AN APPROACH TO PUBLISH STATISTICS FROM OPEN-ACCESS JOURNALS USING LINKED DATA TECHNOLOGIES

M.Hallo[1], S. Luján-Mora[2], J.Trujillo[3]

[1]*National Polytechnic School (ECUADOR)*
maria.hallo@epn.edu.ec
[2]*Visiting teacher at the National Polytechnic School, University of Alicante (SPAIN)*
sergio.lujan@ua.es
[3]*University of Alicante (SPAIN)*
jtrujillo@ua.es

Abstract

Semantic web encourages digital libraries that include 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 structured 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, under open access policies and use several ways to disseminate information. Open access journals collect, preserve and publish scientific information in digital form using a peer review process. The evaluation of the usage of this kind of publications needs to be expressed in statistics and linked to external resources to give better information about the resources and their relationships. The statistics expressed in a data mart facilitate queries about the history of journals usage by several criteria. This data linked to another datasets gives more information such as: the topics in the research, the origin of the authors, the relation to the national plans, and the relations with the study curriculums.

This paper reports a process to publish an open access journal data mart on the web using Linked Data technologies in such a way that it can be linked to related datasets. Furthermore, methodological guidelines are presented with related activities. The proposed process was applied extracting data from a university open journal system data mart and publishing it in a SPARQL endpoint using the open source edition of the software OpenLink Virtuoso. In this process the use of open standards facilitates the creation, development and exploitation of knowledge. The RDF data cube vocabulary has been used as a model to publish the multidimensional data on the web. The visualization was made using CubeViz a faceted browser filtering observations to be presented interactively in charts. The proposed process help to publish statistical datasets in an easy way.

Keywords: Linked Data, university institutional repositories, semantic web, statistical data, RDF data cube vocabulary, SDMX, data modeling, data transformation, knowledge management.

1 INTRODUCTION

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 [1].

A growing number of scholarly journals are using Open Journal Systems (OJS), an open source software platform, specially designed to manage articles trough author submission, peer review, editing and publication. This system provides the journal manager with the ability to extract year-by-year statistics about the history of the journal's usage for data on submissions, editorial practices, and users grouped by editor and reviewer [2].

Statistical data from open journal systems are important for policy definition, planning and control with relevant impact into the society. Libraries routinely collect statistics on digital

The last version of this paper was published at: INTED2015 Proceedings, pp. 5940-5948*., 2015.*

collection use for assessment and evaluation purposes. Libraries then report those statistics to a variety of stakeholders. However, statistics from bibliographic data are dispersed, without relationship between resources and data sets making difficult their discovery and reuse for other information systems. Moreover, there exists no simple way for researchers, journalists and interested people to compare statistical data retrieved from different data stores on the web because of lack of standardization.

To address these issues, we propose a process to publish a data mart, a part of a data warehouse, of statistical data from OJS following the Linked Data principles.

The proposed approach was developed based on best practices and recommendations from several authors [3,4,5] and tested with data from the electronic version of the journal "Revista Politécnica" edited by National Polytechnic School of Quito-Ecuador. 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 and high number of visits, or organizations sponsoring research in specific subjects, statistical indicators below national standards, etc.

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 [6].

The URI (Uniforme Resource Identification) is used to identify a web resource. In addition, RDF (Resource Description Framework) is used for modeling 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 [7]. 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 [8], which provides additional metadata terms for the description of complex models, which are referred to as "ontologies".

Some existing vocabularies and ontologies are used, such as FOAF (Friend of a friend), BIBO (Bibliographic Ontology), ORG (Organization Ontology) and DC (Dublin Core).

The RDF data cube vocabulary has been proposed to describe statistical data organized in a multidimensional model, helping to publish, discover and link statistical data in a uniform way. The model has been proposed based on the ISO 17369:2013 [9].

There are a number of benefits to being able to publish multi-dimensional data, such as statistics, using RDF and the linked data technologies. The W3C recommendation about the RDF data cube vocabulary, present the following benefits:

- The individual observations, and groups of observations, become (web) addressable. This allows publishers and third parties to annotate and link to this data; for example for fine grained provenance trace-back.
- Data can be flexibly combined across datasets. The statistical data becomes an integral part of the web of linked data.
- For publishers who currently only offer static files then publishing as linked-data offers a flexible, non-proprietary, machine readable means of publication that supports an out-of-the-box web API for programmatic access.
- It enables reuse of standardized tools and components.

2 LINKED DATA PUBLICATION PROCESS FOR STATISTICAL DATA

The proposed process allows to publish an statistical data mart in the RDF format using common vocabularies such us RDF data cube from a OJS data mart.

Our approach proposes five main activities:

- Data source analysis.
- RDF data modeling.
- RDF generation.
- Linking.
- Publishing.

2.1 Data source analysis

In this activity the data mart for the publication is selected and the licensing and provenance information is defined. Following we describe those steps:

a) Data source selection

In this step we chose a data mart to publish RDF format, considering that linked to another datasets will give us better knowledge.

The data mart is a subset of the data warehouse that is usually oriented to a specific business topic, is represented with the multidimensional model, or else cube model, comprised of three basic components: dimensions, measures and attributes.

A data mart is represented with a multidimensional model. In this case we have worked with a data mart about OJS article visits.

The open journal selected for analysis uses the open source OJS[1] for the management of peer-reviewed academic journals. Several universities have adopted the OJS, software that provides an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Endpoint.

The data mart used for the test of this proposal is shown in the fig.1.

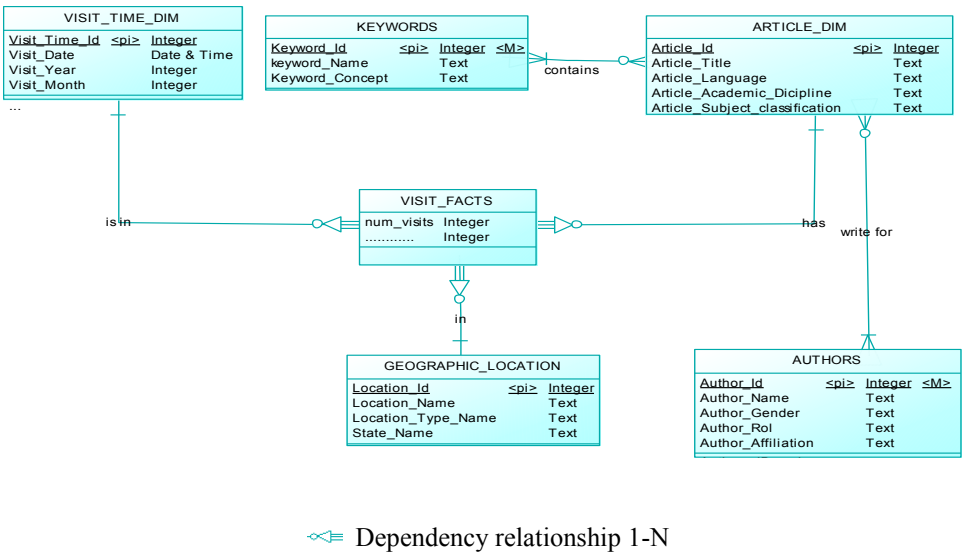


Fig. 1. OJS visits data mart

The used dataset is stored in a MySQL database.

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.

Identification of the licensing and provenance information

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

There is general information about the licensing in the analyzed open journal. It is possible to get statistical information from the online journal and reproduce citing the source.

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 documented using terms of the PROV[2] data model. PROV defines a data model building representations of the entities, people and processes involved in producing a piece of data or thing in the world.

In the future the source data mart will be documented using the SDMX (Statistical Data and Metadata Exchange) standard, which in turn define a set of cross-domain concepts, code lists and categories in order to provide compatibility and interoperability across institutions.

2.2 RDF data modeling

The goal of this activity is to design and implement a vocabulary for describing the statistics datasets 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 [10]. We use the following controlled vocabularies and ontologies for modelling statistical datasets in RDF:

- RDF data cube vocabulary[3] is an standard that provides a means to publish multi-dimensional data, such as statistics, on the web.
- 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. 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.
- SKOS[9] (Simple Knowledge Organization System) for concepts and concept schemes.

The namespaces used are shown in Table 1.

b) Vocabulary development and Documentation

The vocabulary was documented using Protégé (Ontology Editor Tool)[10].

[2] Prov: http://www.w3.org/TR/prov-primer/

[3] RDF data cube vocabulary: http://www.w3.org/TR/vocab-data-cube/

[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(ORG): http://www.w3.org/TR/vocab-org/

[9] Simple Knowledge Organization System(SKOS): http://www.w3.org/2004/02/skos/

Table 1. Vocabularies and Namespaces

Vocabulary/Ontology	Namespaces
QB	http://purl.org/linked-data/cube#
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#

The reduced RDF data cube model used in this work is presented in fig. 2.

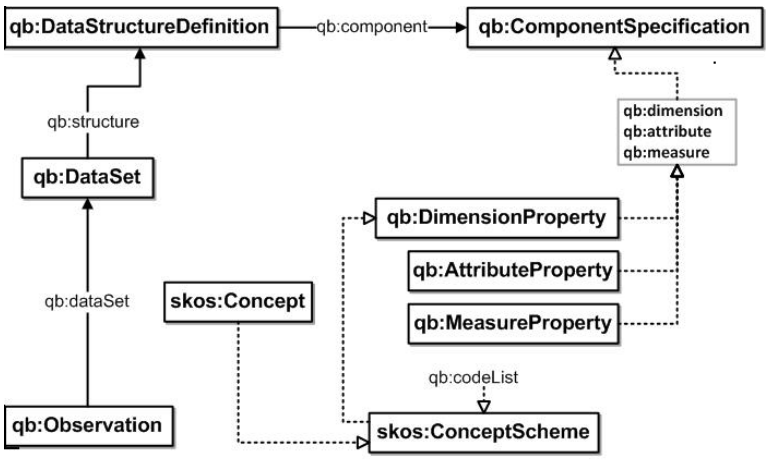


Fig 2. A reduced RDF data cube vocabulary

Following the RDF data cube vocabulary we define the data model corresponding to the selected data mart with the dimensions, measures and attributes components.

The reduced RDF data model developed is presented in fig. 3

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

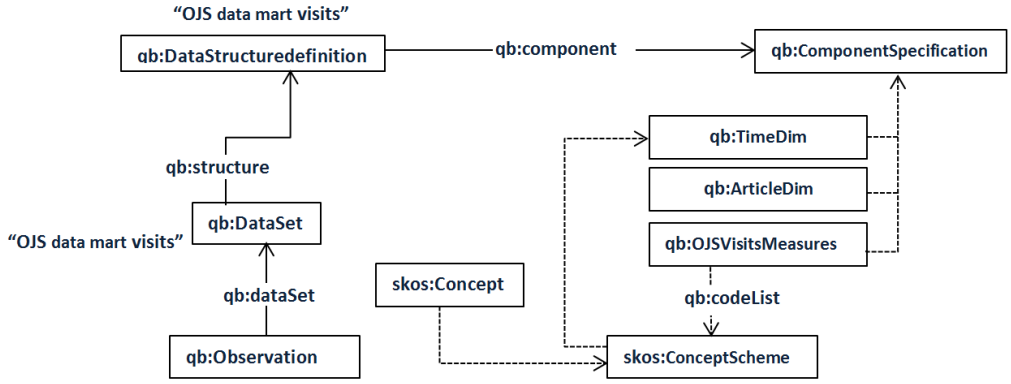


Fig 3. The reduced OJS RDF data model

Each concept is mapped with the corresponding concept of the multi-dimensional model, such as dimension, measure, code list, etc.

The URI structure was defined by:

(1) Schema components (dimensions, measures, and attributes), which are identified by a URI of the form *{Base_URI }/dc/cube_name/prop/{dimension_name* | *measure_name* |*attribute}*.

(2) Datasets are identified by *{Base_URI}/}/dc/cube_name/dataset/{DatasetName}* and the dataset component specification by *{ Base_URI}/dc/cube_name/dccs /{dimension_name*|*measure_name}* respectively, (c) Concepts and their values reused across multiple datasets are identified by *{ Base_URI}/concept/ {ConceptName}* and *{ Base_URI }/concept/{ConceptName}/{value}*.

c) Specify a license for the dataset

The license to publish the RDF datasets is Creative Commons Attribution-ShareAlike 4.0 International [11].

2.3 RDF Generation

The goal of this activity is to define a method and technologies to transform the source data into 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 Open Refine[12] tool has been used to perform the transformation from the multidimensional model stored in a relational database to RDF.

b) Mappings from data sources to RDF

Mappings were defined from the multidimensional data base to RDF.

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

[12] Open Refine: http://openrefine.org/

This step involves mapping of dataset's concepts to the RDF data cube elements, e.g., dimensions as qb:DimensionProperty, measures as qb:MeasureProperty or attributes as qb:AttributeProperty, the identification of the data (observations) as qb:Observation instances. Concepts within the datasets may be mapped with another concepts and code lists providing compatibility and interoperability. The mappings are used to create the dataset's structure, the dataset itself and the observations, using the appropriate URI Scheme for each type of resource. A default URI scheme has been designed as an input to this step to easily map the instances of the data cube vocabulary and the resources. The code lists that are used to give a value to any of the components are also defined using SKOS vocabulary. The data are then exported as RDF in an RDF compliant serialization, such as RDF/XML and validated. Fig.4 shows part of the data cube generated.

```
@prefix qb: <http://purl.org/linked-data/cube#>

<http://opendata.epn.edu.ec/dc/ojsvisits/prop/articleDim> a
qb:DimensionProperty ;

rdfs:label "ArticleDim"@en .

<http://opendata.epn.edu.ec/dc/ojsvisits/prop/timeDim> a
qb:DimensionProperty ;

rdfs:label "TimeDim"@en .

<http://opendata.epn.edu.ec/dc/ojsvisits/prop/OJSvisitsMeasures> a
qb:MeasureProperty ;

rdfs:label "OJSVisits"@en .

<http://opendata.epn.edu.ec/dc/ojsvisits/dataset/dataset-ojsvisits> a
qb:DataSet ;

rdfs:comment "OJS Visits Data Set"@en ;

a qb:DataStructureDefinition .

<http://opendata.epn.edu.ec/dc/dataset-ojsvisits/dccs/articleDim> a
qb:ComponentSpecification ;

qb:dimension <http://opendata.epn.edu.ec/dc/ojsvisits/prop/articleDim> .

<http://opendata.epn.edu.ec/dc/ojsvisits/dataset/dataset-ojsvisits>
qb:component <http://opendata.epn.edu.ec/dc/dataset-
ojsvisits/dccs/articleDim> .

<http://opendata.epn.edu.ec/dc/dataset-ojsvisits/dccs/timeDim> a
qb:ComponentSpecification ;

qb:dimension <http://opendata.epn.edu.ec/dc/ojsvisits/prop/timeDim> .
```

Fig. 4 Partial turtle code generated for the data cube

c) Transformation of data

The process of transformation was run with the software Open Refine getting RDF triples stored in RDF/XML format using RDF data cube vocabulary.

2.4 Interlinking

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

The different versions of code lists coming from the same resource are interlinked with each other using the appropriate linking property, e.g. skos:exactMatch for concepts

The tasks corresponding to this activity are:

a) Target datasets discovery and selection

For this task we used the website "the Datahub"[13] to find some datasets useful for linking. We found several open linked statistics datasets from scientific journals.

b) Linking to external datasets

The open source software Silk[14] 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 code lists coming from the resources are interlinked with another similar using the appropriate linking property, e.g. skos:exactMatch for concepts.

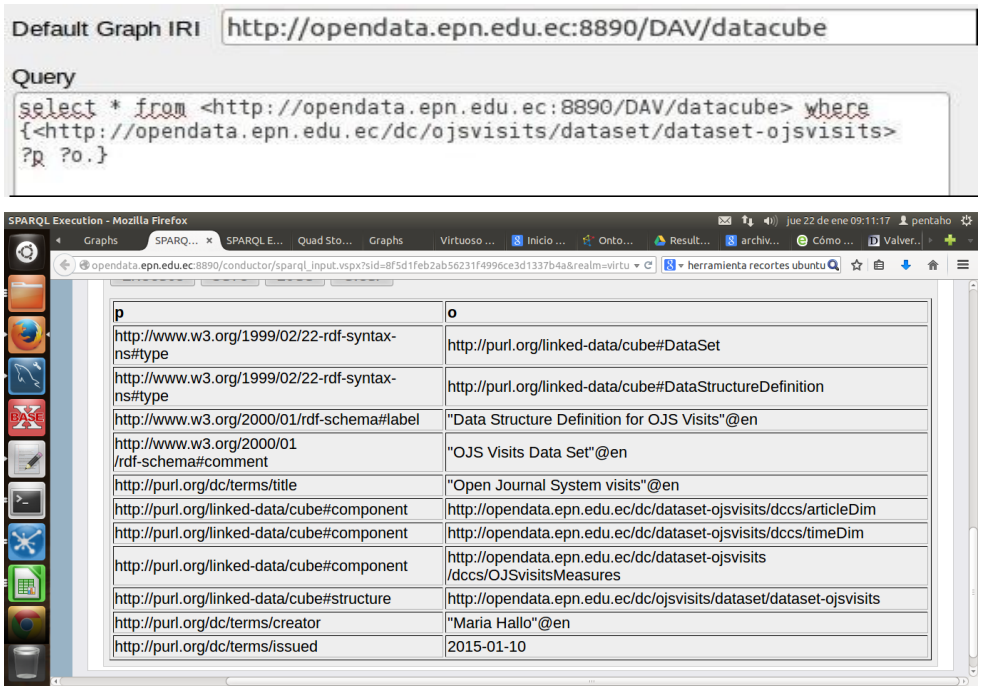
2.6 Publication

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[15], 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; Cubeviz[16] is used as a Linked Data interface to the RDF data cube [11]. Datasets may be further "announced" to the public, to be more discoverable, by publishing the data to international or national open data portals.

Fig. 5 shows a view of the SPARQL endpoint with a partial result of the query about the structure of the OJSvisits data cube.



p	o
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.org/linked-data/cube#DataSet
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.org/linked-data/cube#DataStructureDefinition
http://www.w3.org/2000/01/rdf-schema#label	"Data Structure Definition for OJS Visits"@en
http://www.w3.org/2000/01/rdf-schema#comment	"OJS Visits Data Set"@en
http://purl.org/dc/terms/title	"Open Journal System visits"@en
http://purl.org/linked-data/cube#component	http://opendata.epn.edu.ec/dc/dataset-ojsvisits/dccs/articleDim
http://purl.org/linked-data/cube#component	http://opendata.epn.edu.ec/dc/dataset-ojsvisits/dccs/timeDim
http://purl.org/linked-data/cube#component	http://opendata.epn.edu.ec/dc/dataset-ojsvisits/dccs/OJSvisitsMeasures
http://purl.org/linked-data/cube#structure	http://opendata.epn.edu.ec/dc/ojsvisits/dataset/dataset-ojsvisits
http://purl.org/dc/terms/creator	"Maria Hallo"@en
http://purl.org/dc/terms/issued	2015-01-10

Fig. 5 . SPARQL endpoint query results about the structure of the OJSvisits data cube.

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

[14] Silk – A link discovery framework for the web of data: http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

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

[16] CubeViz : http://cubeviz.aksw.org/

b) Metadata definition and publication

The metadata about the dataset produced was published in the site Datahub using DCAT (Data Catalog Vocabulary)[17], an RDF vocabulary designed to facilitate interoperability between data catalogues published on the web [12]. In addition provenance data were added.

The whole architecture used in this project is shown in the fig. 6.

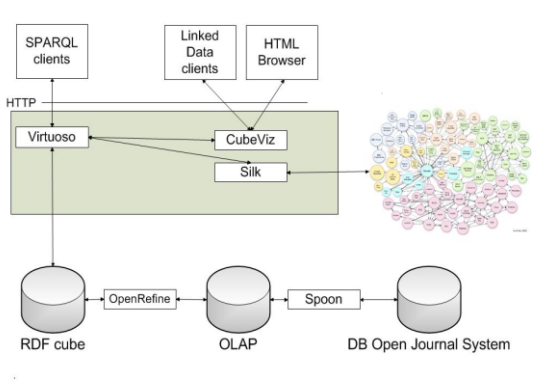


Fig 6. Architecture of statistical RDF publication

3 CONCLUSIONS AND FUTURE WORK

In this paper we analyse and use a process for publishing statistical 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 a data mart from "Revista Politécnica" to RDF. For publishing we use OpenLink Virtuoso and CubeViz. The RDF data cube vocabulary and the Open Refine software were successfully applied for the RDF generation process.

The process could be also applied to data marts build from bibliographic metadata harvested through the OAI-PMH Protocol.

In the future to get statistical data from OJS we can restrict site- and article-level access through the user registration interface. The advantage of selecting these options is that anyone wanted to read the content will need to register, providing reliable readership statistics. In addition, the Logging and Auditing option enables logging of submission actions and user emails sent by the system. This is a very useful feature to make available to the readers. Furthermore we will explore to look for similar properties and class from open linked dataset catalogs to link projects results.

ACKNOWLEDGMENTS

This work has been partially supported by the Prometeo Project by SENESCYT, Ecuadorian Government.

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

REFERENCES

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

[2] Brian, D. Willinsky, E. (2010), A Survey of Scholarly Journals Using Open Journal Systems, *Scholarly and Research Communication*, 1(2),1-22.

[3] Salas, P. E. R., Mota, F. M. D., Martin, M., Auer, S., Breitman, K., & Casanova, M. A. (2012). Publishing Statistical Data on the web. *In Proc. IEEE ICSC 2012* (Sept. 19th-21st, 2012). Palermo, Italy, 285-292.

[4] Ermilov, I., Martin, M., Lehmann, J., & Auer, S. (2013). Linked open data statistics: Collection and exploitation. In *Knowledge Engineering and the Semantic Web,* 242-249.

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

[6] Berners-Lee, T. (2006). Linked Data - Design Issues. Available at: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed Jan 15, 2015].

[7] Guha, RV., Brickley, D. (2004).: RDF vocabulary description language 1.0: RDF Schema.W3C Recommendation, W3C.. Available at: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.[Accessed Jun 15, 2015].

[8] 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 Jan 10, 2015].

[9] Cyganiak, R., Reynolds D., & Tennison, J. (2014): The RDF Data Cube Vocabulary, *World Wide Web Consortium*. Available at: http://www.w3.org/TR/vocab-data-cube/. [Accessed Jan 2, 2015].

[10] 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, 823-827.

[11] Mader, C., Martin, M., & Stadler, C. (2014). Facilitating the Exploration and Visualization of Linked Data. In *Linked Open Data--Creating Knowledge Out of Interlinked Data*, 90-107.

[12] Cyganiak, R., Maali, F., & Peristeras, V. (2010). Self-service linked government data with dcat and gridworks. In *Proceedings of the 6th International Conference on Semantic Systems*, 37-39.