Computers in Human Behavior 30 (2014) 674–688

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Empowering the access to public procurement opportunities by means of linking controlled vocabularies. A case study of Product Scheme Classifications in the European e-Procurement sector

Jose María Alvarez-Rodríguez [a,*], José Emilio Labra-Gayo [b], Alejandro Rodríguez-González [c], Patricia Ordoñez De Pablos [d]

[a] The South East European Research Center, Thessaloniki, Greece
[b] WESO Research Group, Department of Computer Science, University of Oviedo, 33007 Oviedo, Spain
[c] Bioinformatics at Centre for Plant Biotechnology and Genomics UPM-INIA, Polytechnic University of Madrid, Madrid, Spain
[d] WESO Research Group, Department of Business Administration, University of Oviedo, 33007 Oviedo, Spain

ARTICLE INFO

Article history:
Available online 24 August 2013

Keywords:
e-Procurement
Product Scheme Classifications
Linked open data
Semantic web
Expert systems

ABSTRACT

The present paper introduces a method to promote existing controlled vocabularies to the Linked Data initiative. A common data model and an enclosed conversion method for knowledge organization systems based on semantic web technologies and vocabularies such as SKOS are presented. This method is applied to well-known taxonomies and controlled vocabularies in the business sector, more specifically to Product Scheme Classifications created by governmental institutions such as the European Union or the United Nations. Since these product schemes are available in a common and shared data model, the needs of the European e-Procurement sector are outlined to finally demonstrate how Linked Data can address some of the challenges for publishing and retrieving information resources. As a consequence, two experiments are also provided in order to validate the gain, in terms of expressivity, and the exploitation of this emerging approach to help both expert and end-users to make decisions on the selection of descriptors for public procurement notices.

1. Introduction

Government bodies and public institutions as a whole are the most important buyers in the European Union (EU), since public procurement spending represents around 19% of EU Gross Domestic Product (GDP) (European Commission, 2010b). Given this situation there is a growing interest and commitment (European Commission, 2010a) to ensure that these funds are well managed and most of inefficiencies are eliminated. Electronic public procurement or e-Procurement (Podlogar, 2007) emerges as an alternative to link and integrate inter-organizational business processes and systems with the automation of the requisitioning, the approval purchase order management, accounting processes among others through the Internet-based protocol.

In this context the European Commission (EC) is trying to unlock this potential, the 2004 Public Procurement Directives 2004/17/EC and 2004/18/EC introduced several provisions and projects (European Commission, 2013c; European Commission, 2013b) aimed at enabling e-Procurement uptake in all Member States. In this light of modernizing the European public procurement sector to support growth and employment the EC also identified (European Commission Internal Market Directorate-General, 2010) both regulatory and natural barriers in the access to public procurement in the EU context, especially for SMEs (European Commission, 2008). According to this evaluation, a real European Single Market (European Commission, 2011) has not yet been achieved causing losses, being cost-inefficient, missing opportunities for society and leading to a situation where more than 27 national markets co-exist instead of an EU-wide public market (Monti, 2010). In this sense, one relevant action to ease the interconnectivity and interoperability in this landscape was the creation of the Tenders Electronic Daily (European Commission, 2013a; European Commission, 2013d) (TED) by the EC. It is the on-line version of the "Supplement to the Official Journal of the European Union", dedicated to European public procurement notices (1500 new announcements every day). Taking into account that the type of contract and the geographical information are two of the main variables that serve to be aware of new business opportunities, the EU has established the use of the "Common Procurement Vocabulary" (hereafter CVP refers to CPV 2008) and

* Corresponding author.
E-mail addresses: jmalvarez@seerc.org (J.M. Alvarez-Rodríguez), labra@uniovi.es (J.E. Labra-Gayo), alejandro.rodriguezg@upm.es (A. Rodríguez-González), patrio-p@uniovi.es (P.O. De Pablos).
URL: http://www.seerc.org (J.M. Alvarez-Rodríguez).

0747-5632/$ - see front matter
http://dx.doi.org/10.1016/j.chb.2013.07.046

the "Nomenclature of territorial units for statistics" (NUTS) as mandatory to annotate public procurement notices. Nevertheless depending on the country and the scope of use distinct classifications have been also created. In the specific case of e-Procurement, the "United Nations Standard Products and Services Code" (UNSPSC) in Australia, the "North American Industry Classification System" (NAICS) in United States or the "Integrated Tariff of the European Communities" (TARIC) in the EU are examples of similar efforts to unify and model procurement-related data. As a consequence a real, standardized and integrated environment for e-Procurement (and business) data to encourage the creation of knowledge-based services cannot be easily deployed. That is why some of the potential advantages outlined by the EC (European Commission, 2010b) such as increased accessibility and transparency, benefits for individual procedures, benefits in terms of more efficient procurement administration and potential for integration of EU procurement markets and systems cannot be reached in a short-term.

Additionally public bodies are also defining knowledge organization systems (KOS), such as the classifications in the Eurostat's metadata server (RAMON), to enable users to annotate information objects providing an agile mechanism for performing tasks such as exploration, searching, automatic classification or reasoning. The structure and features of these systems are usually very heterogeneous, although some common aspects can be found in all of them: hierarchical relationships between terms or concepts and multilingual character of the information. However the lack of mappings among them makes almost impossible a proper reuse of these systems out of their scope. As motivating and explanatory examples of the necessity of linking knowledge organization systems in the e-Procurement context, four common situations are presented below:

- An American SME that usually manages the NAICS vocabulary for being aware of public procurement notices in United States is looking for new business opportunities in Europe. Although they perfectly know that the CPV is used in European public procurement notices they are not able to retrieve any announcement due to the fact there is not mapping between the CPV and NAICS classification systems.
- A Spanish family-owned company is also looking for new business opportunities in Bulgaria. They have been informed that the national government is requiring providers for low-value procurement opportunities that match their services but they are finding three main issues: (1) Bulgarian civil servants annotate these procurement opportunities with different codes as Spanish ones do; (2) these notices are not published in TED (because of their value) and (3) the language is a simple barrier they cannot overcome.
- A department of statistics in some international organism, such as the WorldBank, is gathering data and information from different public institutions and scopes to compare some economical indicators. Due to the fact that data and information are managed using distinct knowledge organization systems, a simple task of integrating this information to create a report is becoming a major challenge.
- A company has implemented a commercial alert service that enables the possibility of getting latest public procurement opportunities according to a customer profile. Although business users perfectly know how contracts are described using some classification scheme, sometimes they are not able to translate a client query to contract descriptors. If a service could suggest these descriptors exploiting some kind of knowledge-base, business users decisions would be supported by a tool, the alert service would be more accurate and customer satisfaction would be increased.

On the other hand and following the principles of the Open Data initiative, the vice-president Neelie Kroes is leading the Open Data Strategy (European Parliament & European Council, 2003) for Europe. The strategy is strongly focused on the commercial value of the reuse of Public Sector Information (PSI). In this case e-Procurement data is considered to be a key-enabler of the new data-based economy enabling a greater transparency; delivers more efficient public services; and encourages greater public and commercial use. Therefore an enriched version of public procurement data to ease the access to business opportunities can be the first action to encourage SME participation and create a real and competitive public market of procurement opportunities.

Moreover, the emerging Web of Data and the sheer mass of information now available make it possible the deployment of new added-value services and applications based on the reuse of existing vocabularies and datasets. The popular diagram of the Linked Data Cloud (Cyganiak & Jentzsch, 2011), generated from metadata extracted from the Comprehensive Knowledge Archive Network (CKAN) out, contains 336 datasets, with more than 25 billion RDF triples and 395 million links. With regards to e-Procurement, a new group of 130 CKAN datasets have been released from the "OpenSpending.org" site. In this realm, data coming from different sources and domains have been promoted following the principles of the Linked Data initiative (Berners-Lee, 2006) to improve the access to large documental databases, e.g. e-Government resources, scientific publications or e-Health records. In the case of KOS, such as thesauri, taxonomies or classification systems are developed by specific communities and institutions in order to organize huge collections of information objects. These vocabularies allow users to annotate (Leukel, 2004; Leukel, 2004; Leukel & Maniatopoulos, 2005) the information objects and easily retrieve them, promoting lightweight reasoning in the Semantic Web. Topic or subject indexing is an easy way to introduce machine-readable metadata for a resource's content description. Indeed, Product Scheme Classifications (also known as PSCs), such as the CPV, are a kind of KOS that have been built to solve specific problems of interoperability and communication between e-Commerce agents (Omelayenko & Fensel, 2001; Schmitz & Leukel, 2005). In fact PSCs are considered to be a key-enabler of the next European e-Procurement domain and other supply chain processes (Alor-Hernández, Gómez Berbís, & Rodríguez González, 2010).

Obviously the public information published by governmental contracting authorities, more specifically PSCs, are a suitable candidate to apply the Linked (Open) Data (LOD) principles and semantic web technologies providing a new data environment. The conjunction of these initiatives provides the adequate building blocks for an innovative and unified pan-European information system that encourages standardization of key processes and systems. In fact economic operators can take advantage of this approach to overcome some technical interoperability issues and make a better reuse of public sector information. According to the aforementioned points main contributions of this paper can be summarized as follows:

- Modeling, promotion and inter-linkage of PSCs, structure and data, following the LOD principles.
- Publication of all information via a SPARQL endpoint providing a public procurement Linked Data node and, more specifically, a new PSCs catalogue as Linked Data.
- Exploitation of the PSCS catalogue and an existing dataset of public procurement notices via a contract descriptor recommendation service.
- Demonstration of the gain (in terms of number of descriptors to retrieve public procurement notices) and the semantic exploitation capabilities of the new PSCs catalogue.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3.1 briefly presents the MOLDEAS lifecycle for promoting raw data to the LOD initiative. Afterwards next section reviews existing methods to code PSCs as RDF/OWL to finally apply the MOLDEAS lifecycle. Next section describes the evaluation of applying the LOD principles to the product schemes through two studies that include a description of the sample, the method, results and discussion. Finally, the paper ends with a discussion of research findings, limitations and some concluding remarks.

2. Literature review

According to the previous section, some relevant works can be found and grouped by the topics covered in this paper:

- The Semantic Web area, coined by Tim Berners-Lee in 2001, has experienced during last years a growing commitment from both academia and industrial areas with the objective of elevating the meaning of web information resources through a common and shared data model (graphs) and an underlying semantics based on different logic formalisms (ontologies). The Resource Description Framework (RDF), based on a graph model, and the Web Ontology Language (OWL), designed to formalize and model domain knowledge, are the two main *ingredients* to reuse information and data in a knowledge-based realm. Thus data, information and knowledge can be easily shared, exchanged and linked (Maali & Cyganiak, 2011) to other knowledge-based systems and databases through the use URIs, more specifically HTTP-URIs. Therefore the broad objective of this effort can be summarized as a new environment of added-value information and services that can encourage and improve B2B (Business to Business), B2C (Business to Client) or A2A (Administration to Administration) relationships. The implementation of new context-awareness expert systems to tackle existing cross-domain problems such as medical reasoning, analysis of social media, etc. in which data heterogeneities, lack of standard knowledge representation and interoperability problems are common factors that the use of semantics can improve. As a practical view of the Semantic Web, the Linked Data (Berners-Lee, 2006; Heath & Bizer, 2011) initiative emerges to create a large and distributed database on the Web. In order to reach this major objective the publication of information and data under a common data model (RDF) with a specific formal query language (SPARQL (Prud'hommeaux & Seaborne, 2007)) provides the required building blocks to turn the Web of documents into a real database or "Web of Data". Research works are focused in two main areas: (1) production/publishing (Bizer, Cyganiak, & Heath, 2007) and (2) consumption of Linked Data. In the first case data quality (Bizer, 2007; Bizer & Cyganiak, 2009; Alvaro Graves Olye Rickson, 2011; Stadler, Guéret, Lehmann, Groth, & Jentzsch, 2011), conformance (Hogan et al., 2012), provenance (Moreau & Missier, 2011; Hartig & Zhao, 2010), trust (Carroll, Bizer, Hayes, & Stickler, 2005), description of datasets (Alexander, Cyganiak, Hausenblas, & Zhao, 2011; Cyganiak, Stenzhorn, Delbru, Decker, & Tummarello, 2008;TaskForce Linking Open Data DataSets, 2011) and entity reconciliation (Araujo, Hidders, Schwabe, & De Vries, 2011; Maali & Cyganiak, 2011) issues are becoming major objectives since a mass of amount data is already available through RDF repositories and SPARQL endpoints.
  On the other hand, consumption of Linked Data is being addressed to provide new ways of data visualization (Dadzie & Rowe, 2011; Hogan et al., 2011), faceted browsing (Pietriga, Bizer, Karger, & Lee, 2006; Tummarello et al., 2010), searching (Hogan et al., 2011) and data exploitation (Harth, Hogan, Kotoulas, & Urbani, 2011). Some approaches based on sensors (Jeung et al., 2010; Calbimonte, Jeung, Corcho, & Aberer, 2011), distributed queries (Hartig, Bizer, & Freytag, 2009; Ankolekar, Krötzsch, Tran, & Vrandecic, 2007; Buil-Aranda, Arenas, & Corcho, 2011), scalable reasoning processes (Urbani, Kotoulas, Maassen, van Harmelen, & Bal, 2012; Bonatti, Hogan, Polleres, & Sauro, 2011), annotation of web pages (Adida & Birbeck, 2008) or information retrieval (Pound, Mika, & Zaragoza, 2010) are key-enablers for easing the access to information and data.
- In the particular case of e-Procurement there are projects trying to exploit the information of public procurement notices like the "Linked Open Tenders Electronic Daily" project (Di Noia, Valle, & Motta, 2010) where they use the RSS feeds of TED. In the European project LOD2 (LOD2 European Project FP7, 2010), there is a specific workpackage, WP9a "LOD2 for A Distributed Marketplace for Public Sector Contracts", to explore and demonstrate the application of Linked Data principles for procuring contracts in the public sector and, the Media Lab research group at the Technical University of Athens has recently published the "PublicSpending.net" (Meimaris et al., 2013) portal to visualize and manage statistics about public spending around the world. Finally the "OpenSpending.org" (Open Knowledge Foundation, 2013) portal also presents some specifications to model public procurement data and visualize, where the money goes. In general, the "LOD Around-the-Clock" (LATC Project FP7, 2013) (LATC) and the PlanetData (PlanetData European Project FP7, 2013) projects are also increasing the awareness of LOD across Europe delivering specific research and dissemination activities such as the "European Data Forum". Furthermore legal aspects of public sector information are being reviewed in the LAPSI project (LAPSI Project FP7, 2013). In the case of vocabularies and datasets, GoodRelations and ProductOntology are two of the most prominent approaches for tagging products and services using semantic web technologies, for instance Renault UK uses GoodRelations in RDFa in their UK merchandise store. Although these previous efforts to apply Semantic Web principles to procurement data, a system to link existing PSCs to enable a better consumption of public procurement notices and interoperability among e-Procurement systems is still missing.

Since an overview of semantic technologies and Linked Data and their application to e-Procurement has been presented some remarks and discussion about related works are outlined below.

1. On the one hand, entity reconciliation is becoming a major challenge in the Linked Data community due to its relevance to enrich data with existing datasets. Existing techniques are mainly based on natural language processing (NLP) algorithms that perform some kind of string comparison (labels (Araujo et al., 2011) or URIs (Maali & Cyganiak, 2011)) to establish a similarity value between two RDF resources under a particular threshold of confidence. The Silk framework (Jentzsch, Isele, & Bizer, 2010) provides an API for discovering relationships between data items within different Linked Data sources. In this sense the DBPedia Spotlight (Mendes, Jakob, García-Silva, & Bizer, 2011) gives also a "tool for automatically annotating mentions of DBpedia resources in text, a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia". Other approaches coming from the ontology mapping and alignment areas try to create links according to the structure (relationships) and naming convention of the RDF resources. Finally some machine learning techniques such as genetic programming (Isele & Bizer, 2011) are emerging to learn linkage rules from existing datasets. As conclusion, the inter-linkage of RDF resources is consider to be a

key-enabler for an enriched data consumption. Nevertheless the main drawback of these approaches lies in the necessity of human validation to ensure the quality of the link. Furthermore these tools have been designed with a general purpose (parameters such as stopwords cannot be easily configured) and based on the assumption that resources are already available in RDF. That is why it is necessary to provide a custom PSC reconciliation service that takes into account the specific characteristics of PSCs descriptors.

2. On the other hand there is an increasing interest in the creation of methodologies, best practices/recipes (W3C Members, Hyland, Terrazas, & Capadisli, 2011; Hyland et al., 2011) and lifecycles (Hausenblas, Villazón-Terrazas, & Hyland, 2011; Ten-Force, 2011) in the Linked Data community. In this sense, some Linked Data design considerations can be found in Bizer et al. (2007) covering from the design or URIs (Sauermann, Cyganiak, & Völkel, 2007; Berners-Lee, 1998; CTO, 2010), design patterns (Dodds & Davis, 2011), publication of RDF datasets and vocabularies (Berrueta, Phipps, Miles, Baker, & Swick, 2008), etc. to the establishment of Linked Data profiles (Arwe, Nally, Speicher, & Le Hors, 2012). Nevertheless all these guidelines present a tangled environment of aspects with different levels of abstraction that prevent a clear application to a specific problem such as the promotion of PSCs to the Linked Data initiative.
3. Finally information retrieval and recommending techniques have been widely studied to tackle existing problems of searching and filtering information according to an user profile in different domains (García, Álvarez Rodríguez, Gayo, & Merino, 2013). In the particular case of Spreading Activation (Collins & Loftus, 1975; Preece, 1981) (SA) it has been demonstrated its application to extract correlations between query terms and documents analyzing user logs (Cui, Wen, Nie, & Ma, 2003) and to retrieve resources amongst multiple systems (Schumacher, Sintek, & Sauermann, 2008) in which ontologies are used to link and annotate resources. In conjunction with ontologies some works have studied this technique to explore concepts addressing two important issues: (1) the selection and (2) the ranking of additional search terms and to measure conceptual similarity (van Rooyen, Gouws, & Engelbrecht, 2010). On the other hand, data mining, more specifically mining socio-semantic networks (Troussov, Judge, & Botvich, 2008), and applications to collaborative filtering (community detection based on tag recommendations, expertise location, etc.) are other potential scenarios to apply the SA theory due to its high performance and scalability. In particular, annotation and tagging (Gayo, de Pablos, & Lovelle, 2010; Álvarez Rodríguez, Gayo, & de Pablos, 2012) services to gather metadata (Gelgi, Vadrevu, & Davulcu, 2005) from the Web or to predict social annotation (Chen, Chen, & Huang, 2007) and recommending systems based on the combination of ontologies and SA (Gao, Yan, & Liu, 2008) are taken advantage of using this technique. Besides semantic search is a relevant area to apply SA following hybrid approaches (Berrueta, Gayo, & Polo, 2006; Rocha, Schwabe, & Aragao, 2004) or user/concept query expansion (Nie, 2003) combining metadata and user information.

3. Promoting Product Scheme Classifications to the Linked Data initiative

3.1. Summary of the MOLDEAS lifecycle

The aforementioned approaches to define Linked Data lifecycles, see Section 2, are based on performing some processes, recipes, methods or tasks using different tools to promote raw data as RDF triples. However, a formal definition of processes, methods and tasks is missing and, in most of the cases, they are based on the author's expertise. The main consequence of this offhand mixing of approaches is the lack of a quantifiable method to measure the quality of the generated RDF. That is why we have selected and applied the lifecycle proposed in the MOLDEAS project (Álvarez Rodríguez et al., 2012) that perfectly defines which the steps to produce, publish, consume and validate linked open data are. Fig. 1 summarizes several tasks to be carried out in order to promote a controlled vocabulary to the Linked Data effort. The formal definition of processes, methods and tasks in this lifecycle enables separation of concerns and the sustainable management of Linked Data. In the e-Government area this issue must be addressed due to the implicit responsibility of the public administrations of delivering high-quality services and data.

3.2. Knowledge organization system conversion methods

This section specifically discusses existing methods to convert knowledge organizations systems distinguishing between RDF/OWL conversions methods for thesauri and product classification systems.

3.2.1. Thesauri conversion methods

A thesaurus is a controlled vocabulary, with equivalent terms explicitly identified and with ambiguous words or phrases (e.g. homographs) made unique. This set of terms also may include broader-narrower or other relationships. Usually they are considered to be the most complex of controlled vocabularies. Thesauri as a KOS system can be converted by means of different procedures. On the one hand, there are methods (Soergel, 2005) that propose specific techniques for thesauri conversions into ontology. However the method does not target a specific output format and it considers the hierarchical structure of thesauri as logical `is-a` relationships. On the other hand, there are some generic methods for thesauri conversions, as the step-wise method defined in van Assem, Malaisé, Miles, and Schreiber (2006). This method selects a common output data model, the SKOS vocabulary, and is comprised of the following sequenced steps: generation of the RDF encoding, error checking and validation and publishing the RDF triples on the Web. In addition, this method has been refined in Polo and María Alvarez (2008) adding three new sub steps for generating the RDF encoding: analyzing the vocabulary, mapping the vocabulary to SKOS properties and classes and building a conversion program. This initial stepwise method can be considered as a previous effort to the aforementioned Linked Data lifecycles in which all tasks are already defined and these steps are embedded as part of the whole lifecycle.

3.2.2. Product scheme classification conversion methods

Product Scheme Classifications (PSCs) have been developed to organize the marketplace (Leukel, 2004; Leukel & Maniatopoulos, 2005) in distinct vertical sectors that reflect the activity (or some activities) of economy and commerce. They have been built to solve specific problems of interoperability and communication in e-commerce (Schmitz & Leukel, 2005) providing a structural organization of different kind of products collected together by some economic criteria. The aim of a PSC is to be used as a standard (Leukel, 2004) de facto by different agents for information interchange in marketplaces (Omelayenko & Fensel, 2001; Fensel, Ding, & Omelayenko, 2011). Many approaches for product classification systems adaptation to the Semantic Web, like (Lonsdale, Embley, Ding, Xu, & Hepp, 2010), present methods with the goal to convert them to domain-ontologies. The tree-based structure between product terms is then interpreted as a logical `is-a` hierarchy. From our point of view and following the discussion in Hepp (2007); Hepp (2006), hierarchical links between the elements of each economic sector have not the semantics of subsumption relationships. The next example taken directly from

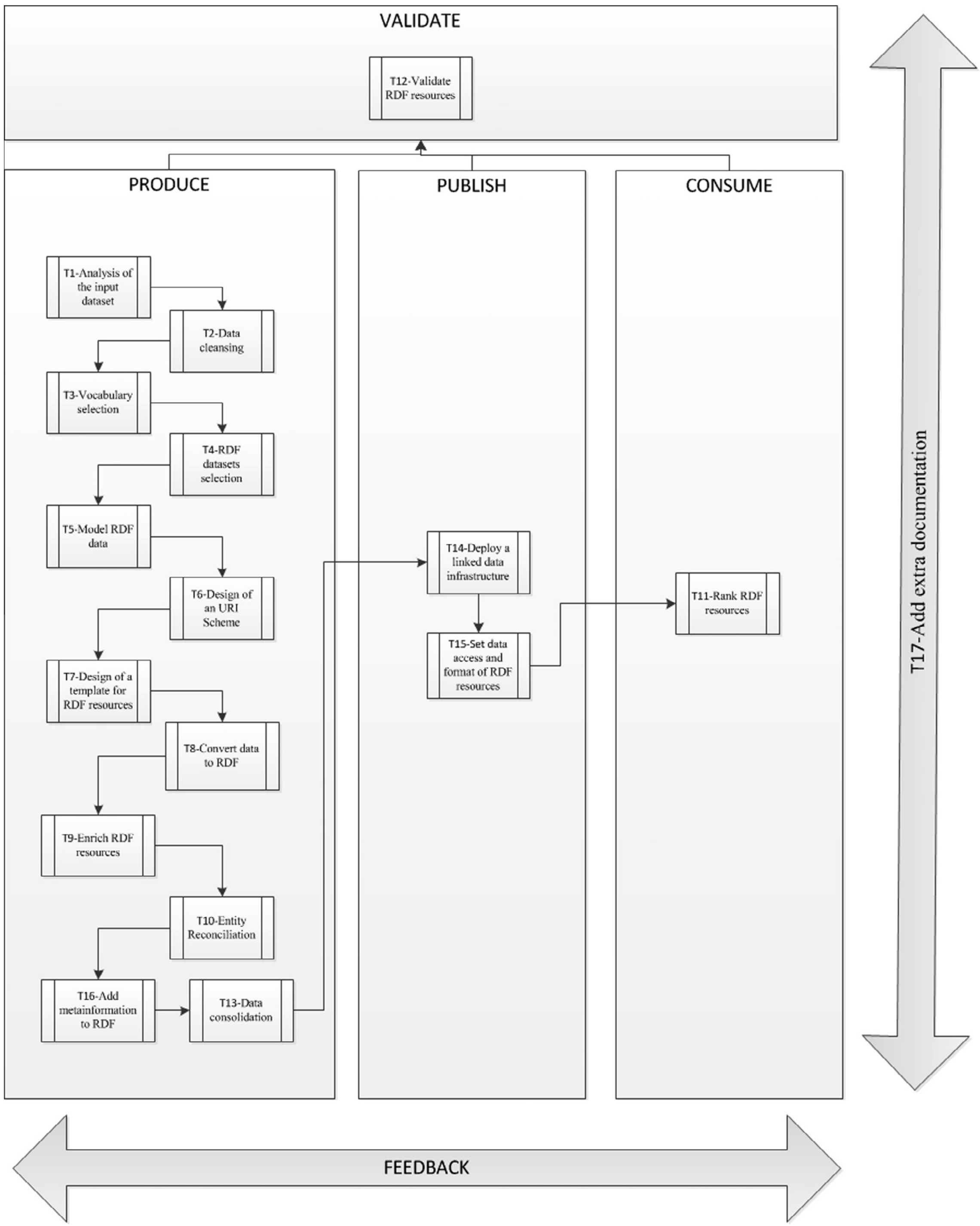


Fig. 1. Summary of tasks of the MOLDEAS Linked Data lifecycle.

the CPV 2008 shows how the relationship between the element "Parts and accessories for bicycles" (34432000-4) and its direct predecessor, "Bicycles" (34430000-0), does not seem an `is-a` relation. In this case, an ontological property for object composition like `hasPart` would be much better. Moreover, there are further remarks against the idea of using the PSCs as domain-ontologies. It is difficult to assert that the CPV element, "Bars, rods, wire and profiles used in construction" (44330000-2), represents any specific product. Rather it should be regarded as a collection of products. To convert correctly this element into a domain ontology concept, it should be considered as equivalent to the union of several concepts (e.g. $Bar \cup Rod \cup Wire \cup Profile$).

Our approach instead will not consider PCSs as domain ontologies, but a specific kind of knowledge organization system. Any PSC is interpreted as a conceptual scheme comprised of conceptual resources. Thus, hierarchical relationships are not considered to be any more logical `is-a` relations, but `broader/narrower` ones.

3.3. Least common multiple of controlled vocabularies

Knowledge organizing systems are used for managing large collections of information objects and efficient information retrieval. Existing controlled vocabularies are currently available in several formats: PDF, MSExcel, CSV or XML. However promoting them to the Semantic Web is not a merely process of RDF/OWL conversions. Transformations need to fulfill some requirements. Firstly, a common RDF/OWL representation is required to ensure: (a) the semantic compatibility between different vocabularies, (b) the processing of vocabularies in a standard way and (c) the sharing of vocabularies for third-parties adoption. In this sense, the W3C vocabulary "Simple Knowledge Organization System" (SKOS) has been especially designed to fully support both structural and lexical features of knowledge organization systems and must be necessarily used in this context. Secondly, although controlled vocabularies have some specific non-shared features, in practice a distinction between them is very hard to draw. We have identified a minimum set of common features for them. The data model should be expressive enough to preserve as much as possible the original semantics of primary sources for these common features. Thirdly, a generic method is needed to ensure quality of data conversions to correct SKOS instances. That is why the MOLDEAS lifecycle has been selected to perform the promotion of PSCs to RDF taking into account their aforementioned special features and making special emphasis in the next points:

3.3.1. URI scheme generation

Controlled structured vocabularies and conceptual resources are interpreted in SKOS as RDF resources, in particular, instances of `skos:ConceptScheme` and `skos:Concept`. Thus they are referenced by means of URIs. Although namespaces are out of the scope of this analysis, one of the tasks is the generation of the `rdf:ID`s of `skos:Concept` and `skos:ConceptScheme` from the original source's information. Usually controlled vocabularies provide unique identifiers for their terms or concepts as follows:

- Generating new identifiers for the concepts of the vocabulary. This option introduces additional management. A mapping between elements of the original source and identifiers should be maintained for updating purposes.
- Using the string of the preferred term. We would like to point out here that multilingual sources introduce a factor of complexity that it is not present in monolingual systems. In multilingual sources, this solution implies selecting a preferred term in a given natural language, thus promoting one language over the others with a possible non-desired political impact. In addition, a control procedure has to be established to ensure URI updating if the source term changes.
- Using the identifier code of an element. This solution avoids the problem of selecting one preferred language to encode concept URIs. Moreover, codes are usually strings composed of a serial number (legal URI characters) and it preserves the original semantics of a multilingual resource, where these codes identify unique terms or concepts and establish mappings between different languages. This last option has been chosen keeping in mind the desired feature of having "Cool URIs" to identify RDF resources.

3.3.2. Hierarchy formalization

From our point of view, one of the common aspects shared by KOS is a hierarchy-based structure, at least by thesauri, taxonomies and by most of classification schemes. Hierarchical relations establish links between conceptual resources, showing that the semantics of a resource is, in some way, more general ("broader") than other ("narrower"). In SKOS, the properties `skos:broader` and `skos:narrower` are only used to assert hierarchical statements between two conceptual resources. By the way, these properties are currently not defined as transitive properties. Nevertheless, third-parties, if they consider valuable, can use an OWL reasoner to infer the transitive closure of the hierarchy by means of using both transitive sub properties `skos:broaderTransitive` and `skos:narrowerTransitive`. From a theoretical point of view, the transitive closure of hierarchical relations of KOS is still an open issue. Transitive logical formalizations (e.g. using a Description Logics-based language such as SKOS/OWL) of broader/narrower properties have some risks. Cycles can appear in the hierarchical-based structured of controlled vocabularies. Even though, transitive closure of these properties can be useful for search applications to expand original user-queries with terms hierarchically related. In our conversion-method, we have followed the recommendation of the current SKOS specification.

3.3.3. Multilingual and lexical features

Regarding controlled vocabularies, multilinguism is a critical issue, especially in European vocabularies such as the CPV. Thesaurus are accessible in 21 (23) official languages of the European Union (Bulgarian, Spanish, Czech, Danish, German, Estonian, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, Finnish and Swedish) and others such as Croatian. In SKOS conceptual resources are labeled with any number of lexical strings, in any given natural language identified by the `xml:lang` attribute, following normative RDF/XML syntax. One of these labels is selected as the `skos:prefLabel` for any given language, and the others as values of `skos:altLabel`.

3.4. Application of the MOLDEAS lifecycle to Product Scheme Classifications

Previous sections have outlined the importance of PSCs to improve the accessibility and other issues in the e-Procurement sector. Common uses of semantic technologies and Linked Data have been also presented as a method to improve interoperability and integration among systems. Furthermore the MOLDEAS lifecycle has been also outlined as well as existing conversion methods and common features of PSCs. In order to carry out the promotion of PSCs to Linked Data following this lifecycle, a detailed summary of the main tasks is presented below.

- Firstly, the criteria to select and promote PSCs, see Table 1, are presented:
  - The use of the CPV 2008 is mandatory for all public procurement notices according to the Regulation (EC) No 2195/2002 of the European Parliament and it is used as a hub classification.
  - European classifications such as CPC or CPA have direct mappings (hand-made) to the CPV so they perfectly fit to the task t_8 of interlinking RDF resources (the SKOS property `skos:exactMatch` can be used for this purpose).

Table 1
Product scheme classifications.

PSC	Acronym	Source
Common Procurement Vocabulary (2003 and 2008)	CPV	European Union
Combined Nomenclature 2012	CN	European Union
Central Product Classification, version 2 (2008)	CPC	European Union
Product Classification by Activity (2008)	CPA	European Union
International Standard Industrial Classification of All Economic Activities, Rev.4	ISIC	United Nations Statistics Division
North American Industry Classification System 2007 y 2012	NAICS	United States
Standard International Trade Classification, Revision 4	SITC	United Nations Statistics Division

- International classifications such as ISIC, SITC or NAICS enable the interoperability with other e-Procurement and e-Commerce systems as well as activities in the field of statistics.
- GoodRelations and Product Ontology (PO) are two of the main references in the e-Commerce sector. That is why we reuse their definitions and instances with the objective of aligning the linkage of PSCs to existing efforts.
- Other classifications such as TARIC, UNSPSC, PRODCOM or NAPCS are ongoing work and they will be also linked to the CPV.

• Tasks t_1 and t_5. There is an implicit structure in each product classification that enables the use of graph definitions (tree and forest), see Fig. 2.
  - Product categories. A PSC is divided into product categories, Cat^k_{psc}, that group different elements according to hierarchy levels generating a disjointed set of elements. Thus, each PSC element or term is defined in only one Cat^k_{psc}.
  - Taxonomy. Apart from categories and hierarchy division, each product sector can be considered as a tree, T_{psc}, and the whole set of trees builds a forest, F_{psc}. Each t^0_{psc} element is the root of a product sector, a t_{psc} is part of only one T_{psc} and the set product sectors are also disjointed by the forest definition.

The use of the SKOS vocabulary to model PSCs is then justified due to the fact this ontology is based on two main classes `skos:Concept` and `skos:ConceptScheme`. Thus all PSC elements are an instance of the class `skos:Concept`. Therefore the interpretation of each element t_{psc} is obvious and avoids the issues presented in Section 3.3 keeping all original PSC semantics. In order to group all PSC elements under a common concept, a new class *PSCConcept* is defined as subclass of `skos:Concept`. Furthermore this class can be divided into different categories to represent the conceptual hierarchy of the PSC, if any. Thus for each Cat^k_{psc} a subclass of *PSCConcept* can be defined, e.g. Eq. (1) represents the CPV structure.

$$\begin{aligned} PSCConcept \ &\equiv Division \sqcup Group \sqcup Class \\ &\quad \sqcup CategoryDivision \sqcap Group \sqcap Class \\ &\quad \sqcap Category \\ &= \bot \end{aligned} \tag{1}$$

As Section 3.2 has outlined, the PSC taxonomy should not be interpreted as a classical `is-a` hierarchy but the SKOS Core ontology defines specific properties such as `skos:inScheme`, `skos:related`, `skos:broader Transitive` or `skos:narrowerTransitive` that perfectly match the requirements of establishing relationships among PSC elements. According to these requirements, Table 2 presents an example of a "semantized" PSC element.

• Task t_3. The Linked Data community has released a lot of vocabularies to model information and knowledge in different domains. One of the key-points to improve interoperability among systems keeping semantics and, therefore, reasoning capabilities lies in the reuse of these definitions. That is why some existing vocabularies have been selected, see Table 3, to model the PSCs.
• Task t_6. In order to get a successful promotion of raw data, the design of a good URI Scheme is a key-task (Heath & Bizer, 2011). Table 4 presents the PSCs URI scheme with the aim of addressing the desired features of providing "Cool URIs", keep the namespaces under our control, etc. and easing the reuse of RDF resources in the "Web of Data".

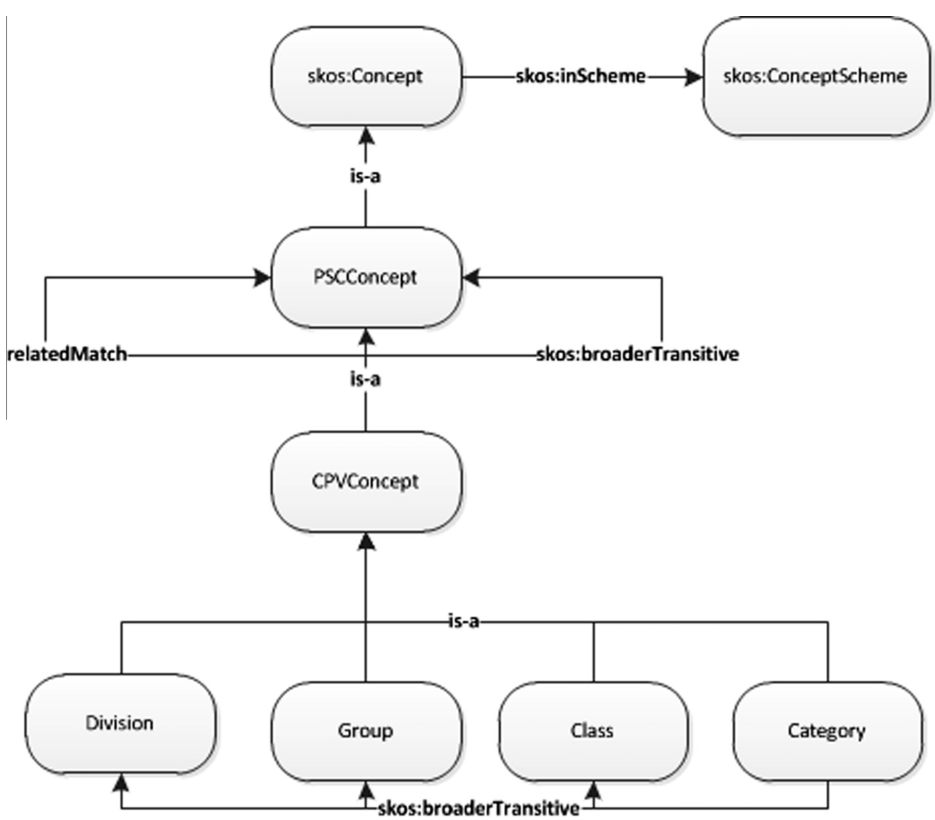


Fig. 2. A partial view of the data model for PSCs in SKOS.

Table 2
Main semantic properties used for describing a PSC element.

Property	Description	Example
`rdf:type`	Type of a PSC element t_{psc}	a `gr:ProductOr ServiceModel`, `PSCConcept`
`skos:inScheme`	URI to the PSC dataset T_{psc}	`skos:inScheme` cpv2008:ds
`dct:identifier`	Id used in the URI of the element t_{psc}	`dct:identifier` "30210000" ∧ ∧ xsd:string
`dct:subject`	Original Id of the element t_{psc}	`dct:subject` "30210000–4" ∧ ∧ xsd:string
`skos:closeMatch` and `pscs:relatedMatch`	URI to a target PSC element t'_{psc} that is partially related to a source PSC element t_{psc}	`po:machine`
`skos:exactMatch`	URI to a target PSC element t'_{psc} that is related to a source PSC element t_{psc}	`cpv2003:30216000`
`pscs:level`	Level in the hierarchy of an element t_{psc} when an explicit hierarchy definition is missing	`pscs:level` "5"
`skos:broader Transitive`	URI to a PSC element t'_{psc} ancestor of a broader PSC element t_{psc}	`skos:broaderTransitive` cpv2008:30200000
`skos:prefLabel`	Multilingual labels and descriptions	"Data-processing machines (hardware)"@EN

Table 3
Main semantic vocabularies selected to model a PSC. These prefixes have been retrieved using the "Prefix.cc" service.

Vocabulary prefix	Description
dbpedia	Reuse of DBPedia concepts and properties
dc and dct	Dubling Core properties for describing metadata
gr	Use of the GoodRelations vocabulary to describe products and services
po	Use of the ProductOntology vocabulary to create links to existing product and service definitions
skos	Taxonomy specification and labeling properties
void	Description of datasets metadata

Table 4
Design of an URI scheme for the PSCs catalogue.

URI	Description	Example
http://purl.org/weso/pscs/	URI base: <base_uri>	NA
<base_uri>/ontology	Common definitions	<base_uri>/ontology/PSCConcept
<base_uri>/resource/ds	Description of the PSCs Catalogue	<base_uri>/resource/ds
<base_uri>/{psc}/{version|year}	PSC Namespace	<base_uri>/cpv/2008
<base_uri>/{psc}/{version|year}/ontology	Specific definitions	<base_uri>/cpv/2008/ontology
<base_uri>/resource/{psc}/{version|year}/id	URI for RDF resources	<base_uri>/cpv/2008/resource/30210000
<base_uri>/resource/{psc}/{version|year}/ds	Description of the PSC dataset	<base_uri>/cpv/2008/resource/ds

- Task t_8. The broad aim of this task is to enable the possibility of querying public procurement notices from any PSC. The mappings can be generated following two approaches: (1) exact mappings created by a domain expert (2387) or (2) related (19328), through the execution of a custom entity reconciliation process based on NLP techniques. To do so a PSC reconciliation service has been designed using the Apache Lucene and Solr libraries in which both filters and the syntactic search engine can be easily customized. Firstly a target PSC is selected and their descriptors (in English) are indexed using Apache Lucene. After that the mappings between elements in a source PSC are automatically discovered and generated through the execution of queries against the indexed PSC. Basically the creation of a mapping lies in a string comparison between the source descriptor and the target descriptor. Nevertheless the personalization of Apache Lucene and Solr allowed us the creation of more accurate mappings. Finally an important issue lies in the mapping confidence, in this case a pairwise human-validation has been carried out to ensure a 80–90% of correct mappings.
- Task t_9. Since a model for representing PSCs is available the real transformation to RDF has been performed using Google Refine and a Java processor. The results of promoting to RDF the selected PSCs are presented in Table 5, more specifically the first four columns specify respectively the vocabulary of a PSC (V_{psc}), the number of elements to be promoted ($\#V_{psc}$), the number of generated RDF triples and the number of links to the CPV 2008. Thus we have promoted more than 42035 product descriptions, generating more than 1842053 triples and creating 21715 links among them.

Table 5
Statistics of promoting to RDF selected PSCs and linking them to the CPV 2008.

$\mathcal{V}_{psc}$	$\#\mathcal{V}_{psc}$	RDF triples	Links to CPV 2008	% Real	Links through PO	% Real trough PO
CPV 2008	10357	803311	0	0	10000	96.55
CPV 2003	8323	546135	462	4.46	8312	80.25
CN 2012	14552	137484	2390	23.08	2390	23.08
CPC 2008	4408	100819	4402	42.50	4403	42.51
CPA 2008	5429	92749	5399	52.13	5410	52.24
ISIC v4	766	18986	765	7.39	765	7.39
NAICS 2007	2328	36292	2300	22.21	2300	22.21
NAICS 2012	2212	70887	2186	21.11	2186	21.11
SITC v4	4017	3811	3811	36.80	3820	36.88
Total ☆	42035	1842053	21715	209.66	29586	285.66

```
cpv2008-res:30210000 a gr:ProductOrServiceModel, cpv-onto:Group;
  skos:prefLabel, gr:description, rdfs:label
  "Data-processing␣machines␣(hardware)"@EN ,
  "Naprave␣za␣obdelavo␣podatkov␣(strojna␣oprema)"@SL ,
  "Databehandlingsmaskiner␣(hardware)"@DA ,
  "Macchine␣per␣l'elaborazione␣di␣dati␣(hardware)"@IT ,
  "Databehandlingsmaskiner␣(maskinvara)"@SV ,
  "Masini␣de␣procesare␣a␣datelor␣(hardware)"@RO ,
  "Machines␣voor␣dataprocessing␣(hardware)"@NL ,
  ... ;
  dct:identifier "30210000"^^xsd:string;
  dct:subject "30210000-4"^^xsd:string;
  pscs-onto:relatedMatch
    <http://www.productontology.org/id/dataprocessing>,
    <http://www.productontology.org/id/hardware> ,
    <http://www.productontology.org/id/machine> ;
  skos:broaderTransitive cpv2008-res:30200000;
  skos:exactMatch
    cpv2003-res:30216000, cpv2003-res:30215000,  cpv2003-res:30213000,
    cpv2003-res:30212000, cpv2003-res:30211000,  cpv2003-res:30214000.
```

Fig. 3. Partial example of CPV 2008 item in RDF.

```
SELECT DISTINCT * WHERE{
   ?product pscs:relatedMatch
    <http://www.productontology.org/id/construction> .
   ?product skos:closeMatch ?cpv .
   ?product skos:prefLabel ?productLabel .
   ?cpv skos:prefLabel ?cpvLabel .
   ?product skos:inScheme ?scheme .
   FILTER (?scheme != <http://purl.org/weso/pscs/cpv/2008/resource/ds>) .
   FILTER (lang (?cpvLabel) ="en" )
} LIMIT 100
```

Fig. 4. Example of SPARQL query.

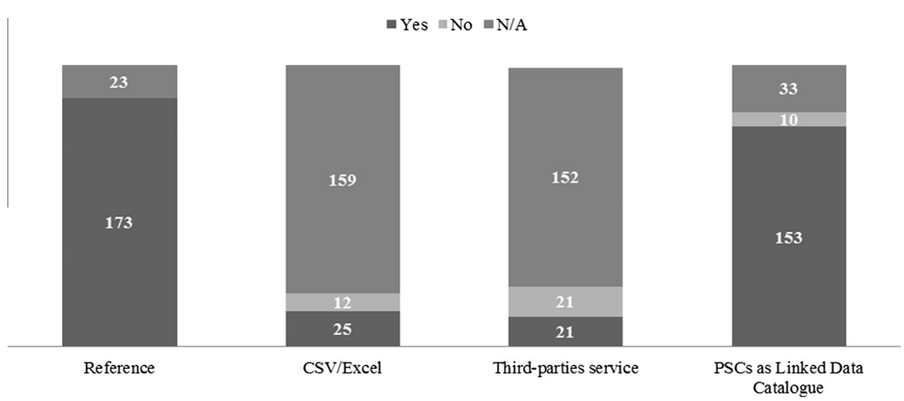


Fig. 5. Validation and comparison of the PSCs catalogue as Linked Data with regards to existing versions.

Finally, Fig. 3 shows a partial example of a generated RDF resource in which multilinguism features, hierarchy relationships and links can be found. Furthermore, data can be now exploited through SPARQL queries such as "Give me 100 products or services related to construction in any PSC that have a mapping with products or services in CPV (descriptions in English)", see Fig. 4, using more descriptors and different vocabularies.

- Other tasks such as "Add metainformation" are considered optional but relevant to mainly ease the process of dataset discovery.
- Task t_{14}. Once the PSCs are available as RDF is necessary to make them publicly available through a Linked Data infrastructure. To do so all RDF triples and the ontology have been deployed in a RDF repository (OpenLink Virtuoso), thus data can be accessed through a SPARQL endpoint. Furthermore and with the aim of making dereferenceable URIs and enabling content negotiation an instance of Pubby (a Linked Data frontend) has been also deployed. Besides in order to ease the creation and execution of SPARQL queries, SNORQL (a tool for exploring SPARQL endpoints) has been also released. Finally, the new PSCs catalogue has been reported to the LOD community, added to the "datahub.io" register and joined the LOD Cloud Diagram under the dataset id "`pscs-catalogue`".
- Task t_{12} (validation process). The MOLDEAS lifecycle also purposes a validation method based on a checklist of 196 points compiled from books, W3C specifications, Open Data and Linked Data principles, CKAN validator, LOD Cloud requirements, etc. which values can be "Applicable and positive (Yes)", "Applicable and not found (No)" or "Not Applicable (N/A)". The objective of this questionnaire is to ensure the principles of the LOD initiative for

easing the reuse, maintenance, governance, coverage and expressivity of data. In order to compare the new version of the PSCs catalogue to existing public versions, we have carried out this survey obtaining the next results, see Fig. 5, in which "Reference" corresponds to the MOLDEAS lifecycle *best* possible evaluation, "CSV/Excel" to the official publicly available versions of PCSs and "Third-parties service" to those versions managed by non-official organizations.

- Task t_{11} (consumption process). The PSCs catalogue as Linked Data has been used in the development of the MOLDEAS project and a public demonstrator can be used to search public procurement notices, see Fig. 6.

4. Evaluation

In order to evaluate if the approach to generate Linked Data from PSCs can improve the access to public procurement data exploiting the advantages of this initiative two experiments have been designed:

1. The first one checks if we can access more public procurement notices when the links between a PSC and the CPV are created. The aim of this study is to compare the initial input vocabulary (containing only the CPV descriptors) and the final input vocabulary (containing all descriptors that have been linked to the CPV). Therefore if we extend the input vocabulary we can ease the access to more public procurement notices because we can jump from any PSC to the CPV. In the same way if we need to make some kind of comparison between systems working with different PSCs we can use the CPV as *lingua franca*.
2. The second study tries to exploit the semantic relationships between the CPV concepts to build a recommender of CPV codes for business users. In order to make a empirical experiment we have used 11 real user queries and their translation into CPV codes made by the company "Euroalert.net".

4.1. Study 1

4.1.1. Research design

The purpose of this study is to compare the size of the initial input vocabulary for retrieving public procurement notices with regards to the final input vocabulary when links between a PSC and the CPV are created. The exploitation of these links can be used to build new SPARQL queries that can take advantage of Linked Data principles to provide a greater input vocabulary. Let's assume the following points about the public procurement notices and the PSCs used in this study:

- All public procurement notices are always annotated by, at least, one CPV element.
- Each PSC is a controlled vocabulary. The CPV is a controlled vocabulary $\mathcal{V}$ comprised of 10357 elements/terms/codes.
- The public procurement notices is a RDF dataset, $\mathcal{D}$, Each document $d \in \mathcal{D}$ is annotated, at least, using a CPV code $v \in \mathcal{V}$. Thus, if a query contains all elements $v \in \mathcal{V}$ the entire dataset of notices will be retrieved.

Once these assumptions are defined, the gain of linking a new finite PSC to the CPV can be calculated as follows:

- The source controlled vocabulary $\mathcal{V}_{psc}$ is comprised of $\#\mathcal{V}_{psc}$ elements.
- The linking of terms between a source vocabulary $\mathcal{V}_{psc}$ and a target vocabulary $\mathcal{V}$ can be carried out in the next ways:
  1. $1-1$ link, there are elements $v^k_{psc} \in \mathcal{V}_{psc}$ that are linked to only one element $v \in \mathcal{V}$.
  2. $1-n$ link, there are elements $v^k_{psc} \in \mathcal{V}_{psc}$ that are linked to some elements $v \in \mathcal{V}$ generating K links.
  3. The result of the previous operation generates links or pairs in the form $p_k = (v^k_{psc}, v^k)$ building a set of pairs $\mathcal{P} = \{p_1, p_2, \ldots, p_k, p_n\}$. Taking into account this situation, the initial vocabulary V is increased in all elements $v^k_{psc} \in \mathcal{V}_{psc}$ that have a link to an element $v \in \mathcal{V}$. The new derivate vocabulary $\mathcal{V}'$ is a controlled vocabulary comprising all elements $v^k_{psc} : v^k_{psc} \in p_i$.

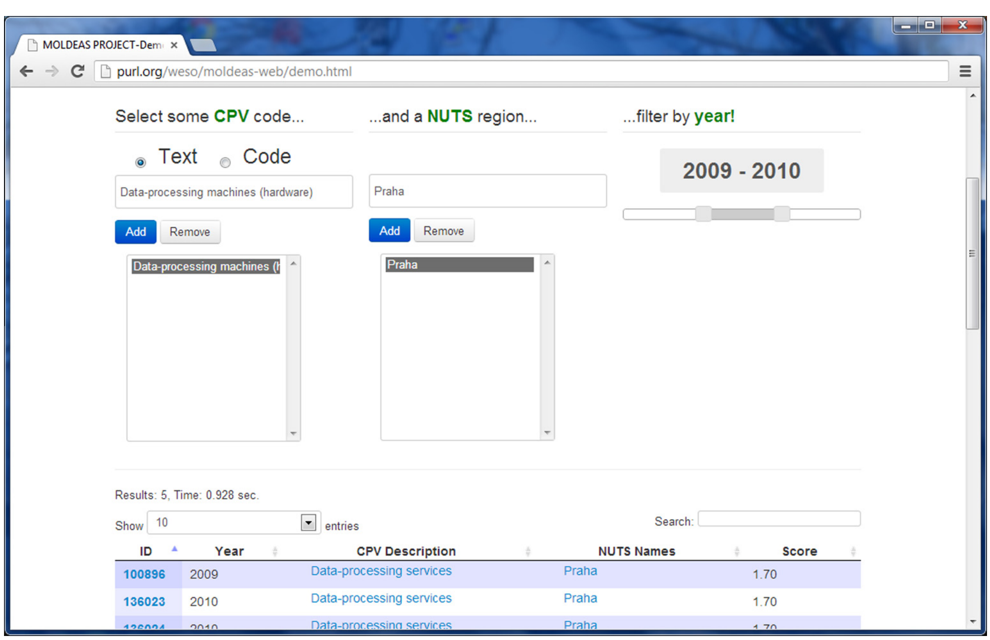


Fig. 6. The MOLDEAS public demonstrator using the PSCs catalogue as Linked Data.

- The percentage of gain in terms of expressivity (number of elements to be used in queries) is related to the number of elements that enables go from $\mathcal{V}_{psc}$ to $\mathcal{V}$:

$$\% = 100 * \left(\frac{\#\mathcal{V}'_{\surd\rfloor} + \#\mathcal{V}}{\#\mathcal{V}} - 1\right) \tag{2}$$

As an example of this approach, two controlled vocabularies and a set of links/pairs are presented to calculate the gain in terms of expressivity:–

Let

$\mathcal{V} = \{A, B, C, D, E\}$ and $\mathcal{V}_{psc} = \{1, 2, 3\}$ the source and target controlled vocabularies.

- Let $P = \{(A,1),(B,2),(C,1)(C,2)\}$ the set of links/pairs between $\mathcal{V}$ and $\mathcal{V}_{psc}$.
- Let $\mathcal{V}'_{\surd\rfloor} = \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$ the set of elements $v^k_{psc} : v^k_{psc} \in p_i$.
- Therefore, the new controlled vocabulary to query the dataset is comprised of $\{\mathcal{V}_{\surd\rfloor} \cup \mathcal{V}'_{\surd\rfloor}\}$ and the percentage of gain in terms of expressivity according to the Eq. (2) is:

$$\% = 100 * \{\langle(3+3)/3\rangle - 1\} = 2 - 1 = 100 \tag{3}$$

- As a consequence the number of final terms to create queries is just two times than the initial set, increasing the expressivity in a 100%.

- Finally, in this study, there is a specific scenario in which elements $v^k_{psc} \in \mathcal{V}_{psc}$ are not directly mapped to elements $v \in \mathcal{V}$ but assuming that there are relations r_k among elements v^j_{psc} and v^k_{psc} new links could emerge between v^j_{psc} and v through r_k. Nevertheless this situation should be avoided in order to prevent an infinite and recursive linking process and keep the advantages of using finite controlled vocabularies, e.g. in the ongoing example the Product Ontology (PO) is used as a bridge classification implying an infinite max gain (∞).

4.2. Sample

The PSCs that have been selected to be promoted to RDF, see Table 1, are also used to check the gain we can get through a direct link to the CPV 2008 or by means of creating a "bridge link" through the ProductOntology. A RDF dataset $\mathcal{D}$ of public procurement notices that contains 919,341 documents annotated with a total number of 1866367 CPV codes (10,062 are different), has been selected as a sample dataset.

4.3. Results and discussion

Table 5 presents the statistics of transforming each PSC as well as the number of terms that have linked to CPV 2008 and the gain (percentage) of new descriptors we can use to query the dataset. Following the same approach as in the previous example, the gain in terms of expressivity is calculated below:

- Let $\mathcal{CPV}_{2003} = \{v^0_{cpv_{2003}}, v^1_{cpv_{2003}}, \ldots, v^n_{cpv_{2003}}\}$ and $\mathcal{CPV}_{2008} = \{v^0_{cpv_{2008}}, v^1_{cpv_{2008}}, \ldots, v^n_{cpv_{2008}}\}$ the source and target controlled vocabularies.
- Let $P = \{\left(v^k_{cpv_{2003}}, v^j_{cpv_{2008}}\right)\}$ the set of links/pairs between $\mathcal{CPV}_{2003}$ and $\mathcal{CPV}_{2008}$.
- Let $\mathcal{CPV}'_{2008} = \{v^k_{cpv_{2003}}\}$ the set of elements $v^k_{cpv_{2003}} : v^k_{cpv_{2003}} \in p_i$.
- Therefore, the new controlled vocabulary to query the dataset is comprised of $\{\mathcal{CPV}_{2008} \cup \mathcal{CPV}'_{2008}\}$ and the percentage of gain in terms of expressivity according to the Eq. (2) is:

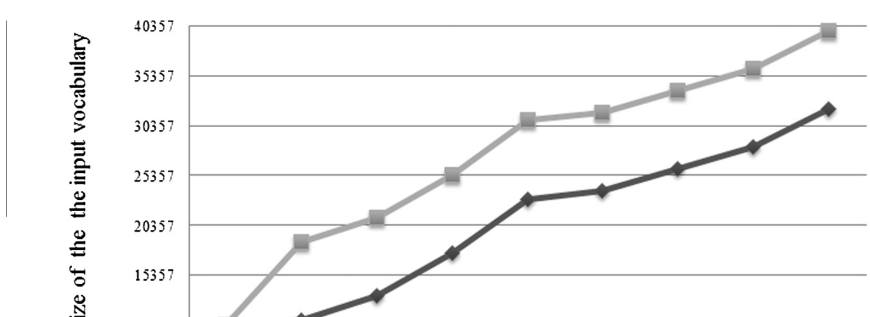
$$\begin{aligned} \% &= 100 * \{\langle(462 + 10357)/10357\rangle - 1\} \\ &= 1.0446 - 1 = 4.46 \end{aligned} \tag{4}$$

- As conclusion the real expressivity gain (in terms of descriptors to query a dataset) with regards to the initial set of 10,357 terms in CPV 2008 has been increased in a 4.46% (462 new elements can be now used to build SPARQL queries).

Taking into account the full results presented in Table 5 the real expressivity gain (in terms of descriptors to query a dataset) with regards to the initial set of 10,357 terms in CPV 2008 has been increased in a 209.66% (21,715 new elements are part of the input vocabulary). Furthermore if the ProductOntology is used as a "bridge" the input vocabulary is then increased in a 285.66% (29,586 new elements). Although this last approach shows a valuable increment in the number of linked elements it should be avoided due to the fact that more "jumps" between classifications imply necessarily less confidence in the mappings. Finally Fig. 7 presents the evolution of the input vocabulary when a new PSC is linked to the CPV 2008.

4.4. Study 2

The aim of this study is to verify if the Linked Data generated from public procurement notices, PSCs, etc. can be exploited to deliver a decision support system that helps business users to select CPV codes from user queries in natural language. This study is motivated by the company "Euroalert.net" that sells an alert service for users (companies, etc.) that want to tender in specific sectors under a certain profile. Usually, the most relevant variables in



	CPV 2008	CPV 2003	CN 2012	CPC 2008	CPA 2008	ISIC V4	NAICS 2007	NAICS 2012	SITC V4
#Links to CPV 2008	10357	10819	13209	17611	23010	23775	26075	28261	32072
#Links through PO	10357	18669	21059	25462	30872	31637	33937	36123	39943

Fig. 7. Evolution of the number of terms when a new PSC is linked to the CPV 2008.

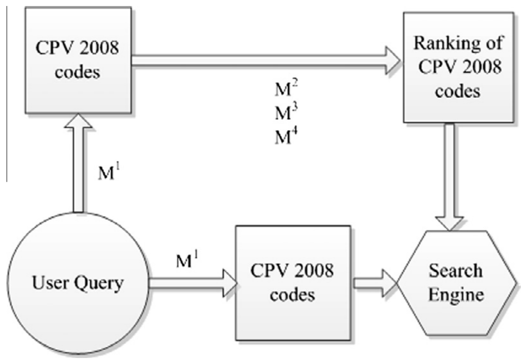


Fig. 8. Workflow to recommend CPV 2008 codes according to an user query.

Table 6
Methods for recommending CPV 2008 codes.

Method	Description	Technology
M^1	CPV English descriptions are indexed and a syntactic search process is performed for each Q_{str}	Apache Lucene y Solr
M^2	CPV codes are selected according to the SKOS taxonomy. Previously, the method M^1 is applied to extract CPV codes from natural language	M^1 + ranking broader/ narrower
M^3	Similar to M^2 but codes are selected using *Spreading Activation*	M^1 + ONTOSPREAD
M^4	Similar to M^2 but codes are selected using the analysis of historical relationships in previous public procurement notices	M^1 + Apache Mahout

public procurement notices are the CPV codes and the location (NUTS codes). In this experiment we simulate the behavior of "Euroalert.net" employees when they have to translate user queries, Q_{str}, into CPV codes for creating a new alert. Basically, the experiment is based on methods for semantic search and concept query expansion, see Fig. 8, that we have designed to tackle the recommendation of CPV codes, see Table 6.

4.5. Research design

In order to validate the accuracy of the different methods to suggest CPV codes the typical measures of precision, see Eq. (6), recall, see Eq. (7) and F1 score (the harmonic mean of precision and recall), see Eq. (5) will be used to compare and select the best method to recommend CPV codes. In particular, *tp* (true positive) is "the set of expected CPV codes that have been retrieved", *fp* is "the set of un-expected CPV codes that have been retrieved", *tn* is "the set of expected CPV codes that have not been retrieved" and *fn* is "the set of un-expected CPV codes that have not been retrieved".

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

$$Precision = \frac{tp}{tp + fp} \quad (6)$$

$$Recall = \frac{tp}{tp + fn} \quad (7)$$

4.6. Sample

This experiment has been carried out using the next datasets:

Table 7
User queries and number of expected CPV codes.

Q_i	User query-Q_{str}	Number of expected CPV codes-$\#Q^i_{cpv}$
Q_1	"Comprehensive overview over all environmental technologies renewable energy products"	463
Q_2	"Tendering of public works: housing, hospitals, roads, housing developments, station drinking water treatment, reforestation"	35
Q_3	"Prefabricated buildings"	7
Q_4	"Games for children (parks swings, slides, land of play equipment in the public sphere"	26
Q_5	"Vital signs monitor"	277
Q_6	"Museum and exhibition and product launch services"	1
Q_7	"Voltmeters, instruments measuring electrical quantities, Ammeters, Instruments for checking physical characteristics, hygrometers, thermometers, measuring equipment and control, leak detector, Analyzers, Cable Splicing insulated cable joints kits, screwdrivers, hand tools, screwdriver"	117
Q_8	"Conservation Maintenance of pavements for roads, airfields, bridges, tunnels"	13
Q_9	"Wood poles, Wooden sleepers, Lattice towers"	10
Q_{10}	"Architectural, construction, engineering and inspection services"	173
Q_{11}	"Medical practice and related services"	13

- A RDF dataset of public procurement notices from 2008 to 2011 that contains 919,341 documents annotated with a total number of 1866367 CPV codes (10,062 are different) and 1438572 NUTS codes (1896 are different).
- A set of 11 real user queries and the expected set of CPV codes created by "Euroalert.net" employees, see Table 7 (queries have been "normalized" according to one CPV descriptor to keep privacy).
- The CPV dataset as Linked Data.

More specifically, Table 7 shows the description of each query where Q_i is the identifier of a user query, Q_{str} is the real query as a string and $\#Q^i_{cpv}$ is the number of expected CPV codes. Obviously, the comparison does not just take into account the cardinality of the output set but the CPV code.

4.7. Results and discussion

According to the results in Table 8, the method that better matches the behavior of an expert user is M^1 with an average precision of 0.28 and recall of 0.26. This implies that a method based on natural language processing techniques is the most appropriate to translate user queries into CPV codes, or, at least is the method that better matches the translation made by business users. On the other hand, although the method M^3 with an average precision of 0.23 and recall of 0.23 obtains the second position it should be able to get a better performance exploiting semantic relationships in the CPV (maybe a training process is required to adjust the Spreading Activation algorithm). Likely, the type of translation that the expert user does is more close to a direct language translation than a real thinking about the relationships in the taxonomy. Finally, the exploitation of the historical statistics using Apache Mahout algorithms, M^4, (after different iterations and tests we have chosen the best configuration based on a collaborative filtering algorithm) does not seem to reflect the expert user behavior. Although search and recommendation should not be compared it is uncommon than a

Table 8
Quantitative measures (P-precision and R-recall) of methods for recommending CPV codes.

M^i	M^1			M^2			M^3			M^4		
Q_i/Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Q_1	0.15	0.08	0.10	0.15	0.15	0.15	0.12	0.06	0.08	0.06	0.06	0.06
Q_2	0.09	0.09	0.09	0.06	0.06	0.06	0.03	0.03	0.03	0.03	0.03	0.03
Q_3	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0	0	0
Q_4	0.19	0.19	0.19	0	0	0	0.12	0.12	0.12	0	0	0
Q_5	0.12	0.01	0.02	0.01	0.01	0.01	0.08	0.01	0.02	0.03	0.03	0.03
Q_6	1.00	1.00	1.00	0	0	0	1.00	1.00	1.00	0.10	0.67	0.17
Q_7	0.20	0.20	0.20	0.09	0.09	0.09	0.15	0.16	0.15	0.03	0.03	0.03
Q_8	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0	0	0
Q_9	0.50	0.50	0.50	0	0	0	0.30	0.38	0.34	0	0	0
Q_{10}	0.39	0.39	0.39	0.42	0.42	0.42	0.34	0.35	0.34	0.16	0.16	0.16
Q_{11}	0.23	0.23	0.23	0.23	0.23	0.23	0.15	0.17	0.16	0	0	0
Average of quantitative measures												
Total	0.28	0.26	0.27	0.11	0.11	0.11	0.23	0.23	0.224	0.03	0.03	0.044

dataset containing the use CPV codes in 919341 public procurement notices could not improve the traditional syntactic search. Nevertheless, there is an open issue that lies in the number of true negative codes that in a second step should be validated by experts, in this situation we strongly believe semantic-based methods could improve their results with regards to M^1.

Finally, after the execution of these experiments and back to the motivating examples presented in the introduction section some assertions can be done:

- The American SME can now use the NAICS classification to search public procurement opportunities in Europe.
- The Spanish family-owned company can access to Bulgarian low-value procurement opportunities because the use of the CPV (through Internet protocols) directly supports a multilingual environment and they can also obtain suggestions of CPV codes.
- The department of statistics within the WorldBank can gather and summarize different economic data due to the link among PSCs. In this case the CPV can be used as *lingua franca* to make data comparisons.
- The company that sells a commercial alert service or more specifically its employees can take advantage of the recommending engine to support their decisions and make a better translation of customers intentions.

5. Conclusions and future work

The main advantage of applying Linked Data to PSCs lies in the possibility of enhancing the number of terms to build enriched SPARQL queries. Thus, a broader input vocabulary can ease the access to more and newer public procurement notices as. In the European e-Procurement context, this work represents a step-forward to a new data realm in which public procurement data must be a key-enabler for the next data-based economy easing the access to business opportunities to companies, specially SMEs, and creating the appropriate environment for a real EU-wide public market. According to the EC, e-Procurement may, by its nature, be more compatible or facilitate the use of procurement budgets in support of the EU 2020 objectives. Some countries have adopted the recommendations of the EU introducing strategies to encourage SMEs participation in the procurement processes but the intended participation of the SMEs is still far from the EU expectations. That is why the e-Procurement sector must take advantage of the LOD initiative to boost and encourage the access to public procurement opportunities.

On the other hand, the exploitation of Linked Data to create a decision support system helping users to select CPV codes from user queries is still an open issue that must be improved applying more advanced techniques: in-depth analysis of the CPV codes, make correlations with other variables, etc. Nevertheless, we have demonstrated that the advantages of promoting public procurement data, with special focus on PSCs, can bridge the gap between human readable formats and machine-processable data. Moreover, this work is part of the MOLDEAS platform in which the application of Linked Data principles to the e-Procurement sector has a broader scope. The PSC reconciliation service has been also applied to the "PublicSpending.net" initiative with the aim of comparing the public spending around the world. Further steps include the application of the life-cycle (and its refinement) to other relevant PSCs such UNSPSC, NACE or TARIC, the improvement of the reconciliation service, the deployment of a PSCs portal similar to the "NCBO BioPortal" and the reporting of results to: (1) the Office for Official Publications of the European Communities; (2) the "Internal Market and Services Directorate General" and (3) the "Information Society and Media Directorate General" of the European Commission. As final future actions, this work must be aligned to current research trends such as data quality (Bizer, 2007; Bizer & Cyganiak, 2009; Alvaro Graves Olye Rickson, 2011; Stadler et al., 2011), provenance (Hartig & Zhao, 2010), trust (Carroll et al., 2005), large dataset processing (Hausenblas, Grossman, Harth, & Cudré-Mauroux, 2012), entity reconciliation (Maali & Cyganiak, 2011), searching (Hogan et al., 2011) or inference (Bonatti et al., 2011) to name a few.

References

Adida, B., & Birbeck, M. (2008). RDFa primer, bridging the human and data webs. W3C working group note, W3C, 2008. <http://www.w3.org/TR/xhtml-rdfa-primer/>.
Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). Describing linked datasets with the VoID vocabulary. W3C interest group note, W3C. <http://www.w3.org/TR/void/>.
Alor-Hernández, G., Gómez Berbís, J. M., Rodríguez González, A., et al. (2010). HYDRA: A middleware-oriented integrated architecture for e-Procurement in supply chains. *Transaction on Computational Collective Intelligence, 1*, 1–20.
Álvarez Rodríguez, J. M., Gayo, J. E. L., & de Pablos, P. O. (2012). An extensible framework to sort out nodes in graph-based structures powered by the spreading activation technique: The ONTOSPREAD approach. *IJKSR, 3*(4), 57–71.
Álvarez Rodríguez, J. M., Gayo, J. E. L., Silva, F. A. C., Alor-Hernández, G., Sánchez, C., Luna, J. A. G., et al. (2012). Publish and search public procurement notices powered by linked open data: The MOLDEAS approach. *International Journal of Software Engineering and Knowledge Engineering, 22*(3), 365–384.
Alvaro Graves Olye Rickson. (2011). LOD data quality vocabulary: LODQ. Working draft, Linking open government data. <http://logd.tw.rpi.edu/lodq>.
Ankolekar, A., Krötzsch, M., Tran, T., & Vrandecic, D. (2007). The two cultures: Mashing up web 2.0 and the Semantic Web. In *Proceedings of the 16th international conference on World Wide Web. 2007 May 7–8*. ACM Press.
Araujo, S., Hidders, J., Schwabe, D., & De Vries, A. P. (2011). SERIMI resource description similarity, RDF instance matching and interlinking. *Work*.
Arwe, J., Nally, M., Speicher, S., & Le Hors, A. (2012). Linked data basic profile 1.0. W3C member submission, W3C. <http://www.w3.org/Submission/2012/02/>.

Berners-Lee, T. (1998). *Cool URIs don't change*.
Berners-Lee, T. (July 2006). *Linked data*. <http://www.w3.org/DesignIssues/LinkedData.html>.
Berrueta, D., Gayo, J. E. L. & Polo, L. (2006). Searching over public administration legal documents using ontologies. In *JCKBSE* (pp. 167–175).
Berrueta, D., Phipps, J., Miles, A., Baker, T., & Swick, R. (August 2008). *Best practice recipes for publishing RDF vocabularies*. W3c working group note.
Bizer, C. (2007). *Quality-driven information filtering in the context of web-based information systems*. PhD thesis.
Bizer, C., & Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics, 7*(1), 1–10.
Bizer, C., Cyganiak, R., & Heath, T. (2007). *How to publish linked data on the web*.
Bonatti, P. A., Hogan, A., Polleres, A., & Sauro, L. (2011). Robust and scalable linked data reasoning incorporating provenance and trust annotations. *Journal of the Web Semantics, 9*(2), 165–201.
Buil-Aranda, C., Arenas, M., & Corcho, O. (2011). Semantics and optimization of the SPARQL 1.1 federation extension. In *Proc. of the 8th ESWC'11*, ESWC'11.
Calbimonte, J. -P., Jeung, H., Corcho, O., & Aberer, K. (2011). Semantic sensor data search in a large-scale federated sensor network. In *SSN* (pp. 14–29).
Carroll, J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th international conference on World Wide Web* (pp. 613–622). ACM.
Chen, A., Chen, H. -H., & Huang, P. (2007). Predicting social annotation by spreading activation. In *Proceedings of the 10th international conference on Asian digital libraries: Looking back 10 years and forging new frontiers. ICADL'07* (pp. 277–286). Berlin, Heidelberg: Springer-Verlag.
Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82*(6), 407–428.
CTO Council's cross Government Enterprise Architecture. (2010). Designing URI sets for the UK public sector. Draft, UK Government.
Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 829–839.
Cyganiak, R., & Jentzsch, A. (2011). The Linking Open Data cloud diagram. <http://richard.cyganiak.de/2007/10/lod/>.
Cyganiak, R., Stenzhorn, H., Delbru, R., Decker, S., & Tummarello, G. (2008). Semantic sitemaps: Efficient and flexible access to datasets on the Semantic Web. In *Proceedings of the 5th European Semantic Web conference* (pp. 690–704).
Dadzie, A.-S., & Rowe, M. (2011). Approaches to visualising linked data: A survey. *Semantic Web, 2*(2), 89–124.
Di Noia, T., Valle, F., d'Aquin M., & Motta, E. (October 2010). LOTED: Exploiting linked data in analyzing european procurement notices. In V. Presutti, V. Svatek, and F. Sharffe (Eds.), *EKAW workshop on knowledge injection into and extraction from linked data (KIELD2010)*, Lisbon, Portugal.
Dodds, L., & Davis, I. (August 2011). Linked data patterns. A pattern catalogue for modelling, publishing, and consuming linked data. <http://patterns.dataincubator.org/book/>.
European Commission. (October 2008). European code of best practices facilitating access by SMEs to public procurement contracts. <http://ec.europa.eu/internal_market/publicprocurement/docs/sme_code_of_%best_practices_en.pdf>.
European Commission. (October 2010). Consultation on the Green paper on expanding the use of e-Procurement in the EU. <http://ec.europa.eu/internal_market/consultations/docs/2010/e-procureme%nt/green-paper_en.pdf>.
European Commission. (November 2010). Public procurement indicators 2009. <http://ec.europa.eu/internal_market/publicprocurement/docs/indicators20%10_en.pdf>.
European Commission. (October 2011). Modernising European public procurement to support growth and employment. <http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1580>.
European Commission. (June 2013). e-Notices. <http://simap.europa.eu/enotices>.
European Commission. (June 2013). eCERTIS. <http://ec.europa.eu/markt/ecertis>.
European Commission. (June 2013). Pan-European Public Procurement Online (PEPPOL). <http://www.peppol.eu>.
European Commission. (June 2013). Standard forms for public procurement. <http://simap.europa.eu/buyer/forms-standard/index_en.htm>.
European Commission Internal Market Directorate-General. (July 2010). *Study on the evaluation of the Action Plan for the implementation of the legal framework for electronic procurement (Phase II)*. Analysis, assessment and recommendation.
European Parliament and European Council. (November 2003). Directive 2003/98/EC of the European parliament and of the council of 17 November 2003 on the re-use of public sector information. <http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/%psi_directive_en.pdf>.
Fensel, D., Ding, Y., & Omelayenko, B. (2011). Product data integration in B2B E-commerce. *IEEE-Intelligent E-Business*, 54–59.
Gao, Q., Yan, J., & Liu, M. (2008). A semantic approach to recommendation system based on user ontology and spreading activation model. In *Proceedings of the 2008 IFIP international conference on network and parallel computing. NPC '08* (pp. 488–492). Washington DC, USA: IEEE Computer Society.
García, A. M., Álvarez Rodríguez, J. M., Gayo, J. E. L., & Merino, M. M. (2013). Towards a journalist-based news recommendation system: The Wesomender approach. *Expert Systems with Applications*.
Gayo, J. E. L., de Pablos, P. O., & Lovelle, J. M. C. (2010). WESONet: Applying Semantic Web technologies and collaborative tagging to multimedia web information systems. *Computer and Human Behaviours, 26*(2), 205–209.
Gelgi, F., Vadrevu, S., & Davulcu, H. (2005). Improving web data annotations with spreading activation. In *WISE* (pp. 95–106).
Harth, A., Hogan, A., Kotoulas, S., & Urbani, J. (2011). Scalable integration and processing of linked data. In *Proceedings of the 20th international conference companion on World Wide Web. WWW '11* (pp. 295–296). New York, NY, USA: ACM.
Hartig, O., & Zhao, J. (2010). Publishing and consuming provenance metadata on the web of linked data. In *IPAW* (pp. 78–90).
Hartig, O., Bizer, C., & Freytag, J. C. (2009). Executing SPARQL queries over the web of linked data. In *8th International Semantic Web conference (ISWC)*.
Hausenblas, M., Villazón-Terrazas, B., & Hyland, B. (2011). GLD life cycle. W3C government linked data group, W3C. <http://www.w3.org/2011/gld/wiki/GLD_Life_cycle>.
Hausenblas, M., Grossman, R., Harth, A., & Cudré-Mauroux, P. (2012). Large-scale linked data processing – Cloud computing to the rescue? In *CLOSER* (pp. 246–251).
Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space* (Vol. 1). Morgan & Claypool.
Hepp, M. (2006). Semantic Web and Semantic Web services: Father and son or indivisible twins? *IEEE Internet Computing, 10*(2), 85–88.
Hepp, M. (2007). Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing, 11*(1), 90–96.
Hogan A., Harth A., Umbrich J., Kinsella S., Polleres A., Decker S., Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine, Web Semantics: Science, Services and Agents on the World Wide Web, 9(4), 2011, 365–401, ISSN 1570-8268, http://dx.doi.org/10.1016/j.websem.2011.06.004..
Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of linked data conformance. *Journal of the Web Semantics, 14*, 14–44.
Hyland, B., Terrazas, B. V., & Capadisli, S. (2011). Cookbook for open government linked data. W3C task force-government linked data group, W3C. <http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook>.
Isele, R., & Bizer, C. (2011). Learning linkage rules using genetic programming. In *OM*.
Jentzsch, A., Isele, R., & Bizer, C. (2010). Silk – Generating RDF links while publishing or consuming linked data. In *ISWC posters & demos*.
Jeung, H., Sarni, S., Paparrizos, I., Sathe, S., Aberer, K., Dawes, N., et al. (2010). Effective metadata management in federated sensor networks. In *Proceedings of the 2010 IEEE international conference on sensor networks, ubiquitous, and trustworthy computing. SUTC '10* (pp. 107–114). Washington, DC, USA: IEEE Computer Society.
LAPSI Project FP7. (June 2013). The European Thematic network on legal aspects of public sector information. <http://www.lapsi-project.eu/>.
LATC Project FP7. (June 2013). Linked open data around-the-clock. <http://latc-project.eu/>.
Leukel, J. (2004). Automating product classification: The. In *Proceedings of the IADIS international conference e-society 2004 (ES 2004)*, Ávila, Spain, 2004.
Leukel, J. (2004). Standardization of product ontologies in B2B relationships, On the role of ISO 13584. In Proceedings of the tenth americas conference on information systems (pp. 4084–4091). New York, United States.
Leukel, J., & Maniatopoulos, G. (2005). A comparative analysis of product classification in public vs. private e-procurement. *The Electronic Journal of e-Government (EJEG)* (4), 201–212.
LOD2 European Project FP7. (August 2010). LOD2 – Creating knowledge out of interlinked data. <http://lod2.eu>.
Lonsdale, D., Embley, D. W., Ding, Y., Xu, L., & Hepp, M. (2010). Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering, 69*(4), 318–330.
Maali, F., & Cyganiak, R. (2011). Re-using cool URIs: Entity reconciliation against LOD hubs. *Library*, 8.
Meimaris, M., Xidias, G., Vafeiadis, G., Vafolopoulos, M., Rodríguez, J. M. A., & Klonaras, M. (July 2013). Publicspending.net: it's time to follow the trillions. In *Samos 2013 summit on digital innovation for government, business and society*, Samos, Greece.
Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents, In *I-SEMANTICS* (pp. 1–8).
Monti, M. (2010). A new strategy for the single market – At the service of europe's economy and society. <http://ec.europa.eu/bepa/pdf/monti_report_final_10_05_2010_en.pdf urlhttp://ec.europa.eu/bepa/pdf/monti_report_final_10_05_2010_en.pdf>.
Moreau, L., & Missier, P. (2011). The PROV data model and abstract syntax notation. W3C working draft, W3C. <http://www.w3.org/TR/2011/WD-prov-dm-20111018/>.
Nie, J.-Y. (2003). Query expansion and query translation as logical inference. *Journal of the American Society for Information Science and Technology, 54*(4), 335–346.
Omelayenko, B., & Fensel, D. (2001). An analysis of B2B catalogue integration problems. In *Proceedings of the international conference on enterprise information systems (ICEIS-2001)*, Setúbal, Portugal.
Open Knowledge Foundation. (June 2013). Open spending: Mapping the money. <http://openspending.org/>.
Pietriga, E., Bizer, C., Karger, D., & Lee, R. (2006). Fresnel – A browser-independent presentation vocabulary for RDF. In *Proceedings of the second international workshop on interaction design and the Semantic Web* (pp. 158–171). Springer.

PlanetData European Project FP7. (June 2013). PlanetData. A European network of excellence on large-scale data management. <http://planet-data.eu/>.

Podlogar, M. (2007). *Chapter III e-Procurement success factors: Challenges and opportunities for a small developing country*. *e-Procurement in emerging economies theory and cases*. Pani and Agrahari, Group Publishing.

Polo, L., Alvarez, J. M., & Rubiera, E. (2008). Promoting government controlled vocabularies for the Semantic Web: The EUROVOC thesaurus and the CPV product classification system. In *Proceedings of the first workshop on semantic interoperability in the european digital library (SIEDL)*.

Pound, J., Mika, P., & Zaragoza, H. (2010). Ad-hoc object retrieval in the web of data. In *Proc. of the 19th WWW, WWW '10* (pp. 771–780). New York, NY, USA,: ACM.

Preece, S. E. (1981). *A spreading activation network model for information retrieval*. PhD thesis, University Illinois, Urbana, IL, USA.

Prud'hommeaux, E., & Seaborne, A. (2007). SPARQL query language for RDF. W3C recommendation, W3C. <http://www.w3.org/TR/rdf-sparql-query/>.

Rocha, C., Schwabe, D., & Aragao, M. P. (2004). A hybrid approach for searching in the Semantic Web. In *Proceedings of the 13th international conference on World Wide Web, WWW '04* (pp. 374–383). New York, NY, USA: ACM.

Sauermann, L., Cyganiak, R., & Völkel, M. (February 2007) Cool URIs for the Semantic Web. Technical Memo TM-07-01, DFKI GmbH. Written by 29.11.2006.

Schmitz, V., & Leukel, J. (2005). Findings and recommendations from a pan-European research project: Comparative analysis of E-catalog standards. *International Journal of IT Standards and Standardization Research (JITSR), Idea Group* (2), 51–65.

Schumacher, K., Sintek, M., & Sauermann, L. (2008). Combining fact and document retrieval with spreading activation for semantic desktop search. In *Proceedings of the 5th European Semantic Web conference on the Semantic Web: Research and applications. ESWC'08* (pp. 569–583). Berlin, Heidelberg: Springer-Verlag.

Soergel, D. (2005). Thesauri and ontologies in digital libraries. In *JCDL* (pp. 421).

Stadler, C., Guéret, C., Lehmann, J., Groth, P., & Jentzsch, A. (2011). *D1.4.1 first deployment of QA module*. Deliverable of European project, LATC project.

TaskForce Linking Open Data DataSets. (December 2011). *Guidelines for collecting metadata on linked datasets in CKAN*. W3c working group note.

TenForce. (2011). *D1.4 – Early LOD2 stack prototype*. Deliverable of European project, LOD2 project. <http://static.lod2.eu/Deliverables/LOD2_D1.4_Early_Stack_Prototype.pdf>.

Troussov, A., Judge, J., & Botvich, D. (2008). Mining socio-semantic networks using spreading activation technique. In *Proceedings of the 9th international conference on knowledge management and knowledge technologies, international workshop on knowledge acquisition from the social web (KASW'08)*.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., & Decker, S. (2010). Sig.ma: Live views on the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web, 8*(4), 355–364.

Urbani, J., Kotoulas, S., Maassen, J., van Harmelen, F., & Bal, H. E. (2012). WebPIE: A web-scale parallel inference engine using mapreduce. *Journal of the Web Semantic, 10*, 59–75.

van Assem, M., Malaisé, V., Miles, A., & Schreiber, G. (2006). A method to convert thesauri to SKOS. In *ESWC* (pp. 95–109).

van Rooyen, G. -J., Gouws, S., & Engelbrecht, H. A. (August 2010). Measuring conceptual similarity by spreading activation over wikipedia's hyperlink structure. In *Proceedings of the 2nd workshop on the people's web meets NLP: Collaboratively constructed semantic resources* (pp. 46–54). Beijing, China.

W3C Members. (2011). Best practices discussion summary. W3C government linked data working group first f2f, W3C. <http://www.w3.org/2011/gld/wiki/Best_Practices_Discussion_Summary>.