LODジェネレータとライフサイエンス辞書のオープンデータ化

慶應義塾大学SFC研究所
清水　昇
2012年 3月 8日(木)

LODとは

◆LODの狙いは、誰もが**自由に使える相互に連携したデータ**を作ることです。
◆誰もが自由に使える様にする為には、データがオープンである事が必要です。
◆誰もが自由に使える様にする為に、もう一つ重要な事は、**データの意味を誰もが、同じ様に理解可能**である事です。
◆データの意味の解釈が使う人毎に異なる様では、困ります。
◆また、データの意味の定義が、特定のメーカや団体に依存する事も、データをオープンにすると言う観点から、望ましくありません。
◆Webに関するオープンな標準を開発しているW3Cは、データの意味を記述する為の言語(正確にはモデルと構文ですが、分かり易くする為に言語と言います)として、**RDF**(Resource Description Framework)を開発しました。
◆RDFは、主語と述語と目的語とから構成される意味モデル(**トリプル**、日本語では「三つ組み」と言います)を有し、意味モデルを記述する為の構文と語彙とを規定しています。
(注)トリプルのモデルは、非常に単純な様に思えますが、RDFでは、開集合、閉集合、トリプルを主語又は目的語とするReification等の複雑なモデルを包含しています。
◆RDFで記述された情報は、RDFの意味モデルと一対一に対応する様になっています。
◆即ち、RDFで記述された情報の意味を、RDFの仕様に則して、解釈するならば、誰もが同じ意味として把握する事が可能となります。

LODとしてデータを公開する場合、RDF形式のデータを公開すべきです。
(注)**OWL**は、RDFの語彙を拡張したものでRDFです。

LODを作成したいが、

① 既存の表データやRDBデータをLODとして公開したい。
② テキストや文書で記述されている情報をLODとして公開したい。
③ 新たな情報をLODとして公開したい。

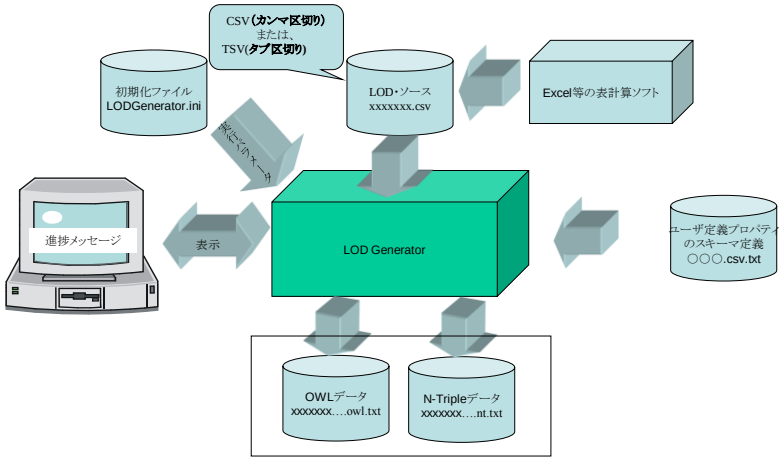
しかし

① RDFやOWLの記述方法や作り方が分からない。
② 勉強やスキルの習得に時間と工数が掛かりそうだ。

そこで

LODジェネレータを使えば、日頃使い慣れたExcel等の表計算ソフトでLODソースを作れ、LODを簡単に間違い無く作れる。

LODジェネレータ



LODジェネレータは、表計算ソフトや関係データベースが出力可能なCSVデータ若しくはTSVデータからLODデータを生成する。

LODジェネレータの主な特徴

1. 表計算ソフトを利用して大規模かつ高度なLODを生成することができます。
   1)木構造
   2)表構造
   3)グリッド構造
   4)SPO構造(主語(subject)、述語(predicate)、目的語(object)の三つの要素で定義)
   5)木構造と表構造との混合構造
2. 既存の関係データベースの中に格納されている情報をLOD化することができる。
3. CSVファイルのデータをLODにすることができます。
4. LODで使われているRDF、OWLおよびトリプルで記述されたデータを簡単に作成することができます。
5. 日本語のLODのデータを作成することができます。
6. 作成したデータをすぐにウェブ上で公開することができます。
7. 作成したLODのデータをセマンティックWeb用ツールで使うことができます。
   ・[例]セマンティックWebエンジンサーバ
2. RDFデータだけでなく、N-Triplesデータも作成可能。
   ・海外のLODでは、N-Triples形式のデータを公開しているケースが多々あります。
9. 複合記述も可能。(和集合、積集合、排他集合、等)

ライフサイエンス辞書のLOD

ライフサイエンス辞書とは、

ライフサイエンス辞書は、ライフサイエンス辞書プロジェクト(URL:http://lsd.pharm.kyoto-u.ac.jp/ja/)が開発したものであり、多くの辞書サービスで使われています。

ライフサイエンス辞書のLOD

ライフサイエンス辞書のLODは、ライフサイエンス辞書プロジェクトと共同で作成しました。
このLODの基データは、ライフサイエンス辞書プロジェクトからご提供頂いています。
LODジェネレータを用いて、OWLとトリプルとに変換し、LODにしたものです。
ライフサイエンス辞書のLODには、次の4種類のデータがあります。
(1).ライフサイエンス辞書の用語階層情報
(2).ライフサイエンス辞書の同義語情報
(3).ライフサイエンス辞書の統制語(Descriptor)情報
(4).ライフサイエンス辞書に於ける共起情報
　量が多いので次の二つに分割しています。
　・共起情報パート1
　・共起情報パート2

ライフサイエンス辞書のLODのURL
http://www.semanticweb.jp/lod/LodOfLsd.html

ライフサイエンス辞書のLOD化の時に生じた問題

①共起情報のLODでの表現方法の問題

親概念1ID	親概念2ID	tf-idf値
・・・		
D015060	D053835	12.9012532207507805
・・・		

共起情報とは、複数の用語が共に使われる程度を表す情報です。
ライフサイエンス辞書の場合、次の様になります。
用語「1,2-ジパルミトイルフォスファチジルコリン(D015060) 」と用語「単層リポソーム (D053835)」とのtf-idf値は、12.9012532207507805です。

これをOWLで如何に表現するかが問題となりました。
何が問題かと言うと、「通常、主語の要素は一つですが、この場合、主語の要素が複数となるので、それを如何に表すか」が問題となりました。
そこで、今回は、次の様に記述する事にしました。
主語を「D015060」と「D053835」との積
述語を「tf-idf値」として
目的語を「12.9012532207507805」としました。

この結果、LODのOWL記述は、次の様になりました。

```
<owl:Class>
  <owl:intersectionOf rdf:parseType="Collection">
    <rdf:Description rdf:about="D053835"/>
    <rdf:Description rdf:about="D015060">
    </rdf:Description>
  </owl:intersectionOf>
  <tf-idf値 rdf:datatype="&xsd;nonNegativeInteger">6.9011099954027817</tf-idf値>
</owl:Class>
```


最後に

ここで説明した「LODジェネレータおよびライフサイエンス辞書のLOD」を**株式会社サイバーエッヂ**のブースでご覧頂けます。
質問やご意見のお持ちの方は、どうぞお立ち寄り下さい。