AN APPROACH TO PUBLISH SCIENTIFIC DATA OF OPEN-ACCESS JOURNALS USING LINKED DATA TECHNOLOGIES

M.Hallo[1], S. Luján-Mora[2] C.Chávez[1]

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

Abstract

Semantic Web encourages digital libraries, including open access journals, 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.

Several universities develop knowledge through scholarship and research with open access policies for the generated knowledge, using several ways to disseminate information. Open access journals collect, preserve and publish scientific information in digital form related to a particular academic discipline in a peer review process having a big potential for exchanging and spreading their data linked to external resources using Linked Data technologies. Linked Data can increase those benefits with better queries about the resources and their relationships.

This paper reports a process for publishing scientific data on the Web using Linked Data technologies. Furthermore, methodological guidelines are presented with related activities. The proposed process was applied extracting data from a university Open Journal System and publishing in a SPARQL endpoint using the open source edition of OpenLink Virtuoso. In this process, the use of open standards facilitates the creation, development and exploitation of knowledge.

Keywords: Scientific data, Linked Data, Open access journals, Semantic Web.

1 INTRODUCTION

Open access (OA) is the free unrestricted online access to digital content. The open access movement began in the 1990s, at the same time the World Wide Web became widely available and Open Access Journal Publishing begin to grow. In 2003, the Budapest Open Access Initiative (BOAI) launched a worldwide campaign for open access to peer-reviewed research [1].

Several universities like Harvard, Stanford, MIT, have adopted guides to good practices for university open access policies. University transmits knowledge through scholarship and research. In both roles there are many initiatives to openly share knowledge and resources. For example, in education there is the Open Education Resource (OER) initiative [2]. On the other hand, in different research fields several forms of open knowledge diffusion have been implemented like open archives, open access journals, blogs and websites, helping users to easily create new developments and new knowledge.

Open access journals collect, preserve and publish scientific information in digital form related to a particular subject. The development of Information and Communication Technologies (ICTs) has increased the number of open access scientific journals in digital format, speeding up dissemination and access to content [3]. However, bibliographic data are dispersed, without relationship between resources and data sets making difficult their discovery and reuse for other information systems. To address these issues, we propose a process for publishing bibliographic data from open journal systems following the Linked Data principles.

The proposed approach has been applied in a case study, “Revista Politécnica”, 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*”) and developed in Ecuador by National Polytechnic School, University of Cuenca and Privated Technical University of Loja.

The last version of this paper was published at: *EDULEARN14 Proceedings*, 2014, pp. 1145-1153.

http://library.iated.org/view/HALLO2014ANA,

Open access journals have a big potential for exchanging and spreading their data linked to external resources using Linked Data technologies, especially in the context of the Open Data Movement [4].

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 [5]. The Linked Data principles are:

- 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 URI is 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. Using a combination of URIs and RDF, it is possible to give identity and structure to data. 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 [6]. 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 [7], 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 [8].

The proposed process for publishing bibliographic linked data was developed based on best practices and recommendations from several authors and tested with data from the electronic version of the journal "Revista Politécnica" edited by National Polytechnic School.

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 publishers giving information, such us authors, publishing papers with similar subjects or organizations sponsoring research in specific subjects, etc.

2 THE PROCESS FOR PUBLISHING OPEN JOURNAL SCIENTIFIC LINKED DATA

Several approaches are being proposed to generate and publish linked data [9, 10], each one represented by activities and each activity composed of several task.

Our approach proposes six main activities: data source analysis, metadata extraction, modelling, RDF generation, linking and publishing. Fig. 1 shows the Life cycle of the Open Linked Data.

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

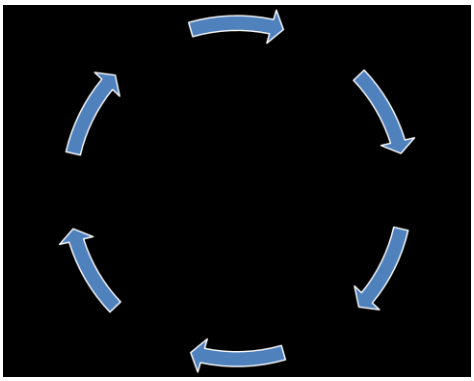

Fig. 1 Open Linked Data Life Cycle

2.1 Data source analyze

The objective of this activity is: to identify data sets that provide benefits to others for reuse. The selected data sets are analyzed looking for attributes useful for answering the 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 publications from a university open journal considering the importance of the diffusion and interlinking of this information.

b) Engaging stakeholders

In this step we explain stakeholders (principals of several universities and a funding organization) the process and benefits of creating and maintaining Linked Open Data related to a scientific information published in the academic journal, after we develop an inter university project for funding.

c) Data source analyze

Several journals affiliated to the open access initiative have adopted the Open Journal System[2], software that provides an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Endpoint. Although this protocol has been widely adopted, it has some problems, including use of non-dereferenceable identifiers and limitations of selective access to metadata [11]. The open journal analyzed uses the open source OJS (*Open Journal Systems)* for the management of peer-reviewed academic journals. The used data set is stored in a MySQL database. In order to have a better knowledge about the scientific publications, the work was focused in the articles stored with Dublin Core metadata, which is a vocabulary for resource description. Following, there are some examples coded with Dublin Core metadata:

- Identifier(dc:identifier): http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf
- Titles (dc:title): “Linked Data - The Story So Far”
- Authors (dc:creator): “Tim Berners-Lee”
- Keywords (dc:subjects): “Linked Data , Web of Data, Semantic Web, Data Sharing”.

[2] Open Journal System: https://pkp.sfu.ca/ojs/

This data linked to another datasets will give us better knowledge about similar subjects published, the authors who work in them, the organizations sponsoring similar research.

d) Identification of the licensing and provenance information

There is general information about the licensing in the analyzed open journal. It is possible to get information from the online journal and reproduce citing the source. This text is added in a dc:rights property.

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 journal, the type of publication (peer-review) and the name of the publisher. In the future, another data about the peer review process could be added.

2.2 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: article, authors and organization.

b) Disambiguation of entities with different values

In the case study we found some problems in the data like typographic mistakes or several authors with similar names, the same for the affiliation data. Additionally, author names were formatted differently. For example, the same author could appear in one document as “Lujan, S.”, in another as “Luján-Mora, S.” and in another as “Luján, Sergio”. A data cleaning process matches the documents of this author and groups these name variants together so that authors, even if cited differently, are linked to their papers. In this step, author names are grouped together under a single identifier number, a process matches author names based on their affiliation, and email address grouping together all of the documents written by that author.

When grouping author names under a unique author identifier number, we should take into consideration last name variations, all possible combinations of first and last names, and the author name with and without initials. As a result, searches for a specific author include a preferred name and variants of the preferred name. This problem is solved in some systems like Scopus, showing potential author matches like in Fig 2. In addition, the authors have the possibility of reporting mistakes. This functionality is planned to be added to our system in a next stage.

An initial pre-processing of the data applying data clean techniques, was performed. Spoon and Silk were used to get the catalogues of authors and author’s affiliations.

Discipline, data delivered by the authors in the study case, and keywords were cleaned by matching data with catalogues of features and thesaurus for linking with SKOS (Simple Knowledge Organization System) concepts.

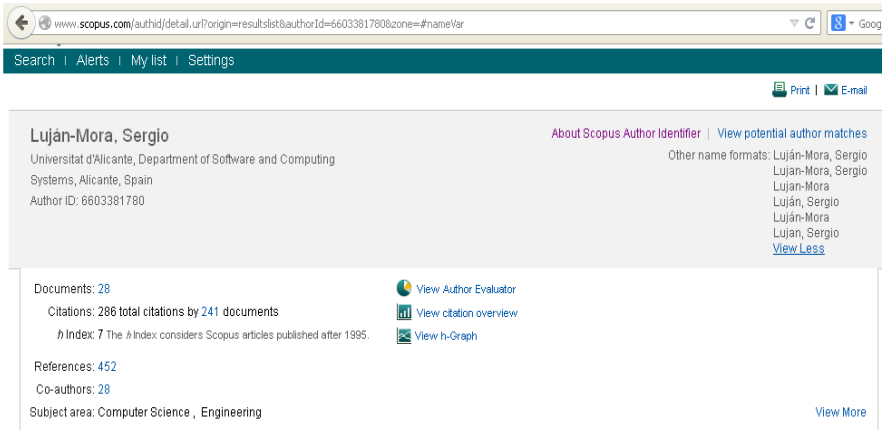

Fig. 2 Scopus author profile with different name formats

2.3 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 [12]. We use the following controlled vocabularies and ontologies for modelling journals, authors, affiliation:

- BIBO[3] (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[4] 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[5]. Dublin Core Metadata may be used to provide interoperability in Semantic Web implementations combining metadata vocabularies of different metadata standards.
- FOAF[6] (Friend of a Friend) is a machine-readable ontology describing persons, their activities and relations to other people and objects in RDF format.
- ORG[7] (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. The namespaces used are shown in Table 1.

b) Vocabulary development and Documentation

The vocabulary was documented using Protégé (Ontology Editor Tool)[8].

---

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

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

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

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

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

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

Table 1. Vocabularies and Namespaces

Vocabulary/Ontology	Namespaces
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/

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 [13].

The generate vocabulary was validate with OOPS![9].

d) Specify a license for the dataset

The license to publish the datasets was Creative Commons[10].

2.4 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 of technologies for RDF generation

For the study case the Triplify[11] 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

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 getting RDF triples stored in RDF/XML format. Fig. 3 shows part of this process.

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

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

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

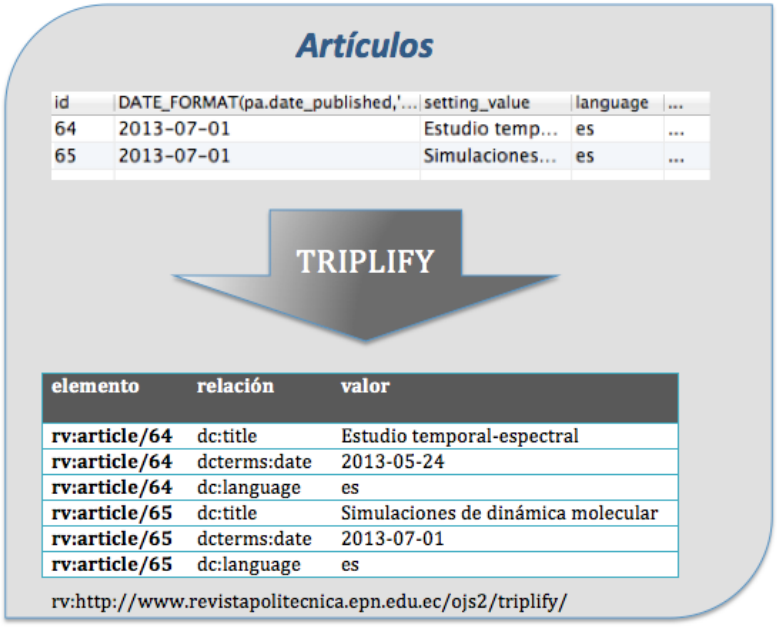

Fig.3 RDF Generation using Triplify

2.5 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 the Datahub[12] to find some datasets useful for linking. We found several open linked datasets from scientific journals.

b) Linking to external datasets

The open source software Silk[13] 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.

2.6 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[14], 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

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

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

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

OpenLink Virtuoso; Pubby[15] is used as a Linked Data interface to 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:

```
select * from <http://192.168.203.128:8890/DAV/home/datasetojs3>
where {<http://www.revistapolitecnica.epn.edu.ec/ojs2/triplify/article/41> ?y ?z}
```

http://purl.org /dc/terms /dateSubmitted	2013-05-21
http://purl.org /dc/elements /1.1/subject	Motor de vce, desulfurizaciorre de enfriamiento, energ exerg, Vortex engine, desulfurization, cooling tower, energy, exergy
http://purl.org /dc/terms/title	Vortex Engine Like New Technology of Flue Gas Discharging with Wet Desulfurization
http://purl.org /dc/elements /1.1/format	pdf
http://purl.org /ontology/bibo/uri	http://www.revistapolitecnica.epn.edu.ec/ojs2/index.php /revista_politecnica2/article/view/41

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 were published in the site Datahub using DCAT (Data Catalog Vocabulary)[16], 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.

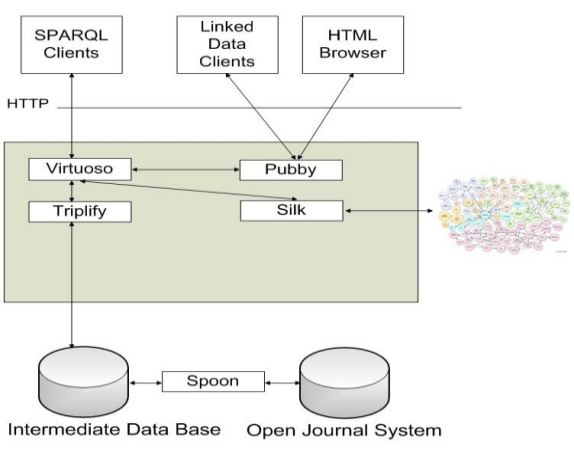

Fig.5. Open Linked Data used architecture

[15] Pubby – A Linked Data Frontend for SPARQL Endpoints: http://www4.wiwiss.fu-berlin.de/pubby/

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

3 CONCLUSIONS AND FUTURE WORK

In this paper we analyze and use a process for publishing scientific data from Open Journal 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 the project development. The process was applied to the transformation of metadata from “Revista Politécnica” to RDF. For publishing we use OpenLink Virtuoso and Pubby.

The process could be also applied to bibliographic metadata harvested through of the OAI-PMH Protocol linking integrated metadata from Open Journal Systems.

The Dublin Core standard used in the source metadata was enough for the integration of data from open journal systems to help answering our questions. For another bibliographic resources it should be important analyze FRBR (Functional Requirements for Bibliographic Records) and RDA (Resource Description and Access) standards to get interoperability with another digital libraries.

For the future a new interface is being developed to ask users to fix errors in author disambiguation, grouping papers par author and organization disambiguation. In addition a team will make the maintenance’s task to be able to publish all the new data with the better quality possible, data curation is the key of the success of Linked Data. Moreover, we will work in using SKOS *(*Simple Knowledge Organization System) to link the papers 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 and by CEDIA (*Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado*) supporting the project: “Platform for publishing library bibliographic resources using Linked Data technologies”.

REFERENCES

[1] Chan, L. et ál (2002). Read the open access initiative. Available at: http://www.budapestopenaccessinitiative.org/read.

[2] Center for Educational Research and Innovation (CERI), 2007. *Giving Knowledge for Free: The Emergence of Open Educational Resources*, Organisation for Economic Co-operation and Development. Available at: http://www.oecd.org/dataoecd/35/7/38654317.pdf [Accessed May 8, 2014].

[3] Harnad, S. (2009). Open access scientometrics and the UK Research Assessment Exercise. *Scientometrics*, *79*(1), 147-156.

[4] Suber, P. (2009). Timeline of the open access movement. Available at: http://www. earlham. edu/~ peters/fos/timeline. Htm. [Accessed May 12, 2014].

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

[6] 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].

[7] 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].

[8] European Library (2013). Linked Open Data. Available at: http://www.theeuropeanlibrary.org/tel4/lod. [Accessed May 11, 2014].

[9] 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.

[10] Hausenblas, M., et al. (2013). Linked Data, Manning Publications Company.

[11] Hakimjavadi, H., Masrek, M. N. & Alam, S. (2012). SW-MIS: A Semantic Web Based Model for Integration of Institutional Repositories Metadata Records. *Science Series Data Report,* 4(11), pp 57–66.

[12] 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.

[13] 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.

[14] Gómez-Pérez, A. Vila-Suero, D., Montiel-Ponsoda, D., Gracia, J. and Aguado-de-Cea, G. (2013). Guidelines for multilingual linked data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (WIMS '13). ACM, New York, NY, USA.