1N4-OS-10b-5

統計 Linked Data 語彙国際標準案とその有用性の検証

Statistical Linked Data Vocabulary Standard Draft and Its Usefulness

文　聞[*1]
Wen Wen

[*1] 東京国際大学大学院商学研究科
Graduate School of Business and Commerce, Tokyo International University

The publication of statistical data in linked data is widely increasing now. The standard vocabulary for statistical linked data has been proposed in order to make it easier to use statistical data from different organizations. This paper explores the problems and solutions that should be considered when integrating statistical data on the basis of this draft standard.

1. はじめに

最近、行政機関、国際組織などから統計データを積極的に公開する動きがある中、容易にデータを 2 次利用できるメリットから Linked Data での統計データ提供が注目されている。しかし、統計データは提供先によって使用する統計用語が異なることと、多次元データの特徴をもつために、どのようにして発見や統合が行えるかが課題となっている。

こうした現状に直面し、統計データを記述するための共通語彙の標準化が進行している。今日、英国政府（data.gov.uk）や欧州統計局（Eurostat）などの機関がこの国際標準案に基づいた統計データの提供と公開を進めている[佐藤 13]。

しかし、国際標準案に沿った統計データの公開は増え続くことを予想でき、PlanetData-Wiki [1]によりそれらの提供元を所在案内できるものの必要性も指摘されている。

ほとんどの統計提供先の Endpoint は SPARQL 文を発行する Web 画面である。一般の統計利用者に使いこなせるとは考えにくいために、より利便性が高いユーザインタフェースを提供すべく検討も始まっている[Fadi 12]。

本稿では、第2章で上記の国際標準案を簡略に紹介し、第3章で標準案を利用する利点を説明するとともに、どのような情報を所在案内として一般の統計利用者に提供したらよいかを探る。第4章で国別ディメンションと時点別ディメンションをもとにした統計データの統合利用の事例を紹介する。最後に、データの統合を行う際、国別と時点別ディメンションの値のマッチング方法とマッピングパターンをまとめる。

2. 統計 Linked Data の標準案の策定

統計 Linked Data 語彙国際標準案は W3C により The RDF Data Cube Vocabulary（略称 QB）として策定されており、現在 Working Draft の段階である。リスト1は、QB の主要概念を使って人口統計データを RDF/Turtle で書いた例である。QB では、ここにあるようにデータセットを始め、データ構造と観測データを記述する[Tennison 12]。

```
1) @PREFIX eg: <http://example.org/ns#> .
2) @PREFIX qb: <http://purl.org/linked-data/cube#> .
3) #  データセット
4) eg:ds-example a qb:DataSet ;
5)     rdfs:comment "都道府県別の人口"@ja ;
6)     qb:structure eg:dsd-example .
7) #  データ構造
8)     eg:dsd-example a qb:DataStructureDefinition ;
9)         qb:component
10)            [qb:dimension eg:年] ;
11)            [qb:dimension eg:都道府県] ;
12)            [qb:measure eg:人口] .
13)  #  観測データ
14)  eg:obs-001 a qb:Observation ;
15)      qb:dataSet eg:ds-example ;
16)      eg:年 "2013-02-01"^^xsd:date ;
17)      eg:都道府県 "東京" ;
18)      eg:人口 "13221566" .
```

リスト1　QB の RDF/Turtle の例

3. 標準案を利用する利点

統計分野では、多数の提供元のデータを組み合わせて利用することが多い。このために、統計利用者は、①情報源の探索、②データファイルのダウンロード、③必要なデータの抽出、④統計間の概念の違いを調整して分析用のデータセットを作成するといったことに大量な時間をとられている。統計 Linked Data に期待されているのは、この状況の改善である[佐藤 2013]。

①、②、③に関していえば、QB に基づいた統計データの提供を前提にすれば、qb:DataSet のタイプをもつデータが公開されているかどうか調べるだけで、そのサイトが統計 Linked Data

RDF Data Cube Vocabulary datasets

	Dataset uri	Measure	Dimension
data.gov.uk - COINS as Linked Data http://data.gov.uk/resources/coins	http://finance.data.gov.uk/coins/coins_fact_table_2009_2010	Amount	RefPeriod, DataType, DataSubtype, DepartmentCode, AccountCode, ProgrammeObjectCode, CounterpartyCode
Published by Linked Open Italia http://www.linkedopendata.it/datasets/istat-immigration.	http://data.linkedopendata.it/istat/resource/dataset-DCIS_MATRIMONISTR	Marriages with at least one foreign spouse	Country of citizenship, Year, Territory, Type of couple, Freq, Obs status
World Bank Linked Data - http://worldbank.270a.info/	http://worldbank.270a.info/dataset/world-bank-finances/sfv5-tf7p	Due-3rd-party, Sold-3rd-party, Cancelled-amount, Undisbursed-amount, Due-to-ibrd, Borrower-s-obligation, Disbursed-amount, Loans-held, Repaid-3rd-party, Original-principal-amount	Loan-type, Closed-date-most-recent, Last-repayment-date, Loan-number, PlanetData, Agreement-signing-date, Borrower, Loan-status, Guarantor, Effective-date-most-recent, First-repayment-date, Region, End-of-period, Board-approval-date, Country

図 1 PlanetData-Wiki のデータセット毎の一覧情報

連絡先：文聞（ブンブン），東京国際大学大学院商学研究科，〒350-1197　埼玉県川越市的場北 1-13-1，E-mail：s11170003bb@tiu.ac.jp

1. http://wiki.planet-data.eu/web/Datasets

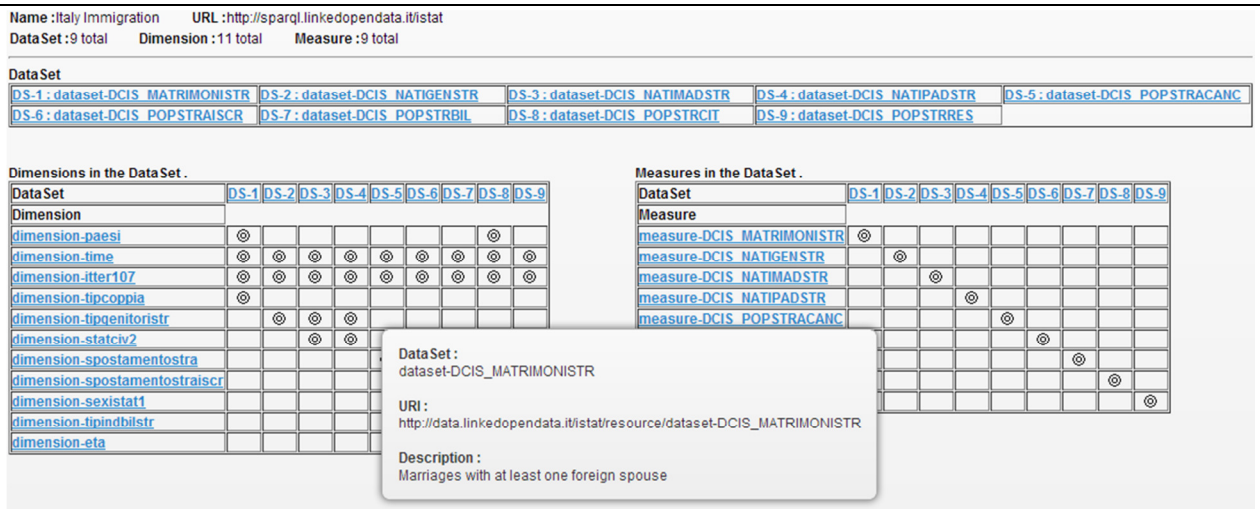

図２　統計データ情報を一覧できるアプリケーションの試作サイト

を提供しているかがわかる。そして、qb:DataStructureDefinition の内容をみることで、どのようなディメンションやメジャーを提供しているかがわかる。これらを利用すれば、公開されているデータセットや、データセット別のディメンションやメジャーの一覧を確認できる所在案内システムを作ることができる。

こういった問題はオーストリア大学の STI [2]も注目しており、26 の Endpoint やデータセットのディメンションとメジャーの一覧を PlanetData-Wiki にリストアップしている。(簡略に図1をまとめた)

図1のテーブルの最初の行は、finance.data.gov.uk/coins には coins_fact_table_2009_2010 というデータセットが提供されており、そのディメンションである RefPeriod(時点別)や DataType(データ種別)及びその他のディメンションが確認できる。メジャーの場合は、Amount のみであった。そして、data.linkedopendata.it や worldbank.270a.info に複数のディメンションやメジャーがあることも同様に確認できる。

しかし、調査したところ、このサイトのリンクが切れていたり、Endpoint が検索不能になったり、同一 Endpoint URL で複数のデータセットを掲載することもある。機械的にこの一覧表を集めたものとは思えない。

ここで、我々は、Endpoint 毎にそこで提供されている統計データ情報を一覧できるアプリケーションを試作した。図 2 は、イタリア移民統計 Linked Data から取得した一覧画面である。世界銀行統計 Linked Data など他の QB 準拠の統計 Linked Data についても同様の一覧を作ることができる。一般の統計 API を利用して、同様なものを作成することもできるが、その場合は統計 API 毎に異なるアプリケーションが不可欠である。

図 2 の「DataSet」欄は、イタリア移民統計 Linked Data にあるすべてのデータセットを取得したものであり、データセット毎にその略称とデータセット名を表示している。

「Dimensions in the DataSet」と「Measures in the DataSet」の欄では、各データセットにあるディメンションとメジャーを示す。更なる詳細を確認するために、ポップアップ画面も用意している。

「Dimensions in the DataSet」欄では、イタリア移民統計 Linked Data で扱っているすべてのディメンションがどのデータセットで利用されているかを示したものである。dimension-pasei の場合は、それぞれ「dataset-DCIS_MATRIMONISTR」と「dataset-POPSTRCIT」に使われている。「Measures in the DataSet」も同様で、データセットに使われているメジャーを一覧できる。

図 2 のようにデータセット毎にディメンションとメジャーの配置情報の一覧を提供することで、統計利用者が利用したいデータがどこにあるのかを確認できるために、①に関する問題を大幅に改善できる。

さらに、もしこの種のツールにそれぞれのデータセットにあるディメンションの値を指定するように機能を拡張していけば、②、③のような個別のインスタンスデータを入手する作業も軽減されると期待できる。

しかし、④については、種々の問題があり、単に QB に従うだけでは解決されない課題がある。これについては次章で述べる。

4. 標準案に沿った統計データの統合利用の事例紹介

複数の情報源のデータを組み合わせた例としては、ここでは、QB に準拠してインスタンスデータまで公開されているイタリア移民統計 Linked Data と世界銀行統計 Linked Data を取り上げて、国別ディメンションと時点別ディメンションの値による統合を試みた。出身国総人口(世界銀行統計 Linked Data)のうち、イタリアに在住している総移民及び性別の人数のデータ(イタリア移民統計 Linked Data)を求めるアプリケーションを試作した。ここではその例として、2008 年のアルバニアに関するものを紹介する。

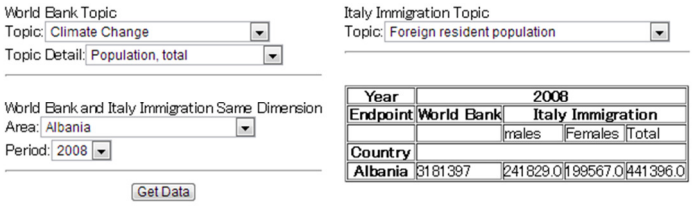

図３　アルバニアの総人口のうちイタリア在住の人口

この２つの Endpoint のデータに共通に含まれる Area と Period というディメンションとその他のディメンションの値を指定すれば、世界銀行統計 Linked Data から Albania の人口が 3181397 人であり、イタリア移民統計 Linked Data から在住している Albania 国籍の総人数は 441396 人、男性 241829 人、女性 199567 人であることを取得できた。

このように、QB に従って統計 Linked Data が提供されていれば、統合可能なディメンションを発見し、同じ観察対象に関する統計観測のデータを取得できる。

次に、どうようにして国別ディメンションと時点別ディメンションの値を、統合できたかを紹介する。

なお、これら 2 つの統計 Linked Data の記述構成の詳細は文献[佐藤 13]の第 4 章を参照していただきたい。

2. http://www.sti-innsbruck.at

4.1 国別ディメンションについて

イタリア移民統計 Linked Data と世界銀行統計 Linked Data は、国別ディメンションの値である国の概念を再定義し、コードリスト化にしている。しかし、いれずもそれらが GeoNames で定義された国の URI(http://sws.geonames.org/783754)と同一であることが示されている。イタリア移民統計 Linked Data の場合は skos:excatMatch 述語(リスト2)を利用し、世界銀行統計 Linked Data は owl:sameAs 述語(リスト3)を使用して同一関係を表現している。

そこで、国別ディメンションの値をマッチングする際、GeoNames で定義された URI と上記同一性述語を利用して、2 つの統計のデータの統合を行った。

```
1) istat:code-paesi-al a skos:Concept ;
      ...省略...
6)    skos:exactMatch <http://sws.geonames.org/783754> ;
```

リスト2 イタリア移民統計 Linked Data の Albania と GeoNames の同一表記。

```
1) classification:country/AL a skos:Concept ;
      ...省略...
28)      owl:sameAs <http://sws.geonames.org/783754/> .
```

リスト3 世界銀行統計 Linked Data の Albania と GeoNames の同一表記。

4.2 時点別ディメンションについて

国別ディメンションと同様に、イタリア移民統計 Linked Data の場合は、統計データの観測年をコードリスト化し、skos:exactMatch を用いて、グレゴリオ標準時にリンクしている。(リスト4)

```
1) istat:code-time-2008 a skos:Concept ;
      ...省略...
8)    skos:exactMatch
<http://reference.data.gov.uk/doc/gregorian-interval/2008-01-
01T00:00:00/P1Y> .
```

リスト4 イタリア移民統計 Linked Data の観測年の表記。

しかし、世界銀行統計 Linked Data の時点別ディメンションの値はコードリスト化されておらず、インスタンスデータに時点の URI が直接に記述されている。(リスト5)

```
1) d-indicators:SP.POP.TOTL/AL/2008 a qb:Observation ;
      ...省略...
5)    sdmx-dimension:refPeriod
   <http://reference.data.gov.uk/id/year/2008> ;
```

リスト5: 世界銀行統計 Linked Data の観測年の表記。

リスト4の 8 行目の観測年の URL 後位にある「/2008-01-01T00:00:00/P1Y」が時間の「2008 年 1 月 1 日から 1 年間」の概念を表している。

リスト5の 5 行目の観測年の URL では「/year/2008」が時点の「2008 年」の概念を示している。

時点別ディメンションの値をマッチングする際、それぞれに異なる URI が利用されていたが、今の場合は同じ概念を表していると解釈できるために、SPARQL 文の正規表現を使用して、同じ年を判別して処理を行った。

5. 機械処理に向けての課題

QB に基づいた統計 Linked Data の Endpoint の構築が行われれば、第 4 章の始めに述べた 4 つの問題のうち、①から③まで容易に機械化による処理が実現できるであろう。

④の統計間の概念調整によって、新たなデータセットを作成することについては、現 QB 標準案では、概念の記述も規定されているが、種々の記法を容認しているために、何を標準的な概念として統合が行われるかは未確定であるという欠点がある。

本節では、イタリア移民統計 Linked Data と世界銀行統計 Linked Data の統合の例を取り上げて、時点別と国別のディメンションによる統合のアルゴリズムを紹介しながら、自動化によるマッピング可能なディメンションの発見・統合の問題点とその解決方法を紹介する。

5.1 統合のアルゴリズム

前章で紹介したイタリア移民統計 Linked Data と世界銀行統計 Linked Data の国別と時点別のディメンションの統合は、人力で調べ、統合できる URI の所在はコードリストとインスタンスデータに記述されていることを明らかにした。

しかし、統計データの統合を行う際に、コードリストやインスタンスデータを目で見てマッピング可能な URI の存在確認を行うことは効率的ではない。機械処理に適するために、以下のアルゴリズムと必要な QB 語彙をここで整理する。

(ア) qb:concept によるマッピング可能なディメンションがあるか。
(イ) qb:codeList を使ったコードリストはあるか。
(ウ) owl:sameAs あるいは skos:exactMatch によるコードは比較可能か。

(ア)についていえば、QB では、ディメンションを定義する際に、コンポーネントプロパティを表す qb:concept 述語を用いて、概念を記述するように規定されている。もしここで、2 つの統計データのディメンションは共通の概念にリンクされていることが確認できれば、統合可能なものを判別できる。詳細なサンプルは、参考文献の[Tennison 12]の 5.3 節を参照してほしい。

(イ)については、qb:codeList 述語を利用して、当該ディメンションに使用する値のコードを列挙することができる。このコードリストを確認すれば、容易に比較可能な項目の発見処理に繋げることができる。コードリストについて詳細な記法は参考文献の[佐藤 13]のリスト 24～26 を参照してほしい。

(ウ)は、QB ではコードリストに記載されたディメンションの値について、その内容に owl:sameAs や skos:mxactMatch を使用して、共通の概念 URI にリンクされているかどうかが確認できれば、統合できるものと判断できる。前章のリスト 2 とリスト 3 の例を参照。

5.2 国別ディメンションの自動統合

本節では、イタリア移民統計 Linked Data と世界銀行統計 Linked Data の国別ディメンションの機械的な統合可能性を検証してみよう。

機械処理で前章のリスト 2 とリスト 3 を辿っていけるまでに、前節に紹介した(ア)と(イ)を確認する必要がある。

まず、(ア)については、イタリア移民統計 Linked Data の国別ディメンション(istat:dimension-paesi)では、qb:concept を用いて独自で定義した概念(istat:concept-paesi)にリンクされている。

世界銀行統計 Linked Data の国別ディメンションは、QB が定義している sdmx-dimension:refArea を流用している。その定義は qb:concept を利用して sdmx-concept:refArea にリンクしている。

sdmx-concept:refArea は、SDMX 規格で定義された標準地域概念である。QB でも、地域別ディメンションを定義する際に、この URI を参照するように推奨している。

2 つの国別ディメンションは、qb:concept を記述しているが、イタリア移民統計 Linked Data の国別ディメンションは、sdmx-concept:refArea にリンクされておらず、機械処理による統合可

能なディメンションであることを発見しにくいといえる。もし、sdmx-concept:refArea を参照していれば、世界銀行統計 Linked Data とのマッピングができる国別ディメンションであることが識別できる。

（イ）では、イタリア移民統計 Linked Data の場合は、qb:codeList を用いてコードリスト（istat:code-paesi）にリンクしている。そうすることで、国別ディメンションに使われる値の統一管理ができるうえ、容易に統合可能な URI の発見にも繋げられる利点がある。

しかし、世界銀行統計 Linked Data の場合は、sdmx-dimension:refArea を直接流用している。QB では、sdmx-dimension:refArea を用いて、プロパティの上位概念を記述するように推奨している。本来、sdmx-dimension:refArea は抽象的なディメンションであり、そこにコードリストを定義することはできない。これを直接にディメンションとして利用することは、統合可能な URI の発見に適していないといえよう。従って、世界銀行統計 Linked Data の国別ディメンションはイタリア移民統計 Linked Data のように再定義すべきである。

（ウ）については、リスト 2 とリスト 3 のように 2 つの統計 Linked Data の国のコードに owl:sameAs や skos:exactMatch 述語を用いて、今日広く使われている GeoNames の URI にリンクしている。これらを利用すれば、2 つの国別ディメンションによるデータの統合ができる。

5.3 時点別ディメンションの自動統合

時点別ディメンションの自動統合については、2 つの統計 Linked Data の記法は、国別ディメンションとほぼ同様である。

イタリア移民統計 Linked Data の時点別ディメンション（istat:dimension-time）では、qb:concept を用いて独自で定義した概念（istat:concept-time）にリンクしている。SDMX 規格の sdmx-concept:refPeriod の概念であることを明示していれば、マッピング可能なディメンションを容易に発見できたはずである。

世界銀行統計 Linked Data の場合は、sdmx-dimension:refArea と同様に sdmx-dimension:refPeriod を利用して、時点別ディメンションとして流用している。時点別ディメンションは再定義の必要がある。

ただし、世界銀行統計 Linked Data の場合は、リスト 5 で示したように、時点のコードリストを定義しておらず、その値を直接にインスタンスデータに記述している。

この時、（ウ）の処理を行うために、時点別ディメンションに利用する値をインスタンスデータから取得し、時点のコードリストを生成すれば比較可能になり、他の統計データと統合できるようになる。

6. まとめ

統計 Linked Data 語彙国際標準案の有用性を検証するために、イタリア移民統計 Linked Data と世界銀行統計 Linked Data を利用して、それぞれの統計観測データの統合を試みた。

6.1 同一と判別できるデータの多様な書き方

イタリア移民統計 Linked Data では、国別ディメンションの値を独自で定義し、コード化にしている。そして、他の統計データとリンクできる URI は、その個別のコードに記述している。

一方、世界銀行統計 Linked Data では、国コードを再定義し、他の統計データとリンクできる URI も記述されていたが、データ構造定義の国別ディメンションの定義は単に SDMX の sdmx-dimension:refArea を流用したものである。

時点別ディメンションについても、イタリア移民統計 Linked Data は国別ディメンションと同様の記法であったが、世界銀行統計 Linked Data では、時点別ディメンションに SDMX の sdmx-dimension:refPeriod を流用している。さらに時点コードも再定義しておらず、インスタンスデータに他の統計データとリンク可能な URI が書かれているだけである。

このように、他の統計データと統合する際に、リンク可能なデータの記述場所がまちまちであることを留意しておく必要がある。

6.2 同義概念を表す述語を用いた標準 URI の参照

本稿の 4 章では、skos:excatMatch と owl:sameAs 述語を用いて、標準的な地域や標準的な時間の URI を取得して、2 つの統計のリンクを行った。しかし、標準的な地域といっても、GeoNames の定義や DBpeida の定義など複数のものがある。何を標準と考えるか明確にすることが重要である。

ここで取り上げた 2 つの統計の例でも、統合に必要な情報がそれぞれ異なる方法で記述されていた。今後 QB 国際標準案に従ったとしても異なる統計データでは書き方も多々現れるであろう。QB 国際標準案は、多種類の記法を容認している利便性がある一方、種々の統計 Linked Data を連携可能にすることは難しい。リンクに有用な情報を発見できるような書き方のパターンを検討すべきであろう。

ユーザインタフェースについても、大量に統計 Linked Data が増え続けてきた際、どういった統計 Linked Data が相互に連携できるかや統合可能な項目が何かを示すことが統計利用者にとってより重要になってくるであろう。その時、いかに機械的に QB に沿った Endpoint を発見し、公開されているデータセットのデータ構造や、統合可能な項目を見つけ出すパターンが不可欠であろう。

筆者は、統計データベースや Linked Data に関する研究や学習を始めたばかりである。筆者が知る限り、統計 Linked Data についての検討は、ほとんど欧州を中心に行なわれているようである。これから、日本やアジアにおいても、統計 Linked Data がより広く知られるように貢献できたら幸せである。

参考文献

[佐藤 13] 佐藤英人, 文聞：統計 Linked Data の現状と課題, 東京国際大学論叢商学部編, 2013.03.

[神崎 12] 神崎正英：Linked Data とデータマッピング, 人口知能学会誌, Vol.27 No.2, pp.163-170, 2012.

[佐藤 88] 佐藤英人：統計データベースの設計と開発, オーム社, 1988.

[Tennison 12] Jeni Tennison, TSO：The RDF Data Cube Vocabulary--W3C Working Draft 05 April 2012, available at http://www.w3.org/TR/2012/WD-vocab-data-cube-20120405/, 2012.

[Fadi 12] Fadi Maali, Gofran Shukair, Nikolaos Loutas：A Dynamic Faceted Browser for Data Cube Statistical Data, 2012.06, available at http://www.w3.org/2012/06/pmod/pmod2012_submission_12.pdf