関連する LOD データセットの発見を目的としたリンク関係の調査

A study on Linked Open Data (LOD) links to identify related datasets

山中勇樹 [1]　永森光晴 [2]　三原鉄也 [2]　杉本重雄 [2]

Yuki Yamanaka[1], Mitsuharu Nagamori[2], Tetsuya Mihara[2] and Shigeo Sugimoto[2]

[1] 筑波大学大学院図書館情報メディア研究科

[1]Graduate School of Library, Information and Media Studies, University of Tsukuba

[2] 筑波大学図書館情報メディア系

[2] Faculty of Library, Information and Media Science, University of Tsukuba

Abstract: Linked Open Data (LOD) datasets and data catalog sites to curate them are proportionally increasing in number. Even though such data catalog sites provide metadata for different LOD datasets, provisions for analyzing the relationship between them are limited. To address this challenge, the authors investigated the property usage trends, based on genre and creation date for exploring relationships between different LOD datasets.

1. はじめに

Web 上で情報の公開・共有を行う仕組みとして Linked Open Data(LOD)が注目されている[1]。Web 上で公開されている LOD の形式に則ったデータセット(LOD データセット)の数は年々増加の一途をたどっており、LOD データセットを用いたアプリケーションの開発なども促進されている。アプリケーションの開発などを行う際、利用者が独自に用意したデータの他、Web 上で公開されている既存の LOD データセットを再利用することもできる。既存の LOD データセットを再利用することで、新規データの用意にかかるコストを削減し、開発を効率化することができる。既存の LOD データセットはデータカタログサイトと呼ばれるサイトで検索することができる[2][3]。データカタログサイトでは LOD データセットのタイトルや概要などの基本的な情報の他、SPARQL エンドポイントやダンプファイルをダウンロードできるページの URL などの情報が提示されている。キーワードや条件などで絞り込みを行い、それらの情報を閲覧することで使用する LOD データセットを決定する。しかし、LOD データセット自身の情報を閲覧することはできても、関連する他のデータセットの情報の閲覧をできるサイトは非常に少ない関連 LOD データセットの情報を閲覧することができれば、類似 LOD データセットの比較や合わせて使える他の LOD データセットの情報の提示などが行えるようになり、より LOD データセットの探索を効率化できると考えられる。

本研究では特定の LOD データセットに関連する別の LOD データセットの発見を目的として LOD データセット同士を結ぶリンク関係の調査を行う。異なる LOD データセット間のリソースを結ぶプロパティに着目し、それらの使用傾向の調査を行う。これにより、ある LOD データセットが他の LOD データセットのどのような情報を参照しているかが分かるようになり、それらの傾向により LOD データセット同士がどのような関係にあるかを推定することができる。LOD データセットで多く使用されているクラスやプロパティと LOD データセットの分野、そのデータセットが作成された年代などの観点から分析を行い、それらの情報が LOD データセットの検索に貢献できるかを考察する。

2. 関連 LOD データセットの発見

Web 上で公開される LOD データセットの数は年々増加の一途をたどっている。LOD チャレンジのように LOD データセットの活用を促進するコンテストなども開催されており、LOD の利用を促進する動きは活発化してきている[4]。Web 上で公開されている既存の LOD データセットはデータカタログサイトで検索し、詳細な情報を得ることができる。しかし、データカタログサイトで得ることができる情報はその LOD データセット自身についてのものだ

けであり、他の LOD データセットとどのような関係にあるかといった情報を得ることができるサイトは非常に少ない。また、データカタログサイトの数も多く、それぞれに異なる LOD データセットが登録されている。そのため、別々のデータカタログサイトに関連する LOD データセットが登録されていた場合、ある LOD データセットを発見してもそれと関連する LOD データセットがどのデータカタログサイトに登録されているかがわからず、関連する LOD データセットにたどり着くことは難しい。

関連する LOD データセットには類似性の高いものと、組み合わせて使いやすいものの 2 パターンが考えられる。類似性の高い LOD データセットの発見が可能になれば、それらのデータ量や更新頻度、メタデータスキーマが公開されているか、それらにはどのようなメタデータ語彙が使用されているかといった情報を比較し、どちらがより良い LOD データセットなのか判断しやすくなると考えられる。また、組み合わせて使いやすい LOD データセットの情報は発見した LOD データセットをどのように利用するかその指針になりうる。さらに、類似する LOD データセットの情報と合わせて閲覧すれば、類似している LOD データセットが他のどのデータセットと関係があるかといった観点から比較することが可能になり、より探索者の目的に適した LOD データセットを選びやすくなると考えられる。関連データセットを提示する先行研究として筆者らは LOD データセットを用いて開発されたアプリケーションに着目して、LOD データセットとそれを用いて開発されたアプリケーションの情報を収集し、それらを検索できるシステムを構築した[5]。同じアプリケーションの開発に使用されている LOD データセット同士ならば、組み合わせて使いやすいと考えられる。しかし、多くの LOD データセットは開発事例が提示されていないため、発見できる LOD データセットは少ない。

関連する LOD データセットの指標にはそのデータセット内のリソースが他の LOD データセット内のリソースを参照する際のプロパティを用いることができる。特定の LOD データセットから他の LOD データセットに対してどのようなプロパティを用いているかその傾向を分析することで、それらの LOD データセットがどのような関係にあるかを推測することができる。一例として、同一ジャンルの LOD データセット同士において「owl:sameAs」などの等価関係を表すプロパティでリンクされたリソースが多ければ、それらの LOD データセットには類似性があると考えられる。また、異なる LOD データセット間において「rdfs:seeAlso」や「foaf:isPrimaryTopicOf」のようなプロパティで繋がれたリソースが多ければ、それのデータセットには関連性の強いリソースが多く含まれているといえ、組み合わせて使いやすいと考えられる。

本研究では異なる LOD データセット同士をリンクするプロパティに着目し、どのようなプロパティが多く使用されているか、それらはどのようなクラスに使用されているか、それらの LOD データセットのジャンルは何かといった観点から LOD データセットのリンク関係を分析する。プロパティの使用傾向を集計することによって、ジャンルごとにどのようなプロパティが使用されることが多いか、LOD データセットが作成された年代によって使用頻度の高いプロパティに変化はあるかといったことを分析する。さらに、プロパティの情報と合わせて、そのプロパティによって繋がれているリソースはどのようなものか、どのようなジャンルの LOD データセット同士をリンクする際に使われているかといった情報を考察することによって、類似する LOD データセットや組み合わせて使いやすい LOD データセット同士をリンクさせる際にどのようなプロパティが用いられることが多いのかを明らかにすることが本研究の最終的な目標である。

3. 既存のプロパティに関する調査

LOD データセットで用いられる主要なメタデータタームは Linked Open Vocabularies(LOV)[6]で検索を行うことができる。LOV には 2019 年 2 月時点でクラスが 27,901 件、プロパティが 35,300 件登録されている。また、それらのタームに関する情報を検索するための SPARQL Endpoint も提供されており、各タームのドメインやレンジ、データセットで使用された数などの情報を検索することができる。LOV に登録されているタームにはそのタームを使用しているデータセットの数とそれらで使用している回数の合計が登録されており、メタデータタームを選出する上での一つの指標になる。しかし、この使用データセット数と合計使用回数が登録されているタームはクラスが 27,901 件のうち 665 件、プロパティが 35,300 件のうち 1,518 件と非常に少ない。また具体的にどのようなデータセットで使用されているかといった情報もユーザ視点で得ることはできないため、データセット間におけるリンクを分析するには適さないと考えられる。

LOD データセットのプロパティの統計情報を分析する研究としては他に LODStats[7]が挙げられる。LODStats では規模が増大するデータの Web を分析することを目的に、データカタログサイトに登録さ

れている RDF 形式のデータセットの統計情報をまとめたデータセットを作成している[8]。このデータセットにはデータカタログサイトに登録されているデータセットの更新情報やトリプル数などの基本的な情報のほか、メタデータタームの使用数や名前空間の出現頻度といった情報をまとめている。また、使用しているメタデータタームの情報を元にリンクの対象となる他のデータセットの識別も可能としている。しかし、このデータセットは 2016 年以降更新がされておらず、最新のデータセットに関する情報を得ることはできない。また、LODStats では登録されているデータセット全体におけるクラスやプロパティの使用頻度などの情報をまとめているものの、データセットのジャンルや年代などで分けてはいないため、ジャンルや年代における使用タームの違いといった情報を得ることはできない。

そのため、本研究では LOD データセットをジャンルや年代でグループ分けし、それぞれにおけるタームの使用傾向の違いや変遷を分析する。

4. LOD データセットをリンクするプロパティの傾向調査

4.1 調査手法

複数の LOD データセットにおいて使用頻度の高いプロパティの傾向を分析する。傾向の分析を行うためには、それぞれのプロパティがどれほど使用されているか調査する必要がある。しかし、データセットの規模は様々であり、大規模なデータセットがあるプロパティを多く使用していた場合、そのプロパティの合計使用回数の値は大きくなり、全体的なプロパティの傾向は大規模なデータセットで使用されているプロパティに大きく影響を受けることが考えられる。よって、合計使用回数とは別に各プロパティを使用しているデータセット数も取得する必要がある。また、プロパティすべての合計使用回数と使用データセット数を計測し、それらの分析を行うにはコストがかかる。そのため、プロパティをそれらの名前空間でグループ分けし、その中で使用されていた各プロパティの数を計測する。調査項目をまとめると以下のようになる。

- ある名前空間を URI に含むプロパティの合計使用回数
- その名前空間を URI に含むプロパティの数
- その名前空間を URI に含むプロパティを使用していたデータセットの数

これらの項目のほか、そのプロパティが外部のリソースと繋がっているかを検証するためにはプロパティの主語と目的語それぞれの URI を取得する必要がある。図 1 の SPARQL 式を実行することであるデータセットにおいてリソース同士をつなぐプロパティとそのプロパティの主語のクラス、それらの組の使用回数を取得することが可能である。単一のデータセットの分析を行うだけならば、SPARQL エンドポイントでこの SPARQL 式を実行するだけでも十分な情報を得ることができるが、複数のデータセットの傾向の分析を行うにはデータセットそれぞれのクラスやプロパティの情報が必要となる。SPARQL エンドポイントにおいてはあるエンドポイントから別のエンドポイントに対して問い合わせを行う Federated Query という手法があり、この手法を用いることで複数の LOD データセットの情報を同時に取得できる[9]。しかし、多くの主要な LOD データセットが SPARQL エンドポイントを公開するために用いている RDF ストア Virtuoso[10]は Federated Query に対応していない。それらのデータセットの情報を得るためには、それぞれの SPARQL エンドポイントにてクエリを実行する必要がある。しかし、すべての LOD データセットの SPARQL エンドポイントから逐一情報を取得するのは、手間がかかる。よって Federated Query を用いずに、複数の LOD データセットの情報を取得できることが望ましい。その方法としては、複数の LOD データセットに登録されているインスタンスを横断的に検索できる LOD Cloud Cache[11]や LOD4ALL[12]のようなサービスを利用する方法とダンプファイルを取得してローカルに準備した RDF ストアに格納し、それを用いて検索する方法が考えられる。

本研究では最初に、Web 上で複数のデータセットのインスタンスを検索可能にしている LOD Cloud Cache を用いた集計を試みた。図 1 の SPARQL 式を用いて使用数の多いクラスとプロパティを取得し、プロパティの主語と目的語やそれらのクラスを集計することで複数のデータセット間で使用されているプロパティの傾向を分析できると考え、クラスやプロパティの情報の取得を行った。しかし、それらの

```
SELECT distinct ?cl ?p (COUNT(?s) as ?num)
WHERE{
        ?s ?p ?o;
        a ?cl.
        FILTER isURI(?o)
}GROUP BY ?cl ?p
ORDER BY DESC(?num)
```

図 1：使用数の多いクラスとプロパティの組を取得する SPARQL 式

表 1: 2014 年以降のデータセットにおける使用メタデータ語彙数

名前空間	ターム数	使用データセット数	合計使用回数
http://www.w3.org/1999/02/22-rdf-syntax-ns#	22	436	6,523,586
http://www.w3.org/2000/01/rdf-schema#	4	55	186,401
https://w3id.org/lio/	236	51	75,104
http://schema.org/	4	45	40,766
http://www.iptc.org/std/Iptc4xmpCore/1.0/xmlns/	3	23	31,204
http://www.w3.org/1999/U2/	1	23	4,899
http://purl.org/dc/elements/1.1/	4	22	16,119
http://www.w3.org/2002/07/owl#	2	16	480,892
http://xmlns.com/foaf/0.1/	13	13	80,914
http://purl.org/dc/terms/	9	10	1,628,281

表 2: 2011~2014 年のデータセットにおける使用メタデータ語彙数

名前空間	ターム数	使用データセット数	合計使用回数
http://www.w3.org/1999/02/22-rdf-syntax-ns#	6	10	391,343
http://xmlns.com/foaf/0.1/	8	10	39,696
http://www.w3.org/2002/07/owl#	5	8	37,673
http://www.w3.org/2004/02/skos/core#	12	8	530,577
http://rdfs.org/ns/void#	3	8	122,557
http://purl.org/dc/terms/	6	3	4,866
http://semanticscience.org/resource/	8	3	237,080
http://purl.org/linked-data/cube#	6	2	58
http://www.aktors.org/ontology/portal#	8	2	72
http://data.nobelprize.org/terms/	9	1	23,355

主語や目的語からなるすべてのトリプルを取得するとデータ量が非常に膨大となり、問い合わせがタイムアウトを起こしてしまうことが多い。そこで SPARQL 式の LIMIT 句を用いて取得する主語と目的語の数に制限をかけ、それらをサンプルとして分析を行おうと考えた。しかし、主語や目的語を問い合わせた場合、結果は主語、または目的語の URI 順に出力される。そのため、大規模なデータセットで対象とするプロパティが多く使用されていた場合、出力結果がすべてそのデータセットに含まれるリソースの URI となってしまうことも多く、サンプルが大規模なデータセット内のリソースに偏ってしまう。そのため、ランダムでサンプリングを行うことが望ましいが、Web 上で公開されている SPARQL エンドポイントでランダムサンプリングを行うのは難しい。

そのため、本研究では Web サイトから取得したダンプファイルをローカルに準備した RDF ストアに格納し、それらを用いて分析を行う。ローカルの RDF ストアならば、Web 上の SPARQL エンドポイントにアクセスするよりタイムアウトの可能性は低くなり、確実に SPARQL 式の実行を行うことが可能になる。

4.2 調査対象データセット

調査の対象となる LOD データセットは LODCloud[13]内から選出する。LODCloud に登録されているデータセットの名称、URL、ジャンル、アウトバウンドリンクの数を Web スクレイピングによって取得する。これらの情報は最新の LODCloud のほか、2014 年 8 月のものと 2011 年 9 月のものを取得する。そして LOD クラウドに登録されているデータセットを 2011 年以前から登録されていたもの、2011~2014 年の間に登録されたもの、2014 年以降に登録されたものの 3 つに大別し、年代ごとの変遷を分析できるようにする。LODCloud に登録されているもののうちすべてのデータセットがダンプファイルを提供しているわけではなく、すべてのデータセットのダンプファイルを取得するにはコストがかかるため、いくつかのデータセットを選出して分析を行う。選出の基準はアウトバウンドリンクの数で決定する。このアウトバウンドリンクの数が多い LOD データセットほど外部の LOD データセットを参照している数が多いため、年代・ジャンルごとにアウトバウンドリンク数上位 5 件のデータセットを確認(アウトバウンドリンク数が同一のものが見られる場合はすべて確認)し、それらのダンプファイルが取得可能なものを取得して、SPARQL エンドポイントにローカルに準備した RDF ストアに格納する。

4.3 調査結果

2011 年 9 月以前に登録されていたデータセット、2011 年 9 月～2014 年 8 月の間に登録されたデータセット、2014 年 8 月以降に登録されたデータセットの 3 つの年代についての調査結果を表 1、表 2、表 3

表 3:2011 年以前のデータセットにおける使用メタデータ語彙数

名前空間	ターム数	使用データセット数	合計使用回数
http://www.w3.org/1999/02/22-rdf-syntax-ns#	3	34	2,432.450
http://www.w3.org/2002/07/owl#	13	17	351,365
http://www.w3.org/2000/01/rdf-schema#	6	10	285,915
http://xmlns.com/foaf/0.1/	6	7	855,058
http://www.w3.org/2004/02/skos/core#	15	7	298,393
http://purl.org/dc/terms/	9	5	1,360,952
http://www.aktors.org/ontology/portal#	15	3	1,210
http://creativecommons.org/ns#	2	2	17,043
http://rdfs.org/ns/void#	6	2	19
http://www.w3.org/2007/05/powder-s#	9	1	515,616

に記す。さらにジャンルごとの調査結果の例として生命科学と言語のジャンルに属するデータセットの調査結果を表 4、5 に示す。表におけるターム数は対応する名前空間を URI に含んでいたプロパティの数、使用データセット数はそれらのプロパティを使用していたデータセットの数、合計使用回数はそれらのプロパティの使用数の合計をそれぞれ示す。それぞれのプロパティの名前空間は LOV に登録されていた語彙の名前空間を URI に含んでいた場合はそれを参照する。LOV に登録されていない語彙を使用していた場合は URI のローカル名を削除したものを名前空間とする。表 1、2、3 はそれぞれ使用データセット数が多い語彙の上位 10 件を記している。一方で表 4、5 は計測された語彙の数が少ないため、計測された語彙を使用データセット数順にすべて掲載する。なお、語彙を使用していたデータセット数は該当するプロパティの主語の URI からドメインを抜き出し、それらの数をカウントしている。

5. 考察と今後の課題

5.1 集計結果から見るプロパティの使用傾向

年代ごとに見ると 2011 年以前に登録されたデータセットで使用されているメタデータ語彙は全体的に合計使用回数が高く、上位 10 件が全て LOV に登録されているものであった。この理由として、2011

表 4: 生命科学分野のデータセットにおける使用メタデータ語彙数

名前空間	ターム数	使用データセット数	合計使用回数
http://www.w3.org/1999/02/22-rdf-syntax-ns#	3	28	799,854
http://www.w3.org/2004/02/skos/core#	3	7	715,051
http://xmlns.com/foaf/0.1/	4	7	834,384
http://www.w3.org/2000/01/rdf-schema#	6	7	412,132
http://rdfs.org/ns/void#	1	7	137,635
http://www.w3.org/2002/07/owl#	4	5	40,771
http://purl.org/dc/terms/	5	4	103,432
http://semanticscience.org/resource/	8	3	297,594
http://eagle-i.org/ont/repo/1.0/	2	2	47,840
http://purl.obolibrary.org/obo/	67	2	51,587
http://eagle-i.org/ont/app/1.0/	14	2	11,226
http://eunis.eea.europa.eu/rdf/	24	1	2,252,244
http://rs.tdwg.org/dwc/terms/	1	1	274,923
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/	29	1	142689
http://www.ebi.ac.uk/efo/swo/	2	1	67

年以前から現在まで提供を続けているデータセットは使用者が多いためであると考えられる。2011 年以前のデータセットはその多くが提供を停止しており、その中で現在まで提供を続けているデータセットは DBpedia[14]や Geonames[15]など使用者が多く規模の大きいものが多い。使用者の多いデータセットは、それだけ汎用性も高く、使用している語彙も知名度が高いものが多いと考えられる。

一方で、2011 年 9 月以降のデータセットでは LOV に登録されていない独自語彙がいくつか見られる。独自語彙は使用するデータセットが限られるため、使用データセット数で見た場合、それほど多くないと予想していた。しかし、表 1 の「https://w3id.org/lio/」など LOV に登録されていない名前空間を持つプロパティにおいても高い使用データセット数を記録す

表 5: 言語分野のデータセットにおける使用メタデータ語彙数

名前空間	ターム数	使用データセット数	合計使用回数
http://www.w3.org/2002/07/owl#	1	3	128,471
http://purl.org/dc/elements/1.1/	2	2	3,472
http://purl.org/dc/terms/	4	2	846,967
http://www.w3.org/2000/01/rdf-schema#	1	2	9,909
http://xmlns.com/foaf/0.1/	2	2	7,878
http://wordnet-rdf.princeton.edu/	26	1	654,992
http://www.w3.org/ns/lemon/	2	1	414,544
http://www.lemon-model.net/	1	1	67,564
http://mlode.nlp2rdf.org/sentiws/sentiws/vocab/	2	1	6,124
http://wiktionary.dbpedia.org/terms/	1	1	3,462

るものがいくつか見られる。また、LOD データセットにおいては名前空間を指定し、各リソースにその名前空間を持つ URI を割り当てる場合が多いため、ドメイン数とデータセット数にはそれほど差はないと予想し、本研究では主語の URI からドメインを抜き出し、それらの数をデータセット数としてカウントした。しかし、2014 年以降のデータセットは全体的に使用データセット数が高い傾向となった。これらの理由として、外部リソースの URI をそのまま主語としたトリプルを含むデータセットが多くなっていることが考えられる。データセットで指定されている名前空間を URI に含まないリソースが多いため、URI のドメインの数とデータセットの数に大きな差が出てしまった。そのため、使用データセット数の計算方法の見直しが必要となる。

ジャンルごとに語彙の使用傾向を見ると、クロスドメインと出版の分野では 30 以上のメタデータ語彙が見られた。一方で、地理情報、政治、生命科学、言語の分野では使用されているメタデータ語彙が 20 以下であり、限られたメタデータ語彙が多く使用される傾向にあると思われる。さらに生命科学では 6/15、言語では 4/10 と LOV に登録されていない独自語彙の比率が高く、これらの分野では知名度の高いメタデータ語彙では表現できないデータを多く取り扱っていると考えられる。全体としてはメタデータ語彙の中で使用されているタームの数や利用しているデータセットの数が合計使用回数と比例しないことがわかる。これは、規模の大きいデータセットがある語彙を多数使用しており、合計使用回数もそれによって大きく増加していることが原因であると考えられる。よって、プロパティの使用傾向を分析する際は単純に使用数が多いだけでなく、多くのデータセットで使用されているかどうかに焦点を当てて分析する必要があるとわかる。

本研究ではアウトバウンドリンク数に着目してデータセットを選出したが、この基準で選出した場合データの規模にばらつきが生じ、分野によってトリプル数に大きく差が出たほか、先述した規模の大きいデータセットの影響も多く見られた。そのため、より正確な分析を行うために、トリプル数が近いデータセットを複数選出して改めて分析を行う必要があると考えられる。

5.2 プロパティの集計を行う際の問題点

本研究では LODCloud に登録されているデータセットを用い、プロパティの使用傾向の分析を行った。しかし、分析を進めるにあたり、いくつかの課題があった。

まず第一に,多くの LOD データセットへのアクセスが困難なことが挙げられる。LOD データセットを利用する方法としては SPARQL エンドポイントを用いた問い合わせとダンプファイルのダウンロードの 2 つがある。しかし、この両方が提供されていないデータセットが非常に多く見られた。そのため、LODCloud に登録されていても、多くのデータセットにはすでにアクセスできないのが現状である。

次の問題として、SPARQL 式の実行時間が挙げられる。LOD データセットにアクセスする際、SPARQL エンドポイントを公開しているデータセットは多いが、本研究のような分析を行う際、データセット全体での頻出クラスやプロパティの問い合わせを行う必要がある。しかし、データセット内のすべてのクラスやプロパティを問い合わせるようなクエリは実行に時間がかかり、Web 上でアクセスする場合、タイムアウトを起こす可能性が非常に高い。そのため、本研究ではダンプファイルが提供されているものに絞り、それらをダウンロードして、分析に使用した。

```
SELECT distinct ?s ?o
WHERE{
        ?s <プロパティ> ?o.
         BIND (MD5(CONCAT(STR(?s),STR(RAND()))) AS ?random)
}ORDER BY ?random
LIMIT 10
```

図 2:ランダムサンプリングを行う SPARQL 式

しかし、ダンプファイルを用いる場合でも大規模なデータセットは容量が非常に大きく、ローカル上に用意した RDF ストアに格納できないケースが多い。また、アップロードをできた場合も SPARQL クエリの実行には非常に時間がかかる。そのため、大規模なデータセットの傾向分析を行うには大きなコストと大容量のレポジトリのほか、リレーショナルデータベースなどSPARQLエンドポイント以外の集計方法を考える必要がある。

5.3 今後の課題

本研究では、ジャンルごとのプロパティの使用数と年代ごとのプロパティの使用数の分析を行った。しかし、これらのプロパティがすべて異なるデータセットに含まれるリソース同士を繋いでいるわけではない。そのため、各プロパティの主語と目的語を集計し、それらの名前空間から異なるデータセットに含まれるリソース同士を繋ぐプロパティを分類すること、またそれがどのようなリソースを繋ぐために使用されているかを調査することが今後の課題である。

プロパティの主語と述語の集計を行う際、それらのインスタンスを取得して分析を行う必要がある。しかし、制限をかけずにすべてのインスタンスを取得した場合、データ量の大きい単一のデータセットに結果が左右されてしまうことが考えられる。そこで、インスタンスのランダムサンプリングを行うことが望ましい。図 2 の SPARQL 式を用いることであるプロパティの主語と述語のランダムサンプリングが可能である。この SPARQL 式ではリアルタイムで乱数を生成し、主語に付与している。この乱数を用いて並び替えを行うことで SPARQL 式を実行するたびに異なる結果を取得することが可能になる。しかし、すべての主語に乱数を付与するため、処理時間は長くなる。そのため、Web 上で公開されている SPARQL エンドポイントでこの式の実行は難しく、ローカル上の SPARQL エンドポイントでも長い処理時間がかかるのが課題である。

6. おわりに

本研究では関連するデータセット同士の関係性を提示することを目的に、複数の LOD データセットで使用されているプロパティに着目し、それらの使用傾向の分析を行った。プロパティの使用傾向から年代による独自語彙の使用の増加やデータセット内のリソースの URI の付与の仕方などについて変化が見られること、分野ごとに使用されるメタデータ語彙の数や独自語彙の比率などに違いが見られることがわかった。

一方で、この集計結果は少数のデータセットを選出して分析を行った結果であるため、より正確な結果を得るために、多数のデータセットを収集して分析を行う必要がある。

本研究ではプロパティの使用傾向の集計を行なったが、それらのプロパティが外部のリソース同士を繋いでいるかどうかといったことや、繋がれているリソースはどのようなものが多いかといったことは検証できていない。データセット同士の関係性を提示できるようにするためは、各プロパティの主語と述語がどのようなリソースか、それらがどのようなクラスなのかといったことを分析することが必要である。よって、今後の課題としてはそれぞれのプロパティの主語や目的語となっているリソースをサンプリングし、どのようなリソースが多いかを検証することが挙げられる。様々な規模のデータセットからより正確な傾向を掴むためにはインスタンスのランダムサンプリングを行うことが望ましく、それを実現できるような環境の構築が必要となる。

謝辞

本研究は JSPS 科研費 JP18K11984 の助成を受けたものです。

参考文献

[1] 国立国会図書館. “OCLC Research、世界の Linked Data プロジェクトの調査結果（2018 年版）を発表". カレントアウェアネス・ポータル. http://current.ndl.go.jp/node/36999, (参照 2019-02-27)

[2] the U.S. General Services Administration, Technology Transformation Service. Data.gov. https://www.data.gov, (参照 2019-02-27)

[3] European Union. European Data Portal. https://www.europeandataportal.eu/en/legal-notice, (参照 2019-02-27)

[4] Linked Open Data チャレンジ Japan 実行委員会. Linked Open Data Challenge 2018(LOD チャレンジ 2018). http://2018.lodc.jp, (参照 2019-02-27)

[5] 山中勇樹, 三原鉄也, 永森光晴, 杉本重雄. アプリケーション開発事例を用いた LOD データセットの探索支援.第 44 回セマンティックウェブとオントロジー研究会発表資料. 2018, 44(5).

[６] Ontology Engineering Group. Linked Open Vocabularies. https://lov.linkeddata.es/dataset/lov/, (参照 2019-02-27)
[７] Agile Knowledge Engineering and Semantic Web (AKSW). “LODStats”. Agile Knowledge Engineering and Semantic Web (AKSW). http://aksw.org/Projects/LODStats.html, (参照 2019-02-27)
[８] Ivan Ermilov, Jens Lehmann, Michael Martin, and Sören Auer. “LODStats: The Data Web Census Dataset”. Proceedings of 15th International Semantic Web Conference(ISWC2016), 2016.
[９] W3C. “SPARQL 1.1 Federated Query”. W3C Recommendation. https://www.w3.org/TR/2013/REC-sparql11-federated-query-20130321/, (参照 2019-02-27)
[１０] OpenLink Software. OpenLink Virtuoso. https://virtuoso.openlinksw.com, (参照 2019-02-27)
[１１] OpenLink Software. LOD Cloud Cache. http://lod.openlinksw.com/fct/, (参照 2019-02-27)
[１２] 株式会社富士通研究所. LOD4ALL. https://lod4all.net/ja/index.html, (参照 2019-02-27)
[１３] John P. McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, Vladimir Andryushechkin. The Linking Open Data Cloud. https://lod-cloud.net, (参照 2019-02-27)
[１４] DBpedia. https://wiki.dbpedia.org, (参照 2019-03-01)
[１５] Geonames. http://www.geonames.org, (参照 2019-03-01)