報，特にプロパティの主語目的語となるクラスの情報が挙げられる．現在，これらの情報を得るために，ライフサイエンス分野の SPARQL エンドポイントにクエリを投げ，その SPARQL エンドポイントで提供される RDF データベースのスキーマを網羅的に取得し蓄積している．しかし，この方法は SPARQL エンドポイントに強い負荷をかける．特にトリプル数が多いデータセットを提供する SPARQL エンドポイントにおいては，LIMIT/OFFSET を利用して細切れで取得する必要があり，その場合は大量の SPARQL クエリを投げることになる．また，SPARQL エンドポイントによっては，トリプル数に比してタイムアウトの時間が短いなどの理由で，かなりクエリを細切れにしてもタイムアウトになり，結果的に取得できないこともある．

一方，プロパティやクラスの関係を記述できる語彙として RDF Schema 1.1 [4] ，より厳密に機械的な意味処理が可能なレベルでプロパティやクラスの関係を記述できる語彙として OWL2 [5]が挙げられる．もし，各データセットにおいて，そのデータセットに含まれる全てのプロパティ，全てのクラス，プロパティの主語目的語となるクラスの情報が適切に記述され提供されていれば，上記のような大量の負荷が高い SPARQL クエリを SPARQL エンドポイントに投げる必要がない．

そこで，本研究では，プロパティの主語目的語クラスに着目し，このスキーマについて RDF Schema 1.1 や OWL2 でどの程度記述されているかをライフサイエンス分野の SPARQL エンドポイントごとに調査した．その結果を網羅的に SPARQL エンドポイントから取得した場合と比較する．

2. スキーマ定義調査

2.1 調査対象

調査対象となる SPARQL エンドポイントは，まず，2018 年 12 月時点で Umaka-Yummy Data[1] のクロール対象となっていた 58 の SPARQL エンドポイントを全て初期の対象とした．Umaka-Yummy Data は，生命科学データを中心とした SPARQL エンドポイントのリストに対し，毎日アクセスを行い，サービスの死活，更新頻度，VoID 等のメタデータ提供状況，クエリ返答速度などを監視し，その結果をスコア付けするシステムである．さらに，それら 58 の SPARQL エンドポイントのうち，プロパティとプロパティの主語目的語クラススキーマ定義を網羅的に取得できた 46 の SPARQL エンドポイントを最終的に調査対象とした．

2.2 調査手法

調査対象となった SPARQL エンドポイントに対し，そこで提供されるデータセットのプロパティの一覧，網羅的に rdf:type を利用して取得したプロパティの主語目的語クラス，RDF Schema 1.1 の rdfs:domain, rdfs:range を利用して取得したプロパティの主語目的語クラス，OWL2 の owl:Restriction, owl:onProperty を利用して取得したプロパティの主語目的語クラスを，SPARQL クエリを投げることで取得した．

利用した SPARQL クエリは以下の通りである．実際には LIMIT/OFFSET を利用しているが，本稿では省く．

(1) プロパティの一覧取得

```
SELECT DISTINCT ?p WHERE{
                    ?s ?p ?o . }
```

(2) プロパティの主語目的語クラス取得 (rdf:type)

```
SELECT DISTINCT ?c1 ?p ?c2 WHERE{
        ?s ?p ?o .
        ?s a ?c1 .
        ?o a ?c2 . }
```

(3) プロパティの主語目的語クラス取得 (RDF Schema 1.1)

```
SELECT ?c1 ?p ?c2 WHERE{
        ?p rdfs:domain ?c1 .
        ?p rdfs:range ?c2 . }
```

(4) プロパティの主語目的語クラス取得 (OWL2)

```
SELECT ?c1 ?p ?c2 WHERE{
         ?c1 rdfs:subClassOf ?s .
         ?s ?p1 owl: Restriction ;
            owl:onProperty ?p .
         { ?s owl:someValuesFrom ?c2 . }
         UNION
         { ?s owl:allValuesFrom ?c2 . }
         }
```

上記 4 つのクエリ結果をそれぞれ P, TYPE_R, RDFS_R, OWL_R と記述することにする．ただし，トリプルストア OpenLink Virtuoso [6]を利用した SPARQL エンドポイントには自動的に独自のメタデータが付き，これらのメタデータは本調査の目的に沿わないため削除した．具体的には URI に”openlinksw.com”の文字列を含むプロパティやクラスを全て削除した．

3. スキーマ定義調査結果と分析

3.1 調査結果

最終的に調査対象になった 46 の SPARQL エンドポイントに対し，P, TYPE_R, RDFS_R, OWL_R の最大サイズ，最小サイズ，平均サイズを表 1 に示す．

表 1: 46 エンドポイントの調査結果サイズ

	最大	最小	平均
P	5606	1	345.1
TYPE_R	450169	0	26066
RDFS_R	2152	0	222.4
OWL_R	387900	0	18574.7

ちなみに, P の最大サイズは NBDC が提供する LSDB アーカイブ，TYPE_R は DBCLS が提供する新着論文レビュー, RDFS_R は EBI が提供する RDF Platform, OWL_R は Patient-Reported Outcome Consortium の SPARQL エンドポイントから得られたものとなる．

TYPE_R は一つのプロパティに対し，複数の主語目的語クラスを割り当てたデータが多くあるため, P のサイズよりかなり大きいものが多い結果が得られている．また，OWL_R はデータセットのプロパティとしては利用されていないプロパティへの定義も含まれるため，P のサイズよりかなり大きいものが多い結果が得られている．そこで，P に含まれるプロパティに対し，TYPE_R，RDFS_R，OWL_R の結果のプロパティが少なくとも一つ存在するものの数について議論する．P(TYPE)，P(RDFS)，P(OWL)はそれぞれ，TYPE_R，RDFS_R，OWL_R に少なくとも一つプロパティの主語クラス目的語クラスが存在するプロパティの集合を表す．このとき, P(TYPE), P(RDFS)，P(OWL)の最大サイズ，平均サイズを表 2 に示す．

表 2: P(TYPE), P(RDFS), P(OWL)のサイズ

	最大	平均
P(TYPE)	767	91.7
P(RDFS)	244	46.0
P(OWL)	11	1.5

P(TYPE)と P(TYPE)の最大サイズは理化学研究所が提供する理研メタデータベース, P(OWL)の最大サイズは Pathway Commons の SPARQL エンドポイントから得られたものである．

3.2 調査結果の分析と考察

表 1, 2 を比較すると, OWL_R と P(OWL)の平均サイズの差が非常に大きいことが分かる．図 1 は 46 の SPARQL エンドポイントに対し，OWL_R のサイズが大きい順に並べたうえで，OWL_R と P(OWL)のサイズをプロットしたものである．SPARQL エンドポイントごとのサイズの差が大きいため，y 軸は対数軸となっている．

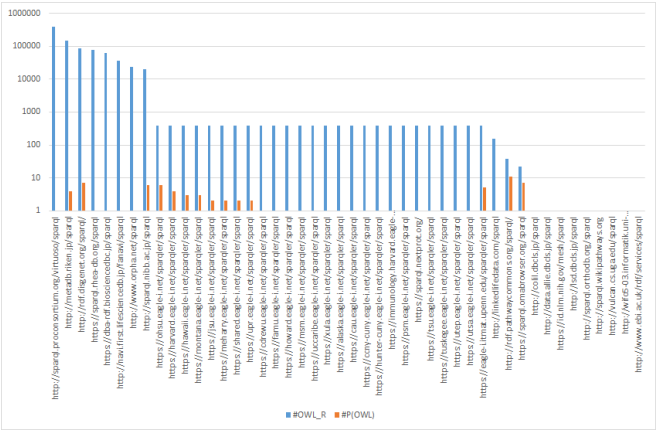

図 1: OWL と P(OWL)のサイズ比較

この図より，OWL2 によって定義されたプロパティは，SPARQL エンドポイントによっては大きな数になるものの，ほとんど利用されていないことが分かる．比較的 OWL_R が大きいデータセット内部を見たところ, OWL/OWL2 で書かれた第三者が作成したオントロジーを複数そのまま，データセットと共にトリプルストアにロードし，SPARQL エンドポイントから提供している例が多く見られた．これらのオントロジーはプロパティの定義のためにロードされたものではなく，他の目的(クラス定義など)のために，データセットと共にロードされていることが推測される．

一方，RDFS_R と P(RDFS)は，OWL_R と P(OWL)ほどの差はなく，RDF Schema 1.1 の rdfs:domain, rdfs:range を用いて定義されたプロパティはデータセット内でも比較的用いられている．データセットによっては，rdf:type を用いて網羅的に取得した P(TYPE)より多く定義されたものも存在した．図 2 は 46 の SPARQL エンドポイントに対し, P のサイズが大きい順に並べたうえで，P, P(OWL), P(RDFS), P(TYPE)をプロットしたものである．SPARQL エンドポイントごとのサイズの差が大きいため，y 軸は対数軸となっている．

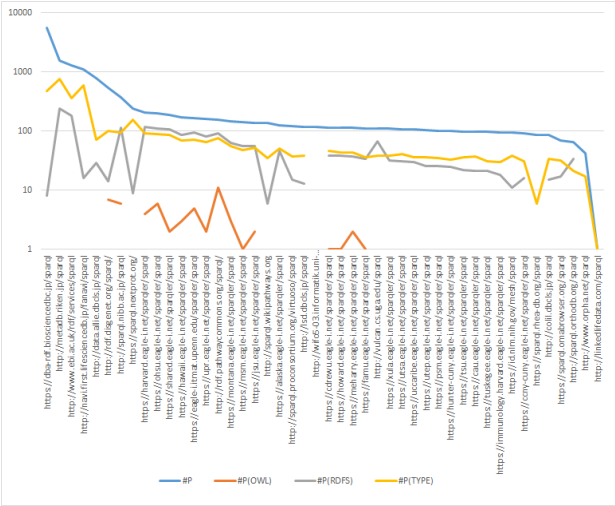

図 2: プロパティに関するスキーマ定義比較

この図により，RDF Schema 1.1 の rdfs:domain, rdfs:range で取得したプロパティの主語クラスや目的語クラスの結果数は，rdf:type で網羅的に取得したものとほぼ数的に差がない SPARQL エンドポイントが多数あることが見て取れる．すなわち，データセットによっては，rdf:type を用いて網羅的に重く大量のクエリを SPARQL エンドポイントに投げなくとも，RDF Schema 1.1 での定義を取得するだけで利用できる可能性がある．

4. まとめ

プロパティの主語目的語となるクラスの情報を SPARQL エンドポイントに網羅的にクエリを投げて取得する代替手段として，RDF Schema 1.1 の rdfs:domain, rdfs:range および OWL2 の owl:Restriction, owl:onProperty の利用可能性について調査を行った．その結果，RDF Schema 1.1 の rdfs:domain, rdfs:range で主語クラス目的語クラスが定義されたプロパティについては，データセット内でプロパティとして多く利用されており，rdf:type を用いた網羅的手法の代替となりうる可能性が示された．一方で，OWL2 の owl:Restriction, owl:onProperty で定義されたプロパティはほとんどデータセット内では利用されていないことがわかった．

今後の課題として，RDF Schema 1.1 の rdfs:domain, rdfs:range が真に rdf:type を用いた網羅的手法の代替となりうるかどうかを示すため，RDFS_R と TYPE_R の関係をより細かく見ていく必要がある．また，今回 OWL2 の owl:Restriction, owl:onProperty で定義されたプロパティはほとんどデータセット内では利用されておらず，TYPE_R の代替とはなりえない一方，データセット自体が OWL2 で記述されている場合など，TYPE_R や RDFS_R では拾えない関係を記述している可能性もある．RDF という同じモデルであっても，異なる記述方法を取るこれらのデータをどのように統合的に利用可能としていくかを考えることも今後の課題である．

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，および科学研究費補助金基盤(C)17K00434，基盤(B) A17H017890 の助成による．

参考文献：

[1] Yamamoto, Y., Yamaguchi, A., Splendiani, A.: YummyData: providing high-quality open life science data. Database, doi: 10.1093/database/bay022, 2018

[2] Yamaguchi, A., Kozaki, K., Lenz, K., Yamamoto, Y., Masuya, H., Kobayashi N.: Semantic Data Acquisition by Traversing Class-Class Relationships Over Linked Open Data, The 6th Joint International Semantic Technology Conference (JIST 2016), LNCS 10055, 136-151

[3] 山本泰智, 山口敦子: より良い生命科学データ利用環境の構築を目指して. 第 31 回人工知能学会全国大会, 1N3-OS-39b-3in1, 2017

[4] RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/

[5] OWL 2 Web Ontology Language Document Overview, https://www.w3.org/TR/owl2-overview/

[6] OpenLink Virtuoso, https://virtuoso.openlinksw.com/