TRANSFORMING LIBRARY CATALOGS INTO LINKED DATA

M. Hallo¹, S. Luján-Mora² J. Trujillo²
¹National Polytechnic School (Quito-ECUADOR)
maria.hallo@epn.edu.ec
²University of Alicante (SPAIN)
sergio.lujan@ua.es
jtrujillo@dlsi.ua.es

ICERI2014, Sevilla
Index

1. Introduction
2. Related Work
3. Linked Open Data Life cycle
4. Open Linked Data Architecture
5. Conclusions
1. Introduction

In this work we present partial results of the project: An Integration Platform for Bibliographic Records using linked data technologies, developed by three ecuadorian universities. The initial goals are:

• To increase the visibility of the bibliographic catalogs

• To enrich data adding links to another sources

At National Polythecnic School we have tested Linked Data metodologies and tools with Bibliographic metadata in MARC21 format taken from a university library with good quality metadata.
Problem

Bibliographic data are dispersed, without relationship between resources making difficult their discovery and reuse for other information systems.
To publish bibliographic data catalogs following the linked data principles.
Research Objectives

• To refine and test a procedure to publish Bibliographic Catalogs using Linked Data Technologies to enhance metadata discoverability and linking.

• To design a vocabulary to convert Marc21 metadata in RDF.

• To develop a technical architecture to generate and publish Bibliographic Catalogs using Linked Data Technologies.
Some definitions

Open access (OA) is the free unrestricted online access to digital content.

http://www.budapestopenaccessinitiative.org/read
Open Bibliographic Data

“Open Data: Data freely usable, reusable, and redistributable”.

“Bibliographic Data: Metadata produced by the library sector”
Marc21

“The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form”.
A MARC record is a **MA**chine-**R**eadable **C**ataloging record.
Semantic Web

“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries”.

http://www.w3.org/2001/sw/
Linked Data

“Linked Data refers to a set of best practices for publishing and interlinking structured data on the Web in a human and machine readable way.”

Tim Berners Lee
Library Linked Data

• "Library Linked Data" is any type of library data that is expressed as Linked Data.
2. Related Work

Best practices:

LOD2: Creating Knowledge out of Interlinkeddata:
http://stack.lod2.eu/blog/

W3C Linked Data cookbook:
http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

LOGD and metadata Lifecycle:
http://www.slideshare.net/OpenDataSupport/the-linked-open-government-data-lifecycle

Linked data Life cycles:
http://www.slideshare.net/mediasemanticweb/linked-data-life-cycle

The linked open government data and metadata lifecycle
www.opendatasupport.eu
3. Linked Open Data Life cycle

- Publishing
- Linking
- RDF Generation
- Modelling
- Metadata Extraction
- Data Source Analyze
3.1 Data sources analysis

The objective of this activity is to identify data sets that provide benefits for others to reuse. The steps in this activity are:

a) Identification of the data source and the attributes to be published and linked to another datasets: In our case the academic authorities have given the authorization to work with a bibliographic data set.
3.1 Data sources analysis

In this study, the test dataset has bibliographic metadata from the Electrical Engineering Department from the National Polytechnic School at Quito-Ecuador, which has a good team, always open to innovation, working on metadata keeping it mostly updated. This test will allows us to develop the procedure to publish de metadata and answer the posted queries. After this test we will try data sets from another departments.
3.1 Data sources analysis

The attributes to be published and linked are a subset of the Marc21 metadata useful to identify entities and for linking them to related resources of information to improve complex libraries queries.
3.1 Data sources analysis

b) Engaging stakeholders:

• Information Unit Director
• Library Coordinator
• Faculty authorities

c) Data sources analysis

Used software: Koha Integrated Library System using mysql database and Marc21 bibliographic metadata.
### 3.1 Data sources analysis

Bibliographic material in the analyzed library

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>Technical catalogs.</td>
</tr>
<tr>
<td>LIEE</td>
<td>Books specialized in Electrical and Electronic Engineering.</td>
</tr>
<tr>
<td>NTEC</td>
<td>Technical standards.</td>
</tr>
<tr>
<td>OLIT</td>
<td>literary Works.</td>
</tr>
<tr>
<td>PTEC</td>
<td>Institutional technical publications.</td>
</tr>
<tr>
<td>RELE</td>
<td>Electronic Resources (CDs, DVDs)</td>
</tr>
<tr>
<td>REV</td>
<td>Journals</td>
</tr>
<tr>
<td>TIEE</td>
<td>Engineering Thesis</td>
</tr>
</tbody>
</table>
3.1 Data sources analysis

Some Marc21 fields used in the test

<table>
<thead>
<tr>
<th>MARC</th>
<th>Descripción</th>
</tr>
</thead>
<tbody>
<tr>
<td>003</td>
<td>Control Number Identifier</td>
</tr>
<tr>
<td>020 a</td>
<td>International Standard Book Number</td>
</tr>
<tr>
<td>041 a</td>
<td>Language code of text/sound track or separate title</td>
</tr>
<tr>
<td>082 2</td>
<td>Edition number</td>
</tr>
<tr>
<td>100 a</td>
<td>Personal name</td>
</tr>
<tr>
<td>245 a</td>
<td>Title</td>
</tr>
<tr>
<td>250 a</td>
<td>Edition statement</td>
</tr>
<tr>
<td>260 a</td>
<td>Place of publication, distribution, etc.</td>
</tr>
<tr>
<td>260 c</td>
<td>Date of publication, distribution, etc.</td>
</tr>
<tr>
<td>856 u</td>
<td>Uniform Resource Identifier</td>
</tr>
</tbody>
</table>
3.1 Data sources analysis

d) Identification of the licensing and provenance information.
The licence of the data in our case study is Creative Commons Attribution-ShareAlike 4.0 International licence. This text is added in a dc:rights property:
Electrical Engineering library from National Polytechnic School, initial load data: 01-09-2014.
3.2 Metadata extraction

Metadata are extracted from the original source and stored in an intermediate database for cleaning.

a) Metadata extraction and storage
Metadata were extracted using the open source software Spoon-Pentaho Data Integration and stored in a relational database. The source entities of metadata extracted were: Work (book, Journal, etc.), Expression (Format, language), Manifestation (editions data) (FRBR model).
3.2 Metadata extraction

b) Cleaning

Each attribute was analyzed finding missing values that were completed taking values from open library datasets.
3.3 Modelling

The goal of this activity is to design a vocabulary for describing the data sources in RDF.
3.3 Modelling

The steps in this activity are:

a) Selection of vocabularies. Dublin Core, BIBO, FOAF, ORG, SKOS, RDFS, OWL, FRBR-RDA.

b) Vocabulary development and Documentation. The vocabulary was documented using Protégé (Ontology Editor Tool).
3.3 Modelling

c) Vocabulary validation.

The generate vocabulary was validate with OOPS!.

d) Specify a license for the dataset.

The license to publish the datasets is Creative Commons Attribution-ShareAlike 4.0 International.
3.4 RDF generation

The goal of this activity is to define a method and technologies to transform the source data in RDF and produce a set of mappings from the data sources to RDF.
3.4 RDF Generation

The tasks in this activity are:

a) Selection or development of technologies for RDF generation

The Triplify tool with some modifications has been used to perform the transformation of the intermediate relational database in RDF.

b) Mappings from data sources to RDF
c) RDF generation
3.5 Interlinking

The objective of this activity is to improve the connectivity with links to external datasets.

The steps in this activity are:

- **Target datasets discovery and selection:** Datahub.io allows to find some datasets useful for linking.

- **Linking to external datasets:** The open source software Silk was used to find relationship between data items of our dataset and the external datasets generating the corresponding RDF links.
3.6 Publication

The goal of this activity is to make RDF datasets available on the Web to the users following the Linked Data principles. The steps in this activity are:

a) Dataset and vocabulary publication on the web.
   The data set was loaded and tested in an Open Link Virtuoso Sparql Endpoint.

b) Metadata definition and publication.
3.6 Publication

Open Link Virtuoso was chosen for publication considering the following features not completely present in other systems:

- Relational Data Management
- RDF Data Management
- XML Data Management
- Free Text Content Management & Full Text Indexing
- Document Web Server
- Linked Data Server
- Web Application Server
- Web Services Deployment (SOAP or REST)
3.6 Publication

```
select * from <http://localhost:8890/DAV/BibliotecaElectricaPrueba>
where {<http://datos.epn.edu.ec/resource/author/8> ?tipo ?descripcion.}
```
4. Open Link Data Architecture
5. CONCLUSIONS AND FUTURE WORK

We test successfully a proposed process and a technical architecture for publishing metadata from Koha Library systems on the Web using Linked Data technologies. For publishing we use OpenLink Virtuoso and Elda. In the future, we will work using SKOS (Simple Knowledge Organization System) to link the subjects and disciplines to another works to offer better queries to the users. Another work for the future is the alignment of the data model with activities of the publication process.
IMAGE CREDITS

- http://www.elsevier.com/societies/open-access
- http://www.loc.gov/marc/bibliographic/
- http://blogs.capita-libraries.co.uk/panlibus/category/semantic-web/
- http://www.w3.org/2005/Incubator/lld/wiki/images/a/ac/LLD-MetadataElementSetTagCloud.png
ACKNOWLEDGEMENT

This research has been partially supported by CEDIA (Consortio Ecuatoriano para el Desarrollo de Internet Avanzado) supporting the project: “Platform for publishing library bibliographic resources using Linked Data technologies” and by the project GEODAS-BI (TIN2012-37493-C03-03) supported by the Ministry of Economy and Competitiveness of Spain (MINECO).
¿Questions?