Query Expansion Methods and Performance Evaluation for Reusing Linking Open Data of the European Public Procurement Notices*

Jose María Álvarez[1], José Emilio Labra[1], Ramón Calmeau[2], Ángel Marín[3], and José Luis Marín[3]

[1] WESO Research Group-Universidad de Oviedo
[2] EXIS TI
[3] Gateway Strategic Consultancy Services
josem.alvarez@weso.es, jelabra@weso.es, ramon.calmeau@exis-ti.com, anmar@gateway-scs.es, josmar@gateway-scs.es
WWW home page: http://www.weso.es, http://www.exis-ti.com, http://gateway-scs.es/

Abstract. The aim of this paper is to present some methods to expand user queries and a performance evaluation to retrieve public procurement notices in the e-Procurement sector using semantics and linking open data. Taking into account that public procurement notices contain information variables like type of contract, region, duration, total value, target enterprise, etc. different methods can be applied to expand user queries easing the access to the information and providing a more accurate information retrieval system. Nevertheless expanded user queries can involve an extra-time in the process of retrieving notices. That is why a performance evaluation is outlined to tune up the semantic methods and the generated queries providing a scalable and time-efficient system. On the other hand this system is based on the use of semantic web technologies so it is necessary to model the unstructured information included in public procurement notices (organizations, contracting authorities, contracts awarded, etc.), enrich that information with existing product classification systems and linked data vocabularies and publish the relevant data extracted out of the notices following the linking open data approach. In this new LOD realm these techniques are considered to provide added-value services like search, matchmaking geo-reasoning, or prediction, specially relevant to small and medium enterprises (SMEs).

1 Introduction

In the European e-Procurement context there is an increasing commitment to boost the use of electronic communications and transaction processing by gov-

* This work is part of '10ders Information Services project' (http://rd.10ders.net/) partially funded by the Spanish Ministry of Industry, Commerce and Tourism with code TSI-020100-2010-919, leaded by 'Gateway Strategic Consultancy Services' and developed in cooperation with 'EXIS TI' and WESO Research Group.

ernment institutions and other public sector organizations. The European Commission outlines the following advantages in the wider use of e-Procurement [4]: 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[5] ('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 [6]) but an unified information system pan-European 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.

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 Knowledge Organization Systems (KOS) provide building blocks for an innovative unified pan-European information system for the benefit of SMEs.

This work aims to apply some semantic-based methods to expand user queries in the e-Procurement sector using semantic web technologies and the LOD approach. In this paper a survey of methods is presented to expand the information variables extracted out of the public procurement notices and a performance evaluation of the user queries is also provided to show how the system works. This study is motivated by the following example: *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 "Construction work for bridges and tunnels, shafts and subways" and a two year duration in the Dutch-speaking region of Flanders in Belgium?.*

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 [7] ("Linked Open Tenders Electronic Daily") where they use the RSS feeds of TED. UK government[8] 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 [9], etc. Most of the public administra-

[4] http://ec.europa.eu/internal_market/consultations/docs/2010/e-procurement/green -paper_en.pdf
[5] http://ted.europa.eu/
[6] http://www.eubusiness.com/tenders
[7] http://loted.eu:8081/LOTED1Rep/
[8] http://data.gov.uk
[9] http://nuts.psi.enakting.org/

tions in the different countries are also betting for LOD approach to make public their information: Spain (Aporta project [10]), USA[11], etc. Regarding the use of LOD and organizations there is a new ontology for modeling the information about organizations [12] and recently it has been released "The Open Database Of The Corporate World" [13].

Product Scheme Classifications (also known as PSCs) like the CPV (Common Procurement Vocabulary available at RAMON, the Eurostat's metadata server) have been built to solve specific problems of interoperability and communication in e-commerce[10]. The aim of a PSC is to be used as a standard *de facto* by different agents for information interchange in marketplaces [2]. Any PSC, as well as other classification systems can be interpreted as: 1) domain-ontologies [9] or 2) conceptual schemes [16] comprised of conceptual resources . Finally, Good Relations [14] is an ontology for the e-commerce developed by Martin Hepp et. al.

The use of semantic methods to exploit the data from the semantic web like Spreading Activation (SA) techniques and Rule Based Systems (RBSs) is widely used. The main application of SA techniques is focus on Document and Information Retrieval [7]. These techniques have been also used in semantic search based on hybrid approaches [13,4], 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 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.

Finally the process of expanding queries is widely accepted to reformulate a seed query and improve retrieval performance in information retrieval operations. In most of the cases the process deals with linguistic issues [3] through the use of controlled vocabularies and taxonomies to find synonyms, spelling errors, etc. In the case of e-Procurement a search engine should be able to process the user query and perform a concept based query expansion process like [4] for legal documents.

3 Survey of Methods to Expand User Queries

The selection of methods to expand the information of a user query about public procurement notices depends on some factors: 1) the type of variable: concept from a taxonomy or ontology, a numeric value in a range or geographical information; 2) the intentions of the user by means of creating a search profile (RDF-based and reusing existing LOD and e-Procurement vocabularies) containing the

[10] http://www.aporta.es/
[11] http://www.data.gov/
[12] http://www.epimorphics.com/web/category/category/developers/organization-ontolog y
[13] http://opencorporates.com/
[14] http://www.heppnetz.de/projects/goodrelations/

initial selected values for the information variables presented in the notices and 3) the statistical information available in previous public procurement notices. Taking into account these factors Table 1 shows a comparison of the selected methods to be applied in the process of query expansion. The current situation

Variable	**Type**	**User Intention**	**Statistical Information**	**Method**	**Tool**
CPV and NUTS codes	Concept	Enhance Codes	Correlation among codes	• Syntactic comparison of descriptions • Dividing an initial value into narrower codes • Recommender • SA • Geo-reasoning	• Apache Lucene • Hard-Coding • Apache Mahout • ONTOSPREAD (API Java for SA) • Geonames and GeoSPARQL
• Total Value • Duration and Publishing year • ...	Numeric	Establish a Range	Correlation with historical information	• Numeric range • FuzzyLogic	• Hard-Coding • JFuzzyLogic
Type of company	Enumerate		Correlation with historical information	Get type of companies for the CPV codes, etc.	Hard-Coding

Table 1. Survey of Methods to Expand User Queries in the e-Procurement sector.

of searching public procurement notices consists on the interaction between a business user and a client that wants to tender with a certain set of restrictions on the information variables. However the intentions of the client do not match with the real information in notices that is why the business user must rewrite client restrictions to convert them in a real query that can retrieve the desired notices. These expansion methods are considered like a decision support system to help business user to rewrite user queries. Following the input SPARQL query of the motivating example including a CPV code, a NUTS region (only coordinates) and some numeric values for total value and duration is presented, see

Fig. 1. After the process of query expansion a new SPARQL query[15] is built, see

```
SELECT * WHERE{
  ?notice rdf:type ppn-def:PublicProcurementNotice .
  ?notice dct:identifier ?id .
  ?notice dct:date ?date .
  ?notice dct:description ?description .
  ?notice ppn-def:hasStatus ppn-def:Active .
  ?notice org:classification <http://purl.org/organizations#SME> .
  ?notice wgs84_pos:lat ?lat.
  ?notice wgs84_pos:lon ?long .
  ?notice ppn-def:totalValue ?totalValue.
  ?amount muo:measuredIn <http://purl.org/weso/units/euro> .
  ?notice ppn-def:duration ?duration.
  ?notice ppn-def:nutsCode ?nutsCode .
  ?duration muo:measuredIn <http://purl.org/weso/units/year> .
  ?notice cpv-def:codeIn2008 ?cpvCode .
 FILTER (
 ((?nutsCode = nuts:BE)) and
 ((?cpvCode = cpv:45221000))
 and (?lat == "50.85") and (?long == "43.49")
 and (?totalValue <= 170,000^xsd:double) and (?duration <= 2) )}
```

Fig. 1. Simple SPARQL query.

Fig. 2. The process of expansion selects new CPV codes (45221100-"Construction work for bridges", 45221110-"Bridge construction work", 45221111-"Road bridge construction work", 45221113-"Footbridge construction work"), new NUTS codes (spreading the geographical scope) and establish a range for the numeric variables according to the historical information available at the database. Currently we are finishing the process of publishing the PSCs and the information extracted from public procurement notices as linked data. Moreover an information retrieval system[16] (implemented using Java technologies) is available to test the process of expansion. On the other hand the result set is sorted according to a rank function. This point is ongoing research due to the fact that a lot of OWA operators [8] and Entity Ranking Functions [12] are available.

4 Performance Evaluation

The evaluation of the system can be carried out from two different points of view: 1) With regards to the validation of the goodness and the improvement

[15] The URI prefixes of this example come from the "Prefix.cc" service.
[16] MOLDEAS-http://moldeas-web.appspot.com

```
SELECT * WHERE{
  ...
  ?notice nuts:containedBy ?place .
FILTER (  ( (?cpvCode =  cpv:45221000) or
            (?cpvCode =  cpv:45221110) or
            (?cpvCode =  cpv:45221111)...)
          ( (?place nuts:containedBy nuts:NL326 ) or
            (?place nuts:containedBy nuts:B3) or
            (?place nuts:containedBy nuts:BE2) or ...)
        and (?duration >= 2 and ?duration <= 4)
        and (?date >= 2008 and ?date <= 2011)
        and (?totalValue > 130,000^xsd:double
             and ?totalValue <= 200,000^xsd:double))}
```

Fig. 2. Expanded SPARQL query.

of the proposed system we have identified, apart from selecting a service to be tested, three main variables: a) the amount of information used; b) the number of tests (execution of prepared user queries) that should be carried out to assess a correct precision and recall of the proposed retrieval system and c) the best combination of expansion methods. From the first variable point of view 1M public procurement notices (provided by Gateway SCS-Eurolert.net[17]) and over 320K organizations [18] are available. On the second one, we have not decided yet how many tests would be appropriate 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. The expected result of this evaluation supposes the first step to validate our approach and select the best combination of expansion methods to improve the access and retrieval of the information about public procurement notices using the LOD approach.

2) In the case of performance evaluation the first tests showed us that the execution of expanded queries involved an extra-time to execute them via SPARQL. Checking existing works in SPARQL optimizations [14,1,6] and efficient querying of triple stores [15] led us to re-think the process of building expanded user queries trying to improve the execution times. In next section the design of the experiment, the steps to accomplish an improvement in the execution of the SPARQL queries and the results of the tests are presented.

4.1 Design of the experiment

In Sect. 3 the methods to expand an user query were presented to show how the systems works to generate enhanced queries using the information variables

[17] http://euroalert.net/
[18] ftp://ftp.ted.europa.eu/META-XML/

of the public procurement notices. In this experiment [19] we will focus on the next variables: CPV and NUTS codes and the publishing year. The CPV is a taxonomy in which concepts are grouped by a category and identified by a code that indicates their category: "Division" e.g. 01000000 , "Group" e.g. 01100000, "Class" e.g. 01110000, "Category" e.g. 01112000, 01112200, 1112210 or 01112211. On the other hand NUTS is the "Nomenclature of Territorial Units for Statistics" established by Eurostat in order to provide a single uniform breakdown of territorial units. Each code begins with a two-letter code referencing the country, which is identical to the ISO 3166-1 alpha-2 and for each EU member country three levels of NUTS codes are established. Finally, the publishing year is just a number indicating when the notice was published. Taking into account the description of these variables the proposed methodology is the next one:

Test/ Feature	F_1	F_2	F_3	F_4	F_5	F_6	F_7
T_1	$\star$						
T_2	$\star$		$\star$				
T_3		$\star$					
T_4		$\star$	$\star$				
T_5		$\star$	$\star$	$\star$			
T_6^1 (n CPV codes and m NUTS codes)		$\star$	$\star$	$\star$	$\star$	$\star$	
T_6^2 ($\equiv$)		$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
T_7^1 (1 CPV code and m NUTS codes)		$\star$	$\star$	$\star$		$\star$	
T_7^2 ($\equiv$)		$\star$	$\star$	$\star$		$\star$	$\star$
T_8^1 ($\equiv$)		$\star$	$\star$	$\star$	$\star$	$\star$	
T_8^2 ($\equiv$)		$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
T_9^1 (1 CPV code and 1 NUTS code)		$\star$	$\star$	$\star$		$\star$	
T_9^2 ($\equiv$)		$\star$	$\star$	$\star$		$\star$	$\star$
T_{10}^1 ($\equiv$)		$\star$	$\star$	$\star$	$\star$	$\star$	
T_{10}^2 ($\equiv$)		$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 2. Description of the tests and optimization features.

- Select the initial CPV codes (with different categories) to build simple and expanded queries.
- Select the initial NUTS codes.

[19] The complete description of the experiment including all tables of selected queries and execution times is available at: http://purl.org/weso/moldeas/papers/caepia2011.pdf.

- Establish the publishing years according to the data in the triple store. Currently, public procurement notices from 2008 to 2011 are stored in the database and grouped by the publishing year using named graphs [20].
- Determine the software and hardware environment.
- Select the datasets stored in the database (e.g. CPV-10K concepts, NUTS-8K codes and Public Procurement Notices-1M of notices altogether about 9 million of triples).
- Build and execute via SPARQL simple and expanded queries with the selected information applying the query expansion methods.
- Combine the different SPARQL and algorithm optimizations, see Table 2.
- Log the execution times and establish the number of replies (e.g. 3) to perform the tests.

According to these steps, all queries (9) use a range between 2008 and 2011 for the publishing year. The software environment is comprised of a Virtual Box (version 4.0.6) virtual machine (Linux 2.6.35-22-server #33-Ubuntu 2 SMP x86_64 GNU/Linux Ubuntu 10.10, 2GB RAM and 30GB HardDisk) in which a Open Link Virtuoso [21] instance (version 06.01.3127) is installed. The virtual machine is hosted in a DELL PC (same configuration as virtual machine) and a regular internet connection is used to execute the queries.

After that it is necessary to define the possible optimizations ("description"-*ID*) that will configure the features of the tests as Table 2 shows. In this case a distinction between "simple queries"-F_1. (1 CPV code) and "enhanced queries"-F_2. (n CPV codes) should be made. Besides there is a list of SPARQL optimizations that can be applied: "LIMIT clause" (value fixed to 10000)-F_3, "Rewrite SPARQL queries" (following the aforementioned works and making the matching and filtering of the triples from the most specific to general)-F_4, "Use of named graphs"-F_5, "Split enhanced queries into simple queries"-F_6 and "Use of an ad-hoc implementation of the Map/Reduce algorithm by Google, with 5 threads to perform the map function and 1 thread to reduce the results of the queries"-F_7. Taking into account these features the different tests are performed in 3 replies using the arithmetic mean to aggregate the execution times.

4.2 Results and Discussion

During the tests about 5751 SPARQL queries have been performed in order to retrieve the data and the execution times of the queries. These results are processed using "bash" scripts that extract out the statistics and generate a spreadsheet. Regarding the comparison of results, the calculation of the gain $(t_{old}/t_{new} - 1 * 100)$ is tackled in two ways depending on the kind of query (simple or enhanced): 1) comparison of test T_1 with T_2 and 2) comparison of test T_3 with tests $T_4 ... T_{10}^2$.

Results show there is no sensible gain when some optimizations are put in action like F_3, F_4 and F_5. In the case of F_3 the use of the "LIMIT clause" fixed

[20] E.g: http://purl.org/weso/ppn/2008
[21] http://virtuoso.openlinksw.com

to (10000) is not representative due to the results of the triple matching process are previously filtered. "Rewriting queries" F_4 usually involves an improvement in the execution time but maybe the information variables used in these queries does not allow minimize the target dataset while the triple matching process is being ran. Also when "named graphs" F_5 are used the execution time of a single query is obviously lower than one query over all public procurement notices dataset but the number queries to be performed is higher implying a slower execution time. On the other hand, the optimizations F_6 and F_7 bring strong improvements in the execution times of tests T_6^2, T_7^1, T_7^2 and T_9^2. Nevertheless tests T_{10}^1 and T_{10}^2 do not get an improvement in execution time due to the fact that the addition of some features does not guarantee a real gain. One of the highlighted outcomes of this study lies on the detection that the number of CPV codes[22] in a query is related to the execution time (it is about 3 sec. per code) thus the use of only one CPV code in a query improves the process of retrieving public procurement notices. Finally, the use of distributed algorithms is widely accepted and proven when scalability problems appear. In conclusion, taking into account these results the best configuration to improve the execution time of expanded queries lies on splitting them into simple queries (only one CPV code) and use distributed algorithms like Map/Reduce. Nevertheless, other actions in the scope of hardware configuration, caching results [5], use of information variables in the queries with more entropy, etc. could improve the behavior of the system.

5 Conclusions and Future Work

The implementation of these expansion methods 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 [11] and we are also interested in reporting 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. On the other hand, the performance evaluation allows us to identify bottlenecks and test the current system at different levels. Now we have used the execution time as target variable to be improved. Nevertheless if the retrieval system is supposed to work off-line (like an alert system of public procurement notices) the execution time should not

[22] Adding or removing NUTS codes does not almost change the execution time.

be a key-factor to deploy a semantic-based platform for e-Procurement in a production environment that takes advantage of the semantic web technologies and the LOD approach. Finally, we are willingness to check the possibility of using other triple-stores and to perform more load, stress, endurance or usability tests following the evaluation points of view presented in Sect. 4.

References

1. Bernstein A., Kiefer C., and Stocker M. Optarq: A sparql optimization approach based on triple pattern selectivity estimation. Technical report, University of Zurich, Department of, 2007.
2. 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.
3. Abdelghani Bellaachia and Ghita Amor-Tijani. Enhanced query expansion in english-arabic clir. In *Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, pages 61–66, Washington, DC, USA, 2008. IEEE Computer Society.
4. D. Berrueta, JE. Labra, and L. Polo. Searching over Public Administration Legal Documents Using Ontologies. In *JCKBSE*, pages 167–175, 2006.
5. R. Blanco, E. Bortnikov, F. Junqueira, R Lempel, L. Telloli, and H. Zaragoza. Caching search engine results over incremental indices. In *Proceeding of the 33rd ACM SIGIR'10 Conference*, SIGIR '10, pages 82–89, New York, NY, USA, 2010. ACM.
6. Arenas M. Buil-Aranda C. and O. Corcho. Semantics and optimization of the sparql 1.1 federation extension. In *Proc. of the 8th ESWC'11*, ESWC'11, 2011.
7. PR. Cohen and R. Kjeldsen. Information Retrieval by Constrained Spreading Activation in Semantic Networks. *Inf. Process. Manage.*, 23(4):255–268, 1987.
8. A. Emrouznejad and GR. Amin. Document similarity: a new measure using owa. In *Proc. of the 6th FSKD*, FSKD'09, pages 186–190, Piscataway, NJ, USA, 2009.
9. M. Hepp. Possible Ontologies. *IEEE-Internet Computing*, (1):90–96, 2007.
10. J. Leukel and V. Schmitz et al. Exchange Of Catalog Data In B2B Relationships - Analysis And Improvement.
11. 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.
12. J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th WWW*, WWW '10, pages 771–780, New York, NY, USA, 2010. ACM.
13. C. Rocha and D. Schwabe et. al. A Hybrid Approach for Searching in the Semantic Web. In *WWW*, pages 374–383, 2004.
14. M. Schmidt, M. Meier, and G. Lausen. Foundations of sparql query optimization. In *Proceedings of the 13th International Conference on Database Theory*, ICDT '10, pages 4–33, New York, NY, USA, 2010. ACM.
15. Y. Yan, C. Wang, and A. Zhou. Efficiently querying rdf data in triple stores. In *Proceeding of the 17th WWW*, WWW '08, pages 1053–1054, New York, NY, USA, 2008. ACM.
16. 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.