Semantic Methods for Reusing Linking Open Data of the European Public Procurement Notices*

Jose María Álvarez[1], José Emilio Labra[1], Ángel Marín[2], and José Luis Marín[2]

[1] WESO Research Group-Universidad de Oviedo
[2] Gateway Strategic Consultancy Services
josem.alvarez@josem.alvarez.es (PhD Student), labra@uniovi.es,
anmar@gateway-scs.es, josmar@gateway-scs.es
WWW home page: http://purl.org/weso,
http://gateway-scs.es/

Abstract. The aim of this paper is to show the activities to be performed in the development of a PhD. about e-procurement using linking open data. The study will be focused on: 1) modeling the unstructured information included in public procurement notices (organizations, contracting authorities, contracts awarded, etc.) using semantic web technologies; 2) enriching that information with the existing product classification systems and the linked data vocabularies; 3) publishing relevant information extracted out of the notices following the linking open data approach; 4) exploiting the information through advanced algorithms providing value added services to users, with special focus on SMEs. Finally an evaluation methodology is outlined to validate the goodness and the improvement of the proposed system regarding to the existing ones.

1 Introduction

The European Commission outlines the following advantages in the wider use of e-Procurement [3]: increased accessibility and transparency, benefits for individual procedures, benefits in terms of more efficient procurement administration and potential for integration of EU procurement markets. TED[4] ('Tenders Electronic Daily') is the on line version of the 'Supplement to the Official Journal of the European Union', dedicated to European public procurement (1500 new procurement notices every day [5] but an unified information system pan-European

* This work is part of '10ders Information Services project' partially funded by the Spanish Ministry of Industry, Commerce and Tourism, leaded by 'Gateway Strategic Consultancy Services' and developed in cooperation with 'EXIS TI' (http://www.exis-ti.com/) and WESO Research Group.

[3] http://ec.europa.eu/internal_market/consultations/docs/2010/e-procurement/green -paper_en.pdf

[4] http://ted.europa.eu/

[5] http://www.ted.europa.eu/TED/main/HomePage.do

dealing with: 1) dispersion of the information; 2) duplication of the same notice in more than one source; 3) different publishing formats; 4) problems regarding to a multilingual environment and 5) aggregation of low-value procurement opportunities, is missing.

On the other hand in the European eGovernment context, there are several conceptual/terminological maps of particular domains available in RAMON[6], the Eurostat's metadata server: Health, Education, Employment or e-Procurement among others. The structure and features of these systems are very heterogeneous, although some common aspects can be found in all of them: 1) hierarchical relationships between terms or concepts; 2) multilingual character of the information. These knowledge organization systems (KOS) enable users to annotate information providing an agile mechanism for performing tasks such as exploration, searching, automatic classification or reasoning.

Obviously one of the most interesting domains to apply the Linking Open Data (LOD) approach is public procurement information published by governmental contracting authorities. In that sense, the growing commitment to the reuse of public sector information (PSI) and initiatives like semantic web, LOD and the use of KOS provide building blocks for an innovative unified pan-European information system for the benefit of SMEs.

Main Contributions

This work aims to apply the semantic web and LOD approaches to public procurement notices: 1) Transforming government controlled vocabularies such as CPV [7], CPC [8], Eurovoc[9] (now available in SKOS), etc. to RDF, RDF(S), SKOS or OWL; 2) Modeling the information inside the public procurement notices as web information resources and enriching them with the aforementioned controlled vocabularies, geographical information (e.g NUTS[10]) and the information now available in the linked data cloud; 3) Publishing the information in a SPARQL endpoint providing a "linked data node" and 4) Providing enhanced services (search and sort, matchmaking, georeasoning, statistics, etc.) exploiting this semantic information through "advanced algorithms" based on Spreading Activation (SA) techniques, rule based systems (RBS), recommending engines, syntactic search engines, fuzzy logic and a mixing of them selecting the best combination that fulfills the intentions of the clients.

E.g: *Which public procurement notices are relevant to Dutch companies (only SMEs) that want to tender for contracts announced by local authorities with a total value lower than 170K € to procure "Transport and Related Services" and a two year duration in the Dutch-speaking region of Flanders (Belgium)?.*

---

[6] http://ec.europa.eu/eurostat/ramon

[7] http://europa.eu/legislation_summaries/internal_market/businesses/public_procurement/l22008_en.htm

[8] http://unstats.un.org/unsd/cr/registry/isic-4.asp

[9] http://eurovoc.europa.eu/

[10] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Existing systems are based on the perfect match of the different parameters of a request applying NLP techniques or query languages like SQL. Moreover they do not take advantage of the semantic information included in the controlled vocabularies used to describe the public procurement notices. That is why the main purpose of the research will be the study, selection, combination, implementation and testing of algorithms able to exploit this semantic information providing enhanced services out of the scope of current systems.

Finally, the methodology used to address these contributions and the PhD is based on STI Research Plan[11].

2 Related work

In the scope of LOD and open government data (OGD) there are projects trying to exploit the information of public procurement notices like LOTED [12] ("Linked Open Tenders Electronic Daily") where they use the RSS feeds of TED. UK government[13] is doing a great effort to promote its information sources using the LOD approach. They have published datasets from different sectors: transport, defense, NUTS geographical information [14], etc. Most of the public administrations in the different countries are also betting for LOD approach to make public their information: Spain (Aporta project [15]), USA[16], etc. Regarding the use of LOD and organizations there is a new ontology for modeling the information about organizations [17] and recently it has been released "The Open Database Of The Corporate World" [18].

Product Scheme Classifications (also known as PSCs) have been built to solve specific problems of interoperability and communication in e-commerce[5]. The aim of a PSC is to be used as a standard *de facto* by different agents for information interchange in marketplaces [8,1]. Any PSC, as well as other classification systems can be interpreted as: 1) domain-ontologies [4] or 2) conceptual schemes [10] comprised of conceptual resources . Finally, Good Relations [19] is an ontology for the e-commerce developed by Martin Hepp et. al.

On the other hand, the main use of SA techniques is focus on Document and Information Retrieval [3]. These techniques has been also used in semantic search based on hybrid approaches [9,2], user query expansion combining metadata and user information to improve web data annotations. RBSs have been used a long time to decision support, diagnosis, etc. in different fields. In the semantic web

[11] http://www.sti-innsbruck.at/uploads/media/STI_Research_Plan_03.12.2008.pdf
[12] http://loted.eu:8081/LOTED1Rep/
[13] http://data.gov.uk
[14] http://nuts.psi.enakting.org/
[15] http://www.aporta.es/
[16] http://www.data.gov/
[17] http://www.epimorphics.com/web/category/category/developers/organization-ontolog y
[18] http://opencorporates.com/
[19] http://www.heppnetz.de/projects/goodrelations/

area and due to the apparition of OWL 2-RL, SPARQL Rules! and RIF these systems are growing in their use to deal with the web of data but a clear approach to mix datasets and RBSs is missing. They can also be applied to SA techniques to handle the activation and propagation of the concepts.

3 Proposed approach

The proposed architecture, see Fig. 1, is based on two main processes: 1) RDFizing. It is the process to transform the data available in the databases about public procurement notices from a XML intermediate format to RDF and enrich them with the vocabularies of the linked data cloud. It also codes the PSCs as linked data. 2) Enhanced services. It is the application of the libraries such as ONTOSPREAD [20], RIFle[21], Apache Mahout [22], Apache Lucene [23] and jFuzzyLogic [24] to exploit the linked database and provide services to the customers.

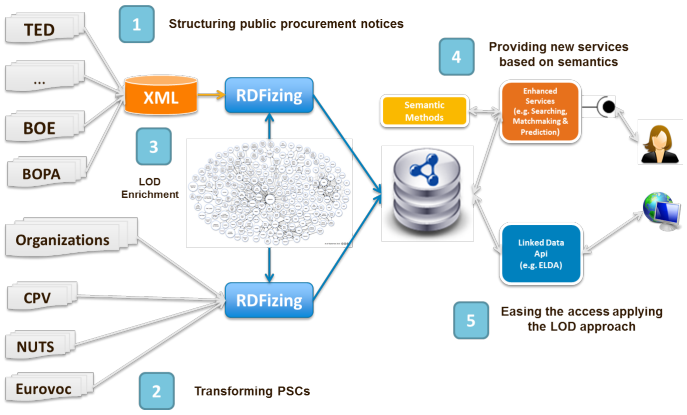

Fig. 1. Proposed Architecture Overview

The combination of these building blocks seeks for creating a new innovative way to exploit the information included in public procurement notices in the context of the semantic web and LOD initiatives reusing the existing technologies, vocabularies, etc. Following, the key points of this approach are summarized: 1) existing PSCs, information about public procurement notices and organizations are published as linked data; 2) the application of SA techniques, RBSs and the abovementioned libraries to provide enhanced services adding value to original information.

[20] http://code.google.com/p/ontospread/
[21] http://rifle.sf.net
[22] http://mahout.apache.org/
[23] http://lucene.apache.org/
[24] http://jfuzzylogic.sourceforge.net/

4 Process to PhD-Results

The main objective of the researching is to seek for the best combination of algorithms to exploit the information about public procurement notices. Some information variables (with a certain relevance fixed by domain experts) of the public procurement notices have been selected to be enriched through the application of the different algorithms: CPV codes, total cost, duration, date, type of company and location. Taking into account that clients provide this information the system is able to enhance their requests: adding new CPV codes, establishing an interval (data mining of the public procurement notices or selected by the client) in which numeric variables like total cost, duration and date can lie or selecting "nearby regions" using georeasoning. More specifically, CPV codes are enhanced with NLP techniques (Apache Lucene), recommending systems (Apache Mahout), SA techniques or dividing into the narrower codes of the input CPV codes a certain value. Finally the enhanced request is translated to SPARQL query language and executed via an endpoint. The results are collected and can be sorted according to different aggregation operators.

Currently the process for publishing the PSCs and the information extracted from public procurement notices is partially finished and the linked data is publicly available at a SPARQL Endpoint. MOLDEAS [25] (Methods On Linked Data for E-procurement Applying Semantics) project is hosting the implementation of the algorithms to enhance the variables of the public procurement notices. We are now focus on the enrichment of the CPV codes and the aforementioned methods have already been implemented but due to the limitations of the current specification of SPARQL (not support of negation, paths, fuzzy matches with ranked results we have reoriented some kinds of queries waiting for the new recommendation of SPARQL 1.1 (actually Virtuoso's SPARQL-BI offers some of the desired features to the proposed system).

Next steps include the implementation of the methods to enrich numeric variables and the definition of the aggregation operators. Besides we are designing the experiment to validate the goodness and the improvement of the system regarding to existing systems. In that sense, the experiment apart from the selected service to be tested depends on two main variables: 1) the amount of information used and 2) the number of tests that should be carried out. From the first variable point of view 1M public procurement notices (provided by Gateway SCS-Eurolert.net[26]) and over 320K organizations [27] are available. On the second one, we have not decided yet how many tests would be appropriated to provide a correct evaluation but the information about how many queries are requested per day in the existing public systems can be a right trail.

On the other hand, taking into account that the service of searching or matchmaking is the most relevant in this kind of system we are preparing a test suite with the aforementioned information(search queries and expected re-

[25] http://moldeas-web.appspot.com/
[26] http://euroalert.net/
[27] ftp://ftp.ted.europa.eu/META-XML/

sults) to compare the precision and recall of existing public systems (free text and advanced key fields search of TED) to the proposed one (LOD+Semantic Methods+SPARQL). The expected result of this evaluation will validate our approach for improving the access and retrieval of the information about public procurement notices using the LOD approach and the best combination of the implemented algorithms.

5 Conclusions and Future Work

The implementation of this work is supposed to afford a new way to exploit the information published inside public procurement notices applying advanced algorithms on LOD. Following we highlight the advantages of this approach: 1) decreasing of the information's dispersion; 2) unification of the data models and formats; 3) implicit support to multilingual and multicultural issues; 4) enrichment of the public procurement notices; 5) alignment with the Digital Agenda for Europe; 6) raise awareness on public procurement opportunities among SMEs and 7) deployment of enhanced services on public procurement notices. Regarding the future work, the results of this study are intended to be exploited by a commercial service like Eurolert.net [7,6] and we are also interested in report the results to *The Internal Market and Services Directorate General (DG MARKT) of the European Commision*, *The Information Society and Media Directorate General (DG INFSO) of the European Commision*, the LOD and OGD initiatives among others.

References

1. G. Alor-Hernández, JM. Gómez Berbís, and A. Rodríguez González et al. HYDRA: A Middleware-Oriented Integrated Architecture for e-Procurement in Supply Chains. *T. Computational Collective Intelligence*, 1:1–20, 2010.
2. D. Berrueta, JE. Labra, and L. Polo. Searching over Public Administration Legal Documents Using Ontologies. In *JCKBSE*, pages 167–175, 2006.
3. PR. Cohen and R. Kjeldsen. Information Retrieval by Constrained Spreading Activation in Semantic Networks. *Inf. Process. Manage.*, 23(4):255–268, 1987.
4. M. Hepp. Possible Ontologies. *IEEE-Internet Computing*, (1):90–96, 2007.
5. J. Leukel and V. Schmitz et al. Exchange Of Catalog Data In B2B Relationships - Analysis And Improvement.
6. JL. Marín. *Web 2.0. Una descripción muy sencilla de los cambios que estamos viviendo.* Netbiblio, 2010.
7. JL. Marín and JE. Labra. Doing Business by selling free services. In P. Ordóñez et al., editor, *Web 2.0: The Business Model*, part 6, pages 89–102. Springer, 2009.
8. B. Omelayenko and D. Fensel. An Analysis of B2B Catalogue Integration Problems. In *Proc. of the International Conference ICEIS-2001*, Setúbal,Portugal, 2001.
9. C. Rocha and D. Schwabe et. al. A Hybrid Approach for Searching in the Semantic Web. In *WWW*, pages 374–383, 2004.
10. JM. Álvarez, E. Rubiera, and L. Polo. Promoting Government Controlled Vocabularies for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System. 2008.