F-034

統計データの RDF 化のためのテンプレート

Template for Converting Statistical Data to RDF

浅野 優[†1] 岩山 真[†1] 武田 英明[†2†3] 小出誠二[†4] 加藤 文彦[†4] 小林巌生[†5]
Yu Asano Makoto Iwayama Hideaki Takeda Seiji Koide Fumihiro Kato Iwao Kobayashi

1． まえがき

近年，各国政府は，オープンガバメントの 1 つの施策として，所有データを機械処理可能な形式で公開するオープンデータ化を進めている．国内では，経済産業省が統計データの Linked Open Data（LOD）化を進めており[武田 2013]，統計データを Resource Description Framework（RDF）で公開し，その検索が可能な SPARQL Endpoint を提供している [1]．LOD 化では，まず，表計算ソフト等で作成された表データを RDF に変換する．

本研究では，任意の表データを RDF に変換するための統一的なテンプレートを提案する．

2．RDF データキューブ語彙を用いた RDF 化

2.1 RDF データキューブ語彙

RDF データキューブ語彙[Cyganiak 2013]は，表データの絞り込み，集計，統合を機械処理することを目的とした，表データの RDF 化に用いる語彙であり，2013 年 6 月現在，World Wide Web Consortium の Working Draft として提案されている． RDF 化を行うためには，各表に対し，以下の 3 つのコンポーネントを定義し，それらを用いてデータを記述する必要がある．

（1）次元とその値：観測値の集合を同定するもの
（2）測度：観測された現象
（3）属性：測度の単位

都道府県別の人口を示す表 1 と，人口と面積を示す表 2 を例に，表頭・表側とコンポーネントの対応を示す．人口は総務省「国政調査」，面積は国土交通省「全国都道府県面積調」から引用した．各表の表頭・表側の内，下線があるものが測度であり，それ以外は次元の値である．

表 1 都道府県別の人口

	2010	2005
埼玉県	7,194,556	7,054,243
千葉県	6,216,289	6,056,462
東京都	13,159,388	12,576,601
神奈川県	9,048,331	8,791,597

表 2 都道府県別の人口と面積

		人口（人）		面積（Km²）
		2010	2005	2010
関東	埼玉県	7,194,556	7,054,243	3,767.92
	千葉県	6,216,289	6,056,462	5,081.91
	東京都	13,159,388	12,576,601	2,102.95
	神奈川県	9,048,331	8,791,597	2,415.86

[†1]（株）日立製作所 中央研究所 [†2] 国立情報学研究所
[†3] 総合研究大学院大学 [†4] 情報・システム研究機構
[†5] Open Community Data Initiative

[1] http://datameti.go.jp/sparql

例えば，表 1 の「2010」は次元「年」の値である．表頭・表側以外の各セル内の数値が観測値である．表にはコンポーネントに対応する表現が省略されている場合もある．例えば，表 1 では，測度「人口」と，各都道府県名を値とする次元「都道府県」と，2010 と 2005 を値とする次元「年」が省略されている．人は，タイトルや表の内容から省略されている次元や測度を推測できるが，機械処理のためには，これらを明示的に記述しておかねばならない．人口の属性は「人」，面積の属性は「Km²」である．表の種別は 2 種類あり，表 1 のように「人口」のみを測度とする単一測度と，表 2 の「人口」と「面積」のように 2 つ以上の測度を持つ複数測度がある．

2.2 表データの RDF 化

データキューブ語彙を用いた表データの RDF 化には，コンポーネントの定義とインスタンスの記述の 2 つが必要となる．前者では，表データの次元，測度，属性を定義する．後者では，セル毎に，次元とその値と，測度と測定値を記述する．例えば，表 2 の埼玉県の 2010 年の人口を示すセル（表 2 の 3 列 3 行目のセル）に関する記述には，図 1 の各矢印に付与された Uniform Resource Identifier（URI）と矢印の先の値を記す．ここでは，LOD 化のために，各コンポーネントを URI で示す．

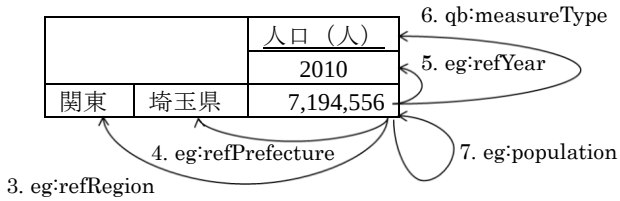

図 1 セルに関する記述内容（埼玉県の 2010 年の人口）

図 1 の RDF 記述を図 2 に示す．表 2 自体の URI を eg:dataset-02，対象セルの URI を eg:dataset-02-000001 とする．1～2 行目は，対象セルが観測値であり，表 2 に含まれていることを記している．3～7 行目は，対象セルが主語，図 1 で矢印に付与された URI が述語，矢印の先が目的語となるトリプルが記述されている．目的語の内，次元の値は URI 化している．図 1 の各リンクに付与された番号は，図 2 の各行の番号に対応している．また，eg と qb は既定義の名前空間である．

```
1: eg:dataset-02-000001 a qb:Observation ;
2:   qb:dataset eg:dataset-02 ;
3:   eg:refRegion eg:region-03 ;
4:   eg:refPrefecture eg:prefecture-11 ;
5:   eg:refYear eg:year-2010 ;
6:   qb:measureType eg:population ;
7:   eg:population 7194556 .
```

図 2 RDF での表現（埼玉県の 2010 年の人口）

3．提案

統計表は数万個の観測値から成る場合もあるため，表データをデータキューブ語彙を用いたインスタンス記述に自動変換するツールが欠かせない．その実現には，各表に対して，次元と測度を明示する必要がある．表を RDF に変換する既存ツールには，RDF Refine[DERI]等が挙げられるが，データキューブ語彙を用いた RDF 化を対象にはしていない．

本研究では，表データの次元と測度を明示するためのテンプレートを提案する．元の表データは人にとって見やすく，編集しやすいため，提案テンプレートでは，元の表構造を可能な限り維持することとする．また，LOD 化を考慮し，次元，次元の値，測度を URI で表すこととする．

3.1 テンプレートを用いた表の作成と URI 入力

元の表から以下の順で提案テンプレートを用いた表に変換する．

【手順①】各表頭の下に 1 行，各表側の右に 1 列追加
【手順②】表頭部と測定値の間に 1 行追加し，表側部と測定値の間に 1 列追加

図 3 は，表 2 に対して，上記の手順で作成し，必要な値を入力した表である．図 3 内の【手順①】と【手順②】は，上記の【手順①】と【手順②】に対応しており，矢印の先が追加された行／列である．図 3 の内，色の付いたセルに URI を入力し，他のセルには何も入力せず，空白とする．

【手順①】で追加された行／列には，上／左の各表頭・表側に対応する次元の値の URI と測度の URI を記述する．図 3 の表頭 1 行目は人口と面積の測度であるため，追加された 2 行目には，人口と面積の URI を入力する．

【手順②】で追加された行／列の各セルには，対応する列／行が次元の値の URI の並びである場合にのみ，次元の URI を入力する．図 3 の 4 行目には，「年」の次元の URI を入力し，2 行目は，測度の URI の並びであるため，空白とする．また，追加された行列の交点（図 3 内の※）には，単一測度の場合の測度の URI を入力する．図 3 の場合は複数測度であるため，※のセルは空白とする．

上記のように提案テンプレートを用いた表に入力を行うことで，表の次元と測度を明示することができる．提案テンプレートを用いた表に入力を行うことで，RDF 化に必要な URI 情報が揃うため，RDF に容易に自動変換できる．データを入力した表と，表の URI と，各セルの URI 基底部を与えると，各セルの URI は自動生成され，例えば，表 2 の 2010 年の埼玉県の人口を示すセルは，図 2 に示すように RDF 化される．

3.2 効果

本提案のテンプレートは， 次の 4 つの特徴および効果を持つ．

表頭・表側と次元・測度の対応を明記：
RDF への変換に必要な情報を元の表に埋め込むことができる．元の表にはない，単一測度の場合の測度の URI を入力する箇所も備えているため，単一測度と複数測度の両方の表に対応している．

元々の表構造を維持：
追加した行／列を隠せば，従来通り，元の表の上でデータの閲覧や編集を行うことができる．

表の次元，次元の値，測度と URI の対応を明記：
対応するもの同士が隣り合っているため，URI の入力忘れや入力誤りが防止できる．また，必要な URI が一覧できるため，URI の一貫性も保ちやすい．

表も次元・測度も単一ファイルに保存：
表の変更を行う際，必要な情報を収集する作業が不要になる．

4. まとめ

本研究では，統計表のような様々な形式の表データを，RDF データキューブ語彙に基づく RDF に変換するための統一的なテンプレートと，テンプレートを用いて RDF 化に必要な情報を入力した表を RDF に自動変換するツールを開発した．今後の展開として，本提案のテンプレートを用いたエディタの開発を考えている．

参考文献

[武田 2013] 武田英明ら, 統計データの LOD 化とデータ間の関係の表現，第 27 回人工知能学会全国大会, 2013.

[Cyganiak 2013] R. Cyganiak and D. Reynolds (Eds.), The RDF Data Cube Vocabulary, W3C Working Draft 12 March 2013, W3C, 2013, http://www.w3.org/TR/2013/WD-vocab-data-cube-20130312/, 参照 2013/06/05.

[DERI] Digital Enterprise Research Institute, RDF Refine, http://refine.deri.ie/, 参照 2013/06/05.

					人口（人）		面積（Km²）	
					eg:population		eg:areas	⇦【手順①】
					2010	2005	2010	
				eg:refYear	eg:year-2010	eg:year-2005	eg:year-2010	⇦【手順①】
	eg:refRegion		eg:refPrefecture	※				⇦【手順②】
関東	eg:region-03	埼玉県	eg:prefecture-11		7,194,556	7,054,243	3,767.92	
		千葉県	eg:prefecture-12		6,216,289	6,056,462	5,081.91	
		東京都	eg:prefecture-13		13,159,388	12,576,601	2,102.95	
		神奈川県	eg:prefecture-14		9,048,331	8,791,597	2,415.86	

⇧【手順①】　⇧【手順①】　⇧【手順②】

図 3　表 2 に対してテンプレートを用いた表