Towards a Knowledge Graph based Platform for Public Procurement*

Elana Simperl[1], Oscar Corcho[2], Marko Grobelnik[3], Dumitru Roman[4], Ahmet Soylu[4,**], María Jesús Fernández Ruíz[5], Stefano Gatti[6], Chris Taggart[7], Urška Skok Klima[8], Annie Ferrari Uliana[9], Ian Makgill[10], and Till Christopher Lech[4]

[1] University of Southampton, Southampton, the UK
[2] Universidad Politécnica de Madrid, Madrid, Spain
[3] Jožef Stefan Institute, Ljubljana, Slovenia
[4] SINTEF Digital, Oslo, Norway
[5] Ayuntamiento de Zaragoza, Zaragoza, Spain
[6] Cerved Group Spa US, Milano, Italy
[7] OpenCorporates Ltd, London, the UK
[8] Ministrstvo za javno upravo, Ljubljana, Slovenia
[9] OESIA Networks SL, Madrid, Spain
[10] OpenOpps Ltd, London, the UK

Abstract. Procurement affects virtually all sectors and organizations particularly in times of slow economic recovery and enhanced transparency. Public spending alone will soon exceed EUR 2 trillion per annum in the EU. Therefore, there is a pressing need for better insight into, and management of government spending. In the absence of data and tools to analyse and oversee this complex process, too little consideration is given to the development of vibrant, competitive economies when buying decisions are made. To this end, in this short paper, we report our ongoing work for enabling procurement data value chains through a knowledge graph based platform with data management, analytics, and interaction.

Keywords: Procurement · Knowledge graphs · Analytics · Interaction.

1 Introduction

Procurement affects virtually all sectors and organisations. In the public sector alone, spending on goods and services across the EU is estimated to exceed €2 trillion per annum[1]. Governments and state-owned enterprises are confronted with massive challenges: they must deliver services with greatly reduced budgets; prevent losses through fraud and corruption; and build healthy, sustainable economies. Managing these competing priorities is notoriously difficult. In times of slow economic recovery and enhanced transparency and accountability, there is a pressing need for better insight into, and management of government spending

* This work is funded by EU H2020 TheyBuyForYou project (780247).
** Corresponding author. Email: ahmet.soylu@sintef.no
[1] http://ec.europa.eu/DocsRoom/documents/20679

(cf. [1]). In the absence of data and tools to analyse and oversee this complex process, too little consideration is given to the development of vibrant, competitive economies when buying decisions are made.

To this end, within an EU project called TheyBuyForYou[2], we work towards enabling the procurement data value chains through a knowledge graph based platform paired with data management, analytics, and interaction. Our objective, from a technical point of view, is to build a technology platform, consisting of a set of modular, web-based services and APIs to publish, curate, integrate, analyse, and visualize a comprehensive, cross-border and cross-lingual procurement knowledge graph, including public spending and corporate data from multiple sources across the EU. In this context, the first challenge is the heterogeneity of the underlying data (cf. [5,10]), which covers structured (e.g., statistics, financial news) as well as unstructured (e.g., text, social media) sources in different languages and using their own terminology and formats (CSV, PDF, databases, websites, APIs etc.). A second challenge will be turning this vast array of information into a semantic knowledge graph [17,18], an interconnected semantic knowledge organization structure using Web URIs and linked data vocabularies, which can be analysed in depth to identify patterns and anomalies in procurement processes and networks. Finally, the last challenge is to support analytics and interaction with the knowledge graph by different groups of stakeholders.

In this short paper, we report our ongoing work. The rest of the paper is structured as follows. In Sect. 2, we present an overview of the related work. We present key challenges and our solution approach in Sect. 3 and Sect. 4 respectively, and finally we conclude the paper in Sect. 5.

2 Related Work

Although there is no single definition of a knowledge graph, from a broader perspective, we consider it as a graph describing real world entities and their interrelations (cf. [12,18]). They have become powerful assets for representing, storing, sharing, and accessing information both in academia, such as Yago [3] and DBpedia [9], and in industry such as Google's Knowledge Graph, Facebook's Graph Search, and Microsoft's Satori [17]. Different approaches and technologies could be used to construct knowledge graphs, such as ad-hoc approaches focusing on graph-oriented data storage and management built on relational and NoSQL databases and semantic approaches built on the Semantic Web technologies [8,18]. The Semantic Web offers relevant standards and technologies for representing, exchanging, and querying knowledge graphs. Typically, semantic knowledge graphs are stored or exported as RDF datasets and queried with SPARQL query language, while the semantics of such datasets are encoded in OWL 2 ontologies allowing logic-based reasoning[3].

There are various initiatives whose purpose is to create de-jure and de-facto standards for electronic procurement, including such as Open Contracting Data

[2] `https://theybuyforyou.eu`

[3] `https://www.w3.org/standards/semanticweb/`

Standard (OCDS)[4] and TED eSenders [5]. However, these are mostly oriented to achieve interoperability (i.e., addressing communication between systems), document oriented (i.e., the structure of the information is commonly provided by the content of the documents that are exchanged), and provide no standardised practices to refer to third parties, companies participating in the process, or even the main object of contracts. This at the end generates a lot of heterogeneity. Procurement domain can take advantage of applying the Semantic Web approach by reusing existing vocabularies, ontologies, and standards [1]. Specifically in the procurement domain, these include among others PPROC ontology [10] for describing public processes and contracts, LOTED2 ontology [5] for public procurement notices, PCO ontology [11] for contracts in public domain, and MOLDEAS ontology [14] for announcements about public tenders. LOTED2 is considered as a legal ontology and is comparatively more complex and detailed with respect to MOLDEAS, PCO, and PPROC. The latter is concerned on reaching a balance between usability and expressiveness.

3 Challenges

The technical landscape for managing contracts is very complex, for example, contracts are handled using different tools and formats across departments. Procurement data has a large scale and covers structured as well as unstructured data sources in different languages, terminology, and formats. By delivering a knowledge graph integrating a variety of procurement related datasets and an open-source platform and APIs for decision support and analytics, a wide range of procurement business cases could be supported. However, there are challenges to be addressed.

3.1 Data Integration

Apart from lack of a common schema, there are several other problems due to ingestion of heterogeneous and often noisy data sources. Firstly, there could be duplicates (e.g., seemingly two different tenders actually refer to the same tender) and missing information. Secondly, mapping entities referred to in data sets is also considerable challenge. For example, linking company data across data sets needs to take into account that not only companies change names frequently, but names are reused; addresses are in multiple forms and often refer to trading or administrative addresses rather than registered addresses; and legal names are represented in multiple forms or languages.

3.2 Data Analysis

Firstly, procurement documentation is typically available in the native language of the issuing organisation. Even in multi-lingual sources such as the Tenders

[4] `http://standard.open-contracting.org/latest/en/`
[5] `http://simap.ted.europa.eu/`

Electronic Daily (TED), the documents in the language of the issuing country are a lot more detailed than their translations. Therefore, comparing and linking documents across languages for various analysis purposes, such as cross-lingual document clustering by topic, is challenging. Secondly, the stream of public spending documentation from governments is a large data source and requires real-time automatic analysis approaches. For example, TED alone is updated with roughly 1500 public procurement notices five times a week.

3.3 Human-Data Interaction

Buyers, suppliers, journalists, and citizens need tools to understand the complex space of procurement data. Potential solutions should go beyond the interactive visualizations offered by the most existing analytics software, better match individual information needs, be integrated with real-time automated analytics, and master the inherent properties of the underlying data, which is very diverse in terms of structure, vocabularies, language, and modality. Finally, there is also a need for improved methods to create interactive visualizations that communicate findings intelligibly. Existing tools often tell an implicit story and are difficult to replicate or integrate with other analytics sources.

4 TheyBuyForYou Approach

We are building an integrated information hub with a semantic knowledge graph for public procurement and company data in order to support: (i) economic development by facilitating better outcomes from public spending for SMEs; (ii) demand management by spotting trends in public spending to achieve long-term goals such as savings; (iii) competitive markets by promoting healthier competition and identifying collusions and other irregularities; (iv) and supplier intelligence by advanced analytics. We will address the aforementioned challenges as described in the followings.

4.1 Knowledge Graphs

We are creating a knowledge graph primarily integrating and linking two large open databases: core company data provided by OpenCorporates[6] and tenders and contracts data provided by OpenOpps[7] in the OCDS format. Once the data is extracted from the aforementioned databases through an extract-transform-load (ETL) process, it needs to be integrated into a common ontology built on existing ontologies and vocabularies, possibly with extensions, and entities need to be linked across data sets. The resulting knowledge graph will be published through open APIs and a SPARQL endpoint. A range of data curation mechanisms need to be applied, including the identification and resolution of duplicates, and the

[6] `https://opencorporates.com`
[7] `https://openopps.com`

completion of missing information by using heuristic and machine learning (ML) approaches as well as manual input from citizens. Additional links between the data sources, for instance between entities in procurement documents and data and the entities coming from business registries, could be discovered through using data reconciliation algorithms, a combination of heuristics, explicit sets of rules, and clustering algorithms (cf. [6,2]).

4.2 Cross-lingual and Real-time Analytics

Cross-lingual document comparison and linking approaches could be adapted to public spending documentation (cf. [7]). In natural language processing, a typical approach to document analysis is to represent each document as a vector of semantic concepts and terms from the document resulting in different vector space for each language. In order to compare vectors from different spaces, we use an approach based on canonical correlation analysis to find a mapping to a common semantic vector space, which preserves document similarities within individual languages while enabling their comparison across languages. As a first application, we are implementing anomaly detection, also known as outlier detection. It is the task of identifying data points or groups of data points that in some sense diverge from normal behaviour (cf. [4]). This is highly relevant in the domain of public spending, where we are interested in both individual exceptional cases as well as large and systematic patterns standing out from the norm, whether they represent examples of good public procurement practice or possible cases of corruption.

4.3 Data Visualisation, Storytelling and Narratives

A visual solution should allow for exploratory navigation of the content space (in reflection to and exploring temporal developments); following links to the used content and their provenance; and providing links to alternative visualizations. Especially, approaches targeting the visualization and understanding of temporal aspects [15] and dynamics, and the visualization of highly networked content [16] drive our work. We employ storytelling methods to create visualizations by using narratives, which are compact, yet coherent natural language summaries of data [13]. A tool is being developed to create machine-readable specification of infographics, deployed using open standards, e.g., JSON and linked data, and that can be easily shared and linked to other media, maintains links to the identifiers of the data and ontologies they refer to, and integrates cross-lingual real-time analytics capabilities. To enhance the natural language generation methods in use, we will compile a collection of data design patterns from the procurement knowledge graph to generate end-user configurable storification templates.

5 Conclusions

We presented key challenges and a solution approach, using knowledge graphs, data analysis and interaction, for enabling procurement data value chains. We

will also build a series of toolkits and portals on top of our solution to allow various stakeholders to explore and understand how public procurement decisions affect economic development, efficiencies, competitiveness and supply chains.

References

1. Alvarez-Rodríguez, J.M., et al.: New trends on e-Procurement applying semantic technologies: Current status and future challenges. Computers in Industry **65**(5), 800–820 (2014)
2. Araújo, S., et al.: SERIMI: Class-Based Matching for Instance Matching Across Heterogeneous Datasets. IEEE Trans. Knowl. Data Eng. **27**(5), 1397–1410 (2015)
3. Biega, J., et al.: Inside YAGO2s: A Transparent Information Extraction Architecture. In: Proceedings of the 22nd International Conference on World Wide Web (WWW 2013). pp. 325–328. ACM, New York, NY, USA (2013)
4. Chandola, V., et al.: Anomaly Detection: A Survey. ACM Computing Surveys **41**(3), 15:1–15:58 (2009)
5. Distinto, I., et al.: LOTED2: An ontology of European public procurement notices. Semantic Web **7**(3), 267–293 (2016)
6. Dorneles, C.F., et al.: Approximate data instance matching: a survey. Knowledge and Information Systems **27**(1), 1–21 (2011)
7. Fortuna, B., et al.: A Kernel Canonical Correlation Analysis For Learning The Semantics Of Text. In: Kernel methods in bioengineering, communications and image processing (2006)
8. Kharlamov, E., et al.: Ontology Based Data Access in Statoil. Web Semantics: Science, Services and Agents on the World Wide Web **44**, 3–36 (2017)
9. Lehmann, J., et al.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015)
10. Muñoz-Soro, J.F., et al.: PPROC, an ontology for transparency in public procurement. Semantic Web **7**(3), 295–309 (2016)
11. Necaský, M., et al.: Linked data support for filing public contracts. Computers in Industry **65**(5), 862–877 (2014)
12. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web **8**(3), 489–508 (2017)
13. Portet, F., et al.: Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence **173**(7), 789–816 (2009)
14. Rodríguez, J.M.Á., et al.: Towards a Pan-European E-Procurement Platform to Aggregate, 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 (2012)
15. Shanbhag, P., et al.: Temporal Visualization of Planning Polygons for Efficient Partitioning of Geo-Spatial Data. In: Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2005). IEEE Computer Society, Washington, DC, USA (2005)
16. Smith, M.A., et al.: Analyzing (Social Media) Networks with NodeXL. In: Proceedings of the 4th International Conference on Communities and Technologies. pp. 255–264. ACM, New York, NY, USA (2009)
17. Suchanek, F.M., et al.: Knowledge Bases in the Age of Big Data Analytics. Proceedings of the VLDB Endowment **7**(13), 1713–1714 (2014)
18. Yan, J., et al.: A Retrospective of Knowledge Graphs. Frontiers of Computer Science **12**(1), 55–74 (2018)