MarcOnt - Integration Ontology for Bibliographic Description Formats

Sebastian Ryszard Kruk
DERI Galway
Tel: +353 91-495213
Fax: +353 91-495541
sebastian.kruk @deri.org

Marcin Synak
DERI Galway
Tel: +353 91-495006
Fax: +353 91-495541
marcin.synak@deri.org

Kerstin Zimmermann
DERI Innsbruck
Tel: +43 512-507-6449
Fax: +43 512-507-9872
kerstin.zimmermann@deri.org

Abstract:

For interoperability between digital libraries a number of bibliographic description standards have been introduced. Some of them, like MARC21 are usually implemented in classic library systems, while new digital libraries tends to support semantically richer formats like Dublin Core or BibTeX. Although it is possible to translate back and forth between these standards, a lot of information in lost while translating from MARC21.

We present MarcOnt an ontology that is based on MARC21, BibTeX and Dublin Core. We elaborate on the purpose and features of MarcOnt ontology. We describe schemata that provide underlying concepts to MarcOnt ontology. We provide an example on differences between those standards and a set of rules that are used to translate to and from MarcOnt ontology-based semantic descriptions. Finally we present the architecture of MarcOnt Mediation Service that enables cooperation.

Keywords:

Digital libraries, Semantic web, ontologies, Dublin Core, MARC21, BibTeX.

1. Introduction

In recent years more and more information has been made available on the Web. However, managing this information and sharing it across distributed and heterogeneous libraries still poses many challenges. New technologies based on research in areas of Semantic Web and Semantic Web Services addresses these challenges promising to resolve most of problems which cannot be solved with standard Web technologies.

In the MarcOnt initiative an ontology is developed for bibliographic description and related tools utilizing Semantic Web technologies. The aim is to deliver an ontology based on the legacy bibliographic description. In this paper we start with a description of the schemes and their history. Then we provide an example of the differences and the challenge of combing them. We define the most important use cases for MarcOnt ontology. We present the first version of the ontology with its mediation service so far.

2. Using Description Formats in Digital Libraries

Diversity of bibliographic description formats for (digital) libraries reflects diversity in target audience. The library users can be divided into 3 groups. Each group representants require different kind of description to make the most of the digital library:

- librarians and library related users - detailed description with MARC21;
- researchers and academia related users citation relations description with BibTeX;
- generic Internet users - compact description like Dublin Core;

Digital libraries that can be found on the Internet are reflecting these scenarios. Very often, a digital library system supports only one of description formats that is appropriate for the given target audience: **Classic (digital) libraries** - library systems for handling physical resources stored in classic libraries, very often provide additionally web interfaces where readers can search and reserve selected books. In many cases MARC21 is the only format of the bibliographic description used for communication. **Publishers' digital libraries** -provide access to publications, conference proceedings, etc. Since this resources are used by researchers mainly most of digital libraries of this type support BibTeX

description format. **Other digital libraries** -provide different types of resources targeted towards a different kind of users or just any Internet users. Many of these web applications use Dublin Core metadata to annotate presented resource. **Semantic digital libraries** - with the dawn of the Semantic Web more and more semantically enabled digital libraries like JeromeDL or DSpace/SIMILE are about to emerge. The core goal of these digital libraries is to provide better retrieval features by enhancing System-Human interaction and providing higher accuracy in distributed search within heterogeneous networks of digital libraries.

3. Bibliographic Description Formats

We present problems with the diversity of known bibliographic description formats like Dublin Core, MARC21 and BibTeX and first consideration on how to overcome this in the MarcOnt ontology.

3.1 Dublin Core (DC)

In 1995 the Dublin Core Metadata Initiative was formed for the development of interoperable online metadata standards that support a broad range of purposes and business models.

Resources are classified according to the DC Metadata Element Set Version 1.1 (2004) which has the following 15 elements: *Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Right*.

Their description (metadata) is made accessible over the Internet for online retrieval. No numbering for the fields/elements is given done longer but the sorting is similar to the 'librarian' style. For the normal user the field name are clear and practical for classifying resources her/himself and providing the own metadata.

For Type the vocabulary with 12 categories is provided, like e.g.: *Collection, Dataset, Event, , Sound, Text, Image, MovingImage, StillImage, Interactive resource, Physical object*. With these categories it is possible to classify textual information as well as art artifacts like sculptures.

3.2 MARC21

MARC the MAchine-Readable Cataloging was introduced in the 60ies when electronic data processing became important for libraries. First of all the catalog cards were converted in an electronic version and were the basis of these format carrying bibliographic data in a specific order. These were and still are: titles, names, subjects, notes, publication data, and information about the physical description of an item. The main purpose was the allocation of a book in the shelves of the library. This information is coded in a signatory.

MARC21 the Format for Bibliographic Data consists of three numerals in the first level (field 001-887). It contains authority, bibliographic, holdings, classification and community data, that can be found in every group of these fields, such as *Control Fields* [00X], *Number and Code Fields* [010-09X], *Main Entry / Primary Name Fields* [1XX], *Title Fields* [2XX], *Physical Description, etc. Fields* [3XX], *Series Statement Fields* [4XX], *Note Fields* [5XX], *Subject Access Fields* [6XX], *Added Entry Fields* [7XX], *Holdings, Location, Alternate Graphs, etc. Fields* [8XX].

In the meantime MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. The related information also covers internal workflow entries which do not appear in the public catalogue. The bibliographic data describe 8 types of material, such as: *Books, Continuing resources, Computer files,* etc. and 7 types of records, like: *Language (textual) material, Manuscript (textual) language material, Cartographic material,* etc.

For a non-librarian this type differentiation is not quite obvious!

3.3 BibTeX

BibTeX was designed by Patashnik and Lamport in 1985 as the LaTeX bibliographic format. LaTex is an open source document preparation system widely used in the academic community. The authors provide their references/citation of other publications entirely character based, so that it can be shared by the community on the Internet.

It provides the following elements, such as:
abbreviation, editor / author, title, booktitle / journal / series, etc. The type of publication can be classified according to 12 different categories, like: *book, inproceedings, article,* etc.

3.4 Correlations and Mappings

As it can be seen from the examples given above each scheme provides different subclasses for type. DC is the broadest with 12 categories including also event and physical object. MARC21 is more flat and gives 8 subclasses for material and 7 for records. BibTeX only focuses on textual information but has again 12 very detailed subclasses for this purpose.

Also in this small example it is evident that a simple one to one mapping even between the two schemes is not possible. For DC to MARC21 it can be

considered: Event can be found in the Community data but Service is not mentioned anywhere in MARC21. The other way round: Maps in MARC21 can be taken as one kind of Physical object, in Visual material the three terms Image, MovingImage and StillImage can be subsumed.

To overcome these different classifications MarcOnt might serve as a bridge in the future. Equal content shall be recognized because information stays information n matter of the given scheme but on its context. So we start with the construction of this new ontology.

4. Construction of MarcOnt Ontology

4.1 Transformation to/from Legacy Bibliographic Formats

Legacy Bibliographic Formats, such as MARC21, Dublin Core or BibTeX may take a form of binary file, text file with specific formatting or (if we are lucky) XML or RDF file. To take advantage of information which they contain, a framework must be created to support importing information from format X into MarcOnt semantic description and exporting information to another format, let's say Y, if needed.

Figure 1 presents a general architecture of tools for format transformations which MarcOnt provides.

4.2 Transformation scenario

Transforming description of a resource in a legacy format such as MARC21 to a semantic description requires few operations. First we have to move the data to RDF format, so the result can be reasoned over and the semantic descriptions can be inferred from them.

An example flow of transforming MARC21 description to MarcOnt semantic description would be:

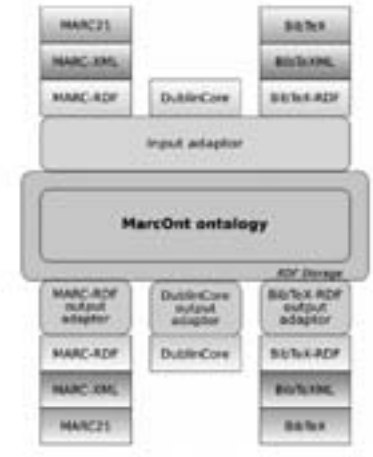

Figure 1: MarcOnt Mediation Service

1. Parse binary MARC21 file and create MARC-XML file
2. Transform MARC-XML file to MARC-RDF file using XSLT
3. Transform the graph to MarcOnt semantic description using inference or other tool

First step is relatively easy, because MARC-XML format is described well in literature and requires only using (or writing your own) parser library.

The second step required creating a new format, which we called MARC-RDF, to translate XML data from MARC-XML file to RDF graph. MARC-RDF does not provide semantic information, it only uses different taxonomy (different terminology) to represent fields and values from MARC-XML. The transformation is easily done using a simple XSLT transform.

The third step represents the most difficult task - translating one RDF graph into another. In other words - it requires specifying a set of rules, where a single rule identifies existence of one set of triples as a requirement of creating another set of triples.

Translating MarcOnt semantic descriptions back into specified legacy format requires going in other direction on the same way - first translate a graph, then XSLT to *-XML format and then parse the xml and write an appropriate text or binary file (which to some extent could also be done using XSLT).

Translation of Dublin Core descriptions back and forth to MarcOnt is much simpler task, because DC is already RDF sand doesn't require additional tasks than graph transformation.

Rule	premises	consequences
1	LDR[06] is "a" or "n"	create individual of type marcont:Book call rules 2 and 3
2	datafield 100 exists	create property marcont:hasCreator pointing to: create individual of type foaf:Person call rule 3
3	subfield 100$a exists indicator[1] is "1"	create property foaf:surname with value from 100$a
4	datafield 245 exists	create property marcont.hasTitle pointing to: create individual of marcont:TitleStatement call rule 5
5	subfield 245$a exists	create property marcont:titleValue with value from 245$a

Table 1 – Mapping rules

Rule premises consequences 1 LDR[06] is “a” or “n” create individual of type marcont:Book call rules 2 and 3 2 datafield 100 exists create property marcont:hasCreator pointing to: create individual of type foaf:Person call rule 3 3 subfield 100$a exists indicator[1] is “1” create property foaf:surname with value from 100$a 4 datafield 245 exists create property marcont.hasTitle pointing to: create individual of marcont:TitleStatement call rule 5 5 subfield 245$a exists create property marcont:titleValue with value from 245$a

4.3 Mapping rules definition examples

Table 1 presents some example rules which could be applied to MARC21 description to create MarcOnt semantic description. Of course, the complete set of rules is to be much larger and its creation will require help from people familiar with MARC21. The example detects if described resource is a book or a manuscript. Then it creates a new description based on information on author and title.

```
<rule name="r1">
  <premise>
   <predicate value="rdf:type"/>
   <object value="marcrdf:Record"/>
  </premise>
  <premise>
   <subject value="{$PS0}"/>
   <predicate value="marcrdf:hasLeader"/>
  </premise>
  <premise>

   <subject value="{$PO1}"/>
   <predicate value="rdf:_7"/>
   <object regexp="[at]"/>
  </premise>
  <consequent>
   <subject value="{marcont:clone($PS0,
'marcont:')}"/>
   <predicate value="rdf:type"/>
   <object value="marcont:Book"/>
  </consequent>
  <call rule_name="r2">
   <param name="IDmarcrdf" value="{$PS0}"/>
   <param name="IDmarcont" value="{$CS0}"/>
  </call>
  <call rule_name="r3">
   ...
  </call>
</rule>
```

Table 2 - Rules coded in XML

4.4 RDF Translator

There is a number of tools which could be used to perform such transformation using a set of rules. We considered logic-driven ones such as TRIPLE, complicated XSLTs and others.

Finally we have developed our own tool called RDF Translator. It operates on two disjoined RDF models. RDF Translator translates RDF triples from an input model to an output model, according to the rules defined in XML syntax similar to Sesame inferencer configuration.

A rule is composed of a number of “premises” or requirements and “consequents” -results. Both premises and consequents are RDF statements (triples) consisting subject, predicate and object (see Table 2). Hundreds of rules will create an input adaptor. Creating output adaptors, i.e. for Dublin Core export requires doing exactly the same operations as creating an input adaptor.

5. Conclusions

In this paper we presented the different schemes which are used to generate electronic meta data and which are not interoperable up to now. The MarcOnt project is going the close this gap by intermediation on building up a general ontology. As it is an ongoing project the status quo of the integration can be found on the website [4].

References

1. S. R. Kruk, A. Mocan, B. Sapkota, M. Zaremba. *Building Semantic Web Services Infrastructure for Digital Libraries*
2. DCMI: http://dublincore.org/
3. MARC: http://www.loc.gov/marc/
4. MarcOnt: http://www.marcont.org/

Acknowledgments

Authors would like to thank Stefan Decker and Lech Zieborak for consultations. The work is supported by SFI (Science Foundation Ireland) under DERI-Lion project and by KBN (State Committee for Science Research, Poland) under grant 4T11C00525