1C-03

オープンデータのLOD化におけるボキャブラリ選定支援

鵜飼 孝典 †　　山根 昇平 †

† 富士通研究所

1 はじめに

2013年のG8サミットにおいて「オープンデータ憲章」が合意されるなど，オープンデータへの注目が急速に高まっている．これらデータの多くはPDFやスプレッドシート，CSVなどの形式で公開されているのが現状である．このようなデータは，人間が読んで理解したり，特定のデータに特化したアプリケーションの開発を行う場合にはそのまま利用できる．しかし，他の様々なデータと組み合わせて機械処理するためには，RDF形式を用いたLinked Open Data(LOD)とすることが好ましい[1].

本研究では，オープンデータのLOD化に置けるRDF形式への変換に際し，述語に統一された語彙を用いるように，既存データから統計的に適切な語彙を提示する技術を開発した．

開発した技術をサンプルデータで試行したところ，オープンデータのLOD化が支援できそうであることはわかった．一方，これらの技術に100%の精度を期待するのは現実的ではなく，完全な自動化は困難である．そこで，専門家の知識を効率的に取り入れる仕組みやツールの開発が，今後の課題であることもわかった．

2 ボキャブラリ統一技術

2.1 目的と課題

Linked Data化の支援ツールとして，Open Refine[2]などを用いることで，CSVやスプレッドシートをRDFに変換することができる．ところが，適切なボキャブラリを選択することは困難であることが課題の一つになっている[3]．ボキャブラリの選択においては，業界で標準的に使われているボキャブラリを使わなければ，再利用が難しくなる[4],[5]．prefix.cc[6]に2014年10月28日現在で1430種類のボキャブラリが登録されていて，これらを適切に活用することが必要である．ところが，この中から適切なボキャブラリを探し出すことは大きな労力を伴う．

本研究では少ない組み合わせでできるだけすべての目的語をカバーするボキャブラリの組み合わせを探して，提示することを目的とする．

2.2 ボキャブラリの統一方法

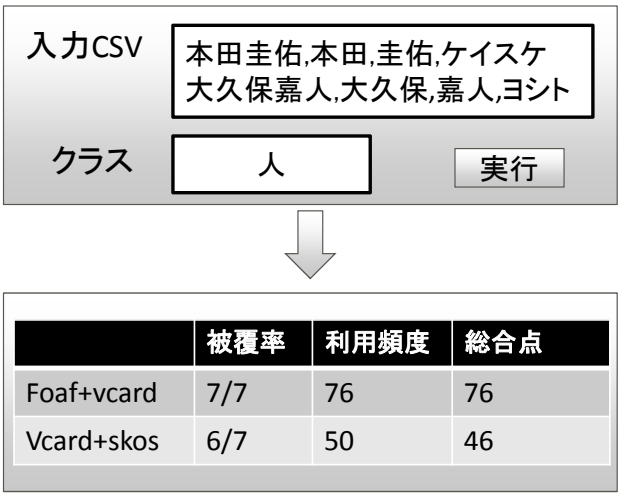
本論文では，図1ようなデータをCSV形式にしたファイルを入力とし，一意に割り当てたID(URL)を主語とし，第1列目以降の値を目的語にするRDF形式のデータを作成するために既存のLODで広く使われているボキャブラリを自動的に見つけることを目的とする．開発した技術のアルゴリズムは以下の通りである．

1. CSV形式のデータと1列目の値に対するクラスを入力とし，このデータのクラスとする
2. 既存のLODから1で得られた主語のクラスが用いているボキャブラリを得る．
3. 2で得られたボキャブラリから，各列の値に対して述語(属性)としてつけられるボキャブラリを得る．
4. 得られたボキャブラリ毎に目的語集合に対する被覆率と使用頻度から最も適切なボキャブラリを提示する
5. すべての行について，ボキャブラリの組み合わせを得て，適切なボキャブラリの組み合わせを提示する

ボキャブラリ提示の動作例を，図1に示す．まず，入力として，CSVデータ及び1列目のクラスである「人 `foaf:person`」を得る．次に，`foaf:person`で使われるボキャブラリの被覆率と利用頻度を計算し，被覆率×利用頻度により総合点を算出する．図1の例では，foafとvcardを組み合わせて使う場合が76点，vcardとskosを組み合わせて使う場合が46点となり，最終的にfoafとvcardの組み合わせが提示される．

図:1に示した例「本田圭佑, 本田, 圭祐, ケイスケ, 大阪府, 1986-06-13, `http://server/KeisukeHonda.jpg`, 男」では，本田圭佑は，`rdfs:label`, `vcard:fn`, `prop-ja:goals`などにマッチし，本田は，`prop-ja:wp`, `rdfs:label`などにマッチする．ケイスケは，`prop-ja:`

Vocabraly selection support making Linked Open Data
†Takanori Ugai　†Shohei Yamane
†Fujitsu Laboratories Limited



	被覆率	利用頻度	総合点
Foaf+vcard	7/7	76	76
Vcard+skos	6/7	50	46

図 1　ボキャブラリの提示.

愛称, foaf:nic, vcard:nicname などにマッチする. 大阪府は, rdfs:label, prop-ja:birthPlace, prop-ja:born, prop-ja:cities などにマッチする. 男は, foaf:gender, prop-ja:性別 などとマッチする.

一致するデータの量から foaf と vcard の組み合わせが推薦される.

2.3　実験と考察

独自に収集した人物データ 100 人分について, 実日本語版 Linked Data クラウド図 [7] に掲載されている LOD を用いて変換すると, http://ja.dbpedia.org/ の prop-ja が推薦されることが多い. これは比較可能なデータの量が多いためである. vcard:gender の範囲が URI で Female, Male のみが定義されているため,「男」が vcard:gender にマッチしない. 共通ボキャブラリとして, vcard に統一して, 再利用性を高めることを考えると, ボキャブラリの定義が多国語に対応することが望ましいと考えられる. また, 実験では人物データを用いたため, 主語のクラスは foaf:person を用いたが, 主語のクラスの候補は, rdf:type などで上記データに 176 種類ある. ここから適切なものを見つけることも必ずしも容易ではない. 本研究では, 既存の Linked Data から頻出のボキャブラリに合わせるために統計的に適当なボキャブラリを推薦する技術を示した. 一方で、各領域でボキャブラリの標準化がすすめられている [8],[4]. LOD への変換対象となるデータが, 標準化がおこなわれた領域のものであれば, 広く使われていない場合でも標準に合わせることが望ましいとも考えられる. 利用者向けのツールとしては, 標準仕様に合わせるか, その時点でのデファクトスタンダードに合わせるか, 選択できるようにすることが望ましいと考えている.

3　おわりに

本論文では, CSV やスプレッドシートで公開されている既存のオープンデータの Linked Data への変換を支援する技術として, 特に他の Linked Data と組み合わせて利用することを考慮し, ボキャブラリの統一を自動化するために, 既存の Linked Data から統計的に適当なボキャブラリを推薦する技術を示した.

今後の課題として, より精緻な定量的な評価が必要であると考えている. そのうえで精度を高めることが挙げられる. また本研究では, 主語のクラスを入力としたが, 主語の入力を不要にしたり, 候補を示す技術が必要であると考える. 一方, ボキャブラリの統一には, データの示す意味を解釈する必要があるため, 最終的な判断は人間が行わなければならない. そこで, ボキャブラリの選定に対して, 効率良く人間の知識を取り込むためのツールや仕組みの開発が必要であると考えている. 本論では用いるべきボキャブラリを名前空間として示しているに留まっているが, マッチングの結果を用いて, 各項目について用いる property の候補を示すことも考えられる.

参考文献

[1] Tim Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html.

[2] Openrefine. http://openrefine.org/.

[3] 加藤文彦. Linked Data 作成支援ツールの現状と課題. 第 24 回セマンティックウェブとオントロジー研究会, No. 03, pp. 1–4. 人工知能学会, 2011.

[4] 林良彦, 楢和千春, Thierry Declerck, Paul Buitelaar, Monica Monachini. 言語サービスオントロジーの開発における国際標準案の適用. 言語処理学会第 14 回年次大会発表論文集, pp. 540–543, 2008.

[5] 伊藤英毅. オントロジーを利用した知識の共有/再利用. *UNISYS Technology Review*, No. 64, 2000.

[6] namespace lookup for rdf developers. http://prefix.cc/.

[7] Linked Open Data Initiative. 日本語版 linked data クラウド図. http://linkedopendata.jp/?p=411 (アクセス:2014 年 10 月 30 日).

[8] 大江和彦. 病名用語の標準化と臨床医学オントロジーの開発. 情報管理, Vol. 52, No. 12, pp. 701–709, 2010.