Organizing Existing Metadata Terms and Structural Constraints to Support Metadata Schema Creation

2015/08/12 @Copenhagen

Tsunagu HONMA, Ph.D. student
University of Tsukuba, Japan
Topics

• Metadata Schema Creation for LOD datasets
  – especially, how to provide infrastructure to support finding and choosing metadata terms
Create LOD Datasets

- When we create datasets as Linked Open Data, we often describe them using RDF
From Spreadsheet to RDF

• There are some services and tools to create LOD Datasets from spread sheets

http://linkdata.org
<table>
<thead>
<tr>
<th>rdfs:label</th>
<th>category</th>
<th>quantity</th>
<th>material1</th>
<th>material2</th>
</tr>
</thead>
<tbody>
<tr>
<td>赤米ごはん</td>
<td>ご飯物</td>
<td>材料（4人分） 米 2合</td>
<td>もち米 30グラム</td>
<td></td>
</tr>
<tr>
<td>あじごはん</td>
<td>ご飯物</td>
<td>材料（4人分） 米 2合</td>
<td>しょうゆ小さじ 2</td>
<td></td>
</tr>
</tbody>
</table>

PREFIX recipe: <http://linkdata.org/work/rdf1s3450i#>
Designer’s Tasks

• Improve Interoperability of Datasets by reusing existing terms
  – Find metadata terms
  – Find metadata instances as examples
  – Find metadata schema of existing datasets
Specific Example Tasks

• What terms are better to describe a book of title?

• What terms are used when someone describe “William Shakespeare” as an author of a book?

• Are there some existing schemas for bibliographic Record? And what terms are used in those schemas?
Goal

• Some (or most) people who create LOD datasets aren’t familiar with Terms, Schemas and Instances
  – Difficult to find
  – Difficult to choose

• It is useful for us to provide more information about terms (e.g. definitions, usages in schema-level, usages in instance-level)

• We have to organize information about terms and provide an infrastructure or a service to support finding and choosing terms.
What terms are better to describe a book of title?
Find Terms

• Linked Open Vocabularies
  – They provide definitions of 515 vocabularies
  – People can find terms by metadata about vocabularies and terms (e.g. label, description, creator and tags)

• Definitions are not enough to find better terms

http://lov.okfn.org/dataset/lov/
Which term is better
to describe a title of a book

Description of Term 1: A name for some things

Description of Term 2: A name given to the resource
Find Terms

• Linked Open Vocabularies
  – They provide definitions of 515 vocabularies
  – People can find terms by metadata about vocabularies and terms (e.g. label, description, creator and tags)

• Definitions are not enough to find better terms
  – foaf:name : *A name for some thing.*
  – dcterms:title : *A name given to the resource.*
  – Many terms have no label and description

• It is hard to find terms with only definitions of terms
  – We would like to know actual usages of terms

http://lov.okfn.org/dataset/lov/
What terms are used when someone describe “William Shakespeare” as an author of a book?
Find Instances

• In the case people have prepared values of attributes like “William Shakespeare”, people would like to find terms by those values.

• In many cases, schemas do not include examples of values, so that we have to accumulate not only definitions of terms but also examples of values
Organize Metadata Instances

• We collect metadata instances from the Datahub
  – Also collect metadata about those datasets

<table>
<thead>
<tr>
<th>subject</th>
<th>subject_class</th>
<th>predicate</th>
<th>object</th>
<th>object_class</th>
<th>Dataset URI</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dataset URI</th>
<th>Name</th>
<th>Creator</th>
<th>Tags in Datahub</th>
</tr>
</thead>
<tbody>
<tr>
<td>http://⋯⋯</td>
<td>Dataset of Bibliographic record</td>
<td>Someone</td>
<td>Publication, ⋯</td>
</tr>
</tbody>
</table>
Mapping Classes and Properties

• Classes of subjects and objects are not enough useful to find terms
  – Users do not find “properties which have foaf:Person as Domain” but find “properties which have Person as Domain”

• Mapping terms into Schema.org’s terms manually

Are there some existing schemas for bibliographic Record? And what terms are used in those schemas?
Extend Definitions by Schemas

- Some documents of metadata schema includes meanings of terms in those contexts

<table>
<thead>
<tr>
<th>attribute</th>
<th>property</th>
<th>domain</th>
<th>range</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>提供元のメタデータへの参照</td>
<td>dc:source</td>
<td>ndlkn:MetaResource</td>
<td>rdf:Resource</td>
<td>メタデータを他のアーカイブから収集して提供する場合に, ...</td>
</tr>
</tbody>
</table>
Find Schema

- Accumulates metadata schemas and make them reusable on the Internet
  - Open Metadata Registry\(^\d\), MetaBridge\(^{\d\d}\), ...

\(^{\d}\) http://metadataregistry.org/
\(^{\d\d}\) https://metabridge.jp/
MetaBridge

• A service of metadata schema registry

• Users can upload and share metadata schema based on OWL-DSP (Description Set Profile)
  – If it is difficult to describe in OWL, users can describe simplified-DSP like spreadsheet

https://metabridge.jp
Dublin Core Description Set Profile

• “A DSP is a way of describing structural constraints on a description set” †

• Description Template
  – “which contain the statement templates that apply to a single kind of description as well as constraints on the described resource”

• Statement Template
  – “which contain all the constraints on the property, value strings, vocabulary encoding schemes, etc. that apply to a single kind of statement”

† http://dublincore.org/documents/dc-dsp/
An Example of Simplified-DSP

template: schema:Book

<table>
<thead>
<tr>
<th>attribute</th>
<th>property</th>
<th>occur</th>
<th>type</th>
<th>relation</th>
<th>value const.</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>dc:title</td>
<td>1</td>
<td>Literal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td>dc:creator</td>
<td>0-n</td>
<td>URI</td>
<td>[foaf:Agent]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>dc:subject</td>
<td>0-n</td>
<td>URI</td>
<td></td>
<td>lcsh:</td>
<td></td>
</tr>
</tbody>
</table>

template: foaf:Agent

<table>
<thead>
<tr>
<th>attribute</th>
<th>property</th>
<th>occur</th>
<th>type</th>
<th>relation</th>
<th>value const.</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>dc:title</td>
<td>1</td>
<td>Literal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Link</td>
<td>rdfs:seeAlso</td>
<td>0-n</td>
<td>URI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How many schemas are published

• There are few explicit schema which we can convert into DSP

<table>
<thead>
<tr>
<th>Data Portal /Data Format</th>
<th>RDF</th>
<th>CSV / XLS</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN Japanese (Data for Japan)</td>
<td>Schemas</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>22</td>
<td>74</td>
</tr>
<tr>
<td>Open Data METI</td>
<td>Schemas</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>LOD Challenge 2013</td>
<td>Schemas</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>44</td>
<td>9</td>
</tr>
</tbody>
</table>

※ As of September, 2013
There are a lot of LOD datasets on the web.
We tried to extract schema from those datasets.

Extraction process:
1. Get class memberships and create Description Templates
2. Get properties and create Statement Templates
3. Get a value type constraint
4. Get other value constraints
   1. Get literal value constraints
   2. Get non-literal value constraints

Tsunagu Honma, et.al. Extracting Description Set Profiles from RDF Datasets using Metadata Instances and SPARQL Queries. DC-2014.
The Number of Resources belonging to each Template

- 10 Datasets, 66,347 Resources, 168 Description Templates
- More than 100 templates have less than 10 resources
- There are similar templates each other
Conclusion

• Improve Interoperability of Datasets by reusing existing terms
  – Find metadata terms
    → Collect existing terms, and map into Schema.org’s terms
  – Find metadata instances as examples
    → Collect LOD datasets and their metadata
  – Find metadata schema of existing datasets
    → Collect metadata schema into schema registry
    → Extract schema from datasets

• Bridge these metadata around metadata terms