Towards Integrating Public Procurement Data into a Semantic Knowledge Graph*

Ahmet Soylu[1,**], Oscar Corcho[2], Elena Simperl[3], Dumitru Roman[1], Francisco Y. Martínez[2], Chris Taggart[5], Ian Makgill[4], Brian Elvesæter[1], Ben Symonds[5], Helen McNally[4], George Konstantinidis[3], Yuchen Zhao[3], 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

Abstract. Public procurement accounts for a substantial part of the public investment and global economy. Therefore, improving effectiveness, efficiency, transparency and accountability of government procurement is of broad interest. To this end, in this poster paper, we present our approach for integrating procurement data, including public spending and corporate data, from multiple sources across the EU into a semantic knowledge graph. We are aiming to improve procurement processes through supporting multiple stake holders, such as government agencies, companies, control authorities, journalists, researchers, and individual citizens.

Keywords: Knowledge graph · Public procurement · Ontology.

1 Introduction

Public procurement accounts for a substantial part of the public investment and global economy. Every year, over 250 000 public authorities in the EU spend around 14% of GDP on the purchase of services, works and supplies[1]. Therefore, improving effectiveness, efficiency, transparency and accountability of government procurement is of broad interest [1]. To this end, European Commission has put several relevant directives forward, i.e., for public sector information (e.g., Directive 2003/98/EC) and public procurement (e.g., Directive 2014/24/EU8), to improve public procurement practices. As a result of these, national public procurement portals have been created, which live together with regional, local as well as EU-wide public procurement portals. However, there is no common agreement across the EU (not even, in many cases, inside the same country) on the data formats to be used for exposing such data sources and on the data models that need to be used for exposing such data, which leads to a large heterogeneity in the data that is being exposed.

* This work is funded by EU H2020 TheyBuyForYou project (780247).
** Corresponding author. Email: ahmet.soylu@sintef.no
[1] https://ec.europa.eu/growth/single-market/public-procurement_en

In Europe, contracting portals like Tenders Electronic Daily[2] (TED) may be seen as a way to homogenise the data that is being provided, but unfortunately this portal is only used for those contracts that are larger than a predefined budget threshold, and hence this does not cover the whole richness of types of public contracts nor does it force the usage of this format for those contracts that do not need to be published there. The only relevant data model that is getting some important traction worldwide is the Open Contracting Data Standard[3] (OCDS). However, it has been mostly developed with a focus on transparency in the public contracting procedures. Though, several ontologies, such as LOTED2 [2], PPROC [3], PCO[4] and upcoming eProcurement ontology[4], are developed with different levels of detail and focus for representing procurement data, there is no solution integrating supplier and procurement data enabling such as matching of suppliers and buyers and advanced analytics and procurement intelligence.

In this poster paper, we present our approach, in the context of TheyBuyForYou[5] project, for integrating procurement data, including public spending and corporate data, from multiple sources across the EU into a knowledge graph. We are aiming to improve procurement processes through supporting multiple stake holders, such as government agencies, companies, control authorities, journalists, researchers, and individual citizens. The proposed solution enables developers to create fully functional, robust, and scalable data integration pipelines, from including sourcing the data, to pre-processing, augmenting, and interlinking it.

2 Data Sources

High-quality company (i.e., legal entities) and procurement (e.g., tenders and contracts) data are needed to form an interconnected knowledge graph for public procurement. However, firstly, in public procurement the vast majority of external government spending (i.e., not government-to-government) is with companies and often there is no explicit and unambiguous reference to the legal entities in the government's own records. Secondly, to truly understand the scope of procurement data across the EU, we must go through a process of identifying and recording data sources that exist alongside the formal TED in Europe, such as procurement transparency initiatives of individual countries including data from tender alert sites, contract registers and spending data. In our context, we collect this information from two main providers, that is OpenCorporates[6] for supplier data (i.e., company) and OpenOpps[7] for procurement data.

OpenCorporates makes data on 140 million legal entities, resulting in the order of 100s of GB data, available through an API. Data is collected from national company registers and other regulatory sources. OpenCorporates uses

[2] `http://ted.europa.eu`
[3] `http://standard.open-contracting.org`
[4] `https://github.com/eprocurementontology/eprocurementontology`
[5] `https://theybuyforyou.eu`
[6] `https://opencorporates.com`
[7] `https://openopps.com`

a variety of methods of data extraction, depending on the format of the source data. Where structured data files are available they are imported, although some scraping is required from less structured sources. OpenCorporates' company data is mapped to its own schema and inactive companies and sole traders are identified and categorised where possible. OpenOpps is gathering tender and contract data from European sites like TED as well as many large national portals, over 300GB of data from over 450 European sources, and makes over 2 million documents, dating back to 2010, available through an API. Included in this data is details on buyers, suppliers (for contracts), titles, descriptions, values and categories. OpenOpps extracts data from these sources using scraper scripts and the extracted data is formatted according to the OCDS. Data is augmented with Common Procurement Vocabulary[8] (CPV) codes where it is not available (i.e., used for classifying the subjects of procurement contracts). Tender notice documents are gathered and referenced whenever possible.

In the context of our work, OpenOpps and OpenCorporates maintain their own code for validating, mapping and monitoring the data. Currently, more company registers from new jurisdictions, such as Germany, Russia, and Portugal, etc., are being added and more scrappers are being built to add more procurement data from other local and national portals by identifying, prioritising, and auditing new sources with respect to some quality criteria (e.g., legal, practical, and technical). OpenCorporates and OpenOpps data is available under their standard share-alike attribution Open Database Licences[9][10].

3 Architecture and Process

The preliminary architecture for data integration is presented in Fig. 1. OpenOpps and OpenCorporates undertake their own processes for gathering, extracting and curating data from distributed sources, including structured, semi-structured, and unstructured data. The data is extracted from OpenOpps' and OpenCorporate's databases through an extract, transform, and load (ETL) process using DataGraft [5], which is a cloud-based service for data transformation and access.

A series of refinements and enrichments are executed over the extracted data, such as normalisation (i.e., data types and formats), and curation (i.e., missing-records and duplicate records). Some of these processes are applied at the data provider side as well. Data will be characterised and integrated through a set of ontologies as discussed above. The entities in the knowledge graph are linked and re-reconciled, that is for example legal entities mentioned in procurement data are linked back to the company record in supplier data for that entity and references to the documents collected in the document store are added.

The knowledge graph is accessed via a linked data enabled REST API. This means that the URIs that are used to identify contracts, tenders, companies etc. are de-referenceable and with content negotiation. SPARQL endpoint access is

[8] `https://simap.ted.europa.eu/cpv`
[9] `https://opencorporates.com/info/licence`
[10] `https://openopps.com/legal`

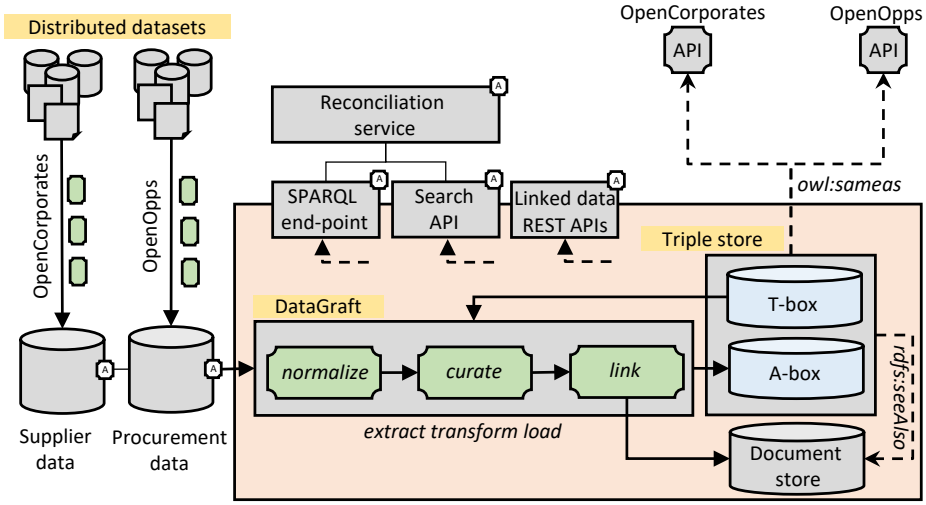

Fig. 1. Preliminary data integration architecture.

provided for those developers willing to make ad-hoc queries to the knowledge graph, as well as a range of additional services to enable search over the knowledge graph and document store, and reconciliation services to facilitate third parties the usage of the URIs. Note that not all data from OpenOpps and OpenCorporates is extracted, but data is linked back to these databases for further access.

4 Conclusions

Our ultimate goal is to ensure that data providers make their procurement data available in their own domains/sites, according to our ontology network. However, since this will not be possible in the short term, we follow a centralised approach in this work. The data in the resulting knowledge graph will be licensed under a combination of CC-BY 4.0 and Open Database License.

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. Roman, D., et al.: DataGraft: One-Stop-Shop for Open Data Management. Semantic Web **9**(4), 393–411 (2018)