View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional de la Universidad de Alicante

TRANSFORMING LIBRARY CATALOGS INTO LINKED DATA.

M. Hallo[1], S. Luján-Mora[2] J. Trujillo[2]

[1]National Polytechnic School, Faculty of System Engineering (ECUADOR)
maria.hallo@epn.edu.ec
[2]University of Alicante, Department of Software and Computing Systems (SPAIN)
sergio.lujan@ua.es
jtrujillo@dlsi.ua.es

ABSTRACT

Traditionally, in most digital library environments, the discovery of resources takes place mostly through the harvesting and indexing of the metadata content. Such search and retrieval services provide very effective ways for persons to find items of interest but lacks the ability to lead users looking for potential related resources or to make more complex queries. In contrast, modern web information management techniques related to Semantic Web, a new form of the Web, encourages institutions, including libraries, to collect, link and share their data across the web in order to ease its processing by machines and humans offering better queries and results increasing the visibility and interoperability of the data.
Linked Data technologies enable connecting related data across the Web using the principles and recommendations set out by Tim Berners-Lee in 2006, resulting on the use of URIs (Uniform Resource Identifier) as identifiers for objects, and the use of RDF (Resource Description Framework) for links representation.
Today, libraries are giving increasing importance to the Semantic Web in a variety of ways like creating metadata models and publishing Linked Data from authority files, bibliographic catalogs, digital projects information or crowd sourced information from another projects like Wikipedia.
This paper reports a process for publishing library metadata on the Web using Linked Data technologies. The proposed process was applied for extracting metadata from a university library, representing them in RDF format and publishing them using a Sparql endpoint (an interface to a knowledge database). The library metadata from a subject were linked to external sources such us another libraries and then related to the bibliography from syllabus of the courses in order to discover missing subjects and new or out of date bibliography. In this process, the use of open standards facilitates the exploitation of knowledge from libraries.

Keywords: Linked Data, Semantic Web, Library Catalogs, RDF.

1 INTRODUCTION

Libraries and other cultural institutions are experiencing a time of changes. The metadata generated through the use of contemporary metadata standards and technical formats is mainly designed for human consumption rather than machine processing failing to interoperate with external information providers [1] . One possible improvement is provided by the standards established by the World Wide Web Consortium (W3C) to build the Semantic Web, a new form of the Web, to increase the visibility and interoperability of the data. Linked data, in particular, is an implementation of these standards useful to work with the metadata produced and maintained by libraries and other cultural institutions [2].

The term 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 [3]. The Linked Data principles are:

The last version of this paper was published at: ICERI2014 Proceedings*, 2014, pp. 1845-185, http://library.iated.org/view/HALLO2014TRA*

- Use Uniform Resource Identifiers (URIs) as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using common standards such as RDF (Resource Description Framework) and SPARQL (RDF query language).
- Include links to other URIs so that they can help to discover more things.

The URIs are used to identify a web resource. In addition, RDF is used for modelling and representation of information resources as structured data. In RDF, the fundamental unit of information is the subject-predicate-object triple. In each triple the "subject" denotes the source; the "object" denotes the target; and, the "predicate" denotes a verb that relates the source to the target. However, using these technologies alone, it is not possible to give semantics to data.
The Semantic Web Stack (Architecture of the Semantic Web) includes two technologies: RDFS (RDF Schema) and OWL (Web Ontology Language). RDFS is an extension of RDF that defines a vocabulary for the description of entity-relationships [4]. RDFS provides metadata terms to create hierarchies of entity types (referred to as "classes") and to restrict the domain and range of predicates. OWL is an extension of RDFS [5], which provides additional metadata terms for the description of complex models, which are referred to as "ontologies".
Some movements like LODLAM (Linked Open Data in Library, Archives and Museums) are working in sharing knowledge, tools and expertise using Linked Data in Libraries[1]. Several national libraries, such as British Library, French Library, Spanish National Library and libraries from universities, such as Michigan, Stanford, Cambridge, etc. have published linked datasets of bibliographic data that they have created. European Library is promoting Linked Open Data innovations in libraries across Europe [6].

This paper reports a process for publishing Library metadata on the Web using Linked Data technologies. The source metadata are Marc21 records from libraries using the Koha software system, based on best practices and recommendations from several authors. The proposed approach has been applied in a case study, "Electrical Enginnering Libray", in the context of the interuniversity project for publishing library bibliographic data using Linked Data technologies, funded by CEDIA ("*Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado*"). The extracted metadata were published in RDF format in a SPARQL endpoint. The library metadata were linked to external sources like another libraries and related to the pensums of the studies to discover missing subjects or out of date bibliography. In this process, the use of open standards facilitates the exploitation of knowledge from Libraries.

Some existing vocabularies and ontologies are used, such as FOAF (Friend of a friend), BIBO (Bibliographic Ontology), ORG (Organization Ontology) and DC (Dublin Core). In addition, the dataset created was linked to external data giving information that goes far beyond the bibliographic data provided by traditional libraries giving information, such us books used by similar educative institutions and faculties, books with similar subjects, authors of similar books, books cited in syllabus that should be replaced etc.

2 LIBRARY METADATA

Zeng and Qin define four kinds of metadata standards used in the library profession [7]:

- Structures like the Dublin Core Metadata Element Set (DCMES).
- Content like the Anglo-American Cataloging Rules, Second Edition (AACR2).
- Values like the Library of Congress Subject Headings (LCSH).
- Exchange like the MARC 21 Format for Bibliographic Data (MARC 21).

The data structure standards "will normally specify the metadata elements that are included in the scheme by giving each of them a name and a definition" [1]. Content standards "specify how values for metadata elements are selected and represented" .Zeng and Qin note that data value standards "include controlled term lists, classification schemes, thesauri, authority files,

[1] LODLAM Linked Open Data in Library, Archives and Museums: http://lodlam.net

and lists of subject headings". Finally, the data exchange standards allows libraries to exchange metadata coherently [7].

"MARC21 was developed by the Library of Congress (LC) in the mid-1960s, primarily to enable the computer production of catalog cards that could subsequently be distributed through the Cataloging Distribution Service" [1].

In addition to metadata standards, the metadata itself falls generally into three categories: descriptive, administrative and structural [8]. "Traditional library cataloging viewed as metadata is primarily descriptive", however, digital resources are more complex and require more than traditional description [7].

In many ways, libraries are traditional and continue to employ a variety of open or proprietary informational models such as MARC21 such as in the Koha integrated library system origin of this work. In contrast, modern web information management techniques related to Semantic Web encourages institutions, including libraries, to collect, link and share their data across the web in order to ease its processing by machines and humans to get better queries and results. Linked Data technologies enable connecting related data across the Web using the principles and recommendations set out by Tim Berners-Lee in 2006 [9]. The convergence between library metadata and linked data is based on the library interests (constructing vocabularies, describing properties of resources, identifying resources, exchanging and aggregating metadata) that are driving the development of Semantic Web technologies [10].

3 THE PROCESS FOR PUBLISHING LINKED DATA FROM MARC21 METADATA

Several approaches are being proposed to generate and publish linked data [11, 12], each one represented by activities and each activity composed of several task.

Our approach use six main activities: data source analysis, metadata extraction, modelling, RDF generation, linking and publishing. This process was also used to extract and publish scientific metadata from Open Journal Systems.

2.1 Data source analysis

The selected data sets were analyzed looking for attributes useful for answering the main queries. The steps in this activity are:

a) Identification of the data source and the attributes of interest to be published and linked to another datasets.

In this study, we have chosen a dataset with bibliographic metadata from the electrical engineering faculty from the National Polytechnic School and will be extended to the integrated library metadata considering the importance of the diffusion and interlinking of this information for the students and teachers.

b) Data source study

The sample library is using the Koha system. Koha is a web-based Integrated Library System, with MySQL database, which provides a simple and clear interface for library users to perform tasks such as searching for and reserving items and suggesting new items. Several national libraries are using the Koha system. In order to have a better knowledge about the books and their relation to scientific curriculums, the work was focused in the bibliographic metadata stored with Marc21 standard.

The analyzed system has the bibliographic material shown in Table 1:

Tabla 1: Bibliographic material in the analyzed library

CAT	Technical catalogs.
LIEE	Books specialized in Electrical and Electronic Engineering.
NTEC	Technical standards.
OLIT	literary Works.
PTEC	Institutional technical publications.
RELE	Electronic Resources (CDs, DVDs)
REV	Journals
TIEE	Engineering Thesis

The data in a MARC bibliographic record is organized into variable fields, each identified by a three-character numeric tag that is stored in the Directory entry for the field. The data fields, shown in the table 2 are grouped into blocks according to the first character of the tag, which with some exceptions identifies the function of the data within the record. The type of information in the field is identified by the remainder of the tag.

Table 2: Field types in Marc21[2].

Field	Description
0XX	Control information, identification and classification numbers, etc.
1XX	Main entries
2XX	Titles and title paragraph (title, edition, imprint)
3XX	Physical description, etc.
4XX	Series statements
5XX	Notes
6XX	Subject access fields
7XX	Added entries other than subject or series; linking fields
8XX	Series added entries, holdings, etc.
9XX	Reserved for local implementation

Within the 1XX, 4XX, 6XX, 7XX and 8XX blocks, certain parallels of content designation are usually preserved. The meanings, with some exceptions, are given to the final two characters of the tag of fields (see table 3):

Table 3: Additional meanings of fields in Marc21

[2] Marc21 Format for Bibliographic Data: http://www.loc.gov/marc/bibliographic/

X00	Personal names	X40	Bibliographic titles
X10	Corporate names	X50	Topical terms
X11	Meeting names	X51	Geographic names
X30	Uniform titles		

This data linked to another datasets will give us better knowledge about another books from an author, books published with similar subjects, the authors who work in a subject, the origin of an author, etc.

The metadata are stored in the biblio, biblioitem, itemtype Tables.

Tabla 4: Some Marc21 Fields used in the test

MARC	**Descripción**
003	Control Number Identifier
020 a	International Standard Book Number
041 a	Language code of text/sound track or separate title
082 2	Edition number
100 a	Personal name
245 a	Title
250 a	Edition statement
260 a	Place of publication, distribution, etc.
260 c	Date of publication, distribution, etc.
856 u	Uniform Resource Identifier

c) Identification of the licensing and provenance information

There is general information about the licensing of the data sets. In our case the data set has Creative Commons Attribution-ShareAlike 4.0 International licence[3] .
Provenance information about a data item is information about the history of the item, including information about its origins. It is a measure about the quality of data.
In our case study, the provenance data are the name of the library: Electrical Enginnering library from National Polytechnic School, initial load data: 01-09-2014.

3.2 Metadata extraction

In this activity the metadata are extracted from the original source and stored in an intermediate database for cleaning. The tasks in this activity are:

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

b) Data cleaning
In the case analized we found absence of data in several fields that could be filled from another open data sets.
An initial pre-processing of the data applying data clean techniques, was performed. Spoon and Silk were used to get the catalogues of authors, works, expressions, manifestations and organizations.

[3] Creative Commons Licences: https://creativecommons.org/licenses/

3.3 Modelling

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

The steps in this activity are:

a) Selection of vocabularies

The most important recommendation from several studies is to reuse available vocabularies as much as possible to develop the ontologies. An ontology represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts [13]. We use the following controlled vocabularies and ontologies for modelling works (books, articles), manifestations (formats, language), expressions (editions), authors, organizations.

- BIBO[4] (The Bibliographic Ontology) provides main concepts and properties for describing citations and bibliographic references (e.g. books, articles, etc.) on the Semantic Web using RDF.
- Dublin Core[5] is a set of terms that can be used to describe web resources as well as physical resources such as books. It consists of fifteen fields, e.g., creator, contributor, format, identifier, language, publisher, relation, rights, source, title, type, subject, coverage, description, and date. The full set of Dublin Core metadata terms can be found on the Dublin Core Metadata[6]. Dublin Core Metadata may be used to provide interoperability in Semantic Web implementations combining metadata vocabularies of different metadata standards.
- FOAF[7] (Friend of a Friend) is a machine-readable ontology describing persons, their activities and relations to other people and objects in RDF format.
- ORG[8] (Organization) is an ontology for organizational structures, aimed at supporting linked data publishing of organizational information. It is designed to add classification of organizations and roles, as well as extensions to support information such as organizational activities.
- FRBR[9] (Functional Requirements for Bibliographic Records) conceptual model relating data from bibliographic records.
- SKOS[10] (Simple Knowledge Organization System) is a OWL ontology that provides a way to represent controlled vocabularies, taxonomies and thesauri.

b) Vocabulary development and Documentation

The vocabulary was documented using Protégé (Ontology Editor Tool)[11].

Table 1. Vocabularies and Namespaces

Vocabulary/Ontology	*Namespaces*

[4] The Bibliographic Ontology: http://bibliontology.com/

[5] Dublin Core Metadata Element Set, version 1.1: http://dublincore.org/documents/dces/

[6] DCMI Metadata Terms: http://dublincore.org/documents/dcmi-type-vocabulary/index.shtml

[7] The Friend of a Friend (FOAF) project: http://www.foaf-project.org/

[8] The Organization Ontology: http://www.w3.org/TR/vocab-org/

[9] Functional Requirements for Bibliographic Records: http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[10] Introductin to SKOS: http://semanticweb.com/introduction-to-skos_b33086

[11] Protégé: http://protege.stanford.edu/

ORG	http://www.w3.org/ns/org#
FOAF	http://xmlns.com/foaf/0.1/
DC	http://xmlns.com/dc/0.1/
DCTERMS	http://purl.org/dc/terms/
BIBO	http://purl.org/ontology/bibo/
SKOS	http://www.w3.org/2004/02/skos/core#
RDFS	http://www.w3.org/2000/01/rdf-schema#
OWL	http://www.w3.org/2002/07/owl#
FRBR-RDA	http://rdvocab.info/uri/schema/FRBRentitiesRDA/

c) Vocabulary validation

Ontology validation is a key activity in different ontology engineering scenarios such as development and selection, that is, assessing their quality and correctness [14].
The generate vocabulary was validate with OOPS![12].

d) Specify a license for the dataset

The license to publish the datasets is Creative Commons Attribution-ShareAlike 4.0 International [13].

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. The tasks in this activity are:

a) Selection of development technologies for RDF generation.
For the study case the Triplify[14] tool with some modifications has been used to perform the transformation from the intermediate relational database to RDF.

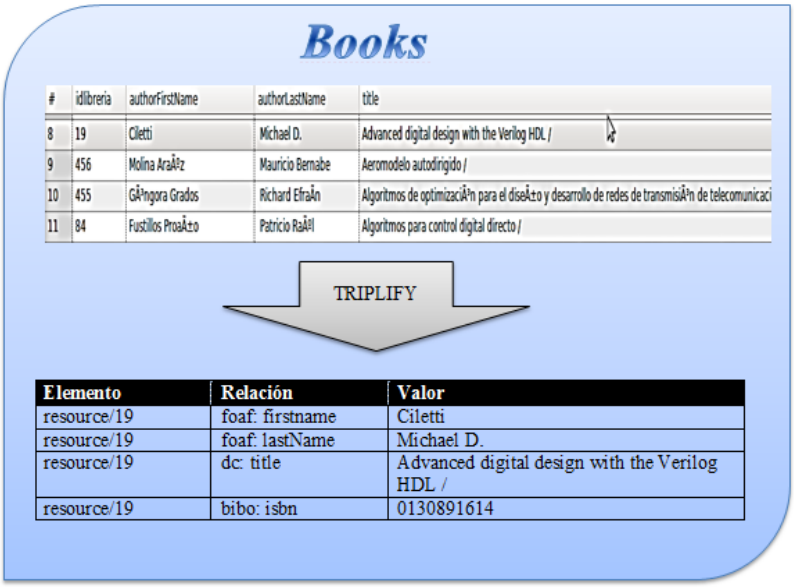
b) Mappings from data sources to RDF.
Mappings were defined from the intermediate data base with metadata extracted from the source system to RDF.

c) Transformation of data.
The process of transformation was run with the open source software Triplify 1.0 getting RDF triples stored in RDF/XML format. Fig. 3 shows part of this process.

[12] Ontology Pitfall Scanner: http://www.oeg-upm.net/oops

[13] Creative Commons: http://creativecommons.org/

[14] Triplify: http://triplify.org/



3.5 Interlinking

The objective of this activity is to improve the connectivity to external datasets enabling other applications to discover additional data sources.
The tasks corresponding to this activity are:

a) Target datasets discovery and selection

For this task we used the website Datahub.io[15] to find some datasets useful for linking. We found several open linked datasets like Open Library[16] with books records useful to help in the cataloguing process, Europeana Linked Open Data[17], Library of Congress Subject Headings [18].

b) Linking to external datasets

The open source software Silk[19] was used to find relationship between data items of our dataset and the external datasets generating the corresponding RDF links that were stored in a separated dataset. The links with test books from the syllabus of the courses in the Electrical Engineering Faculty at National Polytechnic School are also generated in order to discover missing subjects and new or out of date bibliography.

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 generated triples were loaded into a SPARQL endpoint (a conformant SPARQL protocol service) based on OpenLink Virtuoso[20], which is a database engine that combines the functionality of RDBMS, virtual databases, RDF triple stores, XML store, web application server and file servers. On the top of OpenLink Virtuoso; ELDA[21] is used as a Linked Data interface to

[15] Datahub: http://datahub.io/

[16] Open Library: http://datahub.io/dataset/open-library

[17] Europeana Linked Open Data: http://datahub.io/dataset/europeana-lod

[18] Library of Congress Subject Headings: http://datahub.io/dataset/lcsh

[19] Silk – A Link Discovery Framework for the Web of Data: http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

[20] Virtuoso Universal Server: http://virtuoso.openlinksw.com/

[21] Elda– the linked-data API in Java: http://www.epimorphics.com/web/tools/elda.html

the RDF data. Fig. 4 shows a view of the SPARQL endpoint with a partial result of the query about an article in a test platform:

Fig. 4 SPARQL endpoint query

b) Metadata definition and publication

Metadata recommended for publishing Linked Data sets are: organization and/or agency, creation date, modification date, version, frequency of updates, and contact email address [14].

The metadata will be published in the site Datahub.io using DCAT (Data Catalog Vocabulary)[22], an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.

The whole architecture used in this project is shown in the Fig. 5.

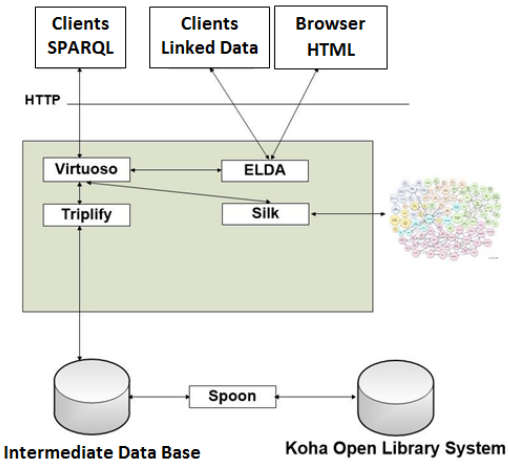


Fig.5. Open Linked Data architecture

4 CONCLUSIONS AND FUTURE WORK

In this paper we refine and use a process for publishing metadata from Koha Library systems on the Web using Linked Data technologies. The process was based in best practices and recommendations from several studies, adding tasks and activities considered important during

[22] Data Catalog Vocabulary(DCAT): http://www.w3.org/TR/vocab-dcat/

the project development. The process was applied to the transformation of metadata from a Koha Libray System to RDF. The source of metadata are bibliographic records in Marc21 format. A preliminary mapping from Marc21 to RDF was made before the generation of RDF. .For publishing we use OpenLink Virtuoso and Elda that was tested like interface to the SQL enpoint. 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. We are also analyzing the best way to validate the generated external links. Another work for the future is the alignment of the data model with activities of the publication process.

ACKNOWLEDGEMENT

This research has been partially supported by the Prometeo project by SENESCYT, Ecuadorian Government, by CEDIA (*Consorcio 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).

REFERENCES

[1] Caplan, P. 2003. *Metadata Fundamentals for All Librarians*. Chicago: American Library Association.

[2] Baker, T., E. Bermès, K. Coyle, G. Dunsire, A. Isaac, P. Murray, M. Panzer, J. Schneider, R. Singer, E. Summers, W. Waites, J. Young and M. Zeng. 2011. *Library Linked Data Incubator Group Final Report*. World Wide Web Consortium, accessed September 18, 2012, www.w3.org/2005/Incubator/lld/XGR-lld-20111025.

[3] Berners-Lee, T. (2006). Linked Data - Design Issues. Available at: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed sept 15, 2014].

[4] Guha, RV., Brickley, D.: RDF vocabulary description language 1.0: RDF Schema.W3C Recommendation, W3C. 2004. Available at: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.[Accessed May 15, 2014].

[5] Hayes, P., Patel-Schneider, PF., Horrocks, I. (2004): OWL web ontology language semantics and abstract syntax. W3C Recommendation, W3C. 2004. Available at: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.[Accessed May 10, 2014].

[6] European Library (2013). Linked Open Data. Available at: http://www.theeuropeanlibrary.org/tel4/lod. [Accessed Sept 2, 2014].

[7] Zeng, M. L. and J. Qin. (2008). *Metadata*. New York: Neal-Schuman

[8] NISO. 2004. *Understanding Metadata*, accessed September 18, 2012, www.niso.org/publications/press/UnderstandingMetadata.pdf.

[9] Berners-Lee, T. (2006). Linked Data - Design Issues. Available at: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed Sept 15, 2014].

[10] Heery, R. 2004. "Metadata Futures: Steps Toward Semantic Interoperability." In *Metadata in Practice*, edited by D. I. Hillman and E. L. Westbrooks, 257–71. Chicago: American Library Association.

[11] Auer, S. and J. Lehmann. 2010. Making the Web a Data Washing Machine—Creating Knowledge Out of Interlinked Data. *Semantic Web Journal*, accessed September 18, 2012, www.semantic-web-journal.net/content/new-submission-making-web-data-wash.

[12] Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In *Linking Government Data*. Springer New York, pp 27-49.

[13] Kim, J. A., & Choi, S. Y. (2007). Evaluation of Ontology Development Methodology with CMM-i. In Software Engineering Research, Management & Applications, SERA 2007. 5th ACIS International Conference, IEEE, pp. 823-827.

[14] Poveda-Villalón, M. Suárez-Figueroa, M., and Gómez-Pérez, A. (2012). Validating ontologies with OOPS!. In Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management *(EKAW'12),* Teije,A., Völker,J., Handschuh,S., Heiner Stuckenschmidt, H., and d'Acquin,M. (Eds.). Springer-Verlag, Berlin, Heidelberg, pp 267-268.