Integrating and Analysing Public Procurement Data through a Knowledge Graph: A Demonstration in a Nutshell

Ahmet Soylu[1], Oscar Corcho[2], Brian Elvesæter[1], Carlos Badenes-Olmedo[2], Francisco Yedro Martínez[2], Matej Kovacic[3], Matej Posinkovic[3], Ian Makgill[4], Chris Taggart[5], Elena Simperl[6], Till C. Lech[1], and Dumitru Roman[1]

[1] SINTEF AS, Oslo, Norway
{firstname.lastname}@sintef.no
[2] Universidad Politécnica de Madrid, Madrid, Spain
[3] Jožef Stefan Institute, Ljubljana, Slovenia
[4] OpenOpps Ltd, London, the UK
[5] OpenCorporates Ltd, London, the UK
[6] King's College London, London, the UK

Abstract. This paper presents a demonstrator of a knowledge graph based approach for integrating and reconciling cross-border and cross-language procurement and company data from distributed data sources. The demonstrator also includes analysis of the resulting knowledge graph, exemplified in anomaly detection and cross-lingual search.

Keywords: Public procurement · Knowledge graph · Linked data.

1 Introduction

The availability of high quality, open, and linked procurement data presents an opportunity to enhance the public procurement processes. In this respect, several directives were put forward by the European Commission (e.g., Directive 2003/98/EC and Directive 2014/24/EU8), which led to the emergence of national and international public procurement portals. However, there is a lack of common agreement across the European Union (EU) on the data formats for exposing such data sources and on the data models for representing such data, leading to a highly heterogeneous technical landscape.

To this end, in order to deal with the technical heterogeneity and to connect disparate data sources currently created and maintained in silos, we developed a platform consisting of a set of modular REST APIs and ontologies, to publish, curate, integrate, analyse, and visualise an EU-wide, cross-border, and cross-lingual procurement knowledge graph (KG). This paper presents a demonstrator for the knowledge graph based platform and end-user tools for integrating and reconciling procurement and company data from distributed data sources, including analytics tools exemplified in anomaly detection and cross-lingual search [4].

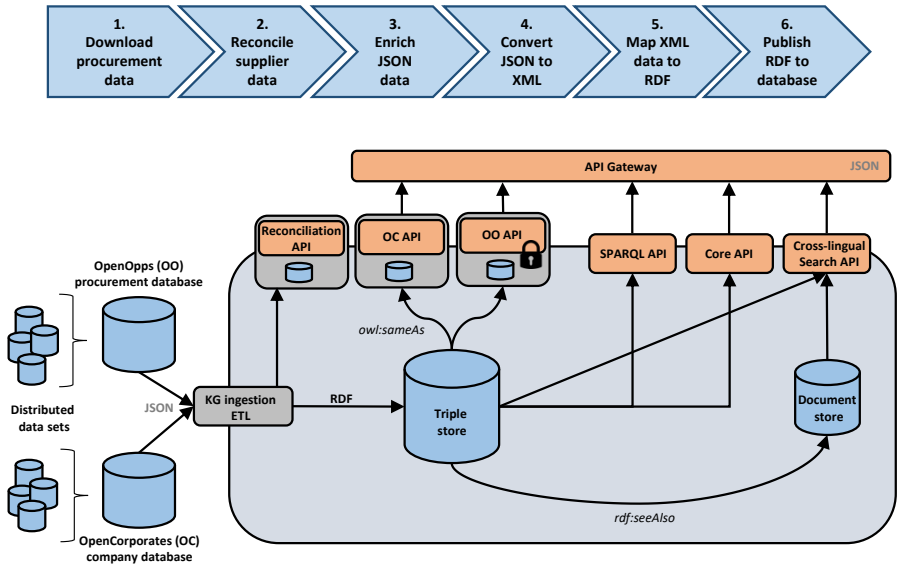

Fig. 1. Data ingestion process and platform architecture.

2 Data Integration

Procurement and company data underlying the KG is provided by two main data providers: OpenOpps[1] for procurement data (e.g., tenders and contracts) and OpenCorporates[2] for supplier data (i.e., companies). OpenOpps has gathered over three million tender documents from more than 685 publishers through Web scraping and by using open APIs, while OpenCorporates currently has 140 million entities collected from national registers. We integrated the two high-quality data sets according to an ontology network to form a knowledge graph. The ontology network includes an ontology for representing procurement data based on Open Contracting Data Standard (OCDS), namely the OCDS ontology[3] [3], and another ontology for representing company data, namely the euBusinessGraph ontology[4] [2].

The data ingestion process (see Fig. 1) comprises several steps using data APIs of both providers, including data curation, matching suppliers appearing in tender data against company data (i.e., reconciliation), and translating data sets into the underlying graph data representation (i.e., RDF) with respect to the ontology network and linked data principles. The platform (see Fig. 1) employs a triple store for the generated RDF-based data, linked to original data sources (using `owl:sameAs`), and a document store for the documents associated with

[1] `https://openopps.com`
[2] `https://opencorporates.com`
[3] `https://github.com/TBFY/ocds-ontology/tree/master/model`
[4] `https://github.com/euBusinessGraph/eubg-data`

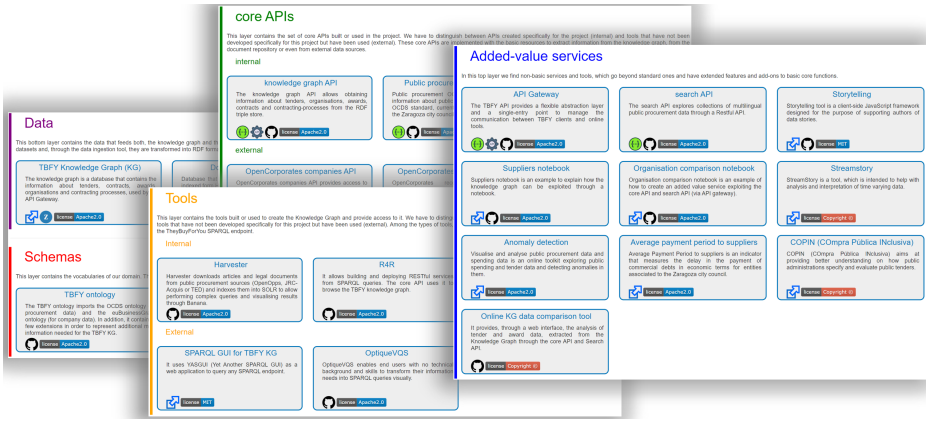

Fig. 2. The catalogue for data, schemas, core APIs, tools, and added value services.

the public procurement data (using `rdfs:seeAlso`). The data is made available through a SPARQL API, a core REST-based API (i.e., KG API), a cross-lingual search API, and an API gateway providing a single entry point to the previously mentioned APIs. The current release of the KG covers data from January 2019 onwards. New data is onboarded on a daily basis. As of August 2020, the KG consists of more than 126 million triples and contains information about 1,31 million tenders, 1,54 million awards, and more than 99 thousand companies[5]. The source data collected from the data providers in JSON and the KG data in RDF are made openly available under the Open Database License (ODbl)[6] on Zenodo[7]. An online catalogue is available[8] providing access to data, schemas, core APIs, tools, and added value services (see Fig. 2).

3 Data Analysis

We implemented a number of analysis techniques on the KG: anomaly detection by using ML techniques for identifying patterns and anomalies, such as fraudulent behaviour or monopolies in procurement processes and networks across data sets produced independently; and, cross-lingual document search for finding documents that are similar to a given one independently of its language.

Public procurement is particularly susceptible to corruption, which can impede economic development, create inefficiencies, and reduce competitiveness. At the same time, manually analysing a large volume of procurement cases is not feasible. Therefore, firstly, we applied several ML techniques, i.e., supervised, unsupervised, and statistical, on top of the Slovenian public procurement data in the KG to

[5] `http://data.tbfy.eu`
[6] `https://opendatacommons.org/licenses/odbl`
[7] `https://github.com/TBFY/data-sources`
[8] `https://tbfy.github.io/platform/`

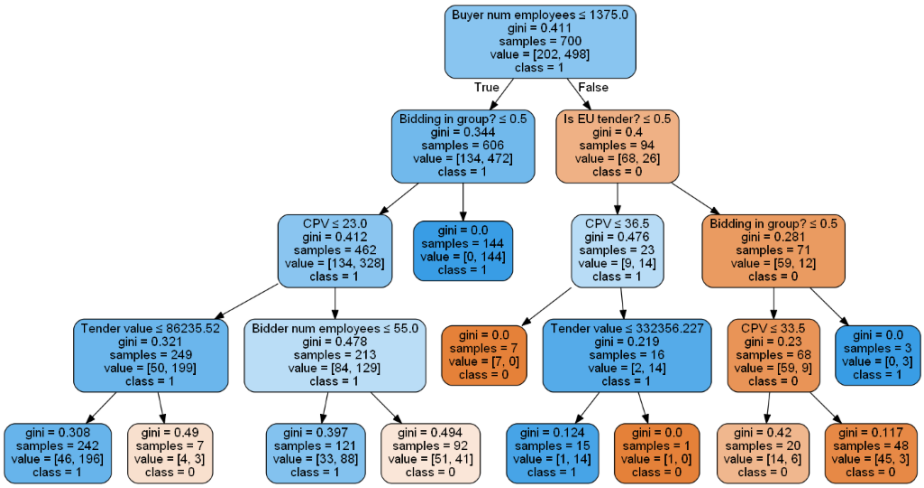

Fig. 3. The decision tree model for identifying successful tenders.

identify patterns and anomalies. We implemented a system and made it available online[9]. The system developed is capable of processing tens of millions of records and allows detecting a large class of anomalies. Fig. 3 depicts the supervised analysis approach implemented in our platform based on a decision tree. Users select parameters by their own choice (e.g., buyer size and bidder municipality), and explore various parameters contributing to the success of public tenders.

Procurement processes are not only creating structured data, but also constantly creating additional documents. These are commonly published in the official language of the corresponding public administrations. Only some of these are multilingual, but the documents in the local language are typically longer. Therefore, secondly, we worked on an added-value service[10] with the possibility of finding documents that are similar to a given one independently of the language in which it is made available [1]. We also generated a Jupyter notebook with some representative examples, so as to facilitate its use[11]. This service (see Fig. 4) is based on the use of unsupervised probabilistic topic models, using cross-lingual labels from sets of cognitive synonyms (synsets) to establish relations between language-specific topics.

4 Conclusions

In this paper, we demonstrated the use of Semantic Web and Linked Data technologies and principles to integrate open procurement and company data sets, and advanced analytics and to unlock their value. The KG enabled easier

[9] `http://tbfy.ijs.si`
[10] `http://tbfy.librairy.linkeddata.es/search-api`
[11] `http://bit.ly/tbfy-search-demo`

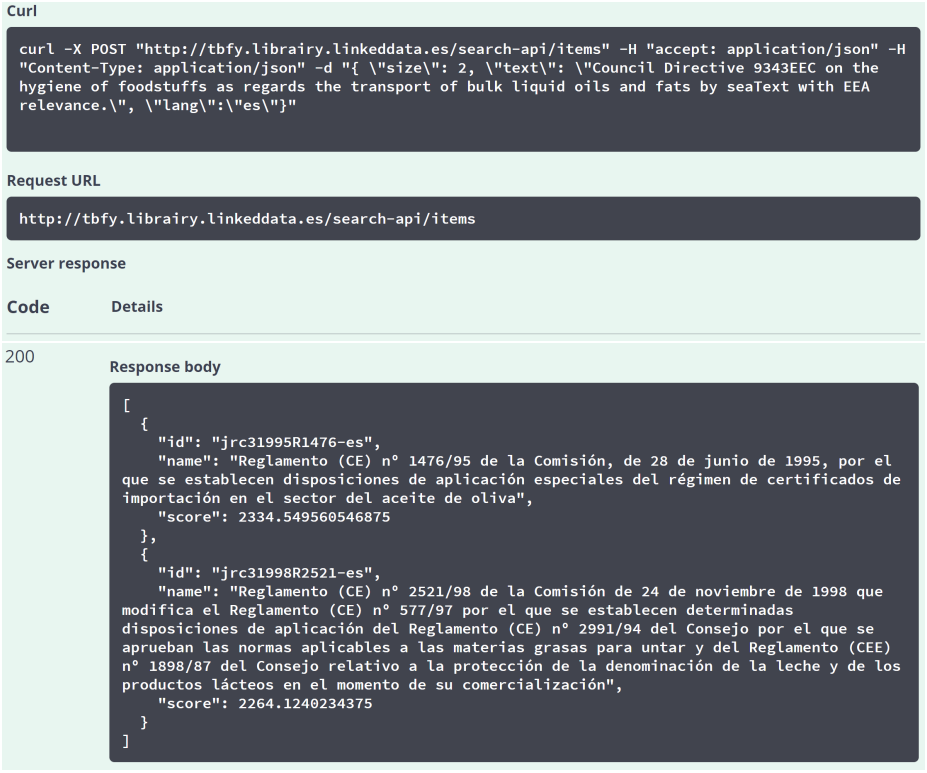

Fig. 4. A call to the cross-lingual search API and the results returned.

and advanced analytics, which was otherwise not possible. However, we also faced a high number of data quality issues, such as missing, duplicate, and erroneous data, even though there are mandates in place for buyers to provide correct data.

Acknowledgements. The work reported in this paper is partly funded by EC H2020 TheyBuyForYou (780247) and euBusinessGraph (grant 732003) projects.

References

1. Badenes-Olmedo, C., et al.: Scalable Cross-lingual Similarity through language-specific Concept Hierarchies. In: Proc. of K-CAP 2019. pp. 147–153 (2019)
2. Roman, D., et al.: The euBusinessGraph Ontology: a Lightweight Ontology for Harmonizing Basic Company Information. Semantic Web **(under review)** (2020)
3. Soylu, A., et al.: Towards an Ontology for Public Procurement Based on the Open Contracting Data Standard. In: Proc. of I3E 2019. pp. 230–237 (2019)
4. Soylu, A., et al.: Enhancing Public Procurement in the European Union through Constructing and Exploiting an Integrated Knowledge Graph. In: Proc. of ISWC 2020 (2020)