# オープンデータにおける RDF 変換の研究
# A study of RDF conversion in Open Data

久永忠範　郭 崇　能登大輔　渕田孝康
(鹿児島大学理工学研究科)
fuchida@ibe.kagoshima-u.ac.jp

## 1 はじめに

近年，ビッグデータやオープンデータの活用が推進され，国や地方自治体をはじめ多くの団体がオープンデータの公開，活用に取り組んでいる．これらの開示されたデータ形式は，ワード形式，エクセル形式や CSV 形式のファイルがいまだに多く，2012 年に策定された「電子行政オープンデータ戦略」に明示されている「機械判読可能で人手を多くかけずにデータの 2 次利用が可能である」というデータ活用までには至っていないのが現状である．開示されている多くのデータを RDF 形式へ簡単に変換できれば，複数のオープンデータの機械的な連携が可能となりオープンデータの活用を促進することが可能となる．本研究では，IPA の推奨する共通語彙基盤の語彙等を活用して，RDF 形式への変換技術の提案を行う．

## 2 研究の流れ

### 2．1 オープンデータの現状

オープンデータは，広く開かれた利用可能なデータである．2012 年の日本政府による「電子行政オープンデータ戦略」[1]や 2013 年の主要８ヶ国首脳会議(Ｇ８)で合意した「オープンデータ憲章」[2]を見ると公共データを広く活用させることにより，国内外の行政，経済活動，情報流通などの活性化が推進されることが謳われている．

日本政府が提供するデータカタログサイト DATA.GO.JP[3]においては，2016 年 7 月末現在，各省庁，国の機関から約 17,000 ファイルのデータがアップロードされている．これらのデータフォーマットをみると PDF ファイルが 9330 ファイルと過半数以上を占め，次に HTML,XLS,CSV の順に 32 のフォーマット形式で提供されている．また日本の各地方自治体の提供するオープンデータは，年々増加してきているが，提供されるフォーマット形式が「機械判読に適さない形式」が多く，これらのデータを利活用するにも「人の手を借りて」データをカスタマイズしなければならないのが現状である．

### 2．2 RDFについて

オープンデータの多くは，PDF,DOC,XLS,CSV 等特定のアプリケーションを介したデータ形式で公開されているが，政府の提唱する「機械判読に適したデータ形式で，二次利用が可能な利用ルールで公開されたデータ」として活用するためには，Tim Berners-Lee の提唱する 5star[4]の 4th stage 以上に該当する RDF,LOD 形式のフォーマットが適している．RDF は，主語，述語，目的語という 3 要素でリソースに関する関係情報を表現する．述語は，主語の特徴や主語と目的語の関係を示す．この述語は URI で表現し，この述語の語彙の URI を共通化することにより，異なるファイル同士の関係を結びつけることができる．RDF は，特にメタデータについて記述することを目的にしており，XML 等で表示される．

### 2．3 共通語彙基盤について

オープンデータの分野を越えた情報交換やデータの二次利用効率化を図るために，IPA を主体として図 1 の共通語彙基盤[5]の構造の整備がなされつつある．同じ単語を違う意味で使うことや違う単語を同じ意味で使うことによる認識の違いや意思疎通の不便さを共通の語彙を整備することによりお互いの連携を図ることができる．これは個々の単語についての表記・意味・データ構造を統一して，互いに意味が通じるような共通語彙を整備する取り組みである．この共通語彙基盤の語彙と RDF 形式の要素を組み合わせることによって多面的なデータ連携が考えられる．

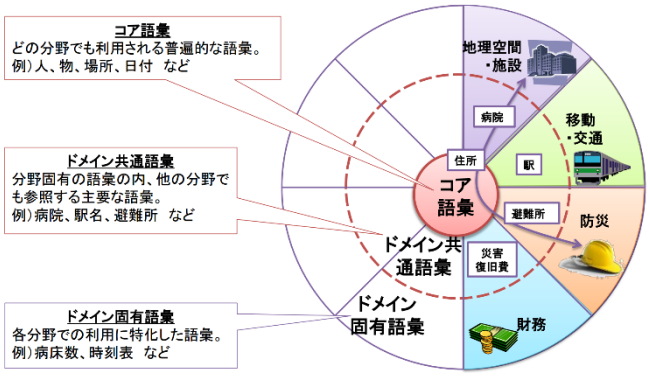


図 1 共通語彙基盤の構造[4]

### 2．4 RDF を取り巻く状況

述語の統一は RDF の活用の本質的な問題であり，近年多くの研究がある．古くは rdf，rdfs，dcterms，dcelements などの名前空間から，最近では dbpedia や prop-ja などの日本語化された名前空間など，様々な形の名前空間が提案されており，それらの名前空間の中で定義されているクラスやプロパティを述語として利用する試みがなされている．

しかし，実際にオープンデータを作成するのは役所や民間組織の担当者であり，そのような担当者が複雑な名前空間の仕組みを理解して正しい述語を使用した RDF を作成することは極めて困難である．

本研究では，すでに公開されている CSV 形式のオープンデータに対して述語の共通化を図り，LOD として活用可能な RDF 形式のデータに変換する手法を提案する．

## 3 提案手法

CSV ファイルのデータでは，第 1 行目に各カラムの意味を表す文字列を書くのが通例である．多くのオープンデータにおいてもこの形式は守られており，鹿児島市が公開しているオープンデータについてはすべてのこの形式である．

ただし，オープンデータを作成する部署によっては同じ意味を表す項目の名前が異なっている場合がある．鹿児島市に限った場合でもこのような状況があること

を鑑みると，一般的なオープンデータにおいてはこの傾向はますます大きくなると考えることができる．

単語の意味と距離に関しては多くの研究成果が報告されており，特に近年，word2vec と呼ばれるニューラルネットワークを応用して，単語をベクトル空間に対応付ける試みが提案されている．word2vec は，自然言語による文章を入力として与えると，その中の単語の連接関係から単語を高次元空間上のベクトルに変換し，ベクトル空間における単語間の距離を算出できるようにする（日本語においては前処理として文章の分かち書きが必要）．word2vec を効果的に利用するためには，入力する自然言語の文章が，求める単語の関係をうまく表現したものである必要がある．

そこで我々は，すでに公開されているオープンデータのタイトル行に現れる文字列（項目名）をキーワードとして Google 検索によりページを検索し，上位に現れたページの中の文章を入力して word2vec により単語を名前空間に写像させ，その結果を使って複数のオープンデータに現れる述語の共通化を図る手法を提案する．図 2 に提案手法の流れを示す．

## 4 まとめ

国や地方自治体等が持つ多くの情報をオープンデータとして 2 次利用可能な形で公開し，情報の利活用を促進する動きが，国の主導のもとで推進されており，様々な活用方法の研究や提案が行われている．しかし，公開されているデータの形式が CSV などの連携性の低い形であり，LOD として活用するためには述語の共通が必要である．

本研究では，公開されている CSV 形式のデータの項目名から Google 検索を用いてページを抽出し，word2vec を用いて単語をベクトル空間に写像することで，複数の CSV ファイルを連携させて LOD として活用可能な RDF の形に変換する手法を提案した．計算機実験の結果については発表時に説明したい．

図 3 に，我々が考えている RDF 生成の流れと活用について示す．本研究で提案した手法によりオープンデータの LOD 化が図れれば，今後はそれらのデータを連携させて新しい価値の創出やビジネスの提案等に活用していきたいと考えている．

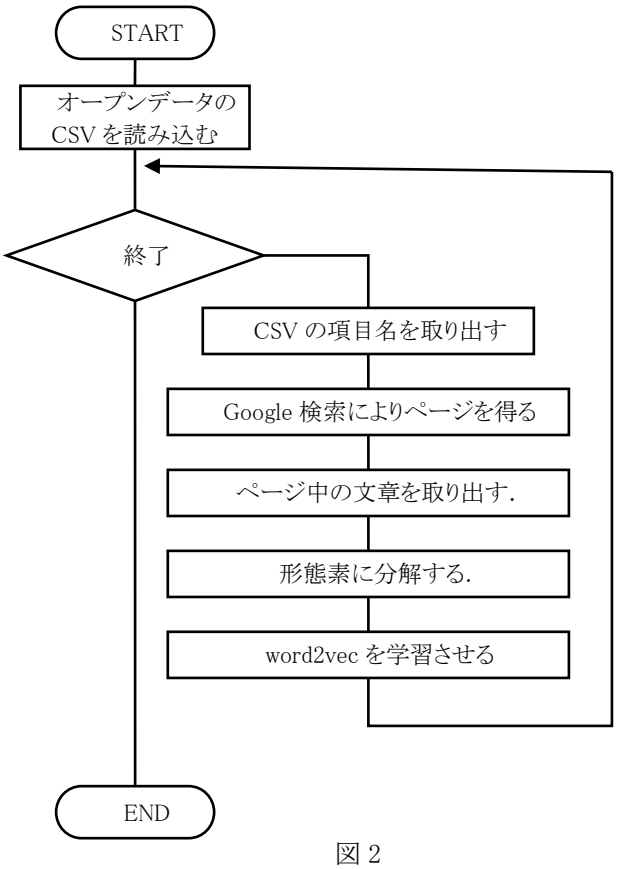


図 2

## 参考文献


[1] 電子行政オープンデータ戦略
http://www.kantei.go.jp/jp/singi/it2/pdf/120704_siryou2.pdf

[2] オープンデータ憲章
http://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf

[3] DATA.GO.JP（データカタログサイト）
http://www.data.go.jp/

[4] Tim Berners-Lee,5star
http://5stardata.info/en/

[5] 共通語彙基盤(IPA 独立行政法人情報処理推進機構)
http://goikiban.ipa.go.jp/


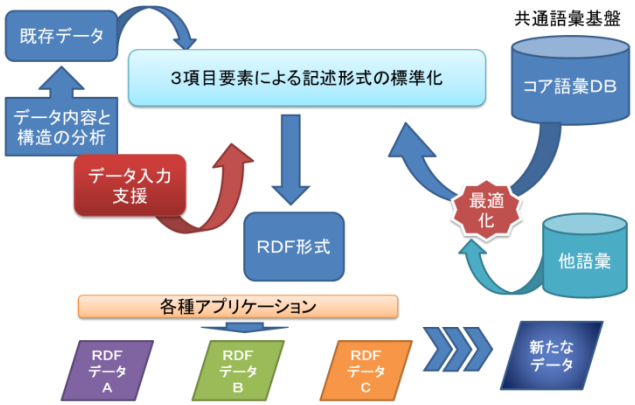


図 3 RDF 生成の流れと活用