Publishing a Scorecard for Evaluating the Use of Open-Access Journals Using Linked Data Technologies

María Hallo
Department of Computer Science
National Polytechnic School
Quito, Ecuador
maria.hallo@epn.edu.ec

Sergio Luján-Mora
Department of Software and Computing Systems
Visiting teacher at the National Polytechnic School
University of Alicante
Alicante, Spain
sergio.lujan@ua.es

Alejandro Maté
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
amate@dlsi.ua.es

***Abstract*—Open access journals collect, preserve and publish scientific information in digital form, but it is still difficult not only for users but also for digital libraries to evaluate the usage and impact of this kind of publications. This problem can be tackled by introducing Key Performance Indicators (KPIs), allowing us to objectively measure the performance of the journals related to the objectives pursued. In addition, Linked Data technologies constitute an opportunity to enrich the information provided by KPIs, connecting them to relevant datasets across the web.**

This paper describes a process to develop and publish a scorecard on the semantic web based on the ISO 2789:2013 standard using Linked Data technologies in such a way that it can be linked to related datasets. Furthermore, methodological guidelines are presented with activities. The proposed process was applied to the open journal system of a university, including the definition of the KPIs linked to the institutional strategies, the extraction, cleaning and loading of data from the data sources into a data mart, the transforming of data into RDF (Resource Description Framework), and the publication of data by means of a SPARQL endpoint using the OpenLink Virtuoso application. Additionally, the RDF data cube vocabulary has been used to publish the multidimensional data on the web. The visualization was made using CubeViz a faceted browser to present the KPIs in interactive charts.

Keywords*—*Linked Data, semantic web, RDF data cube vocabulary, knowledge management.

I. INTRODUCTION

Open access journals collect, preserve and publish scientific information related to a particular subject in digital form [1]. Open access (OA) is the free unrestricted online access to digital content. A growing number of scholarly journals are using Open Journal Systems (OJS), a software platform, designed to manage articles through author submission, the peer review process, editing and publication [2]. While such system fosters the publication process, little attention has been paid to analyse the impact of digital libraries (DL).

Libraries routinely collect statistics about the use of their digital collection for evaluation purposes. However, these statistics are dispersed, stored across data stores lacking a standard structure, and unrelated to the business objectives. As a result, it is difficult for researchers and users to compare statistical information, while for DL it becomes a challenge to develop policies, assess the impact of OJS in society, and share their discoveries.

In order to tackle this problem, this paper proposes a scorecard for evaluating and comparing digital libraries based on statistics suggested in the ISO 2789:2013 standard [3], as well as a technical architecture for publishing them based on Linked Data technologies. The proposed approach was developed based on best practices and recommendations from several authors [4, 5] and tested with data extracted from the electronic version of the journal "Revista Politécnica"[1], edited by National Polytechnic School of Quito (Ecuador). In addition, the dataset created was linked to external data providing information that goes far beyond the bibliographic data supplied by publishers, such as: number of papers in similar subjects, number of visits, statistical indicators below national standards, etc. The results of these evaluation strategies can have a number of significant implications for the continued development of digital libraries.

The remainder of this paper is structured as follows. Section II presents the background on Linked Data technologies. Section III describes the metrics used for evaluating DL. Section IV presents our proposal for defining and publishing a scorecard for the evaluation of DL. Finally, Section V describes the conclusions and sketches future works.

[1] Revista Politécnica: http://www.revistapolitecnica.epn.edu.ec/

The last version was published at: *CITS2015 Proceedings, 2015,* 105-109
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7297730&isnumber=7297712

II. BACKGROUND

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]. It is based on the URI (Uniforme Resource Identification) and RDF (Resource Description Framework) specifications.

URI is used to identify a web resource, whereas RDF is used for modeling and representing 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 only these technologies, it is not possible to add semantics to data.

The Semantic Web Architecture 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 and relationships [7]. OWL is an extension of RDFS [8], which provides additional metadata terms for the description of "ontologies".

For our work, 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 to these standards, it is necessary to describe the KPIs in a multidimensional model in order to enable its analysis, with this purpose we use the RDF data cube vocabulary to publish, discover, and link statistical data organized in a multidimensional model.

Using these technologies we are able to publish scorecards as multidimensional data using RDF and Linked Data technologies, obtaining a number of advantages as described by the W3C recommendation [9]:

- The individual observations, and groups of observations, become (web) addressable. This allows publishers and third parties to annotate and link to this data.
- Statistical data can be combined across datasets.
- Publishing scorecards as Linked Data offers a flexible, nonproprietary, machine readable means of publication.
- It enables reuse of standardized tools and components.

III. EVALUATION OF THE USE OF DIGITAL LIBRARIES

The evaluation approaches, methods, and criteria vary among the existing DL evaluation studies [10, 11, 12, 13]. The majority of the studies adopt Information Retrieval (IR) evaluation approaches at a restricted level (either at the system or the user level) while employing traditional criteria, such as precision, search time, error rate, etc. Very few address the benefits of a DL on the user. Furthermore, there are few metrics devised specifically for this goal interlinked with external information.

A. Scorecards

A scorecard is a tool to monitor strategic objectives in a business. The Balanced Scorecard is one of the best corporate scorecards, it is used to help organizations to align them with their strategic objectives [14].

B. Scorecards and libraries

Performance metrics and indicators should be related to institutional and library mission and objectives [15]. But, analyzing a random sample of OJS from DOAJ2 (Directory of Open Access Journals), few of them publish their vision, mission, strategic objectives, or statistics.

A primary purpose of using library performance indicators is self-diagnosis, including comparisons within the same library in several years [16]. We focus our study mainly on this requirement using Linked Data technologies to allow future analysis based on interlinked indicators.

The ISO 2789:2013 standard defines statistics for "evaluation and comparison of libraries as well as for promoting, marketing and advocating the value that libraries provide for their population and for society". The objectives of the library statistics defined in the ISO 2789:2013 standard are summarized as follows:

- to monitor operating results against standards and data of similar organizations;
- to monitor trends over time;
- to provide a base for planning, decision making, improving service quality, and feedback of the results;
- to inform national and regional organizations in their support, funding and monitoring roles;
- to demonstrate the value of library services obtained by users, including the potential value to users in future generations.

For our work, we have developed a scorecard to: monitor use trends over time, make self-diagnosis, and use the results in marketing. The proposed model can be used as a strategic scorecard which can also be navigated. We have used a subset of indicators of the ISO 2789:2013 and ISO 11620:2014 standard [17], for the use of electronic documents, based on interviews with librarians, local authorities and the data that was possible to retrieval from the OJS records.

The indicators are: (i) number of visits, (ii) number of rejected accesses, (iii) number of downloads, (iv) number of internet accesses, (v) % external users, (vi) % of items not

[2] DOAJ: http://www.doaj.org

used, (vii) user satisfaction, (viii) number of downloads by document, (ix) number of digital documents stored, (x) number of digital documents added. Along with these indicators extracted from the standard, we have included several dimensions of analysis that help in aggregating or disaggregating the information at hand: (i) visit time, (ii) article, (iii) author, (iv) geographic location, (v) keywords, (vi) objective.

IV. LINKED DATA PUBLICATION PROCESS FOR A SCORECARD

In order to publish and feed a scorecard from an OJS data mart transformed into RDF format we propose five main activities:

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

A. Data source analysis

In this initial activity, we analyzed the information provided by the OJS data source that could be useful for the proposed scorecard. This data source has the information about publications, which we needed to link with another datasets to give us better knowledge about the use of publications. First, we represented the OJS data source in the form of a multidimensional model, comprised of three basic components: dimensions, measures, and attributes. This allowed us to approach the data source as a data mart, a subset of the target data warehouse for DL evaluation. Data marts are usually oriented to specific business topics (the topic in this case would be publications), and they allow us to build specialized scorecards for each area.

The data mart obtained as a result of this activity for testing our proposal is shown the Fig.1, and is implemented in MySQL.

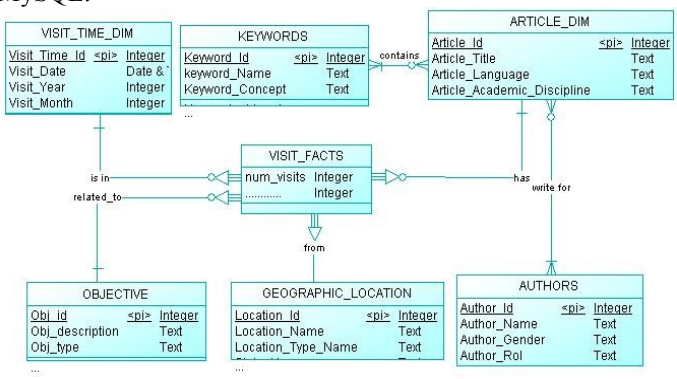

Fig. 1. OJS use datamart

This data linked to other datasets will give us better knowledge about: similar subjects, the authors who work in them, the objectives accomplished related to national goals. However, in order to be able to link this data, we need to transform it into RDF.

B. RDF data modeling

The goal of this activity is to design and implement the vocabularies for describing the datasets in RDF. 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 [18]. To this aim, we use the following controlled vocabularies and ontologies for modelling statistical datasets in RDF:

- RDF data cube vocabulary[3] is a standard to publish multi-dimensional data, such as statistics, on the web.
- BIBO[4] (The Bibliographic Ontology) provides concepts and properties for describing citations and bibliographic references on the semantic web using RDF.
- Dublin Core[5] is a set of terms that is used to describe web resources as well as physical resources. Dublin Core Metadata may be used to provide interoperability in semantic web implementations.
- FOAF[6] (Friend of a Friend) is an ontology describing persons, their activities and relations to other people and objects in RDF format.
- ORG[7] (Organization) is an ontology for describing organizations, roles and organizational activities.
- SKOS[8] (Simple Knowledge Organization System) an standard for sharing and linking concepts and concept schemes.

The reduced RDF data cube model obtained as a result of this step is presented in Fig. 2. In this RDF model, each concept is mapped with the corresponding concept of the multi-dimensional model, such as dimension, measure, code list, etc.

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

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

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

The URI structure was defined by:

- Schema components (dimensions, measures, and attributes), which are identified by: {Base_URI}/dc/cube_name/prop/{dimension_name| measure_name |attribute}.
- Datasets are identified by: {Base_URI}/}/dc/cube_name/dataset/{DatasetName}
- The dataset component is specified by: {Base_URI}/dc/ cube_name/dccs /{dimension_name| measure_name}
- Concepts and their values reused across multiple datasets are identified by: { Base_URI}/concept/ {ConceptName} and { Base_URI }/concept/{ConceptName}/{value}.

C. *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. For the case study we have used Open Refine[9] tool to perform the transformation from the multidimensional model stored in a relational database to RDF data cube vocabulary.

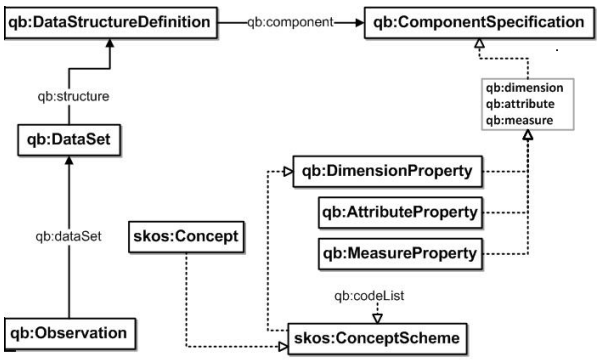

Fig. 2. A reduced RDF data cube vocabulary

Mappings were defined from the multidimensional database to 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 other concepts and code lists (controlled vocabularies) 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 [19]. The code lists that are used to give a value to each of the components are also defined using SKOS vocabulary. The data are then exported as RDF in a RDF compliant serialization, such as RDF/XML.

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

D. *Interlinking*

The objective of this activity is to improve the connectivity to external datasets enabling other applications to discover additional data sources. For this task we perform two steps: (i) discovery, and (ii) linking.

Discovery comprises finding new target datasets. For this step we used the website "the Datahub"[10]. We found several open linked statistics datasets from scientific journals.

Linking allows us to relate external sources for additional information. For this step we used the open source software Silk[11] to find relations between data items in our datasets and the external datasets generating the corresponding RDF links that were stored in a separated dataset.

E. *Publishing*

The goal of this activity is to make RDF datasets available on the web to the users, following the Linked Data principles. For this activity, we need a RDF server, usually in the form of a SPARQL endpoint. In our case the generated triples were loaded into a SPARQL endpoint (a conformant SPARQL protocol service) based on OpenLink Virtuoso[12], 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 top of OpenLink Virtuoso, Cubeviz[13] is used as a Linked Data interface to the RDF data cube [20]. Datasets may be further "announced" to the public, to be more discoverable, by publishing the data to international or national open data portals. Fig. 3 shows a view of the SPARQL endpoint with a partial result of the query on the OJS visits data cube, giving the number of visits by subject and by article.

Fig. 3. Query example on the OJS visits data cube.

The architecture used in this case is shown in the Fig. 4.

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

[11] Silk : http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

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

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

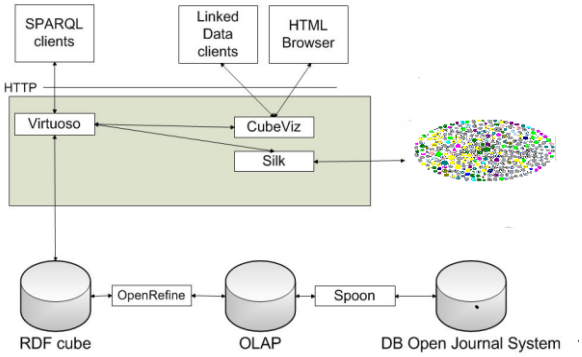

Fig. 4. Architecture of scorecard RDF publishing

V. CONCLUSIONS AND FUTURE WORK

In this paper we described a process for publishing a scorecard about the use of scientific data from Open Journal systems on the web using the principles of Linked Data. The process is based on best practices and recommendations from several studies, adding tasks and activities considered important during the project. The process was applied to the development and the transformation of a scorecard from "Revista Politécnica" into RDF using the RDF data cube vocabulary. For publishing we used OpenLink Virtuoso, Ontowiki and CubeViz applications. The Open Refine software was applied for the RDF generation process. As a result, the developed process fulfilled the requirements of the study.

In the future, we will develop a user registration interface, to be accessed before downloading the articles, in order to get more data for analyzing and comparing search history data. Moreover, we will design metrics to evaluate the performance of the proposed process for the development of new scorecards oriented to other strategic objectives. Finally, we will look for the possibility of finding related open linked dataset catalogues to link projects results.

ACKNOWLEDGMENTS

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

REFERENCES

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

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

[3] ISO, "ISO 2789:2013: Information and documentation—International library statistics", pp. 71, 2013.

[4] F. Maali, R. Cyganiak, and V. Peristeras, "A publishing pipeline for linked government data". In The semantic web: Research and applications, Springer Berlin Heidelberg, pp. 778-792, 2012.

[5] I. Ermilov, M. Martin, J. Lehmann, and S. Auer, "Linked open data statistics: Collection and exploitation". In Knowledge Engineering and the Semantic Web, pp. 242-249, 2013.

[6] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the web" . In Proceedings of the 17th international conference on World Wide Web, ACM. pp. 1265-1266, 2008.

[7] RV. Guha, D. Brickley, "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 Feb 15, 2015].

[8] P. Hayes, P. Patel-Schneider, and I. Horrocks, "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 Feb 10, 2015].

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

[10] Reeves, T., Apedoe, Woo, Y. Evaluating digital libraries: A user friendly guide, University Corporation for Atmospheric Research, 2005. Available at:http://www.dpc.ucar.edu/projects/evalbook/EvaluatingDigitalLibraries. pdf. [Accessed Jan 15, 2015].

[11] Z.Ying Zhang, "Developing a holistic model for digital library evaluation", J. Am. Soc. Inf. Sci. Technol. vol.61(1) , pp. 88-110, 2010.

[12] C. Klas, et al, "A Logging Scheme for Comparative Digital Library Evaluation", Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, pp. 267-278, 2006.

[13] L. Pinto, P.Ochôa, and M. Vinagre, "Integrated approach to the evaluation of digital libraries: an emerging strategy for managing resources, capabilities and results". Library statistics for the 21st century world, pp. 273-288, 2009.

[14] R. Banker, H. Chang, and M. Pizzini,(2004), "The balanced scorecard: Judgmental effects of performance measures linked to strategy". The Accounting Review, vol 79(1), pp.1-23, 2004.

[15] A. Maté, J. Trujillo, and J.Mylopoulos, "Conceptualizing and Specifying Key Performance Indicators in Business Strategy Models", In Proceedings of the 2012 Conference of the Center for Advanced Studies on Collaborative Research pp. 102-115, 2012.

[16] L. Melo, and C. Pires, "Performance evaluation of academic libraries: implementation model". In 17th Hellenic Conference of Academic Libraries, Ioanina, vol 2, pp. 2012, 2008.

[17] ISO, "ISO 11620:2014: Information and documentation-Library performance indicators", pp. 1-99, 2014.

[18] D. Sánchez, M. Batet, D.Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach". Expert Systems with Applications, vol 39(9), pp. 7718-7728, 2012.

[19] M. Hallo, S. Luján-Mora, J. Trujillo, "Transforming Library Catalogs into Linked Data", Proceedings of the 7th International Conference of Education, Research and Innovation, Seville, Spain, pp. 1845-1853, 2014.

[20] C. Mader, M. Martin, and C. Stadler, "Facilitating the Exploration and Visualization of Linked Data". In Linked Open Data-Creating Knowledge Out of Interlinked Data, pp. 90-107, 2014.