Towards the Integration of a Research Group Website into the Web of Data

Mikel Emaldi, David Buján, and Diego López-de-Ipiña

Deusto Institute of Technology - DeustoTech, University of Deusto
Avda. Universidades 24, 48007, Bilbao, Spain
{m.emaldi, dbujan, dipina}@deusto.es

Abstract. This work describes our efforts towards making the website of our research group, namely MORElab, comply with the principles of Linked Data. As a first attempt, we have made the information related to our research publications available in RDF. Such data is published according to the Bibliographic and Dublin Core Ontologies and interlinked with instance data of the FOAF and DBLP vocabularies. Our main contribution has been to adapt the Joomla! CMS so that it can be used for the automatic generation of the semantic metadata about publications.

1 Introduction

The Semantic Web can be seen as a large knowledge-base formed by sources that serve information as RDF (Resource Description Framework) [1] files or through SPARQL endpoints [2]. The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web [3]. Today, more and more initiatives publish their data as Linked Data [4]. Although there are tools for publishing the data stored on relational databases into RDF files and for providing SPARQL endpoints like D2R Server [5] and tools for creating Linked Data interfaces from SPARQL endpoints like Pubby[1], publishing data of a website as Linked Data is usually a non-trivial task. Some Content Management Systems (CMS), e.g. Drupal[2], provide tools for publishing the data of a website as RDF. Other CMS such as RDF Tools for Wordpress[3], allow publishing their content as RDF files. As far as we know, there is a lack of tools for publishing Joomla![4] content as Linked Data despite of being the most used CMS after Wordpress[5]. On this paper, we describe our initial efforts towards migrating the information embedded in our research group's site into the Web of Data. For that we will: 1) describe the ontological vocabularies chosen to export part of our site (concretely the publications part), 2) explain the adaptation we have

[1] http://www4.wiwiss.fu-berlin.de/pubby/
[2] http://drupal.org/
[3] http://bnode.org/blog/2008/01/15/rdf-tools-an-rdf-store-for-wordpress
[4] http://www.joomla.org/
[5] Usage statistics at May 2011. http://trends.builtwith.com/cms

carried out over Joomla! so that it outputs correct Linked Data and 3) the tools used to implement our solution.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 presents our extension for RDF generation. Section 4 referes to tools used to publish semantic data as Linked Data. Finally, Section 5 concludes and outlines the future work.

2 Related Work

There are two significant tools for publishing Linked Data from a traditional HTML and database-based website: Pubby and D2R Server. Pubby adds Linked Data interfaces to SPARQL endpoints that allow DESCRIBE queries, such as Virtuoso Universal Server[6], Joseki SPARQL Server[7] or Drupal SPARQL Endpoint module[8]. It manages dereferenceable URIs and generates HTML and RDF views of Linked Data. Pubby is implemented as a Java web application and can be deployed in Tomcat and Jetty servlet containers.

The D2R Server [5] provides a powerful tool for publishing the content of relational databases as Semantic Data. Mapping of data between relational database and a RDFS schema is achieved with D2RQ language [6]. The D2R Server provides both XHTML and RDF representations through dereferenceable URIs. D2R Server also enables querying relational databases using SPARQL protocol against a virtual RDF graph representing an entire database. The results of these queries can be retrieved both in XML and JSON formats, among others.

Drupal has its own module for publishing its entities as RDFa[9]. It also provides extra APIs to retrieve results in additional formats such as RDF/XML, NTriples and Turtle. It has an administration interface to configure proper RDF mappings. It also allows to import vocabularies to map entities. This data can be accessed through SPARQL Endpoint module. As far as we know, there is not a tool for Joomla! which enables publishing, even partially, the contents of a website neither as plain RDF nor as RDF following the Linked Data principles.

3 RDF Generation Plugin

According to the Joomla! documentation, an article is "some written information that you want to display on your site. It normally contains some text and can contain pictures and other types of content. For many Joomla! sites, articles form the majority of the information presented in the website"[10]. Joomla! manages published articles through the *com_content* component. By default, this component only allows the inclusion of title, alias, category, text and a few more

[6] http://virtuoso.openlinksw.com/

[7] http://www.joseki.org/

[8] http://drupal.org/project/sparql_ep

[9] http://drupal.org/project/rdfx

[10] http://docs.joomla.org/Understanding_sections,_categories_and_articles#Articles

Tool	Features	Limitations
Pubby.	Adds Linked Data interface to SPARQL endpoints. Provides RDF representation and HTML interface	It needs a SPARQL endpoint provided by another tool.
D2R Server.	Relational data to RDF data. SPARQL endpoint. Provides XHTML representation.	Mappings are done by user through D2RQ language.
Drupal + RDFx & SPARQL endpoint modules.	Provides RDF, XHTML representations and SPARQL endpoint.	It does not provide Linked Data interface.
Joomla! Linked Data plugin.	Adds Linked Data interface to SPARQL endpoints. Provides RDF representation and Joomla! default view as XHTML interface. Published data is fully customizable.	It needs a SPARQL endpoint. The Linked Data datasets are hardcoded.

Table 1. Comparative between different semantic publication tools.

things. This represents a problem for publishing data as Linked Data through ontologies as Dublin Core [7] or Bibliographic Ontology [8], because there are not enough fields to fill the information required to make useful this representation. To solve this issue, we may develop a component which adds proper fields to the content administration panel, but rewriting all content could be a difficult task if there are a lot of articles or news already written. Alternately, in this paper an automatic parsing mechanism is used based on custom regular expression patterns and HTTP-mediated queries to third party data sources.

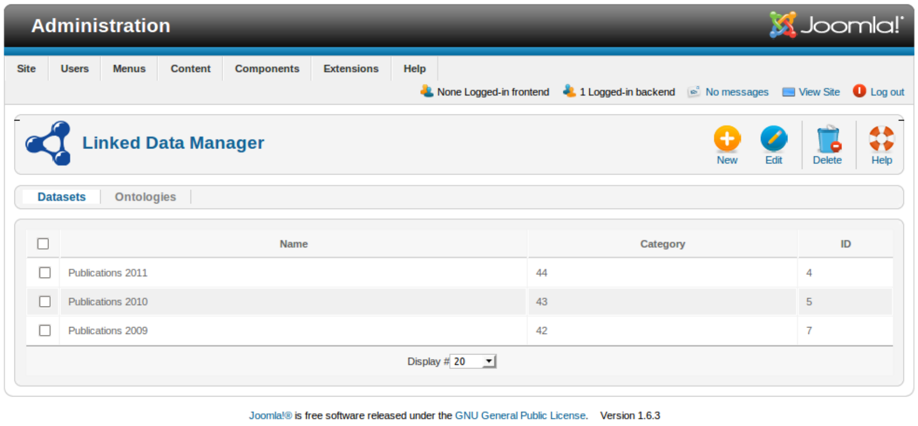
We assume that the majority of organizations have a person responsible of the maintenance of a web site and its contents. According to that premise, we can suppose that almost all contents on those sites are written following the same format, determinate pattern can be applied to extract different fields of the content. Text at Code 1 describes the data of a publication of Morelab:

Code 1. Text of description of a publication.

```
Aitor Almeida, Pablo Orduna, Eduardo Castillejo, Diego Lopez-de-Ipina,
    Marcos Sacristan. Imhotep: an approach to user and device conscious
    mobile applications. Personal and Ubiquitous Computing (Journal).
    Vol. 15. No. 4. PP. 419-429. Springer. Impact Factor (2009): 1.554.
    ISSN: 1617-4909. DOI: 10.1007/s00779-010-0359-8. January 2011.

As the dependence on mobile devices increases, the need for supporting a
     wider range of users and devices becomes crucial. [...] and the
    usability of the proposed platform.
```

As can be seen at Code 1, the administrator wrote the authors of paper and conference or journal where the paper was published. In this concrete case, the



□	Name	Category	ID
□	Publications 2011	44	4
□	Publications 2010	43	5
□	Publications 2009	42	7

Display # 20

Joomla!® is free software released under the GNU General Public License. Version 1.6.3

Fig. 1. RDF Generation Plugin main administration panel.

title of paper can be extracted from title field of Joomla! database and the rest of data can be extracted by the following regular expression pattern (Code 2):

Code 2. RegEx pattern that wraps the data field embedded into HTML code.

```
{dc:creator, sep(,)}\. {dc:title}\. {dc:title, urn:issn}\( {rdf:type,
    urn:issn} \) . Vol. {bibo:volume}\. No\. {bibo:issue}\. PP. \{bibo:
    startPage}-{bibo:endPage}\. {bibo:publisher, urn:issn}\. {dummy}\.
    ISSN: {urn:issn}\. DOI: {bibo:doi}\. {dc:date}\.
{bibo:abstract}
```

If every content of the publications section is written in the same way by the same person the data of all articles can be wrapped without any additional effort. Otherwise, if an article that follows a different pattern is found, this pattern becomes useless.

In order to overcome, make more reliable the process the RegEx-based process, Google Scholar web scraping is used. Taking as input the title of a paper, an automatic search is done through Google Scholar web page including a cookie that requests the BibTeX[11] representation that Google Scholar offers. If Google Scholar has indexed the article, the whole publication metadata can be extracted from the corresponding BibTeX. Another source used by the extension is the DBLP SPARQL endpoint[12] and the FOAF files of our research group's (MORElab) memebers. These FOAF files are manually generated and made available by the labmembers from their personal webpages. Finally, the generated RDF can be seen at Code 3.

Code 3. Example of generated RDF.

```
@prefix bibo: <http://purl.org/ontology/bibo/> .
```

[11] http://www.bibtex.org/

[12] http://dblp.l3s.de/d2r/snorql/

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://www.morelab.deusto.es/index.php/publications-1879995610/2011/A-
    Semantic-Resource-Oriented-Middleware-for-Pervasive-Environments> a
    bibo:article ;
    owl:sameAs <http://dblp.l3s.de/d2r/page/publications/journals/puc/
        AlmeidaOCLS11> ;
    dc:title "Imhotep: an approach to user and device conscious mobile
        applications" ;
    dc:date "January 2011" ;
    dc:isPartOf <urn:issn:1617-4909> ;
    bibo:doi "10.1007/s00779-010-0359-8" ;
    bibo:volume "15" ;
    bibo:issue "4" ;
    bibo:startPage "419" ;
    bibo:endPage "429" ;
    dc:creator <http://www.morelab.deusto.es/index.php/members/35-aitor-
        almeida/foaf.rdf> ;
    dc:creator <http://www.morelab.deusto.es/index.php/members/69-pablo-
        orduna/foaf.rdf> ;
    dc:creator <http://www.morelab.deusto.es/index.php/members/89-
        eduardo-castillejo/foaf.rdf> ;
    dc:creator <http://www.morelab.deusto.es/index.php/members/37-dr-
        diego-lopez-de-ipina/foaf.rdf> ;
    dc:creator "Marcos Sacristan" ;
    bibo:abstract "As the dependence on mobile devices increases, the
        need for supporting a wider range of users and devices becomes
        crucial. [...] and the usability of the proposed platform." .

<urn:issn:1617-4909> a bibo:Journal ;
    dc:title "Personal and Ubiquitous Computing" ;
    bibo:publisher "Springer" .
```

4 Publishing data as Linked Data

Once the RDF files are generated, the next step is offering this data as Linked Data. Linked Data is simply about using the Web to create typed links between data from different sources [12].

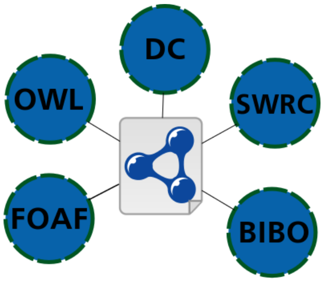


Fig. 2. Used vocabularies to create Morelab's RDF files: OWL [9], DC [7], SWRC [10], FOAF [11] and BIBO [8].

To do that, the extension follows the 6^{th} recipe described on [13]: using 303 redirect, multiple HTML and SPARQL endpoint, combinining Joseki SPARQL endpoint and Pubby Linked Data interface. We choose these tools because of their simplicity compared to other tools like Virtuoso. To link our own data with another datasets, the extension queries the existing SPARQL endpoints of known datasets as DBLP. It also queries MORElab's SPARQL endpoint to retrieve members FOAF data. This SPARQL endpoint can be queried by third party applications that want to consume our Linked Open Data. Another way to consume MORElab's Linked Open Data is doing a HTTP request including `application/rdf+xml` field into `Accept` header, to URL of any resource of Morelab's website[13]. This request will return an RDF graph with the metadata of this article. Used vocabularies (BIBO, DC and OWL) are hardcoded into the source code of the extension.

5 Conclusions and Future Work

This paper presents our efforts towards making the information stored in our research group's site available as Linked Open Data. For that, we have contributed with an extension for Joomla! CMS that allows scientific article related data available as Linked Data. We think that the usage of this extension provides an accessible way to increase the Linked Data provided by organizations that uses Joomla! as their CMS.

For our future work, we will enhance the administration panel allowing the inclusion of custom RegEx pattern for selected articles and custom ontologies, not only Bibliographic Ontology, Dublin Core and OWL[9]. Moreover, we must increase the flexibility of RDF generation patterns letting administrators adding their own data sources and linking the datasets with well-known sources such Geonames[14] or DBPedia [14]. Although the current extension limitation is focused on publications, but there are many more content types that can be published such Linked Data, such as projects or member information.

References

1. "RDF - semantic web standards." http://www.w3.org/RDF/.
2. G. Tummarello, R. Delbru, and E. Oren, "Sindice. com: Weaving the open linked data," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, p. 552–565, 2007.
3. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Int. J. Semantic Web Inf. Syst.* **5**(3), p. 1–22, 2009.
4. F. Peset, A. Ferrer-Sapena, and I. Subirats-Coll, "Open data y linked open data: Su impacto en el área de bibliotecas y documentación," *El profesional de la información* **20**(2), pp. 165–174, 2011.
5. C. Bizer and R. Cyganiak, "D2r server-publishing relational databases on the semantic web," in *5th International Semantic Web Conference*, p. 26, 2006.

[13] Example in http://www.morelab.deusto.es/pubby/resource/a-semantic-resource-oriented-middleware-for-pervasive-environments

[14] http://www.geonames.org/

6. C. Bizer and A. Seaborne, "D2RQ-treating non-RDF databases as virtual RDF graphs," in *Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*, p. 26, 2004.
7. "Dublin core metadata element set, version 1.1: Reference description." http://dublincore.org/documents/dces/, Dec. 2004. The reference description, version 1.1 of the Dublin Core Metadata Element Set.
8. "Bibliographic ontology specification | the bibliographic ontology." http://bibliontology.com/specification.
9. D. McGuinness, F. Van Harmelen, *et al.*, "Owl web ontology language overview," *W3C recommendation* **10**, pp. 2004–03, 2004.
10. Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann, and D. Oberle, "The swrc ontology–semantic web for research communities," *Progress in Artificial Intelligence* , pp. 218–231, 2005.
11. D. Brickley and L. Miller, "Foaf vocabulary specification," 2005.
12. L. Sauermann, R. Cyganiak, and M. Völkel, "Cool URIs for the semantic web," *Working draft, W3C* , 2008.
13. "Best practice recipes for publishing RDF vocabularies." http://www.w3.org/TR/swbp-vocab-pub/.
14. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," *The Semantic Web* , pp. 722–735, 2007.