日本の Open Data 活用を目的としたデータセットのスキーマ分析とリンク関係の調査

西出頼継[†1]　本間維[†1]　永森光晴[†2†3]　杉本重雄[†2†3]

近年，誰もが利用・アクセスを行え，再利用や再配布が許可されたデータセットである Open Data が増えてきている．しかし，Open Data の構造は複雑なものが多く，スキーマ定義がないと再利用が難しい．また日本において LOD として公開されている Open Data の数は世界全体と比べると少ない．本稿では，日本で公開されている Open Data を LOD として活用することを目的に，CKAN 日本語などで公開される日本のデータセットを対象にスキーマの分析やリンク関係などの調査を行った．その結果日本の Open Data ではスキーマ定義やデータセット間のリンクが少ないということが分かった．そして，Open Data を LOD として活用するためにはスキーマ定義を行うためのメタデータ語彙の推薦や，リンクで結ぶためのリソースの同定が必要であると考察した．

An Investigation of Japanese Open Data Schemas and Links to Improve the Use of Datasets

YORITSUGU NISHIDE [†1]　TSUNAGU HONMA[†1]
MITSUHARU NAGAMORI[†2†3]　SHIGEO SUGIMOTO[†2†3]

In recent years, there has been an increase in the use of Open Data. These datasets are freely available to everyone to reuse and republish. Unfortunately, many Open Data structures are complicated and are difficult to reuse without a schema definition. Compared with the rest of the world, Japan releases relatively little Open Data as LOD. We investigated Japanese Open Data, focusing on CKAN-*Nihongo* schemas and links to improve the use of datasets as LOD. We found that there are few Japanese Open Data schemas and links. Therefore, we recommend that metadata vocabularies and identification of links to connect resources should make use of Open Data released as LOD.

1. はじめに

政府，学会，研究機関といった様々なコミュニティが再利用と再配布を行えるように公開しているデータセットを Open Data と呼ぶ [1]．Open Data は世界の各地域で公開されており，その数は急速に増え続けている．近年，日本でもその公開の試みは広がっており，CKAN 日本語 [2]や Open Data METI [3] といった，Open Data 公開のためのプラットフォーム上で様々な日本のコミュニティが作成した Open Data を取得することができる．その中でもデータセットのメタデータ間を型付きリンクで結ぶことにより，人だけでなく機械もデータの内容が可読となった Open Data を Linked Open Data（LOD）[4]と呼ぶ．Open Data を関連する他の分野のデータセットにリンクで結び LOD として公開することにより，そのデータセットの利用者は異なる分野間においても機械によるメタデータの探索が可能となる．

しかし，現在の日本の Open Data では LOD として公開されているデータセットが少ないため，リンク関係を用いた LOD 活用の幅は狭い．また，Open Data はデータ構造が複雑なものが多く，データ構造のスキーマ定義が与えられていない場合，第三者が新たに Open Data を LOD として再利用することは難しい．そこで本研究は，日本の Open Data を LOD として利用することを目的とし，CKAN 日本語などにおいて日本で公開されている Open Data を対象にスキーマの分析やリンク関係の調査を行う．

2. Open Data の現状

2.1 日本と世界全体における Open Data の現状と比較

近年，Web 上では多くのデータセットが公開され利用されるようになった．その中でも Open Data としてデータセットを公開する試みが増加しており，政府が公開している産業ごとのエネルギー消費実態の統計書や生命科学分野における微生物の遺伝子情報など，様々な分野においてデータセットが公開されている．また Open Data では，メタデータ間をリンクで結びつけた LOD として公開されるデータセットも増えてきており，異なる分野間における関連付けも行われている．関連付けが行われているデータセットの例として，米国の New York Times 紙において用いられている件名標目を LOD 化した New York Times – Linked Open Data [5]がある．このデータセットの地理に関するリソースは，地理情報のデータセットである GeoNames Semantic Web [6]のリソースとリンクで結ばれて

†1 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies. Univ of Tsukuba.
†2 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Studies. Univ of Tsukuba.
†3 知的コミュニティ基盤研究センター
Research Ctr for Knowledge Communities. Univ of Tsukuba.

おり，メタデータ間の探索が可能である．

また，主要な LOD 間のリンク関係を表現したものとして LOD クラウドがある [7]．図 1 に示しているのが LOD クラウドの状態を表した図（2011 年 9 月現在）で，2013 年での LOD クラウドのデータセット数は 336 である．一方，2013 年における日本の LOD 間の相互関係を表した日本版 LOD クラウドが加藤によって作成されている（図 2）[8]．図 2 での日本における LOD クラウドのデータセット数は 17 であるが，これらの中でオリジナルの LOD クラウドとして含まれているのは Web NDL Authorities [9]の 1 件のみである．

2.2 海外の Open Data 公開の試み

Open Data は，国際的に様々な地域で公開されている．海外における LOD を含めた Open Data はそれぞれのポータルサイトで取得することができる．例えば米国なら Data.gov State Data Sites [10]，英国なら data.gov.uk [11]といったサイトでその国に関係する Open Data が公開されている．また，公開されているデータセット数も非常に多く，data.gov で 10 万件以上，data.gov.uk で 9 千件以上のデータセットが利用可能である（2013 年 8 月）．Open Data の公開は国単位のみで行われているわけではない．例えばフランスなら，data.gouv.fr [12]で国全域を対象にした Open Data を扱っているが，他にもレンヌ [13]やモンペリエ [14]といった都市単位での公開も行われている．また，世界全体で見ると，約 50 の国と機関によって 200 以上の Open Data 公開のためのポータルサイトが運営されている[15]．

2.3 日本の Open Data 公開の試み

日本でも，海外と同様に複数の団体によって Open Data 公開の取り組みが行われている．その代表的な試みとして以下の四つがある．

(1) CKAN 日本語（http://data.linkedopendata.jp/）
(2) Open Data METI（http://datameti.go.jp/）
(3) LOD チャレンジ Japan（http://lod.sfc.keio.ac.jp/）
(4) LinkData（http://linkdata.org/）

CKAN 日本語 は，data.gov.uk などのデータポータルと同じく CKAN のプラットフォームを用いて，行政が作成したデータセットや個人が作成したデータセットなど様々なデータセットを公開している．その多くは地方行政のデータセットで，神奈川県の横浜市，福井県の鯖江市などが積極的にデータ公開を行っている．2013 年 8 月現在で公開しているデータセット数は 193 件である．

国の行政が独自に Open Data を公開している例として Open Data METI がある．Open Data METI は，経済産業省が 2013 年より運営を開始しており，Open Data の実証用サイトとして公開を行っている．この取り組みは経済産業省が保有しているデータを民間で活用してもらうことを目的としており，主に白書や統計書などが公開の対象になっている．2013 年 8 月現在で公開しているデータセット数は 201 件である．

上記の二つとは異なり，一般の人から Open Data を募集して評価を行っているコンテスト形式の取り組みとして，LOD チャレンジ Japan [16]がある．LOD チャレンジ Japan は 2011 年より毎年開催されており，データセット部門，アイディア部門，アプリ部門，ビジュアライゼーション部門，基盤技術部門（本部門は 2013 年開催より新設）の 5 つの部門ごとに Open Data に関する作品を一般から募集している．データセット部門の応募作品は，LOD チャレンジ Japan の Web ページ上で公開されており，取得してデータの内容を見ることも可能である．さらに LOD チャレンジ Japan ではそれらの作品を審議し，優秀な作品に対しては表彰を行ったりするなど，より多くのデータセットを作成してもらうような促進も行っている．また，2012 年の開催におけるデータセット部門では，2011 年の 21 データセットと比べ，4 倍以上である 89 データセットもの応募があった．

この他にもテーブルデータを RDF 化して Open Data として公開する LinkData [17]が 2011 年から始まったことなど，日本でも Open Data を公開するための手段が増えつつある．しかし，上記で述べた取り組みごとのデータセット数からも分かる通り，日本における Open Data の数は米国や英国などの Open Data 主要国と比べて少ない．また，公開され

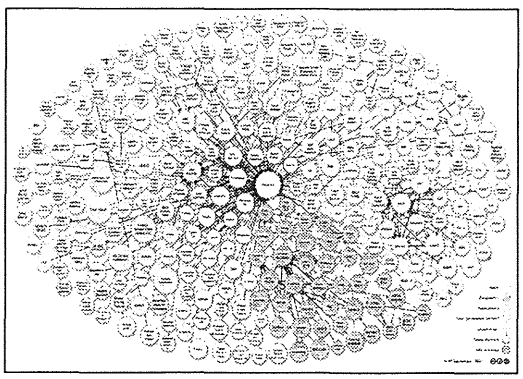

図 1 LOD クラウド

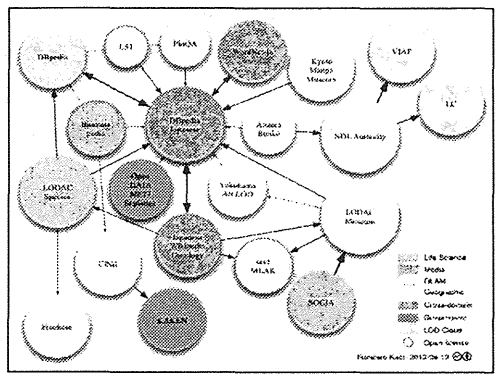

図 2 日本版 LOD クラウド

ている Open Data を活用したアプリケーションなどの事例は，データセットの数に比べまだ多くない．

3. 関連研究

本研究では Open Data の中でも RDF データにおいて用いられているタームについての分析を行う．本稿で述べるタームとは，メタデータ記述に用いる個々のクラスとプロパティのことを指す [18]．クラスは共通の特徴をもつリソースのグループに URI で名前を付けて表現したもので，プロパティはリソースが持つ特徴の一つに URI で名前をつけて表現したものである．例えば，人物や組織のためのタームを定義しているメタデータ語彙として FOAF [19]がある．FOAF では，人物のグループを表すクラスとして foaf:Person を定義していたり，リソースの名前を表すプロパティとして，foaf:name を定義していたりする．この章では，タームの分析を行った関連研究として LODStats [20]と Linked Open Vocabraries（LOV）[21]の二つの研究を述べる．

LODStats は，the Datahub [22]が公開している Open Data の中で，RDF もしくは SPARQL[23]で公開しているデータを対象に統計情報の作成を行っている研究である．統計項目は，データセット中のトリプル数やタームの URI と数，使用されている名前空間などで，これらを算出し統計結果を RDF データとしてデータセットごとに公開している．LODStats では the Datahub の Open Data に対してはタームの統計情報を出しているが，CKAN 日本語といった日本の Open Data には統計を出していないため，本研究で扱うためには日本の Open Data を対象にした統計情報を新たに算出する必要がある．

LOV では，既存のメタデータ語彙として 364 の語彙（2013 年 8 月）を収集しており，メタデータを記述する利用者が最も適切なタームを使用するための支援を行っている．タームの選択にはキーワードで検索する以外にも，“Media” や “Science” といった指定されているドメインから検索することも可能である．また LOV では，「LOV Stats」という統計情報も提供している．LOV Stats ではタームごとに “LOV popularity” などの LOV 独自の統計項目を算出しており，そのタームがどれほど利用されているかの指標を提示している．このように，LOV ではタームごとの統計を公開しているが，データセットや分野ごとのタームの利用頻度などは算出していないため，本研究では調査項目として別に算出する必要がある．

4. 日本の Open Data の調査

4.1 調査目的

Open Data を LOD として扱うためには，そのデータが RDF で記述されていることが望ましい．しかし，現在は CSV や XLS など，RDF 以外のファイルフォーマットでデータセットを公開している Open Data が非常に多く存在している．このような Open Data を LOD として利用するためには，データの構造を機械が識別できるようにデータ構造を定めたスキーマが必要になる．また，第三者がデータセットを編集して再利用する場合にも，データ構造を把握する必要があるためスキーマは必要になる．

Open Data を LOD として活用する利点として，データセット間のリンク関係を利用して異なるデータセットの情報を利用することができる，という点が挙げられる．このような活用を行うためには Open Data をただ RDF として公開するだけではなく，メタデータを関連するメタデータにリンクで結ぶことが求められる．

よって，Open Data を LOD として活用していくためには主に二つの点が重要であると分析した．

(1) スキーマの有無とスキーマ定義が可能かどうか
(2) データセット間でのリンク関係の有無

本研究では上記の二点に関して既存の Open Data を対象に調査を行う．

また，Open Data を第三者が利用するためには，そのデータのライセンスが明記されていることが望ましい．ライセンスの明記がない Open Data は，どの程度までデータ原作者が権利を主張しているかが分からないため，第三者がデータを再利用して公開するといった活用ができなくなってしまう．よって本調査では，ライセンス有無の調査と，明記されているならばそのデータがどの程度まで権利を主張しているかの分類も行う．

さらに本研究では，Open Data を LOD として活用していくためには Open Data をどのように作成していくべきであるか，さらに Open Data 作成のための支援としてどのようなことが考えられるかの考察を，上記の調査結果をもとに行う．

4.2 調査対象

調査対象とするデータセットは，2 章で述べた，CKAN 日本語，Open Data METI，LOD チャレンジ Japan で公開している Open Data を対象とする．CKAN 日本語，Open Data METI では，公開されている中でも分野ごとにグループ化されているデータセットを対象とし，CKAN 日本語で 151 のデータセット，Open Data METI で 201 のデータセットに対してそれぞれ調査を行う．LOD チャレンジ Japan では，LOD チャレンジ Japan 2011 と 2012 において公開された，計 120 のデータセットを対象に調査を行う．ただし，極端にリソースの数が少なく，データとして不完全と判断した，CKAN 日本語の 5 データセットと LOD チャレンジ Japan の 38 データセットに関しては調査対象には含めない．

また，調査対象となる Open Data を LOD クラウドで用い

られているカテゴリで分類を行い，全体的な傾向を探る（Open Data METI は政府情報のみの公開を行っているので割愛する）．

4.3 調査手法

調査手法として，4.1 で述べた二点の調査項目を，4.2 で述べた CKAN 日本語，Open Data METI，LOD チャレンジ Japan におけるそれぞれのデータセットで分析する．スキーマの分析では，データセットに対してのスキーマが定義されているかどうかを調査し，定義されていなければそのデータセットの構造に対してスキーマを定義することが可能かの分析を行う．リンク関係の調査では，リソースを解析し，他のデータセットへのリンクがあるかの調査を行う．

調査対象である 3 つの取り組みに対して事前調査を行ったところ，公開されているデータのファイルフォーマットは，(1)RDF，(2)CSV，XLS，(3)PDF，(4)XML，(5)HTML の五つに大きく分類できることが分かった．よって，そのフォーマットごとにスキーマの分析とリンク関係の調査を行っていく．また，上記の二つの調査項目以外にも，Open Data を LOD として活用するためにフォーマット独自の調査項目が考えられる．以下にそのフォーマットごとの調査項目と調査する理由を述べる．

(1) RDF では，他のフォーマットと異なり，タームに URI が用いられる．タームは，同じ記述対象に対して複数候補が考えられることがあるため，第三者によるデータの利用が想定される場合，RDF データの作成者はタームの記述項目を含めたスキーマの定義をしておくことが重要になる．またタームは，これまでに語彙定義がされていて汎用的に用いられている既存語彙と，データの作成者が独自に定めた独自語彙に分けることができる．そこで今回の調査では，RDF データにおいてどのようなタームを用いて記述しているかの統計を既存語彙と独自語彙でそれぞれ算出し，その傾向を探る．

(2) CSV，XLS では，どちらもテーブルデータとして扱うことができるため，今回の調査では同じグループとする．テーブルデータを LOD として扱うためには，まず RDF データに変換して，カラムを表す語を URI で表現するための変換を行う必要がある．しかし，これまでに Open Data として公開されているテーブルデータにおいて，使用されている語の分類を行った研究はない．よって今回の調査では，Open Data のテーブルデータで用いられる語として，どのような語が用いられているか，またそれらはどのようなメタデータ語彙に分類できるか調査する．

(3) PDF を LOD として活用していくためには，まずそのデータを構造化できるか調べる必要がある．よって PDF データに対しては，Open Data として公開されている PDF データを内容により分類し，その内容ごとに構造化可能かを分析する．

(4) XML では，テーブルデータでの調査理由と同様に，タグで使用されている語の分析を行う．

(5) HTML では，各 HTML が表す Web ページごとの内容で分類し，LOD として活用できそうかの分析を行う．また， RDFa や microformats などで Web ページ中のリソースに対してメタデータによる外部へのリンクがされていないかも調べる．

ライセンスに関する調査では，クリエイティブ・コモンズが提供している基準によって分類を行う．また，調査対象である 3 つの取り組みごとで，ライセンスの定め方に違いがあるかの比較も併せて行う．

4.4 調査結果

本研究の調査結果として，まずファイルフォーマットごとの分類，特徴を提示してから，スキーマ分析とリンク関係の調査結果を述べる．

(1) RDF データに対する調査では，Open Data METI が RDF データを公開していなかったため，CKAN 日本語と LOD チャレンジ Japan の RDF データに対して分析を行った．RDF データは基本的に構造化データであるため，スキーマの作成は可能である．しかし，データ作成者によってスキーマが公開されている例は，66 データセット中 4 データセットであった．また， 9 データセットでのみ他のデータセットへのリンクが結ばれていた．以下がそのデータセット名である．

- スキーマを公開しているもの [24][25][26][27]
  1) 青空文庫 Linked Open Data
  2) 京都国際マンガミュージアム書誌情報 LOD
  3) LODC Works
  4) 落語家 LOD
- 他のデータセットへのリンクがあるもの [28][29][30][31][32][33][34]
  1) ヨコハマ・アート・LOD
  2) 横浜ごみ分別情報
  3) 青空文庫 Linked Open Data
  4) 京都国際マンガミュージアム書誌情報 LOD
  5) LSJ2013（2012）
  6) 日本語 Wikipedia オントロジー
  7) Allie
  8) Biomasspedia
  9) saveMLAK

RDF に対して行ったタームの調査結果を表 1 に示す．タームの種類は全ての項目において既存語彙よりも独自語彙の方が多かった．また，LOD チャレンジ Japan におけるプロパティでの統計において，既存語彙のタームの総数が独自語彙より多かったのは，既存語彙を主として使用しているデータセットのリソース数が他と比べて多かったためである．

(2) XLS, CSV での調査において，テーブルデータに使用されている語彙を集計した結果，同じ意味を表す語でも異なる表記で記述されているものがあることが分かった．例えば，電話番号を表す語には，“電話番号”以外にも，“連絡先”や“Phone”，“Tel”などが用いられていた．これらの語に対しては表記が違っていても同じメタデータ語彙を付与することができる．その例を，調査結果を一部用いて表 2 に示す．スキーマに関する調査では，スキーマが公開されているデータセットはなかったが，表形式であるためスキーマの作成は可能である．しかし，テーブルの値が何であるのかをデータ中に記述していないデータセットもあったため，スキーマの作成が困難なデータも存在した．また，他のデータセットへリンクが結ばれているデータセットはなかった．

(3) PDF での調査では，LOD チャレンジ Japan が PDF データを公開していなかったため，CKAN 日本語と Open Data METI の PDF データに対して分析を行った．その結果を表 3 に示す．Open Data として公開されている PDF データは，(1)文書，(2)テーブルデータ，(3)画像の 3 つのタイプに主に分類された．(1)文書タイプは，文書作成用ソフトで，(2)テーブルデータは表計算ソフトで作成したデータを，PDF として出力したタイプのものを意味し，(3)画像タイプは，対象であるリソースを地図にマークするといったような，画像にマッピングしたタイプのものを意味する．また，今回調査した PDF データに対してスキーマを定義したデータセットはなかった．これらのタイプに対してスキーマの作成を考えた場合，(1)文書と(3)画像タイプではデータとしての構造が複雑になるため，スキーマの作成はできない．ただし，(2)テーブルデータに関しては PDF の解析ツールでテーブルデータの項目名や値を取得できれば XLS などと同様に扱えるので，スキーマの作成は可能である．また，他のデータセットにリンクしたデータセットはひとつもなかった．

(4) XML での調査では，タグに用いられる語彙で分類を行えたが，すでに行った XLS，CSV の分析と特徴が同様なため割愛する．またスキーマの分析では，CKAN 日本語の 2 つの Open Data においてスキーマ定義のあるデータセットがあった．一つは福井県坂井市企画情報課による“避難所”というデータセットで，このデータは主に地理情報のための標準である JSGI Cyber Japan Profile 2003 [35]の定義をもとに作成されていた．もう一つは気象庁による“横浜市の気象情報”というデータセットで，このデータは気象庁防災情報 XML フォーマット Ver1.0 [36]の定義をもとに作成されていた．その他のデータセットではスキーマ定義のされたものはなく，また他のデータセットへとリンクで結ばれたデータセットもなかった．

表 1 RDF データにおけるタームの統計

名称	クラス／プロパティ	既存語彙／独自語彙	タームの種類数	タームの総数
CKAN 日本語	クラス	既存語彙	2	129
		独自語彙	5	370
	プロパティ	既存語彙	13	4581
		独自語彙	137	12810
LOD チャレンジ Japan	クラス	既存語彙	2	1815
		独自語彙	9	39085
	プロパティ	既存語彙	54	347739
		独自語彙	94	272192

表 2 テーブルデータの語の表記による
メタデータ語彙の割り当て例

意味	語	ターム例
名前	名称，作品名，タイトル，項目名，書名	dc:title, rdf:label
電話番号	電話番号，連絡先，Tel，Phone	v:tel, foaf:phone
地名	地名，所在地，町・大字名	gn:name
緯度	緯度，lat，緯度（北緯），緯度（世界測地系）	geo:lat, v:latitude
日時	公開日，投稿日時，発言日，発売日，収拾日時	dc:date

表 3 PDF の内容による分類

名称	文書	テーブルデータ	画像	(NOT Found)
CKAN 日本語	3	6	4	2
Open Data METI	125	6	0	0

(5) HTML での調査では，Web ページの内容として，データセットを公開しているプロジェクトのホームページ，データセットのダウンロード用のページ，データセット内の情報検索用のページの 3 種類あることが分かった．しかし，その全てでスキーマを定義したものはなく HTML 内の構造が複雑なため，スキーマの作成も困難である．また RDFa などで HTML 文書にリンクを表すメタデータを埋め込んだものもなかった．

以上のフォーマットごとにおける(1)スキーマの分析，(2)リンク関係の調査をまとめたものが表 4 である．表 4 では，スキーマ定義と他のデータセットへのリンクがされている Open Data の数を取り組みごとに表している．また Open Data METI の元のデータセット数は 201 であるのに対し，表 4 でのデータセットの合計が 209 になっているのは，同じデータセット内で全く内容の違うデータが異なるファイルフォーマットで公開されていることがあったためである．

ライセンスの調査結果は表 5 のようになった．表 5 では，調査対象である 3 つの取り組みに対し，ライセンスに対応するデータセットの数を，クリエイティブ・コモンズが定めるデータの再利用に対する自由度の高い順に左から提示している．Open Data METI と LOD チャレンジ Japan ではライセンスの明記がデータセット登録の条件のため，ライセンスが全てのデータセットで明記されていた．しかし，CKAN 日本語ではライセンスの明記がデータセット登録の条件になっていないため，三分の二以上がライセンス不明になっていた．またライセンスの分類では，データの原作者のクレジット（氏名，作品のタイトルなど）を表示することを条件にしているデータセットが全体的に多いことが分かった（表 5 の“表示”の列を参照）．一方 LOD チャレンジ Japan では，全ての権利を主張したデータセットもあり，全体として利用における制限が厳しいものもいくつか見られた．

また，LOD クラウドのカテゴリ分類を用いた，データセットの分類結果を表 6 に示す．傾向として，CKAN 日本語では公開しているデータセットが政府情報に偏っていることが分かった．LOD チャレンジ Japan では，政府情報のデータセット数が最も多いが，どのカテゴリにおいてもデータセットが公開されていることが分かった．

5. 考察

今回の調査の結果から，公開されているスキーマや他のデータセットへのリンクが行われている Open Data が非常に少ないことが分かった．よって，このままでは Open Data を LOD として利用するのは難しいと考えられる．

Open Data を LOD として公開するにあたっての段階を示した例として，Tim Berners-Lee が提示した 5 star deployment scheme がある [37]．この例では LOD を含めた Open Data としての利用しやすさを次の 5 段階で表している．

(1) オープンライセンスである（例:PDF）
(2) 構造化データである（例:XSL）
(3) オープンフォーマットである（例:CSV）
(4) リソースに URI を使用している（例:RDF）
(5) 他のデータセットへのリンクがある（例:LOD）

今回の調査結果をこの指標に当てはめて考えると，スキーマ定義が少なかったこと，他のデータセットへのリンクが少なかったことから，特に(3)から(4)，(4)から(5)の段階へと変換する過程が重要であると考えられる．

(3)から(4)の段階へとデータを移行するには，CSV などのテーブルデータで使用される語に対して URI を割り当てる必要がある．また，今回の調査からそのようなメタデータに割り当てることができそうなメタデータ語彙があることも分かった．よって，メタデータの作成時にタームを推薦してくれるようなシステムがあれば，CSV ではなく RDF

表 4 スキーマとリンクの調査結果

名称/フォーマット		RDF	CSV XLS	PDF	XML	HTML	合計
CKAN 日本語	**Schemas**	**0**	**0**	**0**	**2**	**0**	**2**
	Links	**2**	**0**	**0**	**0**	**0**	**2**
	Datasets	22	74	11	6	11	114
Open Data METI	**Schemas**	**0**	**0**	**0**	**0**	**0**	**0**
	Links	**0**	**0**	**0**	**0**	**0**	**0**
	Datasets	0	180	12	6	11	209
LOD チャレンジ Japan	**Schemas**	**4**	**0**	**0**	**0**	**0**	**4**
	Links	**7**	**0**	**0**	**0**	**0**	**7**
	Datasets	44	9	0	4	2	59
Total Dataset		66	263	23	16	24	392

表 6 LOD クラウドのカテゴリを用いたデータセットの分類

カテゴリ／名称	CKAN 日本語	LOD チャレンジ Japan
政府情報	120	15
ユーザ生成コンテンツ	9	6
書誌情報	2	6
クロスドメイン	1	2
地理情報	0	11
メディア情報	0	5
生命科学	0	5

表 6　クリエイティブ・コモンズのライセンス情報を用いた分類表

名称＼ライセンス	パブリックドメイン	表示	表示 継承	表示 改変禁止	表示 非営利	表示 非営利 継承	表示 非営利 改変禁止	全ての権利を主張	ライセンス明記なし
CKAN 日本語	4	35	3	2	0	0	0	0	107
Open Data METI	0	171	0	30	0	0	0	0	0
LOD チャレンジ Japan	40	25	6	2	3	4	3	3	0

として公開されるデータセットが増えると考えられる.

(4)から(5)の段階へとデータを移行するには，データをRDF として記述するだけでなく，データ中のメタデータを他のデータセットへリンクで結ぶことが求められる．そのためには，メタデータのリソースが他のメタデータのリソースと同じであることを知らなければならない. つまり，LOD として公開される RDF データを増やすためには，リソースの同定を行うような支援が必要であるということが考察できる.

ライセンスに関する調査結果からは，ライセンス明記のない Open Data が，再利用や再配布が行えるか明確でないため, 最も問題であると考えられる. この問題に関しては，Open Data METI や LOD チャレンジ Japan の結果から伺えるように，ライセンスの明記を義務付けるような仕組みが予め必要になると考えられる．また，Open Data としてのライセンスは，クレジット表示とライセンスの継承を主張する程度の条件であるべきだと言われている [38]. 今回の調査でライセンスの明記をしているデータセットのほとんどがこの範囲内であったが，改変禁止を主張しているものも多数あることが分かった．この改変禁止に関しては，Open Data METI を提供している経済産業省においても，Open Data の条件として適用範囲を広げるべきかどうかの議論を行っている [39]. このような Open Data としてのライセンス条件の知識は，今後は一般的に認知させていく必要があると考える.

6. おわりに

本研究では，日本で公開されている Open Data を LOD として活用することを目的に，日本の Open Data のスキーマの分析とリンク関係の調査をファイルフォーマットごとに行った．スキーマの分析ではデータセットとしてのスキーマ定義を公開している Open Data が少なかった．また，他のデータセットへのリンクをしているものもほとんどなかったため, 日本の現状の Open Data を LOD として利用することは難しいということが分かった．そこで，本研究で提示したような，テーブルデータで利用されている語彙などにURI を割り当てるといったメタデータ語彙の推薦が今後は重要になると考察した．また，ライセンスに関する統計を取り，現在の Open Data がどの程度自由度を有しているかを分析した．特にライセンスの明記を義務としていない場合，明記のないまま多くのデータセットが公開されてしまうことがある．このような事態を防ぐためにも，ライセンスを明記する環境作りが大切である.

本研究では, 日本の Open Data を調査し, 現状の Open Data を LOD として利用するために必要となる要件を考察した. 今後は今回の調査をもとに，Open Data を LOD として利用するために，一般語彙に対するメタデータ語彙の推薦や，リソースを同定してリンク付けを行うシステムの構築を行う.

謝辞　本研究の一部は平成 23 年度日本学術振興会科学研究費補助金（課題番号:23500295）による.

参考文献

1) Open Data Handbook オープンデータとは何か？, http://opendatahandbook.org/ja/what-is-open-data/
2) CKAN 日本語, http://data.linkedopendata.jp/
3) Open Data METI, http://datameti.go.jp/data/
4) Linked Data, http://linkeddata.org/
5) New York Times - Linked Open Data, http://data.nytimes.com/
6) GeoNames Semantic Web, http://www.geonames.org/ontology/documentation.html
7) State of the LOD Cloud, http://lod-cloud.net/state/
8) 加藤文彦 オープンデータの技術よりの話, http://www.slideshare.net/fumihiro/20130620-23239372
9) Web NDL Authorities, http://id.ndl.go.jp/auth/ndla
10) Data.gov State Data Sites, http://data.gov/
11) data.gov.uk, http://data.gov.uk/
12) data.gov.fr, http://www.data.gouv.fr/
13) Rennes métropole en acces libre, http://www.data.rennes-metropole.fr/
14) Montpellier Terrtoire Numérique, http://opendata.montpelliernumerique.fr/
15) Open Data Sites, Data.gov, http://www.data.gov/opendatasites
16) Linked Open Challenge Japan 2013,

http://lod.sfc.keio.ac.jp/challenge2013/
17) LinkData, http://linkdata.org/
18) メタデータ情報基盤構築事業　メタデータ情報共有のためのガイドライン
http://www.soumu.go.jp/main_content/000132512.pdf
19) FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/
20) LODStats, http://stats.lod2.eu/
21) Linked Open Vocaburary, http://lov.okfn.org/dataset/lov/
22) the Datahub, http://datahub.io/
23) SPARQL 1.1 Property Paths,
http://www.w3.org/TR/sparql11-property-paths/
24) 青空文庫 Linked Open Data,
http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/
25) 京都国際マンガミュージアム書誌情報 LOD,
http://mdlab.slis.tsukuba.ac.jp/lodc2012/kmm/
26) LODC Works
http://mdlab.slis.tsukuba.ac.jp/lodc2012/lodcworks/
27) 落語家 LOD, http://mdlab.slis.tsukuba.ac.jp/lodc2012/rakugo/
28) ヨコハマ・アート・LOD, http://archive.yafjp.org/artsearch/
29) 横浜ごみ分別情報,
http://data.linkedopendata.jp/dataset/yokohama-gomi
30) LSJ: Location Site of Japanimation, http://cheese-factory.info/
31) 日本語 Wikipedia オントロジー,
http://www.wikipediaontology.org/
32) Allie, http://data.allie.dbcls.jp/
33) Biomasspedia, http://biomasspedia.net/
34) saveMLAK, http://savemlak.jp/wiki/saveMLAK
35) CSV->JSGI 電子国土プロファイル形式変換,
http://cyberjapan.gsi.go.jp/csv_converter/csv_converter.html
36) 気象庁防災情報 XML フォーマット, http://xml.kishou.go.jp/
37) Linked Data - Design Issues,
http://www.w3.org/DesignIssues/LinkedData.html
38) Open Definition, http://opendefinition.org/okd/
39) 平成 24 年度 DATA METI 構想の成果報告（案),
http://www.meti.go.jp/committee/kenkyukai/shoujo/it_yugo_forum_data_wg/pdf/005_06_01.pdf