人工知能学会研究会資料
SIG-SWO-A1101-03

Linked Data作成支援ツールの現状と課題

Current issues of software tools to create Linked Data

加藤 文彦 [1*]

[1] 国立情報学研究所
[1] National Institute of Informatics

Abstract: The Web of Linked Data has been growing and already contained hundreds of interconnected data sources. An easy way to transform various raw data into Linked Data to publish and link to existing data sources would be needed to evolve further the Web of Linked Data. This paper introduces software tools to support a user to generate Linked Data from raw data and discusses about current issues of these tools.

1 はじめに

近年，World Wide Web 上で異なる情報源のデータを型付リンクによって結びつける，Linked Data [1, 2] というコミュニティ活動が注目されている．Wikipedia を Linked Data にした DBpedia [3] を中心として，辞書情報や地理情報，政府情報，学術情報，音楽情報，写真情報など，2010 年 9 月の時点で 203 のデータセットによるネットワークが形成されている．Semantic Web が普及しない原因の一つとして基盤となるデータの欠如があったが，Linked Data の活動はその問題を解決する有望な手段の一つであると考えられている．

Linked Data の基礎となるデータモデルは，Semantic Web と同様に RDF である．そのため，Linked Data を作成して提供するというのは，データを RDF モデルに沿う形式にして Web 上で公開するということになる．加えて Linked Data では，RDF データの公開方法に一定のルールを設けている．これにより，Linked Data は Web アーキテクチャと整合性を保ったデータ共有の仕組みとなっている．

著者らが行っている LODAC Museum [5, 6] では，博物館情報や位置情報を主な対象とした RDF データを作成しており，それを Linked Data として Web で公開している．しかし，基としているシソーラスや各博物館の情報等は元々RDF データではなく，Excel や HTML といった形式で提供されていることが殆どである．現状ではこれら一つ一つのデータについて，RDF に変換して必要な箇所にリンクを張るプログラムを用意しなければならない．

LODAC Museum の将来像としては，生のデータを保持している博物館側も Linked Data の提供に参加することで，博物館情報が Linked Data Cloud 上の様々な異種情報と緩やかに繋がっていくことができるようになり，それによって新たな知が発生することを期待している．しかし，Linked Data を作るにはプログラミングが必要であるとなると，元々公開システムを作っているところであれば対応できるかもしれないが，そうではない一般のユーザには取り組みにくい．生データを作成しているユーザは，Excel のようなスプレッドシートや関係データベースでデータを管理している場合が多いため，これらから手軽に Linked Data を作成して提供するための支援ツールが，Linked Data の普及に重要な役割を持つと考えている．

そこで，著者らが運営している LinkedData.jp [4] の第一回勉強会のテーマとして，生データから Linked-Data を作成するためのツールについて取り上げることにした．本稿では勉強会の内容を踏まえて，現在存在するツールをいくつか紹介し，どのような問題を抱えているかをまとめる．

2 LinkedData作成支援ツール

LinkedData 作成支援ツールには，大きく分けて生データから Linked Data を作成するツールと，新しくデータを入力することで Linked Data を作成するツールがある．前節で述べたとおり，我々が想定しているのは既に生データがある場合なので，本節では生データから Linked Data を作成するツールを対象とする．

*連絡先：国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: fumi@nii.ac.jp

生データから Linked Data を作成する方法は，データの種類によって以下のように分類できる．この中では，我々の想定するユーザは主に 2 と 5 が必要であると考えられる．そこで勉強会では，2 の構造化されたデータの変換ツールに該当するものについて，主に取り上げた (表 1).

1. テキストデータの変換
2. 表や XML 等，構造化されたデータの変換
3. WebAPI のラッパー
4. CMS の拡張
5. データベースのラッパー

ReDeFer [7] は，RDF に関する変換ツール群である．XML を RDF に，XMLSchema を OWL へ変換することができる．また，RDF から HTML+RDFa や SVG への出力も同時に提供している．Garca ら [8] は ReDeFer を用いることで，B2B で使われている様々な XML を RDF にしてデータ統合をする手法を提案している．

XLWrap [9] は，Excel や OpenOffice Calc といったスプレッドシートで表現されているデータを RDF に変換するツールである．各セルや範囲に対して RDF へのマッピングルールを記述することによって，変換を行う．DBpedia へのマッピングルールも提供している．単独で HTTP サーバとしても動作可能で，スプレッドシートファイルに書かれているデータをそのまま SPARQL Endpoint として提供することも可能である．

Google Refine [10] は元々Freebase [11] で用いられていたデータクリーニングツールである．ローカルの Web サーバとして動作する．Google Refine に CSV データをアップロードした後に，各項目のデータの表記揺れをクラスタリングしてまとめるなど，クリーニングの機能が充実している．データを RDF に変換するためには RDF 用の拡張 [12] を用いれば可能である．

irON (Instance Record Object Notation) [13] は，RDF ではないデータを RDF にするための記法や語彙を定めている．JSON 用の irJSON，XML 用の irXML，CSV 用の commON が存在する．irON の各形式で記述したデータを OpenStructs システムに入力することで，RDF として出力することができる．

3 課題

本節では，Linked Data 作成支援ツールの課題について，国際化対応，語彙，外部へのリンクという観点から述べる．

ツール	入力形式
ReDeFer	XML,XML Schema
XLWrap	Excel,OpenOffice Calc,CSV
Google Refine	CSV
irON	XML,CSV,JSON

表 1: ツールと対応する入力形式

3.1 国際化対応

Auer ら [14] は，韓国語の DBpedia を作成した時の事例を基に，いくつかの Semantic Web ツールについて，以下の点を確認することで，国際化対応の現状をまとめている．

1. パーセントエンコーディング
2. アンダースコア
3. UTF-8/IRI サポート
4. IRI で問題のある文字列
5. RDF 出力形式

現在非 ASCII の文字列を URI で扱うためには，1 のパーセントエンコーディングを行うのが一般的である．しかしパーセントエンコーディングされた URI は可読性を損なう．DBpedia が Linked Data 普及のきっかけとなった理由の一つとして，URI の構成が人間にわかりやすいということが挙げられる．また，機械的な文字列マッチだけで存在が確認できるため，DBpedia へのリンク作成が容易である．別の問題として，RDF の形式の一つである RDF/XML においては%文字が XML タグ内で使用不可能なため，パーセントエンコーディングで表現されたプロパティの URI を RDF/XML 形式では使用することができない．ソースコード 1 は通常の RDF/XML パーサではエラーとなる．

```
1 <rdf:Description about="#example1">
2   <ex:%97%e1 xmlns:ex="http://example.org/"
3   >'例'のパーセントエンコーディング</ex:%97%e1>
4 </rdf:Description>
```

ソースコード 1: %問題

2 のアンダースコアは，パーセントエンコーディングされたプロパティの問題を回避するために Auer らが提案している方式である．まず，URI の最後に必ず'_'を付けるという約束事にする．そしてパーセントエンコーディングした URI を prefix として記述しておき，プロパティを表現する XML タグを prefix:_という形にする (ソースコード 2)．これにより，ツールによってはうまく処理できるという主張をしている．しかし，URI

に余計な文字を付加してしまうこの方法は，一般的な対応ではない．各プロパティ毎に prefix を割り当てる必要があることも問題である．

```
<rdf:Description about="#example2">
  <ex:_ xmlns:ex="http://example.org/%97%e1"
  >アンダースコア</ex:_>
</rdf:Description>
```

ソースコード 2: アンダースコア

上記のパーセントエンコーディングによる問題は，本来は URI の国際化版である IRI [15] を用いることで解決できるはずである．しかし，IRI をどの程度サポートしているかどうかは，まだプログラミング言語のライブラリやツール自体に依存するのが現状である．

Auer らと同じ基準で，2 節で述べたツールを検証した結果が表 2 である．IRI で問題ある文字に関しては，空白,{,},<,>,(,) を対象とした．

ReDeFer は RDF/XML のみを扱うため，パーセントエンコーディングの処理はできない．アンダースコアを追加して扱う形式では，ReDeFer が prefix の最後に勝手に'#' を付与してしまうため，URI 全体では'#_' を追加したことになってしまう問題がある．IRI は問題なく使用可能である．

XLWrap でもパーセントエンコーディングした URI は使用できなかった．しかし，アンダースコアを追加する形式では，パーセントエンコーディングありの URI を扱える．IRI は問題なく使用できる．

Google Refine では RDF のマッピング作成において問題があった．prefix ex を `http://example.org/ex#` としている場合，プロパティとして ex:test を入力すると `http://example.org/ex#test` が割り当てられるが，ex:テストを割り当てようとすると内部で`<ex:テスト>`に変換されてしまう．これは ex という scheme の IRI だと解釈されてしまうので，問題となる．`http://example.org/ex#`テスト と入力すると，正しく ex:テストと扱われる．また，IRI に空白文字がある場合は削除されてしまうという問題もあった．

irON については，パーセントエンコーディングの取り扱いは問題なかったが，IRI を使用することはできなかった．また，空白以外の記号は問題なく扱うことができたのも特徴的である．

3.2 語彙

生データを Linked Data にする際には，データの各項目を適切な語彙に割り当てることが望ましい．語彙が既存の良く知られたものであると，より再利用性が高まると考えられる．しかし，どの語彙が適切であるか，よく知られたものであるあるかの判断は難しい．Google Refine の RDF 拡張では，prefix uri の先に RDF Schema や OWL のファイルが存在する場合は，そこから語彙を認識して入力時に補完する機能がある．また，`http://prefix.cc` を利用して，良く利用されている prefix uri を推薦する機能もある．

一方，データの項目を URI と接続して，そのまま独自の語彙としてマッピングするという方法もある．今回取り上げたツールでは ReDeFer が該当する．この方法は機械的に変換できるため，手間をかけずに Linked Data として公開することができる．特に必要な語彙が大量な場合や未知の場合は既存の語彙に対応付ける作業が煩雑になるため，取り敢えず Linked Data にしてしまうほうが良いという場合は十分にあり得る．まず Linked Data として公開してあれば，利用者側が語彙の同意性などを判断してオントロジや SPARQL Construct クエリ，ルール言語などで，後から関係を追加することも考えられる．また，提供者側が後から既存の語彙や語彙の関係を追加することもできる．

3.3 外部へのリンク

あるデータがどこか別のデータへリンクしたり，どこかのデータからリンクされることによって，初めてそのデータは Linked Data としての価値が出てくる．そのためには，作成したデータを Web 上に公開して，関連する外部のデータへリンクをすることが望ましい．例えば XLWrap は DBpedia へのリンクを張るための簡便な方法を提供している．

Linked Data のデータセット数は年に 2-3 倍のペースで増加しているため，各自で関連するデータを発見してリンクをはるのが今後は困難になってくると考えられる．そのため，どこにリンクを張るべきかを推薦したり，自動的にリンクを張るなどの仕組みが今後重要となってくるであろう．特に DBpedia のようなデータのハブとなる Linked Data に，容易にリンクを張る方法があることが望ましいと考えられる．

4 おわりに

本稿ではスプレッドシートや XML のようにある程度構造化された生データから Linked Data を作成するための支援ツールを紹介して，その問題点について考察した．特に国際化からの視点での検討は，日本での Linked Data の普及には重要である．Web が普及した理由の一つは，誰もが手軽に文書を公開して，互いに勝手にリンクを張ることができたことである．同様に，Linked Data が普及するためには，Linked Data を誰もが手軽に作成して公開できる手段が必要となるであろう．我々も今後様々な分野で活用されることを目指して，有用なツールの開発に積極的に取り組んでいく．

ツール	パーセントエンコーディング	アンダースコア	UTF-8/IRIサポート	IRIで問題ある文字	RDF出力形式
ReDeFer	-	-	+	空白,{,}<,>,(,)	1
XLWrap	-	+	+	{,},<,>,(,)	1,2,3
Google Refine	◯	-	+	空白,{,},<,>,(,)	1,2
irON	+	+	-	空白	1,2,4

表 2: 国際化対応表
+: できる; ◯: できるが少し問題あり; -: できない
1: RDF/XML 2: Notation3 3: Turtle 4: N-Triples 5: RDF/JSON 6: HTML

謝辞

第 1 回 LinkedData 勉強会にて発表して頂いた清水智公氏，大澤昇平氏，中尾光輝氏，藤澤貴智氏，嘉村哲郎氏，佐久間勇樹氏，並びにご参加頂いた方々に深く感謝致します.

参考文献

[1] Berners-Lee, T.: Design Issues: Linked Data `http://www.w3.org/DesignIssues/LinkedData.html`, (2006)

[2] Linked Data - Connect Distributed Data across the Web `http://linkeddata.org/`

[3] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data *In Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, (2007)

[4] LinkedData.jp `http://linkeddata.jp/`

[5] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Open Data による多様なミュージアム情報の統合人文科学とコンピュータシンポジウム じんもんこん *2010,* 情報処理学会, (2010)

[6] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: LOD.AC: Linked Open Data によるミュージアム情報の結合第 *3* 回知識共有コミュニティワークショップ, 情報社会学会, (2010)

[7] ReDeFer `http://rhizomik.net/html/redefer/`

[8] Garca, R., Gil, R.: Facilitating Business Interoperability from the Semantic Web *In Proceedings of 10th International Conference on Business Information Systems*, Vol. 4439, pp. 220-232. Springer-Verlag (2007)

[9] Langegger, A., Wöß, W.: XLWrap - Querying and Integrating Arbitrary Spreadsheets with SPARQL. *In Proceedings of the 8th International Semantic Web Conference*, (2009)

[10] Google Refine `http://code.google.com/p/google-refine/`

[11] Freebase `http://www.freebase.com/`

[12] RDF Extension for Google Refine `http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/`

[13] irON: Instance Record and Object Notation Specification `http://openstructs.org/iron`

[14] Auer, S., Weidl, M., Lehmann, J., Zaveri, Amrapali J., C., Key-Sun: I18n of Semantic Web Applications *In Proceedings of the 9th International Semantic Web Conference*, (2010)

[15] Duerst, M., Suignard, M.: Internationalized Resource Identifiers (IRIs) *IETF RFC3987*, (2005)