1N3-4

Linked Open Data を利用したメタデータ語彙の用法理解のための用例作成

Making Examples for Understanding Usage of Metadata Vocabulary using Linked Open Data

二十歩 亮介[*1] 本間 維[*1] 永森 光晴[*2 *3] 杉本 重雄[*2]
NIJUBU Ryosuke HONMA Tsunagu NAGAMORI Mitsuharu SUGIMOTO Shigeo

[*1] 筑波大学大学院図書館情報メディア研究科
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
Faculty of Library, Information and Media Science, University of Tsukuba.
[*3] 筑波大学知的コミュニティ基盤研究センター
Research Center for Knowledge Communities, University of Tsukuba.

Linked Open Data (LOD) has been accepted and the amount of data published as LOD is rapidly increasing. However, it is not always easy for dataset developers to appropriately use metadata vocabularies to link datasets because definitions of metadata vocabularies which help users understand semantics and use cases of metadata terms are generally not given appropriately. This paper proposes an approach to create usage examples of metadata terms using statements collected from LOD datasets. It shows an evaluation of the proposed method and discusses some lessons learned.

1. はじめに

近年，情報公開の必要性が国際的に高まっており，政府や学術機関を中心とした様々な組織やコミュニティによって作成されたデータセットが Web 上で公開されている[*1*2]．公開されるデータはオープンデータと呼ばれ，現在は Excel や CSV など多様な形式で公開されているが，再利用が容易である Linked Open Data（LOD）としての公開が望まれる．LOD は URI で表現された事物間を型付きリンクで結んだオープンデータである．

LOD 実現のための枠組みとしてしばしば Resource Description Framework（RDF）が利用される．RDF においてデータの記述にはメタデータ語彙という専用の語彙が使用される．データの記述者はメタデータ語彙を独自に定義することが可能だが，既存の語彙を使用することでオープンデータの相互運用を向上させることができる．既存の語彙を使用する場合，データの非互換をさけるために語彙の定義からその用法を理解し正確に使用する必要があるが，定義の内容やデータ記述者の知識・経験の不足から用法理解が困難なことがある．その際，語彙が実際に使用されている例を確認することで用法理解が容易になる．そこで本研究は，Web 上で公開されている既存の LOD を利用したメタデータ語彙の用法理解のための用例作成手法を提案する．なお，今回はメタデータ語彙のうち，プロパティを対象とした用例作成を扱う．

2. メタデータ語彙の用法とその理解

2.1 Resource Description Framework

RDF は Web 上のリソースに関するデータを機械判読可能かつ再利用が容易な形式で記述するための枠組みである[*3]．主語（記述対象）・述語（属性）・目的語（属性値）の組み合わせ（トリプル）でデータを記述する．RDF データはラベル付き有向グラフで表現することが可能であり，リソースは楕円，リテラル（文字列）は矩形のノードで表される．図 1 において破線で囲われたトリプルは「http://danbri.org/という URI で識別されるリソースのタイトルは”Dan’s home page”である」という情報を表す．このトリプルを組み合わせて RDF グラフを構成することによって，複雑なデータでも柔軟に記述することが可能となっている．

2.2 メタデータ語彙と語彙定義

メタデータ語彙とは RDF の記述に用いられるタームの集合である．タームはプロパティとクラスの総称であり，URI で表現される．代表的なメタデータ語彙として，Web 上のリソースに対して汎用的な属性を記述するための DCMI Metadata Terms[*5] や，人物や組織の活動や関わりを記述するための FOAF[*6]，場所や物体に対し位置情報を記述するための Basic Geo Vocabulary [*4] などがある．プロパティはデータの記述対象が持つ属性を表し，トリプルにおける述語の記述に使用される．クラスは記述対象自身の型を表し，トリプルにおいて「http://www.w3.org/1999/02/22-rdf-syntax-ns#type（rdf:type）」というプロパティの目的語として記述される．これらのプロパティとクラスを定義したものがメタデータ語彙定義である．プロパティの

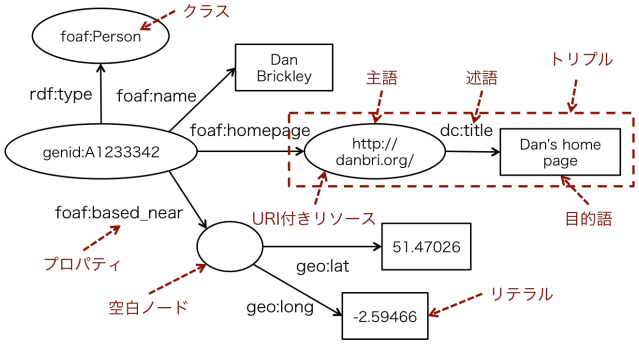


図 1 Basic Geo Vocabulary の用例[*4]

*1 https://datahub.io/
*2 http://dataforjapan.org/
*3 https://www.w3.org/RDF/
*4 https://www.w3.org/2003/01/geo/
*5 http://dublincore.org/documents/dcmi-terms/
*6 http://xmlns.com/foaf/spec/

連絡先：二十歩亮介，筑波大学大学院 図書館情報メディア研究科，〒305-8550 茨城県つくば市春日 1 丁目 2，s1521635@u.tsukuba.ac.jp

場合，記述する属性の用途，記述対象と属性値のクラスに対する制約などが主な定義内容である．

図 2 は Basic Geo Vocabulary のプロパティ geo:lat の定義である．この定義から，geo:lat がプロパティ（rdf:Property）であること，記述対象（rdfs:domain）が地理的物体（geo:SpatialThing）であること，人間が読むための名称が”latitude”であること（rdfs:label），世界測地系における緯度を十進表記で記述するための属性であること（rdfs:comment）が読み取れる．

既存の語彙や語彙定義の入手方法として，独自に語彙を収集・公開している Linked Open Vocabularies[Pierre-Yves 14]や，語彙定義をはじめとした各種スキーマを管理する MetaBridge[Nagamori 11]の利用などが考えられる．しかし，語彙や語彙定義を蓄積し，流通させる標準的な仕組みは確立されていない．

```
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
geo:lat
   rdf:type rdf:Property ;
   rdfs:domain geo:SpatialThing ;
   rdfs:label “latitude” ;
   rdfs:comment “The WGS84 latitude of a SpatialThing (decimal degrees).” .
```

図 2　geo:lat の定義

2.3 メタデータ語彙の用法とその理解

既存のメタデータ語彙を使用する場合，データの再利用時に混乱をきたさぬよう，語彙定義からその用法を理解し正確に使用する必要がある[メタ基盤 11] [秋山 16]．筆者らのメタデータ作成経験からメタデータ語彙（プロパティ）の用法を理解するために確認しなければならない事項は以下の通りである．

(a) 記述する属性の用途
　データ記述者が期待する用途に合うか確認する．
(b) 記述対象（ドメイン）と属性値（レンジ）のクラス
　まず属性値がリソースとリテラルのどちらであるか，さらに記述対象やリソースである属性値が人物か文書かといったクラスを確認する．
(c) 属性値の制約
　属性値がリテラルの場合は文字列か数値か日付かなど，リソースの場合は特定の統制語彙を使用するといった制約があるかを確認する．統制語彙とは意味や使用法などを限定することで曖昧さを排除した用語の集合のことで，例えば MIME はデータ形式を記述するための統制語彙であり，text/html や application/xml といった Media Type を定義している．
(d) 空白ノードを利用した構造化
　記述対象や属性値を匿名のリソースとし，まとめて記述する属性を確認する．図 1 では空白ノードを利用して緯度（geo:lat）と経度（geo:long）のデータをまとめている．
(e) 属性の繰り返し記述の方法
　同じ属性のデータを複数記述する場合の方法を確認する．記述の方法は大きく 2 つあり，1 つは同じプロパティを複数記述する方法，もう 1 つは空白ノードを利用しまとめる方法である．
(f) 共起関係
　用法確認の対象としているプロパティが使用された際に，高い頻度で共に使用されているプロパティがあるかを確認する．特定のプロパティの組が頻出することを共起するといい，本研究ではそのプロパティの組を共起関係にあるとする．

しかし，語彙定義が存在しない，あるいは発見できない場合，定義の内容が不足している場合，データ記述者の専門知識やメタデータ作成経験が不足している場合などに(a)~(f)の事項を十分に確認できないことがある．その際，語彙が実際に使用されている例を確認することで用法理解が容易になる．例えば，図 1 の W3C Semantic Web Interest Group[*7] が公開している Basic Geo Vocabulary の用例を確認することで，図 2 の geo:lat の定義からは読み取れない「位置を表すプロパティの先に空白ノードを記述し，そのノードを主語として geo:long と共に使用する」という用法を得ることができる．だがこうした用例が語彙定義と共に公開されているケースは多くないため，データ記述者は膨大な情報資源から目的の語彙の適切な用例を確認しなければならず，大きな負担となっている．そこで本研究は，Web 上で公開されている既存の LOD から，メタデータ語彙の用法理解のための用例を作成する手法を提案する．

3. LOD を利用したメタデータ語彙用例作成手法

本研究では，用例作成の方針として独自に作成した作例ではなく実例からの引用である引用例を採用する．実例から引用することである程度自然で典型的な文脈での用例を作成できる[千葉 12]．引用元は Web 上で LOD として公開されている RDF 形式のデータセットとする．データセットは事前に収集し，データベースに保存して検索可能としておく．また，収集したデータセットに含まれる全てのプロパティの組み合わせ（P_a, P_b）に対して共起関係にある確率（= P_a が P_b と共起するデータセット数 / P_a が使用されているデータセット数）を求めておく．ここで共起するとは，P_a が使用された場合に必ず同じ記述対象に対して P_b が使用されることを意味する．

今回作成する用例の要求定義を以下にまとめる．

要求1. (a)~(f)の事項が可能な限り確認できる
要求2. 自己完結的である
　引用部分のみで記述された情報が理解できる．
要求3. 必要最小限である
　当該プロパティの用法理解を容易にするため，必要以上の情報を含めない．

2.1 節で述べた通り RDF の記述の最小単位はトリプルであるが，実際の記述では複数のトリプルが組み合わさり RDF グラフをなしている．要求 3 を満たすため，RDF グラフ全体から要求 1, 2 を満たす必要最小限の部分 RDF グラフを引用することで用例を作成する．

手順1. まず，用例作成の対象となるプロパティ（以下，当該プロパティ）を述語としたトリプル，及び事項(f)の確認のためそのトリプルと主語を同じくし当該プロパティと共起関係にある確率が閾値を超えるプロパティ（以下，共起プロパティ）を述語としたトリプルをデータベースから取得する．SPARQL[*8] クエリにおけるグラフパターンを示す．

*7 https://www.w3.org/2001/sw/interest/
*8 https://www.w3.org/TR/sparql11-query/

{?subject <当該プロパティの URI> ?object_1}
UNION {?subject <共起プロパティの URI> ?object_2}

手順2. 次に事項(d), (e)の確認のため，手順 1 で取得したトリプルに空白ノードが含まれていた場合，その空白ノードを主語あるいは目的語としたトリプルを取得する．取得したトリプルに空白ノードが含まれていた場合，さらにその空白ノードを主語あるいは目的語としたトリプルを取得する．取得したトリプルに空白ノードが含まれなくなるまで手順 2 を繰り返す．匿名のリソースである空白ノードの情報を用例に加えることで要求 2 の自己完結性を保証する．

取得したトリプルの目的語が空白ノードの場合：
{<空白ノード ID> ?predicate ?object}
取得したトリプルの主語が空白ノードの場合：
{?subject ?predicate <空白ノード ID>}

手順3. 事項(b)の確認のため，手順 1, 2 で取得したトリプルに含まれるリソースを主語とし，rdf:type を述語としたトリプル（クラスの情報）を取得する．

{?subject rdf:type ?class}

手順 1~3 で取得したトリプルからなる RDF グラフを用例とする．事項(a), (c)の確認は作成した用例から行う．

図 3 は当該プロパティを foaf:based_near，共起プロパティを foaf:homepage として，図 1 の RDF グラフから引用した用例である．

4. 評価実験

4.1 LOD を利用した用例作成と結果

3 章で提案した用例作成手法を用いて，メタデータ語彙の用例を作成した．まず，引用元となる RDF 形式のデータセットを収集した．収集の対象としてデータセット間のリンク関係を表現した図である LOD cloud diagram[*9] に 2015 年 9 月時点で登録されているデータセットを採用した． LOD cloud diagram へのデータセット登録のためにはデータの量・質に対する厳しい条件があり，用例の引用元としてふさわしいと考えたためである．収集は手作業で行い，73 件のデータセットから約 1 億 4 千万件のトリプルを得た．データベースには代表的なトリプルストアの 1 つである Virtuoso[*10] を用いた．収集したデータセットには

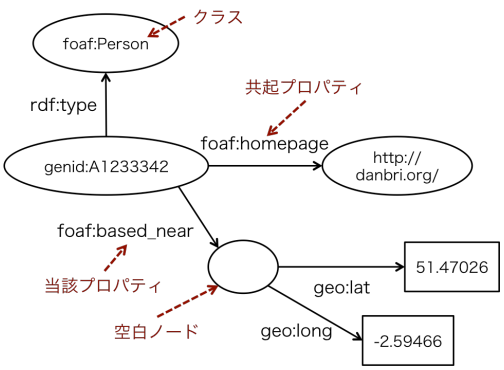


図 3　提案手法による foaf:based_near の用例

*9 http://lod-cloud.net/
*10 http://virtuoso.openlinksw.com/

約 2,300 種類のプロパティが含まれており，それぞれのプロパティについて用例を作成した．データセットや記述対象と属性値のクラスによって用法の違いがある可能性を考慮し，それぞれのプロパティについてデータセットごと，記述対象と属性値のクラスごとに 1 件の用例を手順 1~3 に従い作成した．結果，約 17,000 件の用例を得た．

作成した用例の一部を図 4 に示す．リソースにクラスが記述されている場合は URI の代わりにクラスを表記している．また URI 付きリソースと空白ノードを区別するため，空白ノードは輪郭が破線の楕円で表現している．図 4(2)のようなシンプルな用例から，(1)のような空白ノードや同じプロパティが繰り返し記述されていたり，(3)のような 1 つのリソースに複数のクラスが記述されている複雑な用例まで，様々な用例が見られた．

4.2 考察と課題

用例から事項(b)，(d)，(e)，(f)が確認できることがわかる．dcterms: references は参照や引用文といった属性を記述するためのプロパティであるが，(1)から「記述対象のクラスはデータセット（dcmit:dataset），属性値のクラスは文書（foaf:Document）であること」，「属性値を空白ノードとし，そのノードを主語として参照先の名称（rdfs:label）と URI（foaf:homepage）の情報をまとめて記述すること」，「複数の参照先を 1 つの記述対象に同じプロパティを繰り返し使用し記述すること」が読み取れる．(2)，(3)からは代替ラベル（skos:altLabel）とラベル（skos:prefLabel）が共起関係にあることが読み取れる．事項(a)，(c)については主に語彙定義から確認する事項だが，用例によってその理解を深めることができる．例えば，skos:altLabel の定義文は「An alternative lexical label for a resource.」だが，(2)や(3)から表記揺れや省略形の情報の記述に使用できることがわかる．

引用元となるデータセットによっては，(2)，(3)のように 1 つのプロパティに対して複数の用例が作成される．実際にデータ記述者が用例を確認することを考えると，用法理解により有用な用例を選出する必要がある．

有用な用例の判断基準の 1 つとして語彙定義との比較がある．主に定義されている記述対象・属性値のクラスと用例におけるクラスとの比較を行う．小学館の独和大辞典・第 2 版[*11] など，語の用法理解のためにあえて誤用例を掲載している辞典が存在する．語彙定義に従って記述されている用例や，典型的な誤用例などは有用であると考えられる．

用例の持つ情報量も判断基準の 1 つであろう．用例が複雑である程多くの情報を持つ．用例の複雑さについては，用例に含まれるクラスの件数，プロパティの件数，ノードの種類（URI 付きリソース，空白ノード，リテラル）などの要素を利用することが考えられる．

また，用例による事項(c)の確認に関して改善点が挙げられる．属性値の具体的な制約内容は用例には表されず，実際の値から推測する必要がありデータ記述者の負担となる．そこで，属性値の記述からデータ型や統制語彙などを機械的に推測し，用例と共に提示する必要があると考えられる．

5. 関連研究

外国人の日本語読解支援のために新聞や Web などから用例を抽出する研究[水野 07]では，用例の抽出基準として文長と難易度を採用している．文長についてはユーザが最大単語数と最小単語数を指定可能となっている．自然言語の文における文長は RDF データにおいてトリプル数に相当する．要求 3 にて

*11 http://www.shogakukan.co.jp/

用例は必要最小限としたが，今後はデータ記述者の指定した範囲での用例作成を行えるようにすることが考えられる．用例の難易度については 5 章で述べた複雑さと関わる基準であろう．日本国際教育支援協会[*12] によって定められた日本の単語，漢字，文法の級（難易度）のようなものがメタデータ語彙において作成可能であれば，より正確にメタデータ語彙の用例の難易度の評価が行える．

また，日本人英語学習者の英作文支援として用例検索に関する研究が行われている[綱嶋 07][高松 14]．用例の引用元（検索対象）を綱嶋らは Web 上のテキスト，高松らは ACL Anthology[*13] 上の主要会議の論文データとし，共に検索結果を分類してヒット数順に提示している．分類はクエリにワイルドカードが含まれていた場合にのみ行われ，ワイルドカード部分を埋めた単語やフレーズ，品詞などの基準で分類される．ヒット数が多い用例ほど妥当性が高いものとしている．メタデータ語彙の用例においても分類と頻度情報は有用であると考えられる．分類基準としては当該プロパティが述語として使用されているトリプルの記述対象と属性値のクラスやノードの種類，属性値に記述された値や制約，引用元となったデータセットのカテゴリなどがある．頻度情報を利用するためには，用例の引用元となるデータセットの収集基準を明確にしなければならない．収集基準作成の際にはデータセットやメタデータ語彙の統計情報を提供している LODStats[*14] などの利用が考えられる．

6. おわりに

本研究ではメタデータ語彙の用法理解のために，Web 上で公開されている既存の LOD からメタデータ語彙の用例を作成する手法を提案した．また提案手法を用いて，独自に収集した LOD から用例を作成した．

今後は作成した用例を評価し，語彙の用法理解を助ける情報と共に提示するシステムを構築する予定である．

謝辞

本研究は科研費（基盤研究 C，課題番号：15K00444）の助成を受けたものである．

参考文献

[Pierre-Yves 14] Pierre-Yves Vandenbussche , Ghislain A. Atemezing, María Poveda-Villalón, Bernard Vatant: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web , http://www.semantic-web-journal.net/system/files/swj1178.pdf, (2014).

[Nagamori 11] Mitsuharu Nagamori , Masahide Kanzaki , Naohisa Torigoshi , Shigeo Sugimoto: Meta-Bridge: A Development of Metadata Information Infrastructure in Japan, Proceedings of International Conference on Dublin Core and Metadata Applications 2011, pp. 63-68, (2011).

[メタ基盤 11] メタデータ情報基盤構築事業: メタデータ情報共有のためのガイドライン，http://meta-proj.jp/A03.pdf, (2011).

[秋山 16] 秋山 梓，加藤 文彦，小出 誠二，海沼 靖夫: 自治体が公開している RDF の現状と問題，人工知能学会研究会資料，http://id.nii.ac.jp/1004/00000791/, (2016).

[千葉 12] 千葉 庄寿: 大規模コーパスを用いた用例の典型性評価―大規模コーパスを利用した学習辞書作成のために―，コーパス日本語学ワークショップ 第 1 回予稿集，pp. 185-194, (2012).

[水野 08] 水野 淳太，大山 浩美，小林 朋幸，坂田 浩亮，Noah Evans，谷口 雄作，松本 裕治: 日本語読解支援のための語義ごとの用例抽出システムの構築，言語処理学会第 14 回年次大会併設ワークショップ，pp. 63-66, (2008).

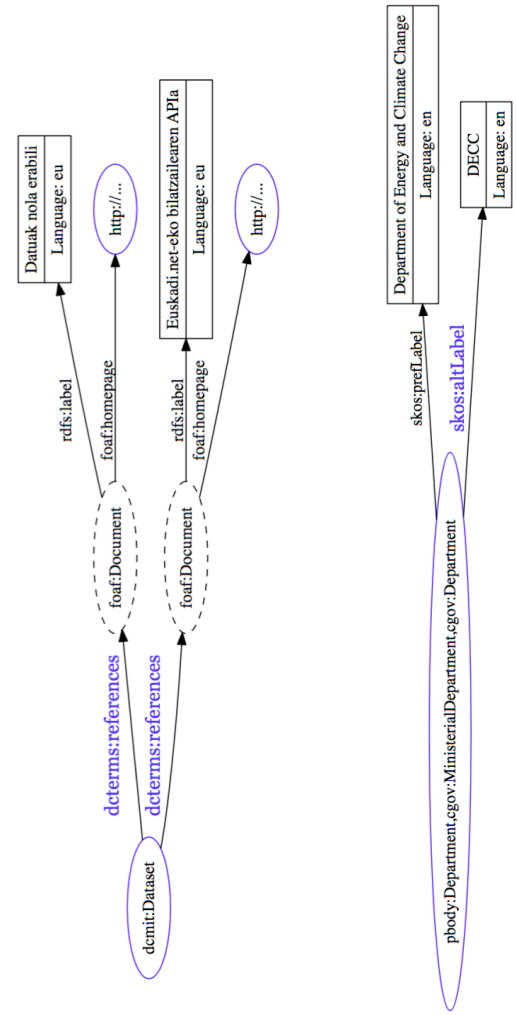
[綱嶋 07] 綱嶋 祐一，川崎 優太，安藤 一秋: 検索エンジンを用いた英作文支援ツール，電子情報通信学会技術研究報告，Vol. 106, No. 583, pp. 87-92, (2007).

[高松 14] 高松 優: 英作文支援のための用例検索に関する研究，東北大学大学院情報科学研究科修士論文，(2014).

*12 http://www.jees.or.jp/

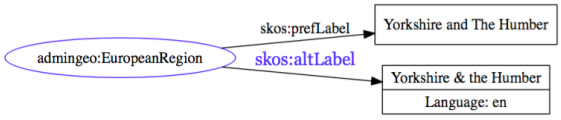
*13 https://aclweb.org/anthology/

*14 http://stats.lod2.eu/



(1) dcterms:references の用例

(3) 省略形の記述に用いた skos:altLabel の用例



(2) 表記揺れの記述に用いた skos:altLabel の用例

図 4 作成した用例の一部