Linked Data Strategy to Achieve Interoperability in Higher Education

Guillermo García Juanes, Alioth Rodríguez Barrios, José Luis Roda García, Laura Gutiérrez Medina, Rita Díaz Adán and Pedro González Yanes
School of Computer Science, University of La Laguna, San Cristóbal de La Laguna, Spain

Keywords: Linked Data, Open Data, Interoperability, High Education.

Abstract: An important challenge in centres of higher education is the use of Linked Data strategy to connect currently existing multiple information systems. These information systems are usually independent from one another, and the ability to obtain information by connecting different sources of data involves, in most cases, unacceptable costs and effort. In this work, we have developed a platform based on Linked Data that permits the interoperability of different sources of data, both internal as well as external. This interoperability is achieved by 1) the use of higher education ontologies, and 2) the use of a process that begins with the analysis of the data sources to be connected, followed by mapping of the closest ontologies, and ends with the generation and publication of data in valid formats for Linked Data. The final product permits stakeholders inside and outside the university to be able to make queries of two or more datasets in different information systems at the same time.

1 INTRODUCTION

The term Open Data concerns offering to society the data collected by public institutions, which, when handled by third-parties, can be of great value for the development of applications, reports, etc. The principal objective of the Open Data strategy is to offer transparency, participation and collaboration in the publishing of information, in standard formats, open and interoperable, facilitating its access and permitting its re-use (Office, 2012). This is nothing more than data that belong to public administration being used, by individuals or companies, with or without commercial ends, provided that use does not constitute public administration activity. There are many institutions that currently publish open data in different formats (Fundación CTIC, 2013) (Bauer and Kaltenböck, 2012).

The objective of Linked Data is to give meaning to connections that are found in different datasets so that machines can obtain more relevant information making use of techniques from the Semantic Web (Berners-Lee, 2006).

The relationship between Open Data and Linked Data was proposed by Tim Berners-Lee who suggested a way of measuring the degree of quality of data published from an Open Data portal, where, if those data were to be published using Linked Data principles, the highest degree that a portal may achieve would be obtained (Berners-Lee, 2006).

The data interoperatibility is achieved through the followig of the Linked Data principles. These vocabularies must cover a wide range of concepts related to the university´s system. From the institution, the departments, the teachers, including the courses, programmes and study material, credits, theory classes, practical classes and laboratory classes, etc., a vocabulary must cover all these aspects to be able to be linked and make full use of the Linked Data strategy.

In this work we present a prototype that was developed at the University of La Laguna (ULL), where various different groups played a part: the Planning and Analysis Office, the ULL Information Technology Service, and the Taro Research Group. The project concerns the demonstration of how Linked Open Data offers a range of benefits for the procurement of information that are not achieved through other conventional methods. The higher education system comprises multiple information systems, some of which are quite complex. This

García Juanes G., Rodríguez Barrios A., Roda García J., Gutiérrez Medina L., Díaz Adán R. and González Yanes P..
Linked Data Strategy to Achieve Interoperability in Higher Education.
DOI: 10.5220/0004853500570064
In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (WEBIST-2014), pages 57-64
ISBN: 978-989-758-024-6

paper is, therefore, concerned with the application of Linked Data strategy, methods and techniques to some of these university systems.

2 RELATED WORK

The development of a platform such as the one intended requires a prior state-of-the-art study, with particular emphasis in the area of ontologies that may be re-used with those that are going to model concepts within the scope of higher education.

In the search for ontologies closer to our problem, we made use of semantic search engines such as Watson[1] or Swoogle[2] apart from search engines for key words or known repositories like Linked Open Vocabularies (LOV)[3].

There is a great set of ontologies related to the field of education, but we limited our search for ontologies to those that could be adapted as far as possible to our case. They were also evaluated for their quality based on whether they were structured, well documented and in current use by other organizations.

From this analysis, candidate ontologies were obtained for re-use in our system. The most relevant ones, related to universities, were Academic Institution Internal Structure Ontology (AIISO)[4], Teaching Core Vocabulary Specification (TEACH)[5] and The Bowlogna Ontology[6]. AIISO describes the organizational structure of the university very simply, showing the hierarchy and relationships between different agencies. In that regard, it provides classes for modelling: teaching staff, subjects, departments, centres, etc. The TEACH and Bowlogna ontologies describe the part more related to teaching, that is, they try to represent the relationship between a program, the subjects that comprise it, the teaching workload and responsibilities of teachers for each of the subjects. Bowlogna is a more special case, representing the organization of the university following the Bologna Plan currently being implemented by European universities (Demartini et al., 2013).

All things considered, a university is still an organization, aside from the academic element, which can be modelled with organizations ontologies like W3C's The Organization Ontology[7] or Buildings and Rooms Vocabulary[8]. Finally, we intend to define the situation for premises and staff, and CTIC's Vocabulary of Localizations[9] or vCard Ontology[10] can be used for that, providing a series of classes with which to model addresses, municipalities, provinces, etc. These ontologies are very important as in many cases they are standard and widely used, as their use is not limited to a specific field.

Another important action was to take examples of the strategies established by other universities as a step towards the publication of Linked Data. They usually follow a basic scheme, re-using as far as possible existing ontologies, and if these do not cover all of the necessary concepts, the ontology itself is created or extended from an existing one. This can be seen in the Open Data sections of the University of Oxford[11], the University of Southampton[12] and the Open University[13], that usually have an Open Data portal from which you can consult information modelled in different forms (navigation, queries, etc.). These portals are usually based on RDF software, Virtuoso being the most popular, although there are other tools such as Pubby [14], Fuseki[15] or D2R[16].

The general situation is that when choosing ontologies that adapt themselves more to universities, there is consensus for the use of several of those named above, but none ever achieves a complete representation of the information, which is why it is necessary to create another to be able to link it with the other ontologies.

To achieve the publication of data, the ontologies must be published in such a way that they are dereferenceable and well documented, permitting the rest of the world to re-use them, thus achieving interconnections and future inferences of information (W3C, 2008) (Heath and Bizer, 2011).

In our case, we do not enter into the creation of ontologies in depth, as that was not one of the principal objectives that we wished to demonstrate, but rather interoperability between systems that

[1] http://watson.kmi.open.ac.uk/WatsonWUI/
[2] http://swoogle.umbc.edu/
[3] http://lov.okfn.org/dataset/lov/
[4] http://vocab.org/aiiso/schema
[5] http://linkedscience.org/teach/ns/#
[6] http://diuf.unifr.ch/main/xi/bowlogna
[7] http://www.w3.org/TR/vocab-org/
[8] http://vocab.deri.ie/rooms
[9] http://purl.org/ctic/infraestructuras/localizacion
[10] http://www.w3.org/TR/vcard-rdf/
[11] https://data.ox.ac.uk/
[12] http://data.southampton.ac.uk/
[13] http://data.open.ac.uk/
[14] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[15] http://jena.apache.org/documentation/serving_data/
[16] http://d2rq.org/

implement Linked Data as a procedure prior to publication of data, bringing it into a real context.

3 MOTIVATION AND GOALS

The University of La Laguna is a large-scale public higher education institution. It currently involves more than 26,000 people including students, teachers and administration staff, distributed between 25 centres, 60 departments and other areas. There are also close links with private entities (companies, foundations and institutes) and public institutions such as the local island administration Cabildo de Tenerife, City Councils and the Government of the Canary Islands.

The organizational structure is decentralized, and many functions are delegated to each of the departments, centres and services. This fact has a special relevance to this work, as we have had to study the functions of the principal organizational units in depth. Each unit works almost independently, each one takes responsibility for handling administrative processes that have been delegated to them, collecting and maintaining the information necessary to operate (accounting, academic administration, libraries, ITC centres, etc.).

With respect to this work, there are many information systems that offer support to the university as a whole. Financial Management and Academic Management are among the main ones that are found. The first system takes care of the management of the administration staff and the teaching staff; while the second is concerned with the enrolment processes and everything related to teaching. Both aggregate a large quantity of data and are more or less controlled forming a fairly homogenous architecture. They use the same group of software development technologies, the same database management system, and, most relevant, they are under the responsibility of the same IT department. Although there is a certain homogeneity between these systems, extracting information across both systems continues to be a complex and costly task.

Apart from larger, older systems, there are smaller, independent systems, of great value to the institution. These have appeared over time according to the needs of services or departments. Examples of these systems are: the research service, the directories of the institution´s staff, quality control, diaries and events, etc. These are usually controlled by different areas and each one can have different software.

It is at this point that this work begins to have meaning. There are many cases where management, statistical or other similar information is requested from other institutions within the university itself. Most of these systems work independently from one another, and when it is necessary to consult information from two or more sources, it is necessary to establish connections between the different systems. Due to the complexity and internal structure of each system, staff have to make a particularly strenuous effort to obtain the data.

The main motivation to work in this environment and to offer a practical solution to these problems based on Linked Data is, on the one hand, the existence of the real problem of access to different sources of data, and on the other, that the university´s staff is quite able to accept new proposals and finally, the existence of a real and concrete problem with which to apply Linked Data methods and techniques (for the re-use of ontologies, generation of RDF links, publication of data and consumption of published data).

The key challenge is to offer the university a solution for offering information obtained from a variety of sources, without modifying the current systems, as many are legacy systems and are so assimilated within the institution that a change to them could cause chaos. This is the solution that we present below.

4 LINKED DATA PROCESS

The process of linking data from the different information systems has followed an iterative and incremental methodology (Suárez de Figueroa Baonza, 2010). This process has allowed us to obtain valuable results from the first iterations and refine them continuously. As a consequence, the Linked Data process is provided with the necessary flexibility to tackle the changes related to the requirements in any phase of the process.

The working methodology comprises six differentiated phases inspired by methodologies commonly used for the publication of Linked Data (Poveda-Villalón, 2012); (Corcho et al., 2013); (Fernández-López, 1997); (Atemezing et al., 2012) and adapted to the needs of the particular working environment. These phases are divided between: specification of data to be published, data modelling, generation of data in RDF, publication, linking and exploitation.

Given the peculiarities of linking data in our

university context, as opposed to that proposed for different methodologies for the publication of data (Poveda-Villalón, 2012), the publication phase is undertaken at an earlier stage to the linking of data.

4.1 Specification

The first task to be dealt with is the specification of the data to be published. ULL has a large amount of data related to administrative, academic or financial activities. With the aim of demonstrating the utility of linking data in the University under the paradigm of Linked Data, and in view of the difficulty of linking all the data in the institution owing to the large volume, two clearly defined organizational units were chosen: the organizational area and the academic area.

The data available in each of these areas were not related with one another, although they made reference to entities that could be easily linked.

The data of the University´s organizational area describe the hierarchy of the institution, the structure of the teaching staff and the distribution of the courses and the teaching areas. Figure 1 presents the organizational domain model:

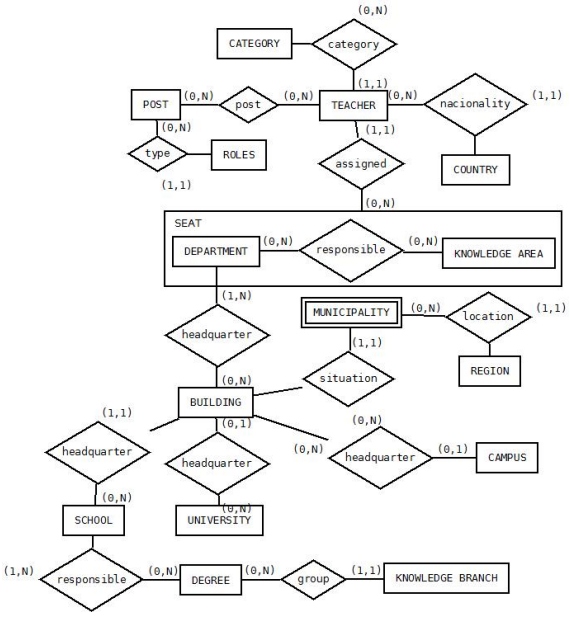

Figure 1: Organizational domain model.

The academic area contains information relating to the courses, which have been adapted to the Bologna Plan. The data from the academic area, Figure 2, contains information relating to the study plan of the courses, provided by the Spanish Government´s Ministry of Education, Culture and Sport, plus information relating to the teaching staff teaching those courses.

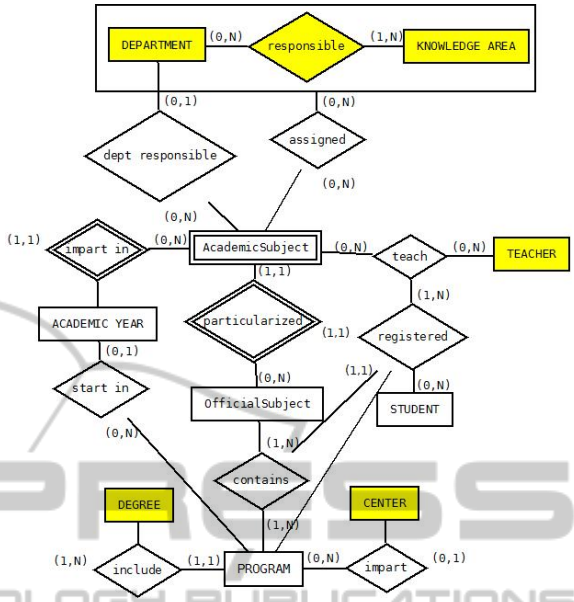

Figure 2: Academic domain model.

4.2 Modelling

Once the data to be linked had been determined, the next step was to analyse them in order to be able to begin modelling it with a set of ontologies.

Due to the peculiarities of the starting dataset, we opted for the development of an ontologies network, which had been re-utilized and created with ontologies with different characteristics. This ontologies network re-utilizes ontologies from the academic environment such as AIISO and TEACH, and more general and frequently used ontologies such as FOAF[17], SKOS[18] and others including LOC, ORG and ROOMS that model localization, and organization and premises, respectively. To complete the network, we created three new ontologies intended to represent the semantics which earlier ontologies could not cover. One of the ontologies was centred on academic information, and the other on organizational information, and the final one modelled general aspects of the institution. The development of the network of ontologies was carried out using NeOn Toolkit[19], an open source tool based on the Eclipse platform. It provides a set of plug-ins that cover many of the needs arising during this cycle of ontology development, such as

[17] http://xmlns.com/foaf/spec/

[18] http://www.w3.org/2008/05/skos

[19] http://neon-toolkit.org/

the generation of documentation, modularization, or the evaluation of ontologies (Zemmouchi-Ghomari & Ghomari, 2013). Figure 3 shows the network ontology obteined.

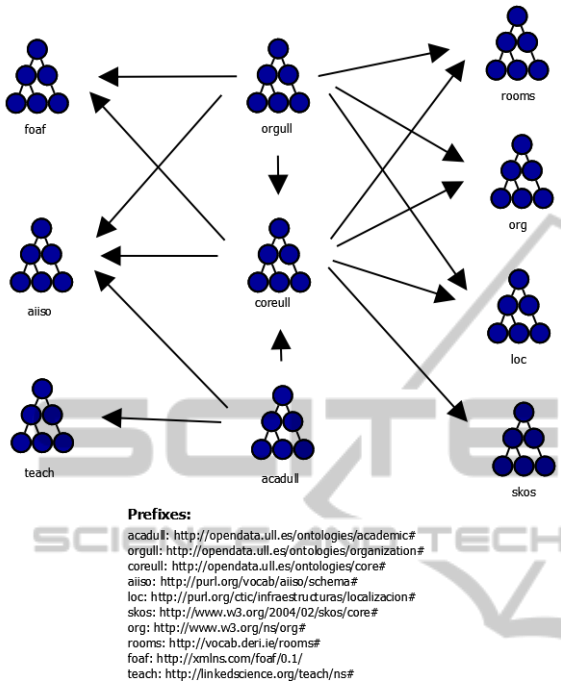

Figure 3: Modelled ontology network.

As part of the modelling phase, the anatomy of the URI is defined under the guidelines of the so called Cool URIs (W3C, 2008). As a result, the URIs of the resources follow the following pattern:

```
http://datos.ull.es/resource/{resource
type}/{resource}
```

Code 1: URI structure.

For example, the "IT Engineering" course is identified by the URI:

```
http://datos.ull.es/resource/programme/
IT_Engineering
```

Code 2: Example of an URI. IT Engineering identifier.

4.3 Generation

Once the ontology network has been defined, the next step is to generate the data in standard RDF format.

The D2RQ platform was used for that process, which allowed for data stored on relational databases to be consulted as if they were RDF graphs, making use of SPARQL language. Using a mapping language, in which the transformations to be made were specified following the defined ontology, this tool makes it possible to obtain data in RDF format by directly consulting a relational database.

As a consequence of using a tool with these characteristics, the results of this phase did not consist of a set of data in RDF, but in a file with the definition of the correspondences or mappings between the different fields in the organizational and academic area databases, and the elements of the ontology network previously defined (see Figure 3).

An extract of the mapping file obtained during this phase can be seen below:

```
map:institutions a d2rq:ClassMap;
d2rq:dataStorage map:database;
d2rq:uriPattern
"institutions/@@UNIVERSIDAD.NOMBRE|urli
fy@@";
d2rq:class ullorg:Universidad;
```

Code 3: Mapping an institution type resource in D2R.

4.4 Publication

The main objective of this phase is to obtain two access points for queries to the RDF format data of the organizational and academic areas.

Thanks to the implementation of a Pubby based interface, through the use of D2R, a front-end for SPARQL endpoints, access can be given to data stored in the institution´s relational databases through an HTML interface which displays them in RDF and at a SPARQL endpoint. This permits us to navigate through the information as well as to make specific queries.

4.5 Interlinking

Throughout this phase, we intended to achieve the objective of the availability of a five star datasets, following Tim Berners-Lee´s recognised classification, that is, a set of data linked with other sets of data (Heath and Bizer, 2012).

In our case, the linking process is divided into two phases, one of which establishes a data link internally, and another that links external data sources.

The internal linking, between our two access points, was achieved thanks to what was already known about how the URIs were going to be formed, as the structure is well defined and the identifiers of each of the resources in most cases are stipulated beforehand, and are common to the whole organization. Due to this, it is not necessary to use tools to discover candidates. Therefore is only

necessary that each system can be able to obtain the identifiers from the other system to create the URIs. In order to do this, the systems retrieve them from the other source using a SQL script. This process is only needed when new resource appears and not each time that a query is done.

If this information was not held, it would be necessary to make public the same resources in both domains (to duplicate the occurrences) and then to use linking tools, with owl:sameAs properties, to indicate that they are the same resource.

The external linking is possible when external data sources exist. For that, we undertook an analysis, and due to the scarcity of reliable sources, we limited ourselves to linking with dbpedia.org and linkeddata.es.

This link entails a process in which candidate links have to be found, and we intended to automate it as far as possible, using a Silk tool[20] and Link Specification Language (LSL) configurations. Once the tool was implemented, results that corresponded with the external resources that could be linked to our data were generated. These links were added to the database *LINKED* table in the following form:

```
http://opendata.ull.es/resource/{resourceid}  owl:sameAs  http://{external resource}
```

Code 4: Structure of sameAs statement.

This table is like a repository or cache where all the links that are found are stored and managed, containing: the URI of the internal resource, the URI of the external resource and the property that links them. In this way we are able to make links not only using owl:sameAs but also other useful properties, as well as management fields for configuration issues. This table also allow us to determine if a link is valid, if has been manually blocked, the date that the link was discovered, etc.

Once the links are added, the information is republished automatically with the external links. This is thanks to the added dynamic property in the D2R map that takes care of reviewing the linked table and showing the properties that exist there:

```
d2rq:belongsToClassMap map:CLASSTOLINK;
d2rq:dynamicProperty
"@@linked.propiedad@@";
d2rq:uriColumn "linked.objeto"; .
```

Code 5: Mapping sameAs statement in D2R.

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

4.6 Exploitation

When addressing this phase, we started from the two SPARQL access points available for the consultation of data from the University´s organizational and academic areas.

The objective of this phase was the availability of a single point of access to the platform that would permit federated queries on the two groups of data. To achieve that objective, a Fuseki server was deployed, allowing serve data in RDF format using HTTP. This service offers data to users through a RESTful API and an enriched SPARQL endpoint, from which so called SPARQL++ queries may be made (Polleres et al., 2007). With this service, complex queries can be made using proprietary databases or external endpoints. Previously, these types of queries were only possible using scripts and by the acquisition of data from sources outside the organization.

Here is an example of a complex query making use of two data sources:

```
PREFIX coreull: < … >
PREFIX orgull: < … >
PREFIX acaull: < … >
SELECT ?NameDept (COUNT(?planning) as
?numberPlanning) WHERE {
   SERVICE <http://orgull/sparql> {
      ?teacher orgull:miembro ?dept .
      ?dept coreull:nombre ?NameDept .
   }
   SERVICE <http://acadull/sparql> {
      ?planning acaull:tieneProfesor
?teacher .
   }
} GROUP BY ?NameDept
  ORDER BY ?NameDept .
```

Code 6: Query example: “How many subjects does each department teach?”

5 FINAL PRODUCT

The project finally developed is presented in Figure 4. We reproduced the complete system on a development server simulating the real system used at ULL. As it can be seen in the figure, we have two independent information systems: Academic and Organizational Management. Both systems reside in different databases supported by MySQL. An architecture based on D2R was developed above each of the two current systems for the mapping of the relational database to RDF files. Each system physically resides in a different server and has,

therefore, a separate portal for each environment.

Each D2RQ server had an SPARQL endpoint added and a web interface (Pubby-type) where corresponding information could be obtained.

A supervised process to discover and update external links was added using the Silk tool. More over, Fuseki service was also added as a junction for the whole system, enabling SPARQL++ queries.

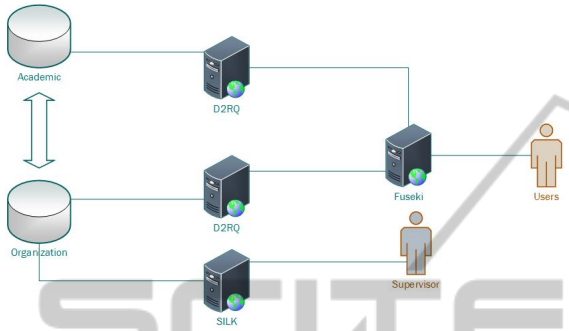

Figure 4: Architecture diagram.

Finally, it was possible to obtain an open data publication portal based on web semantics, meeting Linked Data requirements. Users will be able to use it to consult University information as well for making queries at several SPARQL points (so called SPARQL Endpoints) that may be linked to the published data. An example of this would be making a query such as "Distribution of students by province and their geo-localization to be shown on a map", represented in SPARQL as:

```
PREFIX coreull:
<http://orgull/ontologies/core#>
PREFIX orgull:
<http://orgull/ontologies/organization#
PREFIX acaull:
<http://orgull/ontologies/academic#>
PREFIX georss:
<http://www.georss.org/georss/>

SELECT ?Region (COUNT(?student) as
?numberStudents) ?geoloc
WHERE {
      SERVICE < http://acadull/sparql >
      {
      ?student a acaull:Alumno;
      coreull:provincia ?reg .
      }
      SERVICE < http://orgull/sparql > {
      ?reg coreull:nombre ?Region ;
      owl:sameAs ?linked .
      }
      SERVICE <
      http://dbpedia.org/sparql > {
      ?linked georss:point ?geoloc .
      }
} GROUP BY ?Region .
```

Code 7: Query with external data sources. Recount of students and their geolocalization.

The final platform enables users with simple SPARQL queries to make cross-queries of data from different sources that were previously inaccessible or of very complex interoperability. The end users, both inside and outside the University, have available a tool based on Linked Data to obtain the information they need. If case of need, it would be possible to add a user interface to make its use more user-friendly.

6 CONCLUSIONS

In this work, we have developed a genuine platform based on the Linked Data strategy in which universities may publish data and link with internal and external sources. The interlinking of the different university data sources will permit greater interoperability and can achieve new knowledge from this relationship.

To deploy our prototype in the university production servers, it is necessary to consider two important aspects: the integration costs of our platform with the university systems and also the university staff training in linked data. We firmly believe that both aspects can be addressed as we have the necessary technical expertise.

Whit this project, the institution has a tool to make complex queries, which hitherto existing systems could only handle with great effort. Moreover, the same tool could be offered to other local, island, and regional institutions in such a way that they could make queries directly, rather than through a service request to university personnel.

The ontologies most related to higher education have been revised, and improvements have been proposed to cover concepts not covered by existing ontologies. The network of ontologies used in this work covers ten ontologies, of which only three have been newly created.

Finally, as a summary of the work undertaken, we would stress that effective analysis of the initial data produces a lower number of errors in the re-use and definition of ontologies. We would also indicate that the process of creation of ontologies was based on the basic terms needed to demonstrate the viability of the project. We will continue to research several concepts that are beginning to be requested, in more depth, once the first version of the platform has been validated. With respect to the technological platform, a viability study will have to be undertaken to compare the effort required to incorporate triple store systems like Virtuoso, etc.

We should not forget that any new system for the organization would also implicate the commitment of human resources and materials for its development and future maintenance.

ACKNOWLEDGEMENTS

We would like to thanks to Andrés Palenzuela from the Planning and Analysis Office of University of La Laguna in the Canary Islands. His interest and support has made this Project viable.

REFERENCIAS

Atemezing, G., Corcho, O., Garijo, D., Mora, J., Poveda-Villalón, M., Rozas, P., Villazón-Terrazas, B. (2012). Transforming meteorological data into linked data. En *Semantic Web*. IOS Press.

Bauer, F., & Kaltenböck, M. (2012). *Linked Open Data: The Essentials A Quick Start Guide for Decision Makers.*

Berners-Lee, T. (2006). *Linked Data*. Recuperado el 11 de 10 de 2013, de http://www.w3.org/DesignIssues/LinkedData.html

Corcho, O., Fernández-Lopez, M., & Gómez-Perez, A. (2003). Methodologies, tools and languages for building. Where is their meeting point? En *Data & Knowledge Engineering* (págs. 41–64). Elsevier.

Demartini, G., Enchev, I., Gapany, J., & Cudré-Mauroux, P. (2013). The Bowlogna Ontology: Fostering Open Curricula and Agile Knowledge Bases for Europe's Higher Education Landscape. *Semantic Web – Interoperability, Usability, Applicability, 4*(1), 115.

Fernández-López, M. y.-P. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *AAAI-97 Spring Symposium Series*, (págs. 24-26). Stanford University, EEUU.

Fundación CTIC. (2013). *Public Dataset Catalogs Faceted Browser*. Recuperado el 11 de 10 de 2013, de http://datos.fundacionctic.org/sandbox/catalog/faceted/

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space* (First ed.). Morgan & Claypool.

Office, C. (2012). *Open Data White Paper: Unleashing the Potential.* TSO (The Stationery Office).

Polleres, A., Scharffe, F., & Schindlauer, R. (2007). SPARQL++ for Mapping Between RDF Vocabularies. En *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS* (págs. 878-896). Springer Berlin Heidelberg.

Poveda-Villalón, M. (2012). A Reuse-Based Lightweight Method for Developing Linked Data Ontologies and Vocabularies. En *The Semantic Web: Research and Applications* (págs. 833-837). Springer Berlin Heidelberg.

Suárez de Figueroa Baonza, M. d. (2010). NeOn Methodology for Building Ontology. *M.C. Doctoral Thesis.* Universidad Politécnica de Madrid.

W3C. (2008). *Best Practice Recipes for Publishing RDF Vocabularies*. Recuperado el 11 de 10 de 2013, de http://www.w3.org/TR/swbp-vocab-pub/

W3C. (2008). *Cool URIs for the Semantic Web*. Recuperado el 11 de 10 de 2013, de http://www.w3.org/TR/cooluris/

Zemmouchi-Ghomari, L., & Ghomari, A. R. (2013). Process of Building Reference Ontology for Higher Education. *World Congress on Engineering 2013, III*, 1595-1600.