自治体が公開しているRDFの現状と課題

Current Issues of RDF Published by Local Governments

秋山 梓[1]　加藤 文彦[2]　小出 誠二[3]　海沼 靖夫[1]

Azusa Akiyama[1], Fumihiro Kato[2], Seiji Koide[3], and Yasuo Kainuma[1]

[1]株式会社 日立システムズ
[1] Hitachi Systems, Ltd.
[2]情報・システム研究機構
[2] Research Organization of Information and Systems
[3]国立情報学研究所
[3] National Institute of Informatics

Abstract: In this paper, we describe the results of survey on 1,200 RDF contents published by 51 local governments in Japan at the time of writing. On the way of analyzing them we have extended and refined the four principles of LOD into 11 principles, and the statistical data are ranked on the base of the new principles. The results show not only the usefulness of the new principles but also plenty of room for the activity of systematization specifically on the *content negotiation* for LOD.

1 はじめに

本稿では，自治体が公開しているRDFについて調査した結果を報告する．自治体では，政府推進のもと，自治体が保有するデータをオープンデータとして公開する取り組みが広がっている．その中でも先進的な自治体においては，5つ星オープンデータであるLinked Open Data (LOD) [1]をめざすところが徐々に増えつつある．

増加の背景には，LODを実現するRDFの作成・公開においてLinkData [2]を代表とする，特別の知識がなくともRDFを作成することが可能なサービスが普及していることが大きな要因だと考える．LinkData等を利用することで技術的・コスト的なハードルは下がったものの，現時点で公開されているデータは，自治体ごとに異なる実装・語彙を利用していたり，LODの目的であるデータのつながりがないものが多数存在したりしている．したがって，公開されているRDFを利用するにはデータごとに再加工が必要となり，CSVデータを公開しているのとあまり変わらない状態である．

本稿では，このような実態に警鐘を鳴らし，本来あるべき姿のLODへ導くため，LODの4つの原則[1]を基本とする「LOD実現のための11の基準」を提案する．それに基づき自治体が作成したRDFの現状を調査することで，自治体がデータを公開するうえで今後留意すべき課題について述べる．

本稿の構成は以下のとおりである．2章で自治体のRDF公開状況を述べ，3章で各基準についての自治体の対応状況と誤りの事例を述べ，4章で課題について述べ，5章でまとめとして今後の展望について述べる．

2　RDFを公開している自治体

福野氏が公開している日本のオープンデータ都市一覧 [3]によると，2015年12月時点において，オープンデータとしてデータを公開する自治体は全体の約10%にあたる約170団体である．そのうち，自治体のカタログサイトやウェブサイト，あるいは外部のサービスを利用して RDF を公開している自治体を調査したところ，表1のとおり自治体数は51団体であり，RDFのデータ数は約1,200件であった．オープンデータを公開する団体のうち，43 団体がLinkDataのサービスを利用している[4]．その43団体のうち，1 団体は自サイトへのリンクのみであっ

たが, 42 団体は LinkData のサービスを利用して RDF を公開している.

本調査では，上記約 1,200 件の RDF を確認し，特徴の抽出を行った．なお 50 件以上の RDF を公開している団体については，同じパターンのデータを除いた後の一部のデータのみを確認した.

表 1 RDF を公開している主な自治体

1	北海道室蘭市
2	北海道森町
3	北海道八雲町
4	宮城県石巻市
5	秋田県横手市
6	福島県会津若松市
7	埼玉県川口市
8	埼玉県和光市
9	埼玉県北本市
10	千葉県流山市
11	千葉県我孫子市
12	東京都品川区
13	東京都杉並区
14	神奈川県鎌倉市
15	新潟県三条市
16	新潟県十日町市
17	新潟県見附市
18	新潟県糸魚川市
19	富山県砺波市
20	福井県
21	福井県福井市
22	福井県敦賀市
23	福井県鯖江市
24	福井県越前市
25	長野県上田市
26	長野県須坂市
27	長野県駒ヶ根市
28	長野県中野市
29	長野県塩尻市
30	長野県軽井沢町
31	静岡県静岡市
32	静岡県三島市
33	静岡県島田市
34	静岡県磐田市
35	静岡県掛川市
36	静岡県裾野市
37	静岡県御前崎市
38	愛知県長久手市
39	三重県津市
40	滋賀県大津市
41	大阪府高槻市
42	大阪府枚方市
43	兵庫県神戸市
44	鳥取県
45	島根県松江市
46	岡山県玉野市
47	広島県呉市
48	山口県宇部市
49	山口県周南市
50	香川県高松市
51	熊本県菊池市

3 自治体 RDF 調査

3.1 調査基準

事前調査として，数件の RDF を確認した結果，現在公開されている RDF の LOD 実現度がさまざまであり，LOD の 4 つの原則を基準とするだけでは実態がつかめないことが分かった. 現在の RDF 公開について実態を把握するため, 本研究では 4 つの原則に, 7 つの基準を追加することで，表 2 に示す「LOD 実現のための 11 の基準」を定義して，調査に用いることとした.

表 2 LOD 実現のための 11 の基準

No	基準名	内容
1	機械判読	機械判読可能であること
2	IRI 付与	事物の名前として IRI が付与されていること
3	HTTP IRI の利用	名前を参照できるよう, HTTP IRI を用いていること
4	共通の語彙の利用	共通の語彙を利用していること
5	語彙の正確な利用	語彙の利用が正確であること
6	IRI のドメイン	リソースのドメインがリソースを管理する自治体の所有するドメインであること
7	IRI の参照	主語 IRI の参照する先が存在すること
8	HTTP コンテントネゴシエーション	HTTP コンテントネゴシエーションにより, 要求に応じた形式で情報提供をしていること
9	アウトバウンドリンク	外部 IRI へのリンクを含むこと
10	実在する IRI へのアウトバウンドリンク	参照解決可能な外部 IRI へのリンクを含むこと
11	RDF リソースへのアウトバウンドリンク	参照解決可能な RDF リソース IRI へのリンクを含むこと

3.2 調査結果

3.2.1 機械判読

多くの自治体が，CSV 等のデータから RDF を作成しているため，基本的には機械判読可能なデータであった．しかし，元データからのデータ変換時のミスと推測されるデータが数件存在した．一つ目の事例は，図 1 に示すように 1 レコード全てを 1 文字列として扱ってしまっている例である．RDF の形式にはなってはいるものの，目的語として 1 つの文字列の中に複数データを含んでしまっており，機械判読がしにくい状態であった.

二つ目の事例は，図 2 に示すように，日付項目に整数値が設定されている例である．Excel では，日付データをシリアル値で管理しているため，作成方法によっては, 該当の RDF のように日付項目にシリアル値が設定されてしまう．そのため，日付の値としては適切ではない状態であった.

3.2.2 IRI 付与

LOD 4 原則の一つ目の原則として定義されるように，事物を特定するためには，事物の名前として IRI が付与される必要がある．本調査では，異なる事物に同じ IRI を付与している事例が複数存在した．それらはデータの作成方法により，2 つのグループに分けられた．

一つ目のグループは，同じデータセット内の異なる RDF ファイルにおいて，同じ IRI を付与しているグループである．この特徴は，データセット内に複数の RDF ファイルを持つ形態において，数件みられた．「データセット名＋1 列目の値」を各リソースの主語として定義しているため，複数の元ファイルにおいて，ファイルの 1 列目の値として通し番号等の同一の値が設定されている場合，異なる事物を指す場合でも同じ IRI が付与されてしまっている．

二つ目のグループは，すべてのファイルに同じ IRI を付与しているグループである．このタイプは独自のカタログサイト上に公開されている RDF において多くみられた特徴である．公開されている RDF のすべてのデータにおいて，ファイルごとに 1 から通し番号が付与されている．そのため，異なる事物を指す場合でも，ファイルが異なった場合は，そのファイル内での出現位置が同じであれば，同じ IRI が付与されている．

2 つのグループはともに，IRI が世界中で一意の情報を提供する手段[5]という理解が不足しているために発生する誤りだと考える．

3.2.3 HTTP IRI の利用

LOD 4 原則の二つ目の原則に定義されるように，Web 経由で情報を取得できるようにするために，事物の名前には HTTP IRI を用いる必要がある[5] ．今回の調査では，IRI に HTTP IRI ではなく urn:スキームを利用しているデータが 1 件存在した．

3.2.4 共通の語彙の利用

共通の語彙が利用できる場合には，できるかぎりデータの記述にそれらを再利用すべきである [5] ．その語彙を利用することで，語彙を対応づけるコストを削減し，利用を促進することができる．多くの自治体が，既存の語彙を利用していたが，独自語彙のみで記述をしている団体も数件見られた．利用されている語彙は，主に以下の語彙であった．

・共通語彙基盤
・W3C Basic Geo
・ダブリン・コア
・FOAF
・Schema.org

3.2.5 語彙の正確な利用

共通の語彙を利用する場合，その語彙の定義に従って利用をしないと，外部からの参照時に混乱をきたすことになる．多くの自治体が既存語彙を定義通りに利用していたが，複数の団体において，誤って利用しているデータがあった．ここでは，代表的な 4 つを例示する．

(1)階層構造を持つ語彙

共通語彙基盤[6]や Schema.org[7]のように階層構造を持つ語彙を，その構造を意識せずに記述している，というものである．特に，共通語彙基盤は，体系化，階層構造化により，正確に物事を表現することで，同じ単語を違う意味で使うことによる誤解や，違う単語を同じ意味で使うことによる意思疎通の不便さを解消することを目的として整備されている[6]．今回調査した団体において，共通語彙基盤を利用していた団体は 8 団体あり，そのうち半分にあたる 4 団体のデータに誤りが見られた．そのため，同じ共通語彙基盤の語彙を利用しているにも関わらず，自治体間で構造が異なり，情報連携ができない状態となっている．

(2)オントロジー・ヘッダ

owl:Ontology は通常，クラス定義をする文書上で該当文書そのものを説明するために用いられるが，インスタンスに関して記述された RDF に対して，単に RDF 文書を説明するために「owl:Ontology」を用いていた．

(3)ダブリン・コア

DCMI Metadata Terms で定義された語彙を図 3 のように誤った IRI として利用していた．created 要素は，DCMES の date を精密化した DCMI Metadata Terms の語彙である．したがって，IRI としては異なるものであることを意識して利用する必要があるが，ここでは DCMES と DCMI Metadata Terms を混ぜ合わせたような IRI になっていた．

(4) 言語タグ

四つ目は，言語タグの付与誤りである．中国語等の日本語以外の言語のリテラルに対し，該当言語に関係なく，すべて同じ誤った言語タグを付与しているため，言語の特定ができない状態であった．言語タグは，ISO639-1 の言語コードに従い，日本語リテラルなら”ja”の言語コードを付与し，中国語リテラルなら”zh”と正しく付与することで，SPARQL での問い合わせ時に，言語を指定して情報を取得することが可能となる．

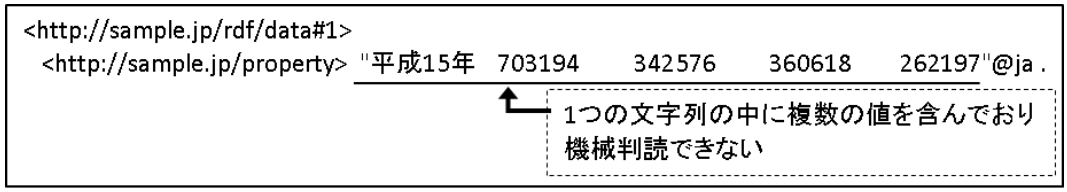

図1 機械判読できない文字列の例

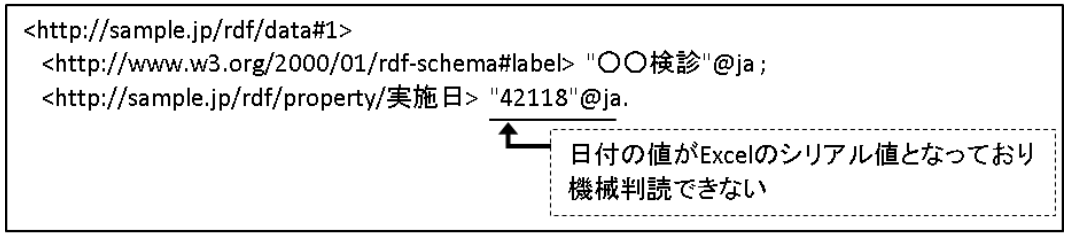

図2 機械判読できない日付の例

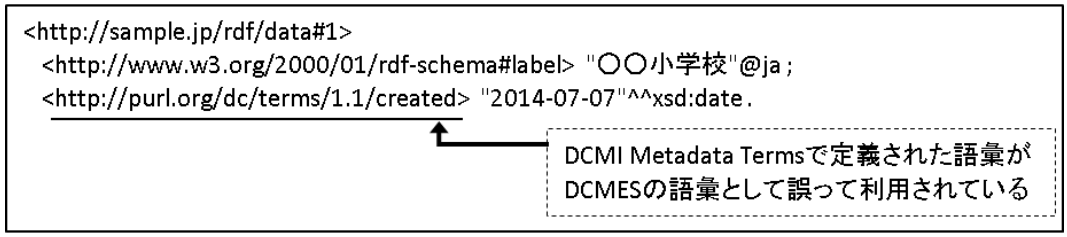

図3 ダブリン・コア語彙の誤った利用例

3.2.6 IRI のドメイン

IRI のドメインは, 事物の管理者を明確にするために，自治体管理下のドメインを利用することで，信頼性を持たせることができる．8 団体が自治体のドメインにて事物に IRI を付与していた．また，LinkData にて公開している自治体が多いことから，多くのデータは LinkData.org ドメインを利用していた．その他，場所を指定するリソースに Wikipedia のページ URI を指定している団体があった.

自治体が管理する事物には, 自治体ドメインで IRI を定義し，owl:sameAs を利用して，Wikipedia のページ URI ではなく，例えば DBpedia のリソース IRI に対し, リンクを含める構成が良い. そうすることで，事物の管理者を明確にし，他のデータとのつながりを持たせることが可能となる. 一方,「http://examples/」というドメインにて URI を利用する自治体もあった. 3.2.7 で後述するが，URI 参照時に何が返されるかをコントロールすることや，永続的にこの URI を維持するためには，管理できないドメインを利用することは避けるべきである[5].

3.2.7 IRI の参照

LOD 4 原則の三つ目の原則は，IRI にアクセスされた際に有用な情報を, WEB 標準技術を用いて提供することである．今回は，範囲を限定し，まずは主語の IRI について参照する先が存在するかの調査を行った. LinkData.org ドメインを利用してデータを公開している自治体は，LinkData により IRI を参照するページが用意されているため, IRI を参照することができる．一方，自ドメインを利用している自治体は, 参照先のサイトを独自で用意する必要があるが, LinkData 上で自ドメインのリソースを定義し，参照先も備えている自治体は 2 団体存在した.

3.2.8 HTTP コンテントネゴシエーション

前述 3.2.7 で主語の IRI の参照先が存在するとして, RDF を解釈できるプログラムからのアクセスには適切な RDF を解決して返す必要がある．また，ウェブブラウザからのアクセスには，ブラウザが理解できる HTML を返すのが望ましい[8]. 3.2.7 で IRI にアクセスされた際に情報提供を実現している団体は 2 団体あったが, いずれも HTML としての情報提供であり，コンテントネゴシエーション機能を有して，情報の提供を行っている自治体は今回の調査では見つけることができなかった.

3.2.9 アウトバウンドリンク

LOD 4 原則の 4 つ目の原則は，外部の IRI へのリンクを含むことであり，データをつなげて多くのことを発見することが LOD の目的である. 内部のデータとのリンクだけでなく，他の機関が提供するデー

タへリンクをしている自治体は，8 団体しかなかった．主なリンク先は以下のとおりである．

・統計局の標準地域コード [1]
・クリエイティブ・コモンズ [2]
・DBpedia Japanese [3]
・GeoNames.jp [4]

3.2.10 実在する IRI へのアウトバウンドリンク

前述 3.2.9 で記述したアウトバウンドリンクのうち，統計局の標準地域コードのリソースは，現時点では実在せず，参照はできない．これは自治体単独で解決できるものではないが，参照できることが望ましい．実在する IRI へのアウトバウンドリンクを実現している自治体は 2 団体であった．

3.2.11 RDF リソースへのアウトバウンドリンク

前述 3.2.10 で記述したアウトバウンドリンクのうち，リンク先がクリエイティブ・コモンズのものは，参照先が web のページである．RDF リソースにリンクしている団体は 1 団体だけであり，リンク先は DBpedia Japanese と GeoNames.jp であった．

3.2.12 集計結果

各基準を満たす自治体数を表 3 に示す．今回調査した 51 の自治体の多くが，基準を満たす RDF と基準を満たさない RDF の両方のタイプを公開している．ここでは，今回調査した RDF について，公開している全ての RDF が基準を満たす自治体の数と，一部の RDF が基準を満たす自治体の数，基準を満たす RDF が 1 つも見つからなかった自治体の数を示す．

表 3 RDF が基準を満たす自治体数（n=51）

No	基準名	全て満たす	一部満たす	全て満たさない
1	機械判読	49	2	0
2	IRI 付与	34	14	3
3	HTTP IRI の利用	50	1	0
4	共通の語彙の利用	19	23	9
5	語彙の正確な利用	30	11	10
6	IRI のドメイン	4	4	43
7	IRI の参照	38	4	9
8	HTTP コンテントネゴシエーション	0	0	51
9	アウトバウンドリンク	5	3	43
10	実在する IRI へのアウトバウンドリンク	0	2	49
11	RDF リソースへのアウトバウンドリンク	0	1	50

[1] http://statdb.nstac.go.jp/system-info/api/api-spec/
[2] http://creativecommons.org
[3] http://ja.dbpedia.org
[4] http://geonames.jp

4 課題

今回，オープンデータに先進的に取り組む自治体の RDF に対し調査を行ったが，11 個の基準を全て満たすデータを見つけることはできなかった．特にシステム的な対応が必要な HTTP コンテントネゴシエーションの普及には課題が残る．また，11 の基準以外に，RDF として誤りではないものの，運用上考慮したほうがよい点が 3 つあった．

一つ目は，IRI の設計についてである．図 4 のように，事物のリソースの IRI 文字列中に「property」の単語が入った RDF が複数あった．IRI 文字列に本来意味は存在しないが，プロパティとして用いないリソースの IRI に「property」が入っているのは，RDF としては混乱するので望ましくない．また「*.ttl」のように，ファイルの拡張子を含む IRI をリソースの主語として利用している団体も存在した．IRI にファイルの拡張子を含んだ場合，その他の形式による返答も実装する場合に利用しづらい IRI となる．データ利用者の利便性やデータの意味を損なわない IRI 設計が必要である．

二つ目は，図 5 のように，ラベル項目「rdfs:label」に通番と名称が定義されているものである．データを利用する際に，一つの IRI に対して異なる二つのラベルがあると，どちらを利用すべきか機械的に判断することが難しい．通番には DCMI Metadata Terms の identifier を利用するなどの検討が必要である．

三つ目は，語彙の揺れである．自治体が公開するデータの種類の多くは，施設情報やイベント情報など各自治体に共通である．現在は，同じ種類のデータでも，各自治体がそれぞれ独自のルールで RDF を公開しているため，語彙の揺れが生じている．形式としては RDF であるが，より再利用可能な LOD を実現するための RDF とはなっていない．自治体が語彙の揺れをなくすために使える語彙として共通語彙基盤があるが，それを用いて RDF を作成するために今後どのような施策が有効か検討したい．

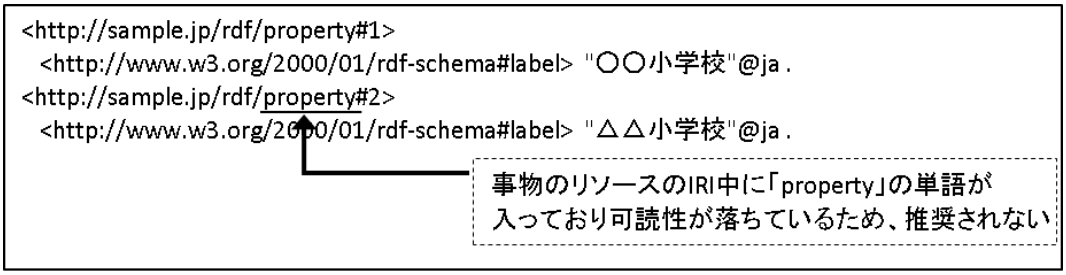

図4 リソース名のIRIにpropertyの文字が入っている例

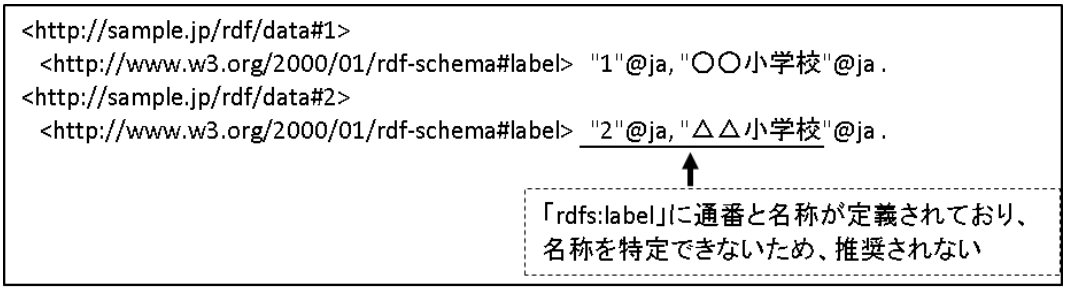

図5 ラベルに複数の値を付与している例

5 おわりに

本稿では，LOD の目的であるデータのつながりを実現できていない RDF が，自治体から多数公開されている実態に警鐘を鳴らし，本来あるべき姿の LOD へ導くため，「LOD 実現のための 11 の基準」を提案し，自治体が作成した RDF の現状を調査した．その結果明らかになった事柄から，自治体が今後留意すべき課題について述べた．自治体が RDF を公開する際は，本稿で紹介した「LOD 実現のための 11 の基準」をチェックリストとして利用することで，LOD の目的に従った RDF を作成，提供することが可能となる．今後は，各自治体がこのチェックリストの理解を深め，正しい LOD の普及・利用が広がるような活動に取り組む．

参考文献

[1] Berners-Lee, T. Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html, (2009)
[2] 一般社団法人リンクデータ, LinkData.org, http://linkdata.org/
[3] fukuno.jig.jp，日本のオープンデータ都市一覧, http://fukuno.jig.jp/2013/opendatamap
[4] 下山 紗代子, LinkData.org によるオープンデータのスタートアップ支援, http://www.sigswo.org/papers/37program, (2015)
[5] Tom Heath, Christian Bizer（武田英明監訳）, Linked Data:Web をグローバルなデータ空間にする仕組み, 近代科学社, (2013)
[6] IPA 独立行政法人情報処理推進機構，共通語彙基盤整備事業, http://goikiban.ipa.go.jp
[7] Schema.org, https://schema.org/
[8] Berners-Lee, T. Cool URIs don't change, http://www.w3.org/Provider/Style/URI.html, (1998)