An Overview of the TBFY Knowledge Graph for Public Procurement*

Ahmet Soylu[1], Brian Elvesæter[1], Philip Turk[1], Dumitru Roman[1], Oscar Corcho[2], Elena Simperl[3], Ian Makgill[4], Chris Taggart[5], Marko Grobelnik[6], and Till C. Lech[1]

[1] SINTEF Digital, Oslo, Norway
[2] Universidad Politécnica de Madrid, Madrid, Spain
[3] University of Southampton, Southampton, the UK
[4] OpenOpps Ltd, London, the UK
[5] OpenCorporates Ltd, London, the UK
[6] Jožef Stefan Institute, Ljubljana, Slovenia

Abstract. A growing amount of public procurement data is being made available in the EU for the purpose of improving the effectiveness, efficiency, transparency, and accountability of government spending. However, there is a large heterogeneity, due to the lack of common data formats and models. To this end, we developed an ontology network for representing and linking tender and company data and ingested relevant data from two prominent data providers into a knowledge graph, called TBFY. In this poster paper, we present an overview of our knowledge graph.

Keywords: Public procurement · Knowledge graph · Ontology.

1 Introduction

In the EU, public authorities spend around 14% of GDP on the purchase of services, works, and supplies every year[7]. Therefore, a growing amount of public procurement data is being made available in the EU through public portals for the purpose of improving the effectiveness, efficiency, transparency, and accountability of government spending. However, there is a large heterogeneity, due to the lack of common data formats and models for exposing such data.

There are various standardization initiatives for electronic procurement, such as Open Contracting Data Standard (OCDS)[8] and TED eSenders [9]. However, these are mostly oriented to achieve interoperability, document-oriented, and provide no standardised practices to refer to third parties, companies participating in the process, etc. This again generates a lot of heterogeneity. The Semantic Web

[7] https://ec.europa.eu/growth/single-market/public-procurement_en
[8] http://standard.open-contracting.org/latest/en/
[9] http://simap.ted.europa.eu/

approach has been proposed as a response [1]. For example, several ontologies have been developed, such as PPROC ontology [3] for describing public processes and contracts, LOTED2 ontology [2] for public procurement notices, PCO ontology [4] for contracts in public domain, and MOLDEAS ontology [5] for announcements about public tenders. Each of these was developed with different concerns in mind (legal, process-oriented, etc.) without significant adoption so far.

To this end, we developed an ontology network for representing and linking tender and company data and ingested relevant data from two prominent data providers into a knowledge graph, called TBFY. In this poster paper, we present an overview of our knowledge graph for public procurement.

2 Knowledge Graph

We integrated two datasets according to an ontology network: tender data provided by OpenOpps[10] in the OCDS format and company data provided by OpenCorporates[11]. OpenOpps has gathered over 2M tender documents from more than 300 publishers through Web scrapping and by using open APIs, while OpenCorporates currently has 140M entities collected from national registers.

2.1 Ontology Network

We are currently using two main ontologies. First, an ontology for tender data (see Figure 1) that we developed using the OCDS' data model[12].

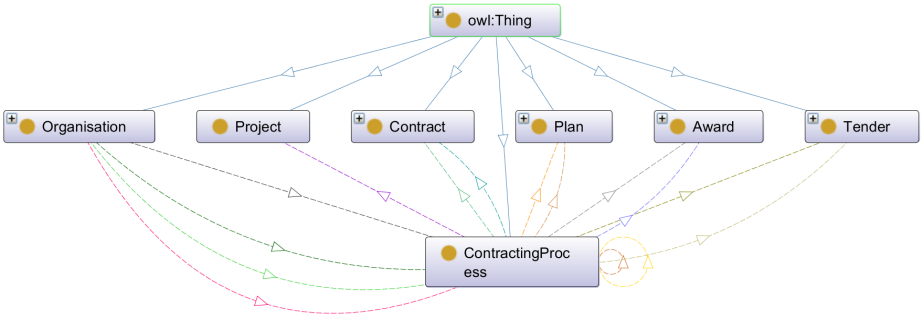

Fig. 1. A fragment of OCDS ontology depicting some of the key classes.

Second, we reused the euBG ontology for company data[13]. Both ontologies reuse other ontologies and vocabularies (FOAF, Dublin Core, etc.).

[10] https://openopps.com
[11] https://opencorporates.com
[12] https://github.com/TBFY/ocds-ontology
[13] https://github.com/euBusinessGraph/eubg-data

2.2 Data Ingestion

The data ingestion process is composed of several steps using data APIs of both providers (see Figure 2). Initially, company data is extracted from OpenOpps for a given period of time and preprocessed primarily to handle null values. Suppliers appearing in tender data are matched against company data provided by OpenCorporates by using the reconciliation service of OpenCorporates. The matched company data is extracted and then supplier data is annotated with the corresponding company identifiers.

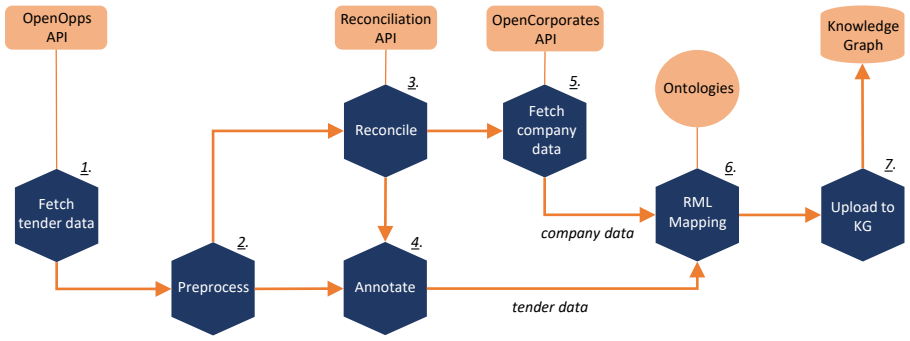

Fig. 2. Data ingestion process.

Finally, the extracted datasets are translated to RDF using the RDF Mapping Language (RML)[14] according to our ontology network. The supplier data and company data is linked through the `owl:sameAs` property. Thereafter data is uploaded to a graph database, namely GraphDB[15].

2.3 Current Release

The current release of the knowledge graph includes 23M triples originating from tender data collected initially for the first quarter of 2019. The knowledge graph is available online[16]. An example query and its results are depicted in Figure 3. The example query lists top ten companies in the Norwegian jurisdiction that have the highest number of supplier role, where the jurisdiction data comes from OpenCorporates and contract data comes from OpenOpps.

3 Open Issues

Currently, one of the main issues concerns data quality including missing, duplicate, poorly formed, and erroneous data. For example, the missing address

[14] `http://rml.io/`
[15] `http://graphdb.ontotext.com/`
[16] `http://data.tbfy.eu`

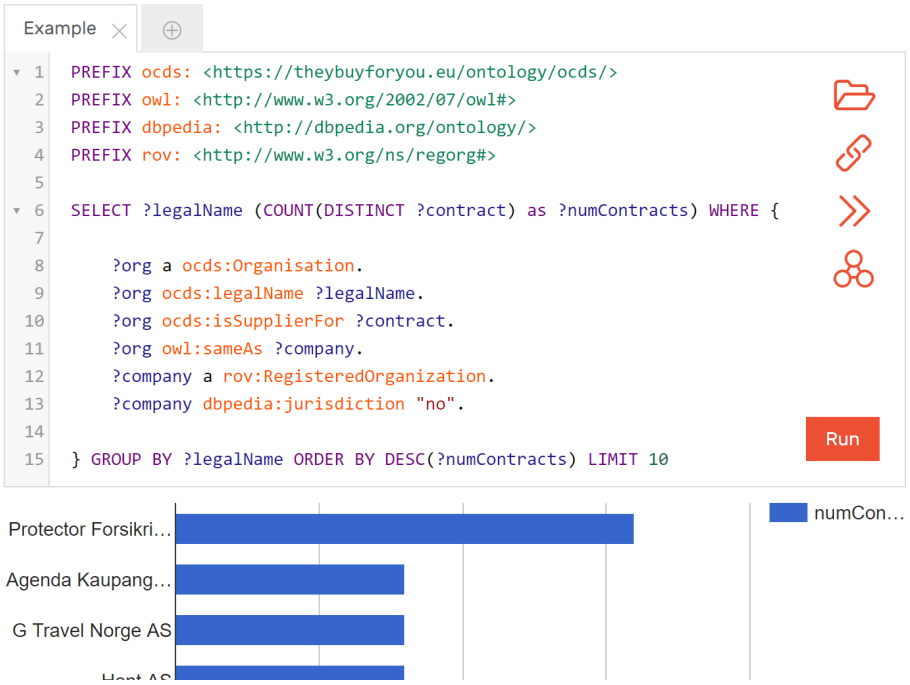

Fig. 3. An example query executed on the TBFY knowledge graph.

information, variations on address format, etc. hinder the quality of reconciliation process. Currently, we are working on various approaches to improve data quality ranging from machine learning to crowd-sourcing.

Acknowledgements. This work has been partly funded by the EU H2020 projects TheyBuyForYou (780247) and euBusinessGraph (732003).

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. Distinto, I., et al.: LOTED2: An ontology of European public procurement notices. Semantic Web **7**(3), 267–293 (2016)
3. Muñoz-Soro, J.F., et al.: PPROC, an ontology for transparency in public procurement. Semantic Web **7**(3), 295–309 (2016)
4. Necaský, M., et al.: Linked data support for filing public contracts. Computers in Industry **65**(5), 862–877 (2014)
5. 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)