Experiences from the Design and Development of an Institutional Linked Open Data Portal

Daniel Alami [1], Isaac Lera [1], Carlos Guerrero [1], Carlos Juiz [1]

[1] *Deptartament de Matemàtiques i Informàtica, Universitat de les Illes Balears, 07122, Palma, Spain*

***Abstract* –Linked Open Data initiative, based on the open publication of datasets, provides new mechanisms for the development of novel services, and through rigorous data analysis new indicators for government strategies. We contribute with a linked open portal to integrate curriculum data of our academic staff. We analyse our experience in the creation of the semantic model and we include some cases of data exploitation. The main contributions are a better management of curricular data, the quality and efficiency of exploitation tasks, and the transparency, dissemination and collaboration with other researchers and with the citizenship.**

***Keywords* – linked open data, semantic web, curriculum data.**

1. Introduction

Linked Open Data Initiative (LOD) is about using Semantic Web technologies for the establishments of relationships between pieces of data to facilitate publication, access and reuse of data. With these best recommendations, private and public institutions can describe their functions, services, their organizational structure, and the results among other operative aspects through small and public pieces of interrelated data. There is an explicit enrichment in resources description when third datasets are related in the pieces of data. This kind of tangled content catalyzes the appearance of new exploitation data models and, as we mentioned, it facilitates the obtaining of multidisciplinary indicators for government policies.

Data publication under LOD initiative has to follow a series of guidelines. In this sense, Tim Berners-Lee, director of the World Wide Web Consortium, defined a series of categories based on a 5-star ranking to classify open data portals. The maximum category is reached with the five stars and corresponds to the semantic representation of the data and to link it with other third data sources of semantic data. This type of representations requires adequate abstractions of the pieces of information because it gives rise to a different perspective of the exploitation of the data, since it opens the option to the citizen and other companies to use them.

In this study, we show the initiative carried out in our university for the creation of a LOD portal with a ranking of 5 stars for the publication of curricular information of the research activity of the academic staff. We discussed technical and operational aspects required to implement this catalogue with 5 years of scientific production with an average number of 1,200 researchers.

Academic curriculum data represent staff activities: teaching, management, professional experience and, of course, research production. In academic ambit, CV management may be more critical than in other professional fields. Multiple services depend on it such as: project grants, scholarships, collaboration with other groups, and accreditation for job offers. Our institution offers a tool to manage the CV that facilitates the creation of reports for various official applications (national projects, Marie Curie actions, etc.). Moreover, CV data is used to obtain indicators to maximize the quality of assignments, control and planning of human and economic resources; for example, the financial assistance for event attendance relies on it. The application was developed by a third company and is used by a group of universities. The

DOI: 10.18421/TEM64-01
https://dx.doi.org/10.18421/TEM64-01

Corresponding author: Isaac Lera ,
Deptartament de Matemàtiques i Informàtica, Universitat de les Illes Balears, 07122, Palma, Spain
Email: Isaac.lera@uib.es

Received: 17 September 2017
Revised: 31 October 2017
Accepted: 07 November 2017
Published: 27 November 2017

information maintained there is not accessible by other people and is stored under the guidelines of the application itself.

The information stored in the application is accessible through well-defined internal reports, which means it is not accessible to general public. However, the research activity of academic staff is accessible but distributed on multiple platforms such as indexing or editorials portals: Springer, Scopus or Google Scholar. On the other hand, the application generates predefined statistics every certain period of time and the navigation is little fluid for the discovery, the analysis and the extraction of information. That is, it is not contemplated that the user can extract information to do studies or to use it as he or she wishes.

This situation is a clear example of how a portal developed for the management of contents with a reduced set of functionalities becomes insufficient for the exploitation of the data. Implementing a new requirement requires the involvement of multiple agents and a considerable number of changes in the current environment. Therefore, if we opt for a highly linked open data portal model for the development of this curricular manager we would obtain a series of advantages. In our opinion, the most important is offer raw data to provide full interoperability to the users that they have the capacity to incorporate and share new services; metadata without ambiguity and without being subject to a specific tool and allowing to relate pieces of information with other external sources of information; and finally, the free extraction of any content without adhering to a predefined model. All this guarantees the scalability of functionalities, reuse and transparency.

We have summarized the advantages of an open CV based on LOD portal that will be exposed throughout the document:

- At the academic-personal level:
  - Discovery of new items
  - Complementing the item definition
  - Establishment and discovery of new collaborative networks at the academic level
  - Export CV in any format
- At institutional level:
  - Obtaining impact and performance indicators
  - Support for strategies and best practices
  - Reduction of administrative barriers
  - Monitoring mechanisms
- At the public level:
  - Transparency
  - Diffusion and technological trends
  - Return of value by obtaining new services
  - Establishment of business networks

Throughout the document, we describe the methodology, the technological aspects, we will discuss the psychological barrier to publish data in open and how this project can contribute to improve the government of the institution. In this proposal, we highlight the main research points: how to publish data in open and in our case specific curricular data in an academic environment and the establishment of relationships with other data sources. On the other hand, the contribution of this work relies on applying the recommendations in a real scenario and in the generation of new exploitation models with the published data.

2. Related work

We have considered two main issues: the creation of LOD portals and the exploitation of curricular data. To cover the first point, we describe some universities with LOD portals to analyse how they publish data under the paradigm of the Semantic Web. The obligation to publish open data, transparency policies, means that many of the entities in possession of LOD portals are public entities. In the second part, we show cases of data exploitation focused on obtaining indicators of scientific productivity. Most of the papers analysed deal with different topics: government, semantic data modelling, data mining, and complex networks.

Briefly, we highlight the portals of two universities such as Southampton[1] and Münster[2]. Southampton portal is characterized by having the largest number of catalogues: telephone directory, institutional services, academic events, catering, locations and map, transport, job offers, room reservation and equipment. It offers a direct query service to any catalogue similar to SQL through the SPARQL protocol [1]. The LOD portal of Münster contains the following catalogues: staff, publications, projects, patents, locations, fees and academic courses. Both portals are characterized by having a ranking between 4 and 5 stars according to the catalogue. The data can be consulted freely, and can be exported in formats like RDF and CSV.

Spreckelsen et *al.* [2] analyse the use of semantic technologies to model curricular data in medical research. They introduce a semantic MediaWiki for the bidding process and the revision of learning objectives based on the catalogue based on the Aachen curriculum model of medicine. The semantic MediaWiki uses a model domain curriculum and offers a structured model of queries, multiple views, semantic indexes, and a set of learning objectives. In this way, students can prepare evaluative tests and

[1] http://data.southampton.ac.uk/

[2] http://lodum.de/

teachers can adapt the course to the multiple disciplines of enrolled students. In their opinion, the approach based on this MediaWiki allowed an agile implementation of knowledge management.

EUROCV platform [3] provides an interesting study on the electronic development of national curriculum information systems in countries such as Norway, Spain, Portugal, United Kingdom, France, Israel, etc. It mentions some future recommendations that in our proposal are included. The first is to identify template fields in PDF, as is done at the National Science Foundation of Switzerland and currently carries out its Spanish counterpart in the digital and standardized curriculum vitae (CVN). This is an advantage for the efficient computation of these documents, but the semantic web still improves the identification, treatment and consultation of these data. Another recommendation is the need for researchers not to devote so much effort to change their CV in each project application or scholarship format. In our case, the creation of different templates is based on the creation of a set of queries about the information that is required in each section.

Thomson Reuters, one of the most important scientific publishers, published a paper on the evaluation of research performance using citation data [4]. Institutions and researchers in several proposals such as accreditations, and internal faculty aids, economic resource planning, among others have used this information. In general terms, editorials have more up-to-date information than the academic institution. In addition, the services generated from these data are not free for the institution and they do not have the option to adapt them to their requirements. In the LOD paradigm, the data are of public use and being able to be linked with other entities can complement the available information. At the same time, it is possible to exploit the data by any entity with the consequent adaptation to the social, cultural and economic context of each institution.

3. Data Modelling

Institutions and companies store their data in structured formats. The modelling of such information requires an effort to be interpreted, either the adaptation of pieces of information that form a curriculum in a model entity relationship. The interpretation of modelling is directly related to the interpretation of metadata. This restricts the number of users who can exploit it due to the complexity and implicit knowledge used in modelling. For example, in our case, there were more than 30 tables of the entity-relationship model that housed the entire curriculum system. Each table contained on average 20 attributes, and the only interpretable element was the name of the attribute, which was usually very sparse. Finally, after considerable effort and with the help of technicians all the metadata could be interpreted. The absence of an easily interpretable scheme hampers the publication and exploitation. Only technical and knowledgeable personnel can exploit it.

The languages proposed by the Semantic Web, RDF and OWL, allow the inclusion of business logic in the data itself. That is, each piece of information has a typing that determines its interpretation. The relationships between pieces of information also contain such typing. The smallest piece of information in SW is a triplet and is self-sufficient. A triplet consists of a subject, a predicate, and an object. This representation resembles simple sentences of the natural language where there is a subject of the action, the verb and the complement. Each element of a triplet is identified with a URI and at the same time is accessible and linkable via that url. Table 1 shows an example with three triplets where each one has a prefix that represents the abbreviation of the URI and is the place where that entity is defined and found. The prefix 'cv:' is symbolic and the prefix 'rdf:' refers to the definition of the RDF schema since the predicate "type" is a constructor defined by the W3C on the typing of the elements.

Table 1. Four triplets regarding CV productivity

cv:Lera cv:coauthor cv:ExperiencesOnLOD
cv:Lera rdf:type cv:Professor
cv:ExperiencesOnLOD rdf:type cv:Journal
cv:Alami cv:coauthor cv:ExperiencesOnLOD

These languages allow to extend this information, for example we can add more definitions to the previous example as is the authorship: "cv:coauthor rdfs:subProperty cv:author". At the same time these languages support inference with which we can discover implicit triplets, for example: "cv:Lera rdf:type cv:author" using the authorship relation.

With this triplet we could infer that any co-author is also at the same time author, inheriting the implications that the author concept may have. This ability to infer information gives a greater expressiveness in the reuse and extension of existing knowledge. In [5] there is an extensive description of descriptive logic as a mechanism for inferring information in the SW languages.

We recommend considering five issues regarding semantic data modelling. The first is about data transformation. The transformation of an entity-relationship model into a semantic model is not straightforward. There are different tools for such a process [6], but the important thing is to keep labels or specific references to entities in the different pieces of information. That is, we cannot use primary identifiers of database records as names of each part of the triplet. This action would hinder the interpretation and, therefore, the use of the information. Therefore, URIs must be easily readable to simplify human interpretation.

The second aspect, semantic modelling should not be oriented to any previous exploitation. The data are represented as they are, with the least number of possible relationships among other sources not to condition them to a target. This allows the user to introduce new knowledge without encountering inconsistencies due to the existence of logical constraints. That is, we can include the ORCID identifier in our field as it complements and provides authoring equalization services. And on the other hand, we should not restrict knowledge to a particular case, introduce the triplet: "author is Academic Member" requires that any author is personal of the institution and in many cases it is not. Therefore, the second recommendation is to avoid restricting the knowledge exposed and offer it with the least number of logical constraints.

The third aspect is to assign at least one type to each individual. The typifying action is the most basic piece of information. Each individual has identified the class to which he belongs and therefore, it is under the logical constraints of the typifying classes.

The fourth consideration lies in the definition of relations of inheritance between relationships. In object-oriented programming, classes subsume other classes. At the semantic level, a property between resources can also subsume other properties. This fact maximizes the vocabulary involved in defining the context and gives expressiveness to the granularity of services. The above example, on co-authorship, concerns this case.

The last consideration is to represent the evolution of the pieces of information in different periods of time. The curricular items are linked to a date. Over a certain period of time, items can be repeated, such as being a member of a program committee or having assigned the same subjects. The identification of pieces of information using Semantic Web languages is unique. In the assumption that we take the approximation of an ER model, if a user has been a member of a program committee, the cardinality of this assignment would carry the date. The ternary relationships in SW require a complex representation of triplets. We have opted for an efficient solution in terms of queries and generation of academic performance indicators. We have separated the data of each year in a different catalogue. In this way, each piece of information belongs to a catalogue with its specific IRIs. All catalogues share the same schema, which facilitates queries, and obtaining a certain period of time just requires filtering the catalogues involved. In the SPARQL query language, such selection consists of reference to the involved graphs.

A. Name entity recognition

A critical step in the semantic transformation of data is the recognition of the names of entities (NER) with their own meaning, such as authors, editorials, patents, etc. This problem has its origin by the free inclusion of text in the corresponding form that inserts it directly in the database. That is, the same fact can be presented in plain text in multiple ways. The classical example is the diverse formats of authors in publications. Some possible cases are: "Lera, I .; Juiz, C." or “Isaac Lera and Carlos Juiz”, and a long list of combinations. In order to select the most likely assignments, we established a system that evaluates the relationships between the researchers in the database (group membership, participation in projects, or co-authoring of other publications without the existence of ambiguity). In our data, the average number of NER is about 7752 per year.

4. System architecture

The application was built on a classic web client-server architecture using AngularJS technology to facilitate the presentation of data between the client and the server application. Messages were exchange using RESTful protocol, which simplifies the publication of database dumps in the browser. The NoSQL database used is AllegroGraph, which is based on graphs and automatically supports RDF files and SPARQL queries.

5. Data exploitation

At this point, any user can analyse the information through two mechanisms: by partial data downloads using SPARQL queries or by downloading the entire catalogue.

SPARQL is the query protocol in the Semantic Web. Similar to SQL can be used to access multiple catalogues using various logical functions. Through this technology, any kind of user has the ability to make queries without having the need to implement that service. This service saves costs and development time. For example, a manager might conduct a query to find out the cities where academic staff is traveling with the goal of establishing new collaboration policies. Table 2. implements this query, as we can observe it set by two triplets. Following, this idea of simplicity, table 3. contains a query to obtain the scientific productivity and his/her job category of each employee.

Table 2. A SPARQL query to obtain the cities more visited by academic staff.

```
SELECT ?country (COUNT(?country) as ?total)
WHERE {
?outcome base:author ?authors .
?outcome base:country ?country
} GROUP BY ?country
```

Table 3. A SPARQL query to obtain the throughput and job category of each researcher.

```
SELECT ?investigator ?employeeCategory ?outcomes
?type
WHERE {
```

```
?investigator model:has ?outcomes
?outcomes rdf:type ?type
?investigator model:has ?employeeCategory
}
```

To facilitate the analysis of the data, all the results of SPARQL queries can be downloaded in the following formats: rdf, xml, json, graphml, and csv.

We present a small case study to show the capacity of this type of portals to offer data for analysis. SW languages can be modelled like graphs enabling the applicability of complex network theory. The transformation is based on mapping the nodes as the resources of the model, both subjects and objects, and as vertices, the properties among classes. Also, we can include attributes (the type of productivity, the category of the revises, the country, etc.) to complement the description in order to facilitate the visual interpretation. In this case, we have load the catalogue through the graphml format into our algorithm using the NetworkX library [7] to compute topological features.

In the network, we analyse the degree of intergroup academic collaboration, that is, between the institution's staff and that of other institutions and other research groups through scientific publications. For this case, we have used the catalogue of 2012. In Figure 1., we show the representation of the graph of this year between researchers and throughputs. The degree of collaboration is estimated by calculating the eccentricity distribution of the graph. Eccentricity is the maximum distance between vertices. On the data set, we calculated the distribution on the selection of authors of the institution and the rest. Figure 2. shows the degree of collaboration between authors. Approximately 4000 publications have a single author. The rest of publications have a higher degree of collaboration from 1 to 6 and from 14 to 24.

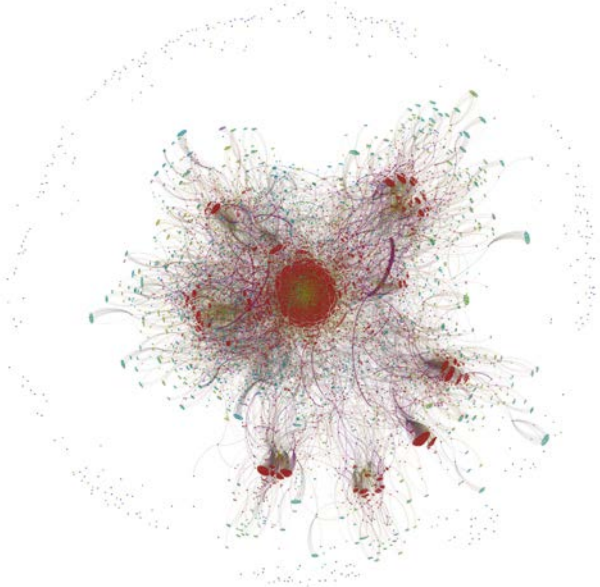

Figure 1. Network of the throughput results of the academic staff along one year.

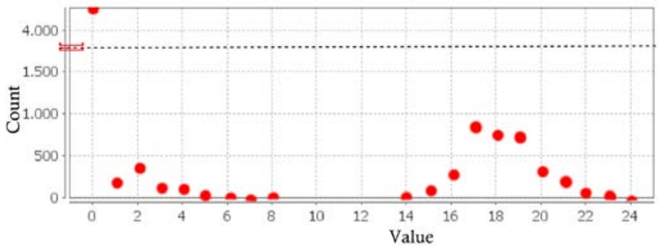

Figure 2. Degree collaboration distribution of authors in journals publications.

Another case of data exploitation has been to take advantage of the linking of information in our catalogue with information from other catalogues using semantic properties. Thus, we compare the information with an external catalogue such as DBLP. This comparison can help to find inconsistencies in the bibliographic entries and offer recommendations for their correction. The DBLP semantic repository contains bibliographic information in computer science and publishes the information in RDF format. Clinging to the area of computer science, we applied this study to 255 academics. Only 67 had publications in conferences or journals indexed by DBLP. The first step is to link both entities using the attributes of authors and the title of the publication. Linking is syntactic using Levenshtein's distance [8] with a tolerance under 7 characters. Of the 67 academic involved, we know that the institution's data manager contains 15015 publications and the DBLP catalogue contains 5893. Using correspondence by publication title, 2621 publications are counted. Therefore, at the level of inconsistencies, we estimate two issues: I) 3272 (55.52%) publications contain a different title; II) Of the 2621 publications, 58.73% have a different number of authors.

6. Discussion

Mapping processes are necessary to establish equivalency relationships among resources. For example, they are crucial to detect that an author of the institution, is also an author of the DBLP catalogue. This process can be improved having a high number of attributes but in many cases only the label is available. This reduces the level of accuracy. To avoid this, we think the only solution is the universal identification of resources through some type of organization. At the academic level, the ORCID proposal is to unequivocally identify each author and several publishers are beginning to use that identifier to manage the identities of the authors. Also, the identification of articles is done through the DOI system

Currently, there are several works related to the academic world that would increase the related knowledge of the entities of the data catalogue. The VIVO project is an initiative to link academic information in any type of degree. Associated entities, among which we can highlight Thomson-Reuters, Elsevier, ORCID, EuroCris, etc. make a series of efforts to offer the above services within the paradigm of LOD. Another initiative to facilitate the publication and services of academic data exploitation is Scholarlydata [9].

Establishing relationships between entities and the metadata used to describe them facilitate search processes. The SemanticScholar[3] search engine performed by the Allen Institute in 2015 clearly demonstrates how this type of linked information offers a high versatility to analyse and discover articles relevant to a text entry.

7. Conclusion

Transparency and release of data in public institutions is an indicator of healthy democratic societies. The nature of the catalogue and the format in which it is released mark the degree of utility and the degree of exploitation, respectively. In our case, we have shown the remarkable advantages of publishing a catalogue of scientific productivity using the W3C standards. This catalogue, regardless of the sensitivity of the interpretations, facilitates the creation of services by anyone that can be useful to the academic staff, the institution and the rest of the citizens. One of the fundamental aspects is the transformation of information already present in traditional formats as entity-relationship models in semantic formats. In the new model, the logic of the data is explicitly reflected so it is not worth a literal transformation of a database. Preserving and adding meaning to relationships is fundamental. In addition, we have demonstrated with various cases the simplicity of obtaining indicators without the intervention of a development team that implements internally the queries inside the whole infrastructure. This fact simplifies the queries on any type of information from listing projects to knowing the countries most frequented, or even filtering subjects that meet certain criteria (i.e. to offer scholarships). The second case of exploitation demonstrates how to link information with other catalogues with the goal of discovering inconsistencies among items.

Acknowledgements

All authors acknowledge financial support from the Departament de Matemàtiques i Informàtica de la Universitat de les Illes Balears. Thanks to Vice-chancellor for Innovation and Dissemination and, especially to Miquel Titos from Oficina de Suport a la Recerca for his support to analyze the data stored in GREC system.

References

[1] Prud'hommeaux, E. & Seaborne, A. (2008). SPARQL Query Language for RDF W3C Recommendation. (http://www.w3.org/TR/rdf-sparql-query)

[2] Spreckelsen, C., Finsterer, S., Cremer, J., & Schenkat, H. (2013). Can social semantic web techniques foster collaborative curriculum mapping in medicine?. *Journal of medical Internet research*, *15*(8). DOI: 10.2196/jmir.2623

[3] Fontes, M., Assis, J., Araújo, E. R., Bento, S., Henriques, L., Pirralha, A., & Santos, L. D. D. (2008). *Building new indicators for researchers' careers and mobility based on electronic curriculum vitae: the case of the DeGóis platform* (pp. 1-13). INETI/Dinâmia.

[4] Pendlebury, D. A. (2008). White paper: Using bibliometrics in evaluating research. Thomson Reuters.

[5] Pan, Z., & Horrocks, I. (2004). *Description Logics: reasoning support for the Semantic Web*. University of Manchester. http://homepages.abdn.ac.uk/jeff.z.pan/pages/pub/thesis.pdf

[6] Farouk, M., & Ishizuka, M. (2012, March). An extensible approach for mapping relational DB to RDF. In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*(pp. 163-167). IEEE.

[7] Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Laboratory (LANL).

[8] Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 1091-1095. DOI: 10.1109/TPAMI.2007.1078

[9] Gentile, A. L., & Nuzzolese, A. G. (2015). cLODg-Conference Linked Open Data Generator. In *International Semantic Web Conference (Posters & Demos).*

[3] https://www.semanticscholar.org/