Research Proposal

International Open Public Digital Library: A Proposal for the Future
Roles of Common Terminology for Improving Metadata Interoperability
(Technical Requirement)

(“Boaz”) Sunyoung Jin
August, 2012
Last Modified January 14, 2014
I. Research Plan

This paper outlines understanding of the research and work plan for the research – Roles of Common Terminology for Improving Metadata Interoperability. This research goal is to fulfill the second requirement, technological requirement, to establish the proposed International Open Public Digital Library (IOPDL). The technological requirement is to achieve and improve interoperability among well-designed digital libraries using a common terminology or using a unified standard metadata format. The proposed IOPDL and the implemented first requirement, General Requirement: evaluating of existing digital libraries, are able to be found on http://courseweb.lis.illinois.edu/~sunjin/Papers/InternationalOpenPublicDigitalLibrary-EvaluatingDLs.pdf.

1. Introduction

Interoperability issues are serious nationwide problems and have been a big barrier in sharing and transferring information among digital libraries and repositories. One of main reasons that cause problems in achieving interoperability is using diverse metadata according to each community’s needs. By the awareness, the project researches a common terminology among several metadata schemas (e.g., MARC, MODS and DC) to achieve and improve interoperability. We will investigate roles of the common terminology to improve interoperability in several metadata interoperability levels. We will implement the Common Terminology (CT) on the semantic Web in RDF and XML achieving several levels’ interoperability. We will evaluate performance of the CT in achieving and improving metadata interoperability among three chosen databases (Harvard Library, MIT DSpace Library, and UIUC library) by mapping experiments. The CT may give a certain innovative solution of troublesome interoperability problems by using various metadata formats. Thus, the research project has the following outlines:

1) Review previous researches of 20C about convertible and compatible problems in sharing data on the bibliography description such as metadata, indexing language, controlled vocabularies, common terms or common languages.

2) Review present researches and methods for achieving interoperability especially on metadata interoperability levels (e.g., metadata instance, metadata schema, metadata language, record level, and repository level).

3) Document problems of existing methods to achieve metadata interoperability. Suggest using a Common Terminology (CT) and suggest processing of creating CT at each metadata interoperability level. Discuss the roles of CT at each metadata level.
4) Design a crosswalk among several metadata schemas (e.g., MARC, MODS and DC). Create a Common Terminology considering lexical, semantic, syntactic, grammatical level metadata interoperability for them.

5) Design a practical crosswalk between three chosen databases refining the CT to improve sharing information among them. The three University libraries are Harvard Library that uses MARC21 bibliographic format, MIT DSpace Library that uses qualified Dublin Core format, and UIUC Library that uses MODS format.

6) Implement the CT on the semantic Web in RDF and XML achieving several levels’ interoperability (e.g., metadata language level, record level, and repository level).

7) Prove performance of the CT through experiments of direct and indirect mappings. The Direct mappings with the crosswalk for MARC, QDC, and MODS include the mapping between Harvard library and MIT library, between MIT and UIUC, and between Harvard and UIUC. Indirect mappings are also experimented through a suggested common terminology (CT) with three different databases of Universities. That is, the mappings are done between Harvard and CT, between MIT and CT, and between UIUC and CT. Moreover, direct mapping and indirect mapping are implemented with Python computer programs language based on a dictionary or RDF graphs formats.

8) Compare and analyze the performances of direct mappings and indirect mappings in reducing losing data rates and in increasing accuracy for preserving data. The research paper reports that the indirect mapping through CT shows much better performance to achieve and improve interoperability. The research paper also presents a meaningful, actionable guidance and implementation strategies (e.g. mappings that use the CT based on RDF format), which will be a remarkable solution to achieve and improve metadata interoperability.

Lastly, the scope of the research proposal covers three different well known metadata schemas and the university libraries. It focuses on suggesting a common terminology for the element names of three chosen metadata schemas. This research will be focused on providing guidance that helps share information among the libraries, and that helps search materials by their search engines with the suggested common terminology (CT).
II. Resources

Professor Dave Dubin suggests some predecessor articles in 20C about compatibility and convertibility, which are relevant to interoperability in mappings these days. Also, I found a number of directly appropriate resources about metadata, metadata schemas, and crosswalks that are provided by Library Congress.

- Articles about compatibility and convertibility in 20C.

- Resources about metadata, metadata schemas, and crosswalks that are provided by Library Congress.
    
  - Metadata Services MIT Libraries: http://libraries.mit.edu/metadata/
  - Dspace An Open Source Dynamic Digital Repository http://www.dlib.org/dlib/january03/smith/01smith.html
- Harvard Library Releases Nearly 100% of Its Records
  http://isites.harvard.edu/icb/icb.do?keyword=k77982&pageid=icb.page498373
- Harvard Library Bibliographic Dataset
  http://openmetadata.lib.harvard.edu/bibdata
- Harvard Library Bibliographic Dataset
- UIUC-Best Practices for Descriptive metadata
  http://www.library.illinois.edu/dcc/bestpractices/chapter_08_descriptive_metadata.html
- Metadata Standards Crosswalk
  http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html
- MARC standards MARC 21 Formats http://www.loc.gov/marc/marcdochtml
- Metadata Object Description Schema (MODS) http://www.loc.gov/standards/mods/mods-conversions.html
- MARC Mapping to MODS Version 3.4 http://www.loc.gov/standards/mods/mods-mapping.html
- Dublin Core Metadata Element Set Mapping to MODS Version 3 http://www.loc.gov/standards/mods/desimple-mods.html
- MODS to Dublin Core Metadata Element Set Mapping Version 3 http://www.loc.gov/standards/mods/mods-desimple.html
- MARC to Dublin Core Crosswalk http://www.loc.gov/marc/marc2dc.html
- Dublin Core to MARC Crosswalk http://www.loc.gov/marc/dccross.html
- ConverterToRdf http://www.w3.org/wiki/ConverterToRdf
  (MARC marcmods2rdf transforms MARC records from Z39.2 format into MODS and then from MODS to an RDF representation of MODS. MARiMbA is a command-line tool, designed with librarians in mind, to transform MARC (MAchine-Readable Cataloging) records to RDF, following Linked Data best practices.)
III. Research Objectives, Approach, Object, and Work Plan

1. Objectives and Approach

The proposal has the following objectives:

- To examine and access the current problems in sharing information, which is caused specially by using various metadata schemas;
- To give a certain solution, a Common Terminology (CT), of troublesome interoperability by using various metadata schemas;
- To implement the suggested CT and to prove performance of the designed indirect mappings with the suggested Common Terminology;
- To develop meaningful, actionable guidance and implementation strategies of mappings that uses a common terminology in order to improve metadata interoperability;

Through the suggested common terminology,

- To reduce losing information rates in mappings with the designed crosswalk;
- To increase accuracy in mappings improving preservation of data;
- To improve metadata interoperability, compatibility and convertibility in mappings; and
- To improve effectiveness of search engines in searching through the designed CT.

2. Objects of the research

The objects of the research are everyone who is involved and interested in metadata, mapping, and interoperability. Especially, the objects of the research are focused on three university libraries and metadata schemas:

- **Harvard Library** that open all metadata so that anyone can research them, and that uses the **MARC21** bibliographic format;
- **MIT DSpace Library** that uses **qualified Dublin Core** format;
- **UIUC Library** that uses **MODS** format.
  (Those are the representative metadata schemas these days)

In order to accomplish these objectives, the work plan is built on the following:

- We will connect the prior researches about compatibility and convertibility in 20C into present researches about interoperability, metadata, crosswalks and mappings.
• The project will focus on researching the element names of three chosen metadata schemas to recommend a common terminology including global issues of control vocabularies, indexing languages, and thesaurus.

• The suggested common terminology will be reviewed by the faculty who are involved in the project.

• We will implement the experiments for the designed direct and indirect mappings by computer programs with Python computer language. We will implement direct mappings and indirect mapping based on dictionaries and RDF graph format.

• We will develop meaningful, actionable guidance and implementation strategies of direct and indirect mappings as a proposed solution to improve materializing interoperability.

• Progress will not be expensive. Recent developments in technology should enable to success to implement the experiments with little cost.

3. Work Plan

Task 1. Proposal Research Project and Background Research

Objective

The objective of Task 1 is to propose the research project and to assemble all background research materials required to support the research about the problems that are caused by using many different metadata schemas. Also, it is to collect the paper which is related to achieve interoperability, metadata interoperability, among the chosen three databases and universities.

Work Steps

• Propose the research project with the work plan.
• Assemble all background research materials required to support the research.
• Literature reviews about the problems of interoperability by using various metadata schemas.
• Identify and assess the existing interoperability methods.
• Document about the found problems of interoperability, evaluate the existing interoperability methods, and identify best way to achieve interoperability.

Deliverables
• Research proposal – Roles of Common Terminology for Improving Metadata Interoperability (Technical Requirement)
• Progress reports as described above and a documentation of literature reviews.

Tools and Technologies

A number of open source tools and technologies could be used in the project, including:

• Google docs: https://docs.google.com/
• Drop box: https://www.dropbox.com/
• Zotero: https://www.zotero.org/groups/

Task 2. Design a crosswalk for MARC, MODS and DC Suggesting Common Terminology

Objective

The objective of Task 2 is to examine and research MARC, MODS and DC metadata schemas. It is to design a crosswalk for them. Through the research about their metadata practically, we will build a crosswalk in detail for three very different metadata schemas. Moreover, in the task 2, we will suggest a Common Terminology for three different metadata schema considering sincerely achieving lexical, semantic, syntactic and grammatical level metadata interoperability among them.

Work Steps

• Research MARC21, QDC, and MODS metadata schemas and their crosswalks that Library of Congress presents (DC to MODS 3.2/3.3/3.4, MARC to MODS 3.4, MODS 3.2 to DC, MODS 3.4 to MARC)
• Build a crosswalk for three metadata schemas considering lexical, semantic, syntactic and grammatical level metadata interoperability based on each above crosswalk.
• Design a frame of Common Terminology for the element names of three different metadata schemas.
• Document each element name of CT (e.g., why we chose it, how it achieves interoperability at each interoperability level).

Deliverables

• Progress reports as described above and document assessments of the chosen metadata schemas in achieving interoperability.
• A built crosswalk for three different metadata schemas.
• A suggested Common Terminology that achieves lexical, semantic, syntactic and grammatical level metadata interoperability.
• A document that describes each element name of CT (e.g., why we chose it, how it achieve interoperability at each interoperability level).

Task 3. Applying the suggested CT to three different databases of Harvard, MIT, and UIUC libraries refining the CT

Objective

The objective of Task 3 is to refine the suggested Common Terminology by applying it into three different databases (e.g., Harvard University that uses MARC 21, MIT DSpace Library that uses DC and UIUC libraries that uses MODS). In Task 3, we will then implement the common terminology into RDF and XML considering achieving schema definition language interoperability. We will publish it on the semantic Web. We will define concepts of the CT using existing SKOS concept or DCMI terms, or creating new concepts for CT, so that we can use it in mappings as the defined SKOS terms and as the defined Common Terminology (CT).

Work Steps

• Research Harvard, MIT and UIUC metadata
• Revise the designed CT with three universities’ metadata to reduce losing info. rates.
• Identify and assess the chosen databases and metadata schemas to achieve interoperability.
• Choose sample metadata in various genres of three different databases
• Apply the suggested CT to the chosen sample metadata in various genres.
• Refine the CT so that it can improve metadata interoperability.
• Define the CT into SKOS terms.
• Implement the CT in RDF and XML to achieve schema language interoperability.
• Publish the CT on the semantic Web.
• Document the refined and published CT.

Deliverables

• Chosen sample metadata in various genres of three different databases
• Refined CT that can improve metadata interoperability in every level.
• Defined the CT with SKOS terms.
• Implemented and published CT in RDF and XML.
• Refined document of the CT
Task 4. Implement the direct mappings and the indirect mappings of the CT

Objective

The objective of Task 4 is to implement actual mappings with Python programming language: the direct mappings between two different schemas’ metadata and the indirect mappings between a metadata and the CT. Each program measures accuracy and losing data rates to assess each performance of the mapping.

Work Steps

- Develop Python programs to implement the direct mappings based on dictionaries or RDF, measuring accuracy and losing data rates:
  - The direct mapping from Harvard metadata format (MARC 21) to MIT metadata format (QDC),
  - The direct mapping from Harvard metadata format (MARC 21) to UIUC metadata format (MODS),
  - The direct mapping from MIT metadata format (QDC) to UIUC metadata format (MODS),
  - The direct mapping from MIT metadata format (QDC) to Harvard metadata format (MARC 21),
  - The direct mapping from UIUC metadata format (MODS) to Harvard metadata format (MARC 21),
  - The direct mapping from UIUC metadata format (MODS) to MIT metadata format (QDC).

- Develop Python programs to implement the indirect mappings based on RDF graph formats, measuring accuracy and losing data rates:
  - The indirect mapping from Harvard metadata format (MARC 21) to the designed Common Terminology,
  - The indirect mapping from MIT metadata format (QDC) to the designed Common Terminology,
  - The indirect mapping from UIUC metadata format (MODS) to the designed Common Terminology.

Deliverables

- The direct and indirect mapping programs.
Task 5. Evaluate their performance and develop the final paper

Objective

The objective of Task 5 is to compare performance of the direct and the indirect mappings, to evaluate their performance results, to describe analyses, and to develop the final paper. The final paper includes common terminology (CT)’s roles, usefulness, effectiveness, and importance to achieve interoperability among different metadata schemas and databases. It also includes what we learn from the project and what are our recommendations or suggestions for the future works to improve interoperability especially metadata interoperability in sharing information.

Work Steps

• Compare and analyze the performance between the direct mapping and the indirect mapping.
• Describe their performance and common terminology (CT)’s roles, usefulness, effectiveness, and importance, including findings.
• Recommend best strategies to achieve and improve interoperability in mapping for the future works.
• Prepare draft research paper and submit to faculty who involve in the project to be reviewed.
• Prepare PowerPoint briefing and present.
• Upon receipt of written comments from faculty, revise the research paper
• Submit the research paper

Deliverables

• Draft research paper of Roles of Common Terminology for Improving Metadata Interoperability
• PowerPoint of Roles of Common Terminology for Improving Metadata Interoperability
• Final research paper of Roles of Common Terminology for Improving Metadata Interoperability
• Suggest expending the project into involving more various metadata schemas and databases so that we can establish interoperability among well-designed digital libraries to realize the proposed National Open Public Digital Library.
IV. Time Requirements

The schedule represents my best approximation of the time requirements and as activity sequence, it can be adjusted to situations and to needs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1. Proposal Research Project and Background Research</td>
<td>6 months</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
</tr>
<tr>
<td>Task 2. Design a crosswalk of MARC, MODS and DC suggesting a Common Terminology of them</td>
<td>6 months</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
</tr>
<tr>
<td>Task 3. Refining the CT with Harvard, MIT and UIUC databases, and implementing the CT in RDF and XML</td>
<td>6 months</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
</tr>
<tr>
<td>Task 4. Implement experiments of the direct mappings and the indirect mappings with Python Programs</td>
<td>12 months</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
</tr>
<tr>
<td>Task 5. Evaluate their performance and develop the final paper suggesting future works</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
<td>`</td>
</tr>
</tbody>
</table>

*The approximated required time for the research project is three years from Sep. Jan. 2016. It can be rescheduled, according to the progressing situations.*
V. Expended schedule (optional)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1. Proposal Research Project and Background Research</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 2. Build a crosswalk for MARC, MODS and QDC &amp; DC</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 3. Build a Common Terminology for the chosen University Libraries</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 4. Implement the direct mappings and the indirect mappings Programming</td>
<td>12 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 5. Evaluate their performance and develop the final paper</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 6. Implement a integrate search engine for Harvard, MIT, and UIUC databases</td>
<td>12 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 7. Experiment performance and develop the final paper and Defense the paper</td>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The approximated required time for the research project is four and half years from Sep. 2012 to Feb. 2017. It can be rescheduled, according to the progressing situations (last updated on Dec. 9, 2013)*