様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２１　日現在

機関番号：　１２１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００４４４

研究課題名（和文）LinkedDataを基礎とした用例に基づくメタデータ語彙発見とスキーマ設計支援

研究課題名（英文）Finding Metadata Vocabularies and Support Designing Metadata Schemas by Existing Usage Examples in Linked Data Environment

研究代表者

永森　光晴（Nagamori, Mitsuharu）

筑波大学・図書館情報メディア系・講師

研究者番号：６０２７２２０９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：　本研究では，既存のLinked Data(LOD)を利用してメタデータスキーマに関する知識を補い，メタデータ語彙の発見とメタデータスキーマの設計支援環境の構築をした．本研究では，(1)Linked Dataを基礎としたスキーマ設計のためのメタデータの用例集の作成，(2)メタデータの用例集とメタデータスキーマレジストリを基礎としたメタデータ語彙発見モデルの提案の2つの目標を設けて研究を進めた．
　研究成果として，LOD データセットを基礎とした用例作成手法の提案と，用例の有用性評価のための基礎的知見を得ることができた．また，LOD の構造理解を目的としたドメインモデル推定手法の提案をおこなった．

研究成果の概要（英文）：　Metadata vocabularies define a set of terms and are widely used to describe LOD. However, the definitions of terms in metadata vocabulary lack specific examples that practically show how to use the term. Therefore, it is not easy to interpret and select the terms in metadata vocabularies. This paper proposes an approach of creating specific usage examples of metadata vocabularies using existing LOD datasets. The results of analysis on our created usage examples show that it is necessary to evaluate them and provide useful information for understanding the usage of metadata vocabularies, e.g., frequency of appearance of the term, category of the term.
　We also considered difficulties to understand metadata schema as cause that datasets are not utilized . In other to solve it we developed a method for estimating a domain model from metadata instances. Then we evaluated our method to compare correct domain model generated manually with one generated by our method.

研究分野：　ディジタルライブラリ

キーワード：　メタデータ　メタデータスキーマ　Semantic Web　Linked Data　Linked Open Data

１．研究開始当初の背景

メタデータは「データに関するデータ」と定義され，情報資源の組織化や検索のためだけではなく，保存や流通のための重要な役割を担うようになって来た. 現在では，学会，博物館，自治体といった様々なコミュニティが目的に合わせたメタデータを作成することも珍しくない．メタデータの記述方法の定義をメタデータスキーマと呼ぶ. メタデータスキーマではメタデータの記述に用いる属性語彙（例えば，タイトル，作者，更新日など），属性値語彙（例えば，件名標目や分類番号など）の定義をおこなう. 2001 年頃から始まったセマンティック Web の活動によってメタデータの利用はさらに広がっている．これまでのセマンティック Web では，主にメタデータスキーマ定義やメタデータの記述に用いるための語彙（例えば，RDF Schema, OWL(Web Ontology Language), SKOS(Simple Knowledge Organization System)など）の標準化がおこなわれてきた．さらに 2007 年頃からメタデータの利用に焦点を当てた W3C による Linking Open Data (LOD)プロジェクトが開始された. LOD プロジェクトはネットワーク上の RDF で書かれたメタデータを結びつけ，メタデータの相互利用性や流通性を高めることを目的としている．この結びつけられたデータセットを「Linked Data（リンクするデータ）」と呼ぶ．Linked Data として公開されるデータセットの数が膨大となるにつれ，メタデータの記述に使われるメタデータ語彙は多様なものとなり，その数も増加の一途をたどっている.

２．研究の目的

新たにメタデータを作成する場合，作成者はメタデータの記述対象である情報資源の特徴を分析し，利用目的に合わせてメタデータスキーマの設計をおこなう．コミュニティの目的に特化したメタデータ記述項目の場合は独自のメタデータ語彙を定義するが，メタデータの相互利用性や流通性を高めるためには，Dublin Core のような広く使われている既存のメタデータ語彙を再利用することが望ましい．メタデータスキーマ設計のコストを削減し，メタデータの相互利用性を高めるためには，目的にあった適切なメタデータ語彙の発見や類似の目的を持つ先行事例を知ることが重要である. しかしながら，多くの場合メタデータスキーマ設計者はコミュニティの扱う領域の専門家であって，メタデータ語彙やメタデータスキーマについての専門知識や設計に関する経験を持っているとは限らない．そのため現在のメタデータスキーマ設計は非常にコストのかかる作業となっている．筆者はこれまで，Dublin Core Metadata Initiative（DCMI）において，様々な種類のメタデータスキーマを登録し，人やソフトウェアに対してメタデータスキーマに関する情報を提供することを目的としたメタデータスキーマレジストリの開発や，総務省「メタデータ情報基盤構築事業」において「メタデータ情報共有のためのガイドライン」作成とメタデータスキーマレジストリ「MetaBridge（http://metabridge.jp/）」の開発に携わって来た. そこで本研究では，既存の Linked Data を利用してメタデータスキーマに関する知識を補い，メタデータ語彙の発見とメタデータスキーマの設計支援環境の構築をおこなう．本研究では，以下の 2 つの目標を設けて研究を進めた.

（1）Linked Data を基礎としたスキーマ設計のためのメタデータの用例集の作成：メタデータを記述するために用いるメタデータ語彙の定義が詳細に記述されていることは多くない．例えば，FOAF というメタデータ語彙の name 属性は「A name for something」と簡潔に定義されているのみである．そのため，メタデータ語彙に関する知識やメタデータスキーマ設計の経験をあまり持たない設計者は, 語彙定義だけでは,目的とする情報資源の特性を適切に表すプロパティやクラスであるかどうかや,どのようにメタデータを構造化すればよいのかを判断することが難しい．そこで本研究では，一般的な辞書において言葉の定義に加えて語義ごとの用例が記載されていることに発想を得て，ネットワーク上に公開されている Linked Data から，様々なメタデータ語彙がどのような領域の情報資源について，どのような型の値を記述するために用いられているかを表した用例集の構築をおこなう. また,実際のメタデータがどのように構造化されているのかを用例として抽出する.

（2）メタデータの用例集とメタデータスキーマレジストリを基礎としたメタデータ語彙発見モデルの提案: (1)において作成した用例集と筆者がこれまでに作成したメタデータスキーマレジストリを基に，メタデータ語彙に関する知識やメタデータスキーマ設計の経験をあまり持たないコミュニティが目的に合ったメタデータ語彙を効率良く発見し，メタデータスキーマを設計するためのモデルの提案をおこなう．本モデルでは，コミュニティが持つメタデータとして記述したい値（例えば「映画のタイトルとして『Harry Potter』や『Star Wars』を記述したい」という要求）に着目し，筆者がこれまでに作成したメタデータスキーマレジストリに蓄積するメタデータスキーマに加えて,用例に基づいたメタデータ語彙の推薦をおこないながらメタデータスキーマを設計する.

３．研究の方法

まず平成 27 年度において Linked Data

として公開されているメタデータの収集・蓄積をおこない，実際に記述されたメタデータの中でのメタデータ語彙の利用状況と，記述されている値およびメタデータの構造についての分析調査をおこなった．そして，適切なメタデータ語彙を選択するために必要な用例の要求要件を検討した．その後，用例集のプロトタイプシステムを作成し，その評価をおこなった．平成 28 年度では，平成 27 年度での評価に基づき，用例集の改善と用例集を利用したメタデータスキーマ設計支援モデルの提案とそのシステムの実装をおこない，限定した利用者にシステムを公開した．平成 29 年度では，前年度までの評価に基づいて，用例集とシステムの改善をおこなった．

４．研究成果

(1) Linked Open Data を利用したメータデータ語彙の用法理解のための用例作成手法の提案

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した．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った．

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．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．

(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．

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した．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た．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した．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（Pa，Pb）に対して共起関係にある確率（= Pa が Pb と共起するデータセット数 / Pa 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，Pa が使用された場合に必ず同じ記述対象に対して Pb が使用されることを意味する．用例は次の手順によって生成した．

手順 1：まず，用例作成の対象となる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．下記は SPARQL クエリにおけるグラフパターンを示す．

```
{?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用例の自己完結性を保証する．

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．

手順 1 から 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．

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．また，利用者のメタデータ語彙の用法理解のためには，用例の有用度を判断するための尺度が求められることがわかった．

（2）SPARQL Endpoint を利用したメタデータインスタンスに基づくドメインモデル推定手法の提案

本研究ではメタデータスキーマ情報の一つであるドメインモデルをメタデーインスタンス（LOD の実データ）から推定をおこなった．ドメインモデルとはメタデータ記述対象とその記述対象間の関係を表すものである．ドメインモデルはデータセット利活用の初期段階に必要な，メタデータスキーマのおおまかな理解に適している．

データセットの利活用にはそのメタデータスキーマを理解する必要がある．一方，現状では多くの LOD データセットが十分なメタデータスキーマ情報を公開していない．そのためデータセット利用者はインスタンスからメタデータスキーマを理解しなければならない．どこまで詳細にメタデータスキーマを理解するのかは利用者の目的に応じて変化するが，まずそのデータセットの主要な記述対象（事物）と大まかな構造が理解できることが望まれる．

LOD 利用者が未知のメタデータスキーマを把握する方法として，メタデータインスタンスのテキストデータを直接参照しながら,SPARQL クエリのやり取りを行う方法がある．この方法は専門性の必要な作業であり，体系化されていない．データセットの規模や内容に依存するが，非常に手間のかかる作業である．

本研究ではこのインスタンスからスキーマを次の手順で推定した．まずデータセット内で出てくる事物を把握するために，全クラスの列挙する SPARQL クエリ処理を行った．一般的に LOD データセットでは事物にクラスをつける．したがってこのクエリで事物を確認することができる．次に各事物の属性を知るために，各クラスに所属するリソース集合ごとのプロパティを取得した．属性の数が多いクラスほどそのデータセット内で主要な役割を担っていることが予想できる．そして，主要な事物間の関係を探るために列挙されたプロパティの中からサンプリングいくつかのレンジを探索してクラスとクラスをつなげるプロパティを発見し，データセット内の事物間の関係を把握した．

本研究ではメタデータインスタンスからドメインモデルを推定するための手法を提案し，検証実験を行った．実験結果から適切なドメインモデル推定には十分な量のインスタンスが必要なことがわかった．

（3）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手法の提案

ウェブ上で分野横断的にデータを利用するための仕組みとして Linked Open Data(LOD)が注目されている．LOD として公開されたデータセットはアプリケーション開発など様々な方法で活用されており，その際利用者が独自に用意したデータに加えて Web 上で公開されている既存の LOD データセットを合わせて利用することで開発の効率を上げることができる．しかし，Web 上で公開されているデータセットの数は膨大であり，その用途は多岐に渡る．

LOD データセットの探索に関する研究はこれまでにもおこなわれているが，データセットがアプリケーションにおいてどのように利活用されているのかについては注目されていない．そこで本研究では，既存のアプリケーションの開発に利用されているデータセットならば，利活用しやすく，共通した情報の項目が多いと考えた．また LOD データセットの情報と開発事例の閲覧が可能になれば，より目的に適した LOD データセットの発見が容易になると考え，それらの情報の検索・閲覧が可能なシステムを開発した．LOD データセットを実際に用いて開発されたアプリケーションの事例に基づいて関連するデータセットや利用方法の指針を示すことでデータセット探索の効率化や活用の幅を広げることが可能になる．検索システムと既存の検索手法の比較実験を行い，開発事例の情報によって LOD データセットの探索を効率化できることを示した．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔学会発表〕（計 4 件）

[1] 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,2017. http://www.sigswo.org/papers/44program

[2] Ryouta Kinjou, Mitsuharu Nagamori, Shigeo Sugimoto. Estimating Domain Models from Metadata Instances to Improve

Usability of LOD Datasets. International Conference on Dublin Core and Metadata Applications 2017, 2017. http://dcpapers.dublincore.org/pubs/article/view/3859
[3] 金城良大，永森光晴，三原鉄也，杉本重雄. SPARQL Endpoint を利用したメタデータインスタンスに基づくドメインモデル推定.第 41 回セマンティックウェブとオントロジー研究会,2017. http://www.sigswo.org/papers/41program
[4] 二十歩亮介，永森光晴，本間維，杉本重雄. 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.2016 年度人工知能学会全国大会(第 30 回), 2016.

６．研究組織

(1)研究代表者

永森　光晴（NAGAMORI, Mitsuharu）
筑波大学・図書館情報メディア系・講師
研究者番号：60272209