Enhancing Public Procurement in the European Union through Constructing and Exploiting an Integrated Knowledge Graph

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
{name.surname}@sintef.no
[2] Universidad Politécnica de Madrid, Madrid, Spain
{ocorcho,cbadenes,fyedro}@fi.upm.es
[3] Jožef Stefan Institute, Ljubljana, Slovenia
{name.surname}@ijs.si
[4] OpenOpps Ltd, London, the UK
ian@spendnetwork.com
[5] OpenCorporates Ltd, London, the UK
chris.taggart@opencorporates.com
[6] King's College London, London, the UK
elena.simperl@kcl.ac.uk

Abstract. Public procurement is a large market affecting almost every organisation and individual. Governments need to ensure efficiency, transparency, and accountability, while creating healthy, competitive, and vibrant economies. In this context, we built a platform, consisting of a set of modular APIs and ontologies to publish, curate, integrate, analyse, and visualise an EU-wide, cross-border, and cross-lingual procurement knowledge graph. We developed end-user tools on top of the knowledge graph for anomaly detection and cross-lingual document search. This paper describes our experiences and challenges faced in creating such a platform and knowledge graph and demonstrates the usefulness of Semantic Web technologies for enhancing public procurement.

Keywords: Public procurement · Knowledge graph · Linked data.

1 Introduction

The market around public procurement is large enough so as to affect almost every single citizen and organisation across a variety of sectors. For this reason, public spending has always been a matter of interest at local, regional, and national levels. Primarily, governments need to be efficient in delivering services, ensure transparency, prevent fraud and corruption, and build healthy and sustainable economies [1, 13]. In the European Union (EU), every year, over 250.000 public authorities spend around 2 trillion euros (about 14% of GDP) on the purchase of

services, works, and supplies[1]; while the Organisation for Economic Co-operation and Development (OECD) estimates that more than 82% of fraud and corruption cases remain undetected across all OECD countries [19] costing as high as 990 billion euros a year in the EU [10]. Moreover, small and medium-sized enterprises (SMEs) are often locked out of markets due to the high cost of obtaining the required information, where larger companies can absorb the cost.

The availability of good quality, open, and integrated procurement data could alleviate the aforementioned challenges [11]. This includes government agencies assessing purchasing options, companies exploring new business contracts, and other parties (such as journalists, researchers, business associations, and individual citizens) looking for a better understanding of the intricacies of the public procurement landscape through decision-making and analytic tools. Projects such as the UK's GCloud (Government Cloud)[2] have already shown that small businesses can compete effectively with their larger counterparts, given the right environment. However, managing these competing priorities at a national level and coordinating them across different states and many disparate agencies is notoriously difficult. There are several directives put forward by the European Commission (e.g., Directive 2003/98/EC and Directive 2014/24/EU8) for improving public procurement practices. These led to the emergence of national public procurement portals living together with regional, local as well as EU-wide public portals [9]. Yet, there is a lack of common agreement across the 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 built 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 [23, 22] (i.e., KG, an interconnected semantic knowledge organisation structure [12, 27]). The knowledge graph includes procurement and company data gathered from multiple disparate sources across the EU and integrated through a common ontology network using an extract, transform, load (ETL) approach [3]. We built and used a set of end-user tools and machine learning (ML) algorithms on top of the resulting knowledge graph, so as to find anomalies in data and enable searching across documents in different languages. This paper reports the challenges and experiences we went through, while creating such a platform and knowledge graph, and demonstrates the usefulness of the Semantic Web technologies for enhancing public procurement.

The rest of the paper is structured as follows. Section 2 presents the related work, while Section 3 describes the data sets underlying the KG. Section 4 explains the KG construction, while Section 5 presents the KG publication together with the overall architecture and API resources. Section 6 describes the use of the KG for anomaly detection and cross-lingual document search, while Section 7 presents the adoption and uptake. Finally, Section 8 concludes the paper.

[1] `https://ec.europa.eu/growth/single-market/public-procurement_en`
[2] `https://www.digitalmarketplace.service.gov.uk`

2 Related Work

We focus on procurement data, related to tenders, awards, and contracts, and basic company data. We analyse relevant related works from the perspective of such types of data. Procurement and company data are fundamental to realising many key business scenarios and may be extended with additional data sources.

Public procurement notices play two important roles for the public procurement process: as a resource for improving competitive tendering, and as an instrument for transparency and accountability [15]. With the progress of eGovernment initiatives, the publication of information on contracting procedures is increasingly being done using electronic means. In return, a growing amount of open procurement data is being released leading to various standardisation initiatives like OpenPEPPOL[3], CENBII[4], TED eSenders[5], CODICE[6], and Open Contracting Data Standard (OCDS)[7]. Data formats and file templates were defined within these standards to structure the messages being exchanged by the various agents involved in the procurement process. These standards primarily focus on the type of information that is transmitted between the various organisations involved in the process, aiming to achieve certain interoperability in the structure and semantics of data. The structure of the information is commonly provided by the content of the documents that are exchanged. However, these initiatives still generate a lot of heterogeneity. In order to alleviate these problems, several ontologies including PPROC [16], LOTED2 [8], MOLDEAS [20], or PCO [17], as well as the upcoming eProcurement ontology[8] emerged, with different levels of detail and focus (e.g., legal and process-oriented). So far, however, none of them has reached a wide adoption mainly due to their limited practical value.

Corporate information, including basic company information, financial as well as contextual data, are highly relevant in the procurement context, not only for enabling many data value chains, but also for transparency and accountability. Recently, a number of initiatives have been established to harmonise and increase the interoperability of corporate and financial data. These include public initiatives such as the Global Legal Entity Identification System—GLEIS[9], Bloomberg's open FIGI system for securities[10], as well as long-established proprietary initiatives such as the Dun & Bradstreet DUNS number[11]. Other notable initiatives include the European Business Register (EBR)[12], Business Register Exchange (BREX)[13], and the eXtensible Business Reporting Language (XBRL)

[3] https://peppol.eu
[4] http://cenbii.eu
[5] https://simap.ted.europa.eu/web/simap/sending-electronic-notices
[6] https://contrataciondelestado.es/wps/portal/codice
[7] http://standard.open-contracting.org
[8] https://joinup.ec.europa.eu/solution/eprocurement-ontology
[9] https://www.gleif.org
[10] https://www.omg.org/figi
[11] http://www.dnb.com/duns-number.html
[12] http://www.ebr.org
[13] https://brex.io

format[14]. However, these are mostly fragmented across borders, limited in scope and size, and siloed within specific business communities. There are also a number of ontologies developed for capturing company and company-related data including the W3C Organisation ontology (ORG)[15], some e-Government Core Vocabularies[16], and the Financial Industry Business Ontology (FIBO) [4]. They have varying focuses, do not cover sufficiently the basic company information, or are too complex due to many ontological commitments [21].

There is so far no existing platform or KG (in whatever form) linking and provisioning cross-border and cross-language procurement and company data allowing advanced decision making, analytics, and visualisation.

3 Data Sets

The content of our KG is based on the procurement and company data that is provided by two main data providers extracting and aggregating data from multiple sources. The first one is OpenOpps[17], which is sourcing procurement data primarily from the Tenders Electronic Daily (TED)[18] data feed and from the procurement transparency initiatives of individual countries. TED is dedicated to European public procurement and publishes 520 thousand procurement notices a year. The second provider is OpenCorporates[19], which is collecting company data from national company registers and other regulatory sources. OpenOpps is the largest data source of European tenders and contracts, while OpenCorporates is the largest open database of companies in the world. Both OpenOpps and OpenCorporates gather relevant data using a range of tools, including processing API calls and Web scraping and data extraction.

Regarding procurement data, in the context of this work, OpenOpps provides gathered, extracted, pre-processed, and normalised data from hundreds of data sources completely openly through an API that can be used for research purposes. OpenOpps currently handles 685 data sources, with 569 of these being from Europe. This totals over 3 million documents dating back to 2010. All of the data for OpenOpps is gathered using a series of over 400 different scripts configured to collect data from each source. Each script is triggered daily and runs to gather all of the documents published in the last twenty-four hours. Each script is deployed on a monitored platform, giving the ability to check which scripts have failed, or which sources have published fewer than expected. Data is collected in the raw form and then mapped to the OCDS format after being cleansed. Where necessary, the data is processed, e.g., splitting single records into several fields, to comply with the data standard. Regarding company data, OpenCorporates provides

[14] https://www.xbrl.org
[15] https://www.w3.org/TR/vocab-org
[16] https://joinup.ec.europa.eu/solution/e-government-core-vocabularies
[17] https://openopps.com
[18] https://ted.europa.eu
[19] https://opencorporates.com

more than 140 million company records from a large number of jurisdictions[20]. OpenCorporates also pre-processes and normalises data collected, maps collected data to its own data model, and makes data available through an API.

The data collected from OpenOpps and OpenCorporates is openly available under the Open Database License (ODbl)[21]. It is available on GitHub[22] in JSON format and is updated on a monthly basis. The data is also made available through Zenodo[23] with a digital object identifier (DOI) [26].

4 Knowledge Graph Construction

The KG construction process includes reconciling and linking the two aforementioned and originally disconnected data sets, and mapping and translating them into Linked Data with respect to an ontology network [24].

4.1 Ontology Network

We developed two ontologies, one for representing procurement data and one for company data, using common techniques recommended by well-established ontology development methods [18, 6]. A bottom-up approach was used, including identifying the scope and user group of the ontology, requirements, and ontological and non-ontological resources. In general, we address suppliers, buyers, data journalists, data analysts, control authorities and regular citizens to explore and understand how public procurement decisions affect economic development, efficiencies, competitiveness, and supply chains. This includes providing better access to public tenders; spotting trends in spending and supplier management; identifying areas for cost cuts; and producing advanced analytics.

Regarding procurement data, we developed an ontology based on OCDS [25] – a relevant data model getting important traction worldwide, used for representing our underlying procurement data. The OCDS' data model is organised around the concept of a contracting process, which gathers all the relevant information associated with a single initiation process in a structured form. Phases of this process include mainly planning, tender, award, contract, and implementation. An OCDS document may be one of two kinds: a release or a record. A release is basically associated to an event in the lifetime of a contracting process and presents related information, while a record compiles all the known information about a contracting process. A contracting process may have many releases associated but only one record. We went through the reference specification of OCDS release and interpreted each of the sections and extensions (i.e., structured and unstructured). In total, there are currently 25 classes, 69 object properties, and 81 datatype properties created from the four main OCDS sections and 11 extensions. The core classes are `ContractingProcess`, `Plan`, `Tender`, `Award`, and `Contract`. A

[20] `https://opencorporates.com/registers`
[21] `https://opendatacommons.org/licenses/odbl`
[22] `https://github.com/TBFY/data-sources`
[23] `https://zenodo.org`

contracting process may have one planning and one tender stage. Each tender may have multiple awards issued, while there may be only one contract issued for each award. Other ontology classes include `Item`, `Lot`, `Bid`, `Organisation`, and `Transaction`. We reused terms from external vocabularies and ontologies where appropriate. These include Dublin Core[24], FOAF[25], Schema.org[26], SKOS[27], and the W3C Organisation ontology [28]. The OCDS ontology is available on GitHub in two versions[29]: one with the core OCDS terms and another with the extensions.

Regarding company data, one of the main resources used during the ontology development was data models provided by four company data providers: OpenCorporates, SpazioDati[30], Brønnøysund Register Centre[31], and Ontotext[32]. The data supplied by these data providers originally came from both official sources and unofficial sources. The need for harmonising and integrating data sets was a guiding factor for the ontology development process, since data sets have different sets of attributes and different representations with similar semantics. The resulting ontology, called euBussinessGraph ontology [21], is composed of 20 classes, 33 object properties, and 56 data properties allowing us to represent basic company-related data. The ontology covers registered organisations (i.e., companies that are registered as legal entities), identifier systems (i.e., a company can have several identifiers), officers (i.e., associated officers and their roles), and data sets (i.e., capturing information about data sets that are offered by company data providers). Registered organisations are the main entities for which information is captured in the ontology. The main classes include `RegisteredOrganisation`, `Identifier`, `IdentifierSystem`, `Person`, and `Dataset`. Three types of classifications are defined in the ontology for representing the company type, company status, and company activity. These are modelled as SKOS concept schemes. Some of the other external vocabularies and ontologies used are W3C Organisation ontology, W3C Registered Organisation Vocabulary (RegOrg)[33], SKOS, Schema.org, and Asset Description Metadata Schema (ADMS)[34]. The ontology, data sets and some examples are released as open source on GitHub[35].

4.2 Data Ingestion

The ingestion process extracts procurement and company data from the data providers, matches suppliers appearing in procurement data against company

[24] `http://dublincore.org`
[25] `http://xmlns.com/foaf/spec`
[26] `https://schema.org`
[27] `https://www.w3.org/2004/02/skos`
[28] `https://www.w3.org/TR/vocab-org`
[29] `https://github.com/TBFY/ocds-ontology/tree/master/model`
[30] `http://spaziodati.eu`
[31] `http://www.brreg.no`
[32] `https://www.ontotext.com`
[33] `https://www.w3.org/TR/vocab-regorg`
[34] `https://www.w3.org/TR/vocab-adms`
[35] `https://github.com/euBusinessGraph/eubg-data`

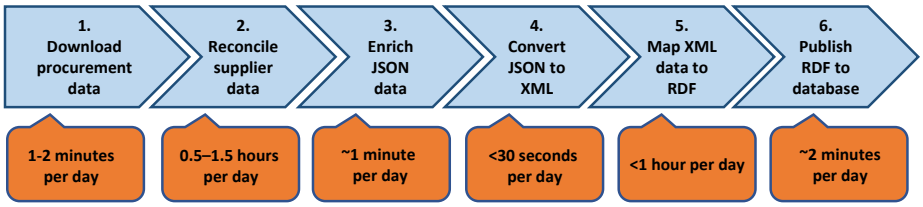


Fig. 1. The daily data ingestion process for the KG – on average 2500 OCDS releases are processed and 2400 suppliers (i.e., companies) are looked up per day.

data (i.e., reconciliation), and translates the data sets into RDF using RML[36]. The daily process is composed of the following steps (see Fig. 1):

(1) **Download procurement data**: Downloads procurement data from the OpenOpps OCDS API[37] as JSON data files.
(2) **Reconcile suppliers**: Matches supplier records in awards using the OpenCorporates Reconciliation API[38]. The matching company data is downloaded using the OpenCorporates Company API[39] as JSON data files.
(3) **Enrich downloaded JSON data**: Enriches the JSON data files downloaded in steps 1 and 2, e.g., adding new properties to support the mapping to RDF (e.g., fixing missing identifiers).
(4) **Convert JSON to XML**: Converts the JSON data files from step 3 into corresponding XML data files. Due to limitations in JSONPath, i.e., lack of operations for accessing parent or sibling nodes from a given node, we prefer to use XPath as the query language in RML.
(5) **Map XML data to RDF**: Runs RML Mapper on the enriched XML data files from step 4 and produces N-Triples files.
(6) **Store and publish RDF**: Stores the RDF (N-Triples) files from step 5 to Apache Jena Fuseki and Apache Jena TBD.

We have been running the ingestion pipeline on a powerful server (see some performance metrics in Fig. 1), with the following hardware specifications: 2x Xeon Gold 6126 (12 Cores, 2.4 GHz, HT) CPU, 512 GB main memory, 1x NVIDIA Tesla K40c GPU, and 15 TB HDD RAID10 & 800 GB SSD storage. Python was used as the primary scripting language, RMLMapper was used as the mapping tool to generate RDF, and finally Apache Jena Fuseki & TDB was chosen as the SPARQL engine and triple store. The Python scripts operate on files (output and input) and services have been dockerised using Docker and made available on Docker Hub[40] to ease deployment. All development work and results towards the creation of the knowledge graph are published and maintained as open source

[36] `https://rml.io`
[37] `https://openopps.com/api/tbfy/ocds`
[38] `https://api.opencorporates.com/documentation/Open-Refine-Reconciliation-API`
[39] `https://api.opencorporates.com/documentation/API-Reference`
[40] `https://hub.docker.com/r/tbfy/kg-ingestion-service`

software on GitHub[41]. The data dumps of the KG, including more than 126M statements as of August 2020, are available on Zenodo [26].

5 Knowledge Graph Provisioning

We developed a platform and core API services for the KG ingestion and provisioning, using recent Linked Data and REST API design practices and principles.

5.1 Platform Architecture

Our platform follows state-of-the-art principles in software development, considering a low decoupling amongst all the software components.

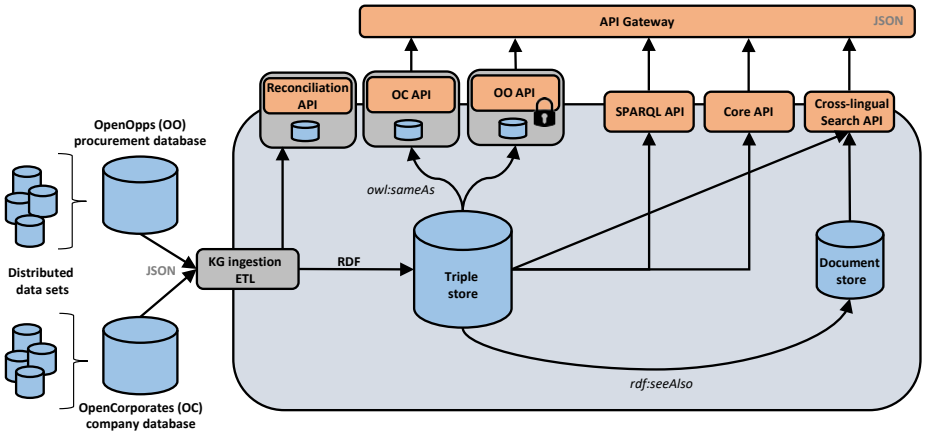


Fig. 2. The high-level architecture for KG ingestion and provisioning.

Fig. 2 provides a high-level overview of the architecture. On the left-hand side, we include the ETL processes that are being used to incorporate the data sources into the KG. On the right-hand side we provide an overview of the main data storage mechanisms, including a triple store for the generated RDF-based data and a document store for the documents associated to public procurement (tender notices, award notices, etc.), whose URLs are accessible via specific properties of the KG (using `rdfs:seeAlso`). For those specific cases where a URI is also available in the original data sources (from OpenOpps and OpenCorporates), such URI is provided in the KG using a statement with `owl:sameAs`. This would allow our data providers to provide additional information about tenders or companies with a different license or access rights (e.g., commercial use).

The KG is accessible via a core REST API. Our API catalogue is mostly focused on providing access mechanisms to those who want to make use of the

[41] `https://github.com/TBFY/knowledge-graph`

knowledge graph, particularly software developers. Therefore, they are mostly focused on providing access to the KG through the HTTP GET verb and the API catalogue is organised around the main entities that are relevant for public procurement, as discussed in Section 4, such as contracting processes, awards, and contracts. Since the KG is stored as RDF in a triple store, there is also a SPARQL endpoint[42] for executing ad-hoc queries. Finally, there is a cross-lingual search API for searching across documents in various languages and an API Gateway providing a single-entry point to the APIs provided by the platform.

5.2 Core API

The core API was built using the R4R tool[43]. This tool is based on Velocity templates[44] and allows specifying how the REST API will look like and configure it by means of SPARQL queries, similarly to what has been proposed in other state of the art tools like BASIL (Building Apis SImpLy) [7] or GRLC [14]. Beyond exposing URIs for the resources available in the KG, it also allows including authentication and authorisation, pagination, establishing sorting criteria over specific properties, nesting resources, and other typical functionalities normally available in REST APIs. The current implementation only returns JSON objects for the API calls and will be extended in the future to provide additional content negotiation capabilities and formats (JSON-LD, Turtle, HTML), which are common in Linked Data enabled APIs.

There is an online documentation[45], which is continuously updated. It provides the details of the resources provided by our REST API in relation to the OCDS ontology. The core resources derived from the OCDS ontology are: (i) `ContractingProcess`, (ii) `Award`, (iii) `Contract`, (iv) `Tender`, and (v) `Organisation`. For all these resources, there is a possibility of paginating (e.g., `GET /award?size=5&offset=1`), sorting (e.g., `GET /contract?sort=-startDate`), and filtering (e.g., by the title of the award: `GET /award?status=active`).

6 Knowledge Graph in Use

We implemented a number of real-life use cases on the platform and KG: anomaly detection and cross-lingual document search.

6.1 Anomaly Detection

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 for detecting possible frauds is not feasible. In this respect, using ML techniques for

[42] `http://yasgui.tbfy.eu`
[43] `https://github.com/TBFY/r4r`
[44] `https://velocity.apache.org`
[45] `https://github.com/TBFY/knowledge-graph-API/wiki`

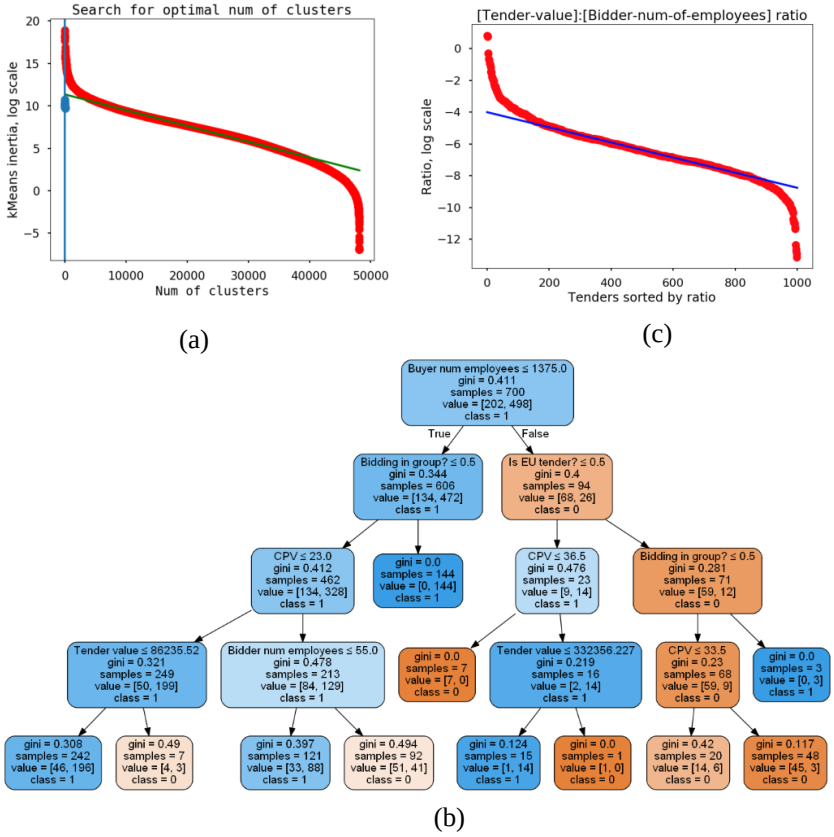


Fig. 3. (a) Anomaly detection in public procurement data with k-Means analysis. (b) The decision tree model for identifying successful tenders. (c) A graph showing interdependence between tender value and number of employees of bidder.

identifying patterns and anomalies, such as fraudulent behaviour or monopolies, in procurement processes and networks across data sets produced independently, is highly relevant [5]. For example, by building a network of entities (individuals, companies, governmental institutions, etc.) connected through public procurement events, one can discover 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.

We applied several ML techniques, i.e., supervised, unsupervised, and statistical, on top of the Slovenian public procurement data in the KG to identify patterns and anomalies. First, clustering was used for anomaly detection (see Fig. 3 (a)), since one could quickly spot deviations with this approach. Every tender was transformed into a feature vector. After determining the optimal number of clusters, public procurement data were clustered with the K-means algorithm. Vectors deviating most from their centroids are identified and ordered

by the deviation value (i.e., Cartesian distance). The approach was used to identify tenders with highest deviations. For example, among others, our method identified public procurement cases with an unusually high tender value. This approach gives a hint on data that "stick out" and are worth of more in-depth scrutiny. Second, supervised analysis implemented in our platform is based on a decision tree (see Fig. 3 (b)). We enabled users to select parameters by their own choice (for instance buyer size, bidder municipality, and the depth of decision tree model), and thus enabling users to compare the importance of subsets of various parameters contributing to the success of public tenders. The success definition is up to decision makers. We define success as a tender that received more than one bid. According to our preliminary analysis, a tender is successful – i.e. there will be competition (more than one bid) - if public institution who opened the tender is small (less than 1375 employees) and if bidding is done in group. Third, statistical approach was used to deal with various ratios between pre-selected parameters (see Fig. 3 (c)). Currently, the ratio between the tender value and the estimated number of employees for a bidder is examined. Bidders are then sorted by their ratio value and every bidder turned into a point: the x value is a consecutive number and the y value is the ration. We developed a visual presentation of interdependence of tender value and the number of employees. The graph shows deviating behaviour at the beginning as well as at the end of the list. On the upper left corner of the graph, we can see big companies with a high number of employees that won small tenders, and on the bottom right corner, there are companies with a small number of employees that won big tenders.

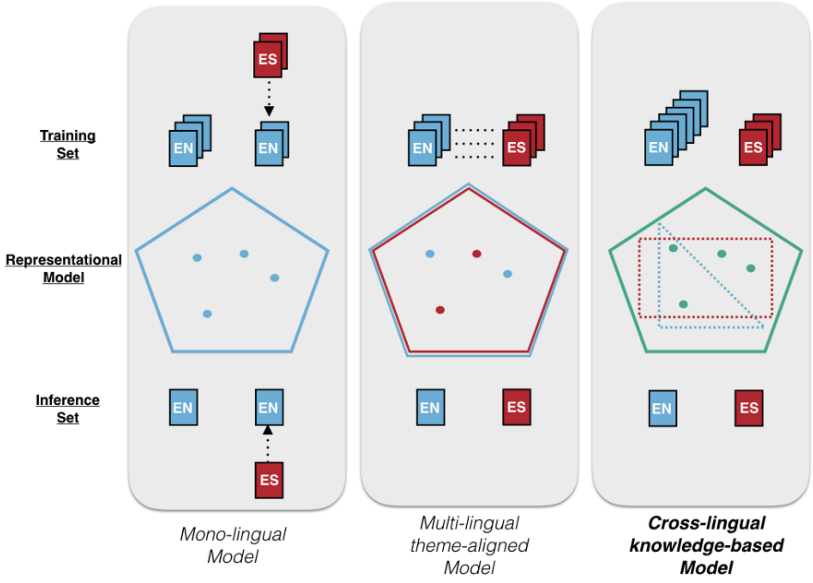
We implemented a system capable of processing tens of millions of records, based on the techniques mentioned and made it available online[46].

6.2 Cross-lingual Document Search

Procurement processes are not only creating structured data, but also constantly creating additional documents (tender specifications, contract clauses, etc.). These are commonly published in the official language of the corresponding public administrations. Only some of these, for instance those published in TED, are multilingual, but the documents in the local language are typically longer and much more detailed than their translations into other languages. A civil servant working at a public administration on a contracting process may be interested in understanding how other public administrations in the same country or in different countries (and with different languages) have worked on similar contexts. Examples may include finding organisations related to a particular procurement process, or search for tenders related to given procurement text.

We worked on an added-value service[47] in order to provide support to these types of users, with the possibility of finding documents that are similar to a given one independently of the language in which it is made available. We also generated a Jupyter notebook with some representative examples, so as to facilitate its use[48].

[46] `http://tbfy.ijs.si`
[47] `http://tbfy.librairy.linkeddata.es/search-api`
[48] `http://bit.ly/tbfy-search-demo`



(a)

Topic3@EN *Communication Systems*	Topic3@ES *Sistema de Comunicación*	Topic26@FR *Systeme de Comunicacion*	Topic10@PT *Communication Systems*
radio	*equipo*	*communications*	*rede*
equipment	*red*	*reseaux*	*comunicação*
network	*comunicación*	*electroniques*	*electrónico*
communication	*espectro*	*acces*	*acesso*
regulatory	*electromagnético*	*telecommunications*	*utilizador*

(b)

Fig. 4. (a) Documents are represented in a unique space that relies on the latent layer of cross-lingual topics obtained by LDA and hash functions through hierarchies of synsets. (b) Theme-aligned topics described by top 5 words based on EUROVOC annotations.

This service is based on the use of unsupervised probabilistic topic models, based on cross-lingual labels from sets of cognitive synonyms (synsets) to establish relations between language-specific topics [2]. Documents are represented as data points in a low-dimensional latent space created by probabilistic topic models for each language separately (see Fig. 4). Topics are then described by cross-lingual labels created from the list of concepts retrieved from the Open Multilingual WordNet. Each word is queried to retrieve its synsets. The final set of synsets for a topic is the union of the synsets from the individual top-words of a topic (top5 based on empirical evidences).

The JRC-Acquis data set[49] was used to build the model relating the documents. It is a collection of legislative texts written in 23 languages that have been manually classified into subject domains according to the EUROVOC[50] thesaurus.

[49] `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis`
[50] `http://eurovoc.europa.eu`

The English, Spanish, French, Italian and Portuguese editions (about 20.000 documents per edition) of the corpora were used for each language-specific model. The EUROVOC taxonomy was pre-processed to satisfy the topic independence assumption of probabilistic topic models, by using hierarchical relations. The initial 7.193 concepts from 21 domain areas such as politics, law or economics were reduced to 452 categories, that are independent and can be used to train the topic models. Documents were pre-processed (Part-of-Speech filtering and lemmatized format) by the librAIry NLP[51] service and projected into the previously created topic space. The method is evaluated in several document retrieval tasks by using a set of documents previously tagged with EUROVOC categories. Results are quite promising across languages with a performance close to 0.8 in terms of accuracy, although a better performance is achieved with English texts.

7 Adoption and Uptake

We have used Semantic Web technologies to integrate disparate open data sources in a standardised way. They enabled us to ingest new data sources and integrate other relevant data sources (e.g., company and procurement data) without major restructuring efforts. Similar solutions could be provided using other technologies; however, without following the Linked Data and Semantic Web principles, they would rather remain ad-hoc and could not be easily scaled, given various independent data publishers and consumers.

The uptake of our platform and KG has been exemplified in four different cases so far. The core API and KG are used by the Spanish company OESIA[52] and by the city of Zaragoza, Spain. OESIA created a commercial tool for tender analysis, which is offered to SMEs. Zaragoza includes economic information in their transparency portal[53], including public procurement. Regarding advanced tools, the anomaly detection tool is used by the Ministry of Public Administration in Slovenia for detecting procurement anomalies, while the cross-lingual similarity search is used by the Italian company CERVED[54] for finding tenders in other countries/languages and offering this as part of their services. The categories of users using the system include civil servants (i.e., Zaragoza and Slovenia), citizens (i.e., Zaragoza), and companies, especially SMEs (i.e., CERVED and OESIA). As of August 2020, over 3.700 queries have been submitted to the system APIs.

We plan to maintain the KG in the context of already funded innovation projects. Maintenance will include ingesting new data and operating the system. Agreements with data providers, i.e., OpenOpps and OpenCorporates, have been established to provide the KG with data on a continuous basis. Furthermore, the data and platform components are made available openly for the community to contribute (a catalogue is available[55]). We are also proposing our ontology

[51] `http://librairy.linkeddata.es/nlp`
[52] `https://grupooesia.com/en/`
[53] `https://zaragoza.es/sede/servicio/transparencia`
[54] `https://company.cerved.com`
[55] `https://tbfy.github.io/platform`

network as the way to publish open data about procurement by governments. An example is the case of Zaragoza, which already adopted our ontology network.

8 Lessons Learned

There are plenty of lessons learned in the context of this work, which may be applicable to the construction of other KGs in similar or different domains. First, we provide a non-exhaustive list of major takeaways related to the whole process:

(i) The KG enabled easier and advanced analytics, which was otherwise not possible, by connecting companies (i.e., suppliers) appearing in the procurement data set to companies in company data set. However, getting and pre-processing the data (e.g., data curation) was a major and time-consuming task, requiring attention from national and EU-wide data providers.
(ii) The existing Semantic Web technologies and tools scaled well for ingesting and provisioning large amounts of data and RESTful approach was useful for bringing the Linked Data to non-Semantic Web application developers. However, more support is required such as visual editors for creating mapping definitions and specifying data transformations.
(iii) The process of building a high-quality KG that can be used extensively by users would be clearly improved if all data sources were providing their procurement data in a more structured manner. Data quality problems are still a relevant issue, as described in the followings, and reduce the result quality of ML processes such as anomaly detection and reconciliation.
(iv) There are still many documents associated to the procurement processes that are provided as PDFs (in some cases even scanned PDFs). Providing all documents in the form of raw texts as well would simplify the processing that needs to be done, and would allow applying more easily the techniques like the ones described for cross-lingual search.
(v) Data providers should also aim at publishing the information of all types of contracting processes that they are handling, independently of their size. Currently, due to many types of regulations across countries, not all contracting processes (especially the smallest ones) are published.

We also faced a high number of data quality issues, even though there are mandates in place for buyers to provide correct data. This particularly applies to procurement data sources. These data quality issues could be classified as:

(i) *Missing data*: It is frequent that data is missing. Among others, the least frequently completed field in the tender and contracting data is the *value* field; it is usually completed in less than 10% of tender notices. One item of data that is particularly important to procurement transparency is the reference data required to link a contract award to a tender notice (very common in the TED data). We found that just 9% of award notices had provided a clear link between tenders and contracts. Subsequently, the majority of contract award notices had been orphaned and there was no link to the source tenders.

(ii) *Duplicate data*: Publishers frequently publish to multiple sources in order to meet the legal requirements of their host country and that of the European Union. This means that all over-threshold tenders are available at least twice. The task of managing duplicates is not always simple. It is common for different publishing platforms to have different data schemas and interoperability between schemas is not guaranteed.

(iii) *Poorly formed data*: Sources are frequently providing malformed data or data that cannot be reasonably parsed by code. The tender and contract value field can often include string values rather than numbers (same goes for the dates). Across the sources, approach to using character delimiters in value data is frequently heterogeneous, with different nationalities using different delimiters to separate numbers and to indicate decimals.

(iv) *Erroneous data*: Structured data such as numeric and date records are frequently a problem. Buyers often submit zero value entries in order to comply with the mandate and the lack of validation on date related data has allowed buyers to record inconsistent date data. There are some contracts where the date of publication exceeds the end date of the contract or the start date of the contract is greater than the end date of the contract.

(v) *Absent data fields*: In some cases, the sources lack core pieces of information, for instance, there is no value field in a number of European sources. A large number of sites also fail to publish the currency of their monetary values. In all cases, if a publisher sought to add the additional information, such as a different currency, there would be no capacity in the system to provide the information required in a structured form.

Most of these problems can be resolved through the use of standards and validation at the point of data entry. Requiring buyers to publish records to a standard would, in turn, require the platform providers to both mandate the field format and validate data entries. The usage of an ontology network for the development of the KG allowed us to inform public administrations willing to provide data on the minimum set of data items that are needed, and some of them are already adapting their information systems for this purpose [9].

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

References

1. Alvarez-Rodríguez, J.M., et al.: New trends on e-Procurement applying semantic technologies. Computers in Industry **65**(5), 797–799 (2014)
2. Badenes-Olmedo, C., et al.: Scalable Cross-lingual Similarity through language-specific Concept Hierarchies. In: Proc. of K-CAP 2019. pp. 147–153 (2019)
3. Bansal, S.K., Kagemann, S.: Integrating Big Data: A Semantic Extract-Transform-Load Framework. Computer **48**(3), 42–50 (2015)
4. Bennett, M.: The financial industry business ontology: Best practice for big data. Journal of Banking Regulation **14**(3), 255–268 (2013)

5. Chandola, V., et al.: Anomaly Detection: A Survey. ACM Computing Surveys **41**(3) (2009)
6. Corcho, O., et al.: Ontological Engineering: Principles, Methods, Tools and Languages, pp. 1–48. Springer (2006)
7. Daga, E., et al.: A BASILar approach for building web APIs on top of SPARQL endpoints. In: Proc. of SALAD 2015. CEUR-WS.org (2015)
8. Distinto, I., et al.: LOTED2: An ontology of European public procurement notices. Semantic Web **7**(3), 267–293 (2016)
9. Espinoza-Arias, P., et al.: The Zaragoza's Knowledge Graph: Open Data to Harness the City Knowledge. Information **11**(3) (2020)
10. The Cost of Non-Europe in the area of Organised Crime and Corruption. Tech. rep., European Parliment (2016), `https://www.europarl.europa.eu/RegData/etudes/STUD/2016/579319/EPRS_STU\%282016\%29579319_EN.pdf`
11. Futia, G., et al.: Removing Barriers to Transparency: A Case Study on the Use of Semantic Technologies to Tackle Procurement Data Inconsistency. In: Proc. of ESWC 2017. pp. 623–637. Springer (2017)
12. Giese, M., et al.: Optique: Zooming in on Big Data. Computer **48**(3), 60–67 (2015)
13. Janssen, M., et al.: Driving public sector innovation using big and open linked data (BOLD). Information Systems Frontiers **19**(2), 189–195 (2017)
14. Meroño-Peñuela, A., Hoekstra, R.: grlc makes GitHub taste like linked data APIs. In: Proc. of ESWC 2016. pp. 342–353. Springer (2016)
15. Miroslav, M., et al.: Semantic technologies on the mission: Preventing corruption in public procurement. Computers in Industry **65**(5), 878–890 (2014)
16. Muñoz-Soro, J., et al.: PPROC, an ontology for transparency in public procurement. Semantic Web **7**(3), 295–309 (2016)
17. Necaský, M., et al.: Linked data support for filing public contracts. Computers in Industry **65**(5), 862–877 (2014)
18. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Tech. rep., Stanford Medical Informatics (2001)
19. OECD Principles for Integrity in Public Procurement. Tech. rep., OECD (2009), `http://www.oecd.org/gov/ethics/48994520.pdf`
20. 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)
21. Roman, D., et al.: The euBusinessGraph Ontology: a Lightweight Ontology for Harmonizing Basic Company Information. Semantic Web **under review** (2020), `http://www.semantic-web-journal.net/system/files/swj2421.pdf`
22. Simperl, E., et al.: Towards a knowledge graph based platform for public procurement. In: Proc. of MTSR 2018. pp. 317–323. Springer (2018)
23. Soylu, A., et al.: Towards integrating public procurement data into a semantic knowledge graph. In: Proc. of EKAW 2018 Poster and Demonstrations. CEUR-WS.org (2018)
24. Soylu, A., et al.: An Overview of the TBFY Knowledge Graph for Public Procurement. In: Proc. of ISWC 2019 Satellite Tracks. CEUR-WS.org (2019)
25. 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. Springer (2019)
26. TBFY: KG data dump (2020). https://doi.org/10.5281/zenodo.3712323
27. Yan, J., et al.: A Retrospective of Knowledge Graphs. Frontiers of Computer Science **12**(1), 55–74 (2018)