Estimating Domain Models from Metadata Instances to Improve Usability of LOD Datasets

Ryota Kinjo[1], Mitsuharu Nagamori[2], Shigeo Sugimoto[2]

[1]Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan. [2]Faculty of Library, Information and Media Science, University of Tsukuba, Japan.

1. Introduction

In this research we extract a domain models[1] from metadata instances in LOD datasets. Domain model is one piece of information about a metadata schema. Domain model is useful for metadata developers/designers to understand the structure of the LOD in an early stages of their development. Fig1 is an example domain model of Aozora Bunko LOD[2]. We developed an estimation method to estimate domain models by extracting well-used metadata terms from metadata instances. We then applied the method to existing datasets and compared estimated domain models with correct domain models which we created manually.

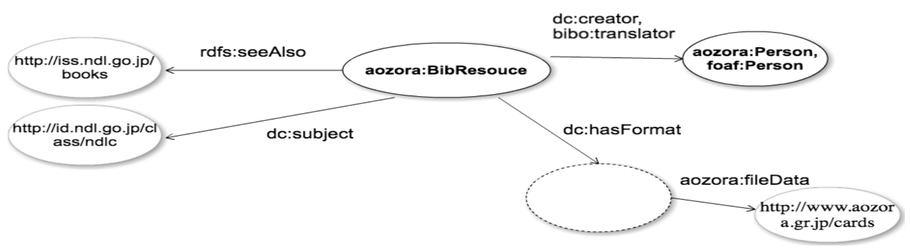


FIG. 1. A domain model of Aozora Bunko

Problem

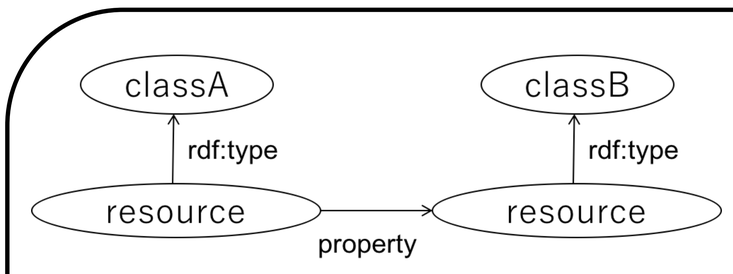
Domain models of existing LOD datasets are useful for metadata designers/developers. However few LOD datasets provide their Domain models.

Solution

To create a method which estimate domain models from metadata instances in LOD datasets.

2. Method

1. Identification of a main class(es)



ncs: number of times the class as a subject
nco: number of times the class as an object

If (*ncs* > *nco*) { the class is regarded as main class. }

Main class(es) and properties which belong to the main class are put into the domain model.

2. Creating a domain model as directed graph

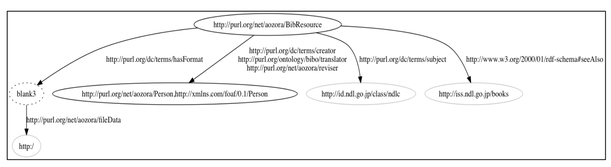
3. Experiment

We prepared 5 LOD datasets and 5 correct domain models of the 5 LODs. Correct domain models were created by us manually. But if existing domain model/s exist, we used it/them as the correct domain model/s. Then we compared estimated domain models with correct ones in RDF format.

Dataset name	Correct domain model	Memo	Number of triples
Aozora bunko LOD	Manually made		50,000
CiNii	Manually made		20,000
Europeana	Existing model	1 / 1000 of Overall	400,000
Kyoto kokusai manga museum [KMM]	Existing model		8,510,000
NDLSH	Manually made		470,000

TABLE.1. datasets used in the experiment

FIG.2. A estimated Domain model



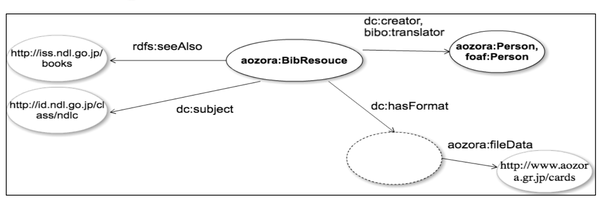
Convert

FIG.3. Estimated triples

```
<BibResource>
  <rdfs:seeAlso> "http://iss.ndl.go.jp/books";
  <dc:creator> <foaf:Person>,<aozora:Person>;
  <bibo:translator> <foaf:Person>,<aozora:Person>;
  <bibo:reviser> <foaf:Person>,<aozora:Person>;
  <dc:subject> "http://id.ndl.go.jp/class/ndlc";
  <dc:hasFormat> [ <dc:reviser> <foaf:Person>,<aozora:Person>;
                   <aozora:fileData> "http://www.aozora.gr.jp/cards"].
```

Compare

Fig.4. A correct domain model



Convert

FIG.5. Correct triples

```
<BibResource>
  <rdfs:seeAlso> "http://iss.ndl.go.jp/books";
  <dc:creator> <foaf:Person>,<aozora:Person>;
  <bibo:translator> <foaf:Person>,<aozora:Person>;
  <dc:subject> "http://id.ndl.go.jp/class/ndlc";
  <dc:hasFormat> [ <dc:reviser> <foaf:Person>,<aozora:Person>;
                   <aozora:fileData> "http://www.aozora.gr.jp/cards"].
```

We calculated precision and recall in accordance with this formula.

$$\text{Precision} = \frac{\text{Estimated triples} \cap \text{Correct triples}}{\text{Estimated triples}}$$

$$\text{Recall} = \frac{\text{Estimated triples} \cap \text{Correct triple}}{\text{Correct triples}}$$

4. Results and discussion

Table2 shows the Results of the experiment.

Dataset name	Overall		All classes		Main classes		External link	
	precision	recall	precision	recall	precision	recall	precision	recall
Aozora bunko LOD	0.85	1	1	1	1	1	1	1
CiNii	0.83	0.83	1	1	1	0.5	0.75	1
Europeana	0.07	0	0.33	0.4	0.33	0.5	0	
KMM	0.23	0.2	0.9	0.53	0.83	0.29	0	
NDLSH	0.63	0.63	1	1	1	1	0.33	0.33

TABLE.2. Results of the experiment

P/R of Aozora bunko and CiNii is good.
Especially, P/R of aozora bunko is good.
If scale of the LOD is small, this method works well.

P/R of Europeana and KMM are not good.
Europeana – amount of metadata instances are not enough.
KKM - unused terms which are defined in existing domain model are not used in metadata instances.

Existing domain models do not include external links.
We made estimation method include external links. but exiting domain models of two datasets do not include external links.

5. Conclusion and Future tasks

Conclusion and future tasks are as follows.

Conclusion

1. We estimated domain models by very simple method
2. A primary problem is the evaluation for validity of our method. Because, Existing domain models are not necessarily intended for LOD users

Future tasks

1. To review evaluation method
2. To increase experiment datasets
3. To improve the decision of what information is put into the domain model

[1] Dublin Core singapore-framework. Retrieved February 19, 2017, from http://dublincore.org/documents/singapore-framework/.
[2] Aozorabunko LOD. Retrieved February 19, 2017, from http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/.