Innovation Action (IA)

ICT-14-2016-2017

H2020-ICT-2017-1

Enabling procurement data value chains for economic development, demand management, competitive markets and vendor intelligence

Deliverable D1.3

Ingested Data v1

Date	28.09.2018
Author(s)	Ian Makgill (OO), Brian Elvesæter (SINTEF)
Dissemination level	Public
Work package	1
Version	Final

Document metadata

Quality assurers and contributors

Quality assurer(s)	Till Lech, Ahmet Soylu
Contributor(s)	Ian Makgill, Brian Elvesæter

Version history

Date	Version	Description
21.09.2018	1	Initial draft outline.
26.09.2018	2	Updated document submitted for internal review.
27.09.2018	3	Addressed comments by internal review.
28.09.2018	Final	Final formatting and layout.

Executive summary

This document represents Deliverable D1.3 "Ingested Data v1", which comprises the initial development effort in WP1 and WP2 related to data ingestion, knowledge graph representation and knowledge graph publication. The goal of the knowledge graph is to be an interconnected semantic knowledge organization structure, which can be analysed in depth to identify patterns and anomalies in procurement processes and networks.

Deliverable D1.3 is of type other (i.e., a software prototype) and the results are published and maintained as open source software on the GitHub repository https://github.com/TBFY/knowledge-graph

This document provides a summary of the initial development results and the intermediate point in the process of the data publication. It introduces:

- the GitHub repository structure and contents;
- the data sources that will be ingested;
- the knowledge graph representation (ontology); and
- the data ingestion process.

Finally, it outlines the development plan for the next phase of the project.

Table of contents

1 Introduction

This document represents Deliverable D1.3 "Ingested Data v1", which comprises the initial development effort in WP1 and WP2 related to data ingestion, knowledge graph representation and knowledge graph publication in the TheyBuyForYou (TBFY) project.

The data in WP1 that has been ingested will be made available to WP2 so that it can be enriched and published as a knowledge graph. The goal of the knowledge graph is to be an interconnected semantic knowledge organization structure, which can be analysed in depth to identify patterns and anomalies in procurement processes and networks. The work reported here relates to the following deliverables:

- **Deliverable D1.3 "Ingested Data v1" (month 9).** This document, which provides a summary of the initial development results and the intermediate point in the process of the data publication.
- **Deliverable D2.2 "Knowledge graph publication" (month 18).** Presentation of the knowledge graph representation and the data ingestion workflow and services to be developed.
- **Deliverable D1.6 "Ingested Data v2" (month 27).** Presentation of the final development results and the continuous knowledge graph publication processes and services.

The deliverables are of type other (i.e., software prototypes). The results are published and maintained as open source software on a GitHub repository.

The remainder of this document consists of the following sections:

- Section 2 presents the GitHub repository structure and contents.
- Section 3 presents the data sources that will be ingested.
- Section 4 presents the knowledge graph representation (ontology).
- Section 5 presents the data ingestion process.
- Section 6 outlines the development plan for the next phase of the project.

2 GitHub repository

All development work and results towards the creation of the TBFY knowledge graph are published and maintained as open source software on the GitHub repository https://github.com/TBFY/knowledge-graph

Figure 1 shows the GitHub repository page for the knowledge graph. In this repository we will keep:

- Ontology model files that defines the schema for the TBFY knowledge graph.
- Source code for the data ingestion services that is being developed for onboarding data to the knowledge graph.
- Modules with references to related ontologies in the public procurement domain.
- Documentation of the knowledge graph schema (ontology) and related services that is being developed.

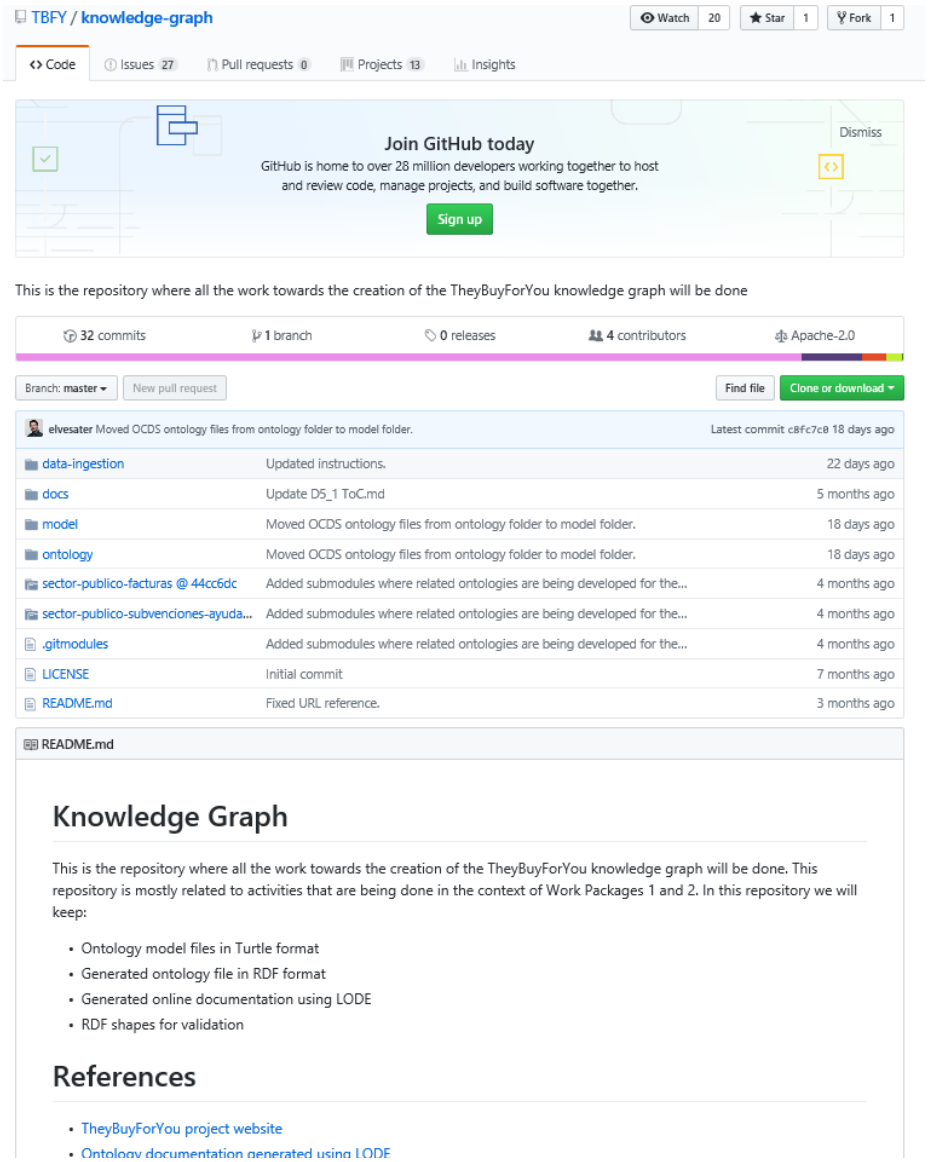

Figure 1: GitHub knowledge graph repository

3 Data sources

The TBFY knowledge graph will be generated from a wide range of data sources (mostly the buyer profiles and transparency portals of public administrations EU-wide). To create the TBFY knowledge graph the following data sources are considered:

- market engagement documents, such as tender notices or specifications;
- post-award documents, such as contract award notices or contracts;
- spending data, including transaction data and budget data;
- tertiary data for the procurement process, such as administrative responsibilities and geospatial data;
- core company reference data - from corporate registers and other official public registers;
- non-core company data - from Official Journals (i.e., government gazettes) and similar, which contain important information both on companies and tenders;
- relevant ontologies, vocabularies, and standards such as the eProcurement ontology;

- existing knowledge graphs such as the euBusinessGraph.

In the first phase of the knowledge graph publication we have focused on onboarding data from OpenOpps and OpenCorporates.

3.1 OpenOpps tender and contracting data

OpenOpps is the largest data source of European tenders and contracts in the world. OpenOpps provides the core TED data as well as thousands of daily opportunities from local and national portals around the EU. Included in this data is details on buyers, suppliers (for contracts), titles, descriptions, values and categories.

OpenOpps will provide gathered, extracted, pre-processed and normalised data from hundreds of European data sources completely openly to TBFY for the duration of the project, through an API. This means TBFY will have open access to over 3 million documents dating back to 2010, while OpenOpps will maintain its own code for validating, mapping and monitoring the data. Further details about the OpenOpps data are elaborated in Deliverable D1.2 "Data gathering, extraction, pre-processing and normalisation components v1".

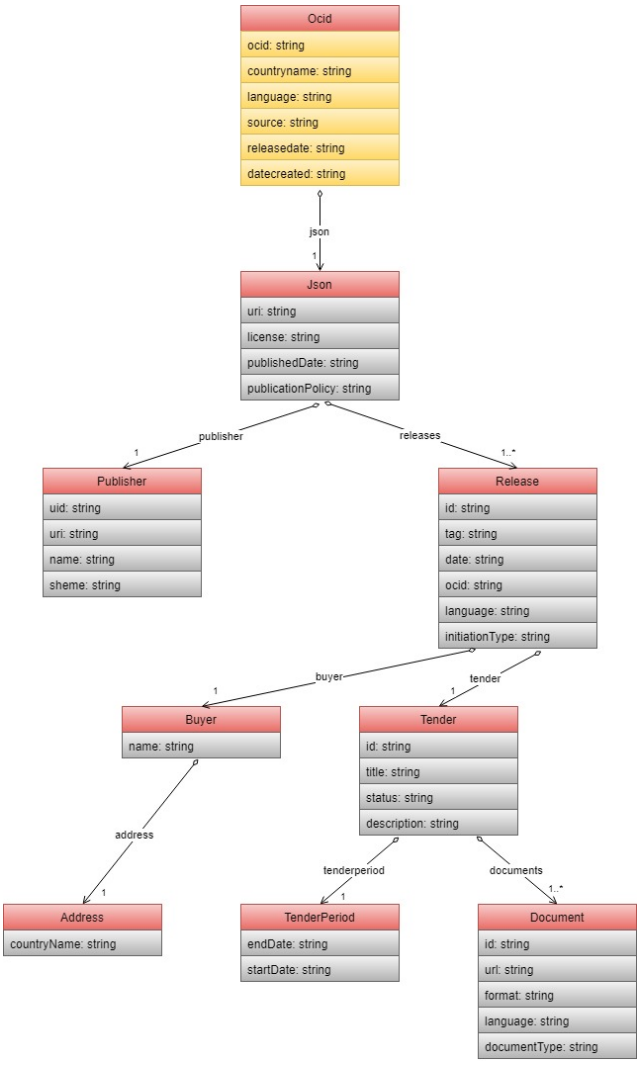

Figure 2: OpenOpps data model (a subset of OCDS)

The OpenOpps API provides access to tender and contract data from a range of European government bodies. The data is formatted according to the Open Contracting Data Standard (OCDS)[1], a data format that is focused on enabling the disclosure of data and documents at all stages of the contracting process. Figure 2 above shows the data model supported by the current OpenOpps API taken from Deliverable D5.2 "Procurement APIs and platform release v1". Further details about the OpenOpps API are found in that deliverable.

3.2 OpenCorporates corporate company data

OpenCorporates is the largest open database of companies and company data in the world, with in excess of 140 million companies in a similarly large number of jurisdictions. The data will be used in TheyBuyForYou as the reference dataset for legal entities. Legal entities mentioned in procurement data will be linked back to the OpenCorporates record for that entity.

OpenCorporates has data from the following European jurisdictions:

Albania	Gibraltar	Liechtenstein	Romania
Belarus	Greece	Luxembourg	Slovakia
Belgium	Guernsey	Malta	Slovenia
Bulgaria	Iceland	Moldova	Spain
Croatia	Ireland	Montenegro	Sweden
Cyprus	Isle of Man	Netherlands	Switzerland
Denmark	Jersey	Norway	Ukraine
Finland	Latvia	Poland	UK
France			

The database contains core company data (i.e. existence and basic attributes) in the world. Figure 3 below shows the data model supported by the OpenCorporates API taken from Deliverable D5.2 "Procurement APIs and platform release v1". Further details about the OpenCorporates API are found in that deliverable.

[1] http://standard.open-contracting.org/latest/en/schema/reference/

Figure 3: OpenCorporates data model

4 Knowledge graph representation

As described in Section 3, the OpenOpps and OpenCorporates data represent the two main data sources for creating the knowledge graph. In order to build a first version of the TBFY knowledge graph we focus on these two data sources and define a TBFY schema that comprises and links the tender and contracting data from OpenOpps and company data from OpenCorporates.

For the OpenCorporates data we use the euBusinessGraph schema (ontology) that is developed in the euBusinessGraph project. The euBusinessGraph ontology for representing company data is released as open source under the GitHub repository https://github.com/euBusinessGraph/eubg-data/tree/master/model

Currently, there are several ontologies and data formats that are being used to represent procurement data as described in Deliverable D5.1 "Report on best practices for software engineering, publishing and using procurement data". The OCDS[2], PPROC[3], the upcoming EU eProcurement Ontology[4] are initiatives that will be taken in the course of the project for the selection of the final ontology network that will be

[2] http://standard.open-contracting.org/latest/en/
[3] http://contsem.unizar.es/def/sector-publico/pproc
[4] https://github.com/eprocurementontology/eprocurementontology/

used for the TBFY knowledge graph representation. We have decided on an incremental approach to developing the TBFY knowledge graph:

- The **first release** will onboard tender data from OpenOpps (based on an OCDS schema) and company data from OpenCorporates (based on the euBusinessGraph schema). The OpenOpps data is currently aligned with OCDS version 1.1.7, which will be the base for the OCDS schema being developed in TBFY.
- The **second release** will interconnect supplier records from contracting data (from OpenOpps) and the company records (from OpenCorporates). OpenOpps has already done some work on linking UK data on suppliers to OpenCorporates data which will be valuable input in developing a general model and approach to interconnect these data also for other jurisdictions.
- The **third release** will evolve the TBFY schema to support additional data, including further data interconnections via crowd sourcing. OpenOpps currently has 100 million lines of spending data (much of which is already linked to OpenCorporates) that can be integrated into the knowledge graph. Where there is a strong case for doing so, we will undertake to align the release with new and relevant procurement schemas and ontologies (e.g. version 2.0 of OCDS and the upcoming EU eProcurement Ontology).

Figure 4 below illustrates the approach to our knowledge graph schema (ontology) representation. In the first release we are developing an OCDS schema based on the OCDS version 1.1.7 standard to represent the OpenOpps tender and contracting data. This will cover the main concepts and attributes of OCDS used by OpenOpps. Later we will make evolve the schema and make the required alignment with relevant procurement ontology to use such. The fragment of OCDS being used by OpenOpps is quite generic and we believe it would fit into any such ontology.

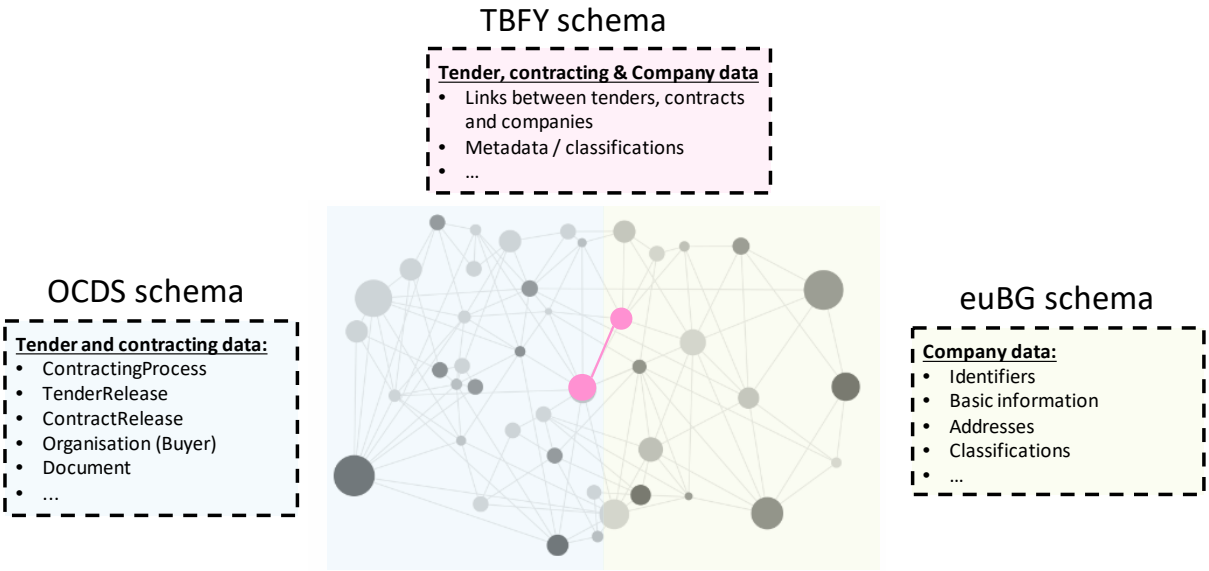

Figure 4: Knowledge graph representation

5 Data ingestion workflow

Figure 5 illustrates the data ingestion workflow being developed for the TBFY knowledge graph. Currently OpenCorporates data are provided as dump files (in CSV format), while OpenOpps data are retrieved from their API (in JSON format according to OCDS). The JSON data is pre-processed and[5]

[5] https://datagraft.io/

converted to a flattened CSV format. The CSV data from OpenCorporates and OpenOpps are then mapped to the TBFY schema by using the Grafterizer cloud-based mapping and RDFization service which is part of the DataGraft platform. The result of the Grafterizer is an executable transformation service that can be used as a standalone service for converting and publishing updated data from OpenCorporates and OpenOpps at regular intervals (e.g. batches). The resulting RDF data from the transformation is published to a graph database. Currently we are using Ontotext GraphDB[6].

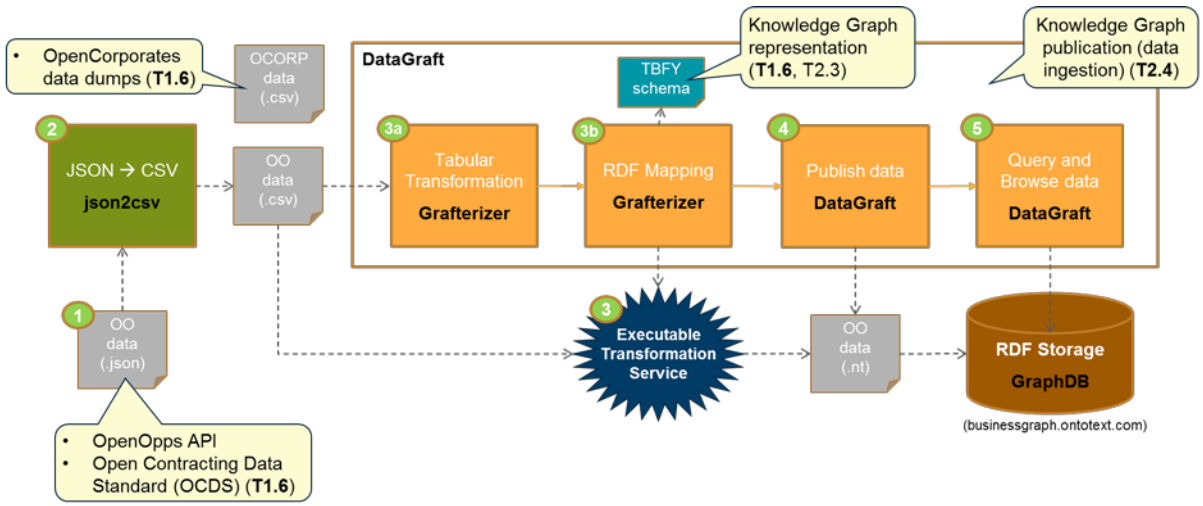

Figure 5: Data ingestion workflow

Grafterizer simplifies the data cleaning and transformation processes for onboarding procurement and company data. It provides suggestion-based data cleaning and transformation, and visual data profiling. When using Grafterizer, two main steps need to be carried out. The first step is the "Tabular transformation" which works on the tabular data (imported from the CSV files). This step mainly consists of cleaning and preparing the original dataset, e.g. deriving new data columns with URIs, properties and values so that is easier to map the data to the knowledge graph schema. Once you have prepared the original dataset, the second step is the "RDF mapping" which is a tree-based editor that allows you to map to the TBFY schema.

6 Conclusions and Future Work

This document has provided a snapshot of the ongoing work in WP1 and WP2 related to data ingestion, knowledge graph representation and knowledge graph publication in the TheyBuyForYou (TBFY) project.

The work follows an iterative and incremental development approach. Development tasks are continuously being added and assigned on the project GitHub as issues (see https://github.com/TBFY/knowledge-graph/issues) set up for the TBFY project. Some of the immediate tasks for the next phase are:

- Refactoring and updating the OCDS ontology model
- Complete the mapping rules for the OpenOpps data model
- Automate the ingestion pipeline with services with batch processing of the OpenOpps API
- Better understand the business requirements, covering business questions/queries and selection/scoping of jurisdictions for the first release of the knowledge graph

[6] https://ontotext.com/products/graphdb/

- Link OpenOpps and OpenCorporates data
- Gather information/data sources on public entities/buyers (WP1) that are not covered by OpenCorporates (which provides corporate entities)

An update of these and other development tasks will be provided in the forthcoming deliverables:

- **Deliverable D2.2 "Knowledge graph publication" (month 18).** Presentation of the knowledge graph representation and the data ingestion workflow and services to be developed.
- **Deliverable D1.6 "Ingested Data v2" (month 27).** Presentation of the final development results and the continuous knowledge graph publication processes and services.