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 D8.3

Data Management Plan

Date	25/06/2018
Author(s)	Ahmet Soylu
Dissemination level	Public
Work package	WP8
Version	2.0

Document metadata

Quality assurers and contributors

Quality assurer(s)	Oscar Corcho
Contributor(s)	Tom Blount, Divna Djordjevic, Brian Elvesæter, Matej Kovačič, Helen McNally, Ben Symonds, Francisco Yedro

Version history

Date	Version	Description
08/06/2018	1.0	First complete version of this document
25/06/2018	2.0	Revised version after internal review

Executive summary

This document describes data management plan for TheyBuyForYou project. During the project two types of data will be generated/collected. One, as the result of main project activities, is a procurement knowledge graph collecting procurement data including such as company and tender data, and second, as the result of core research activities, concerns experimental data for cross-lingual document comparison and linking (WP3) and user data including user experiments, surveys, interviews for developing data interaction components. In this document, first data is summarised, and then we present our plan for making data findable, openly accessible, and interoperable; data re-use; resource allocation; security; and ethics. Note that this is a live plan and hence this document will be updated during the project.

Table of contents

1 Data Summary

1.1 The purpose of the data collection/generation

The primary purpose of data collection in TheyBuyForYou project is to make procurement across the EU more efficient, competitive, accountable, and fair. This will be realised through gathering, normalising, curating, and integrating procurement data across EU28 countries from different primary sources into a knowledge graph (WP1 and WP2) and then providing technologies and online tools for data management, analytics, and interaction for supporting buyers and suppliers in the public-sector chain.

In addition to procurement data collection/gathering activities, research data will be generated as a result of our efforts for cross-lingual real-time monitoring and analytics (WP3) primarily for providing means to compare and link documents across languages, and interaction design and storytelling (WP4), in which interviews and survey data will be used to inform the development of interaction and visualisation tools, which will then be further refined and evaluated through the use of user studies.

1.2 Relation to the objectives of the project

The resulting knowledge graph (WP1-WP2) will focus on four key procurements areas per the objectives of the project: *(i)* economic development: facilitating better economic outcomes from public spending for SMEs; *(ii)* demand management: spotting trends in public spending to achieve long-term goals such as savings; *(iii)* competitive markets: promoting healthier competition and identifying collusions and other irregularities; and *(iv)* supplier intelligence: optimising supply chains.

The research data generated as a result of cross-lingual real-time monitoring and analytics activities (WP3) will be used to contribute to the development of a robust methodology and tools for cross-lingual comparison and linking of public spending documentation and a comprehensive real-time monitoring and analytics framework, and to discover and extract common spending patterns.The data revealed as a result of interaction design and storytelling will be used to inform the design and development of a framework combining user interaction design concepts and components, new data interfaces for data publication, compelling stories and reports, dashboards for data comparison, and visualisations of patterns and anomalies.

1.3 Types and formats of data generated/collected

Public procurement data is quite heterogeneous and covers structured (e.g., statistics, financial news) as well as unstructured (e.g., text, social media) sources in different languages and using their own terminology and formats (CSV, PDF, databases, websites, APIs etc.). Therefore, data collected will include documents in a wide variety of formats, as well as semi-structured and structured data such as JSON, CSV, RDF, PDF documents, Word documents, and HTML documents. Collected data will be integrated in a knowledge graph by applying several processes such as curation, linking, and reconciliation. The knowledge graph generated will be made available as linked data in several formats particularly RDF and JSON-LD through APIs and query end-points (i.e., SPARQL).

As part of WP4, data will be collected in the form of in-person interviews and web-based surveys (including demographic data), as well as logging data about use of the investigated systems (such as hyperlinks explored, time taken per page, etc.) as part of usability studies. Data will be made available in tabular format such as through CSV. Types and formats for data generated and collected as part of WP3 will be described later in the project.

1.4 Existing data is being re-used

Existing data that we will use mainly will come from two types of resources. First from two complementary data publisher companies, namely OpenOpps and OpenCorporates; and, secondly, from public data sources including the city council of Zaragoza, Slovenian public procurement data, and award notices in Italy.

OpenOpps is the largest data source of European tenders and contracts in the world. This data will all be made available to TheyBuyForYou via API[1]. 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 currently augments data with CPV codes where this data is not available.

OpenCorporates has the largest open database of core company data (ie existence and basic attributes) in the world. 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. To increase the accuracy and coverage of the linking OpenCorporates will also be using sources of peripheral company data which provide more identifiable attributes to entities, e.g. charity registers, VAT numbers, etc.

The city council of Zaragoza already publishes public procurement data and more generally, data about the economic information handled by the council as part of its "open data by default" principles. This includes the availability of public contract data since 2014 in its SPARQL endpoint[2][3][4] The former (public procurement data) is structured according to the Public Procurement Ontology[5], while the latter (invoices) is only available in JSON and CSV, with an ad-hoc format.

JSI has already collected data about Slovenian public procurements. These data came from Ministry of Public Administration. We also collected some additional ("support") data from Agency of the Republic of Slovenia for Public Legal Records and Related Services. These are, Slovenian business registry, public posting of company annual reports, Transaction account numbers. We have collected financial data about public spending from Erar application (maintained by Slovenian Commission for the Prevention of Corruption).

Cerved is already collecting data on award notices in Italy i.e. ANAC data. This data is gathered as part of the provisions for the prevention and repression of corruption and illegality in the Public Administration in Italy. The Contracting Authorities must comply (Law 190/2012 article 1, paragraph 32) with the obligation to publish, in open format, data on the procedures for the selection of the contractor of public contracts (calls for tenders), where public contracts are defined as contracts that commit public economic resources.

[1] https://openopps.com/api/tbfy/
[2] Examples: http://zaragoza-sedeelectronica.github.io/sparql/queries/perfil-contratante/
[3] Instructions for the end-point: https://www.zaragoza.es/sede/portal/datos-abiertos/sparql
[4] The provision of some additional data about invoices in its REST API: https://www.zaragoza.es/docs-api_sede/#/Ayuntamiento:_Facturas
[5] PPROC: http://contsem.unizar.es/def/sector-publico/pproc

1.5 Origin of the data

The data to be collected and integrated will come from the following resources: 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; and existing knowledge graphs such as the euBusinessGraph.

OpenOpp's data comes from a variety of open European tender and contract portals and currently comprises the largest source of European tenders and contracts in the world. Data from more portals will be added over the course of the project in line with the requirements of the business cases. A full list of all identified European tender portals, with an indication of whether the portal is an existing data source or is a priority to be added as a data source, is available here:

https://sintef.sharepoint.com/:x:/r/teams/work-2368/_layouts/15/doc.aspx?sourcedoc=%7B55AAE3EF-4871-40C2-AAF5-A413643C1020%7D&file=urls3%20-%20prioritisation.xlsx&action=default

OpenCorporates' data comes from a variety of official sources. Mainly company registers in each jurisdiction but also other regulatory sources like the US SEC.

The City of Zaragoza has already unified its internal treatment of public procurement data, which was previously available in a range of formats and schemata, including Excel files, relational databases and Lotus Notes. Slovenian use case data could all be obtained through Access of Public Information Act. Data sources are Ministry of Public Administration, Agency of the Republic of Slovenia for Public Legal Records and Related Services and Commission for the Prevention of Corruption. Slovenian data comes from Ministry of public administration (data about public procurements), Agency of the Republic of Slovenia for Public Legal Records and Related Services (business registry, public posting of annual reports, transaction account numbers), Commission for the Prevention of Corruption (financial data about public spending from Erar application). Cerved consolidated the Italian award notices from the open data catalogue of the National Anti-corruption Authority[6].

The data generated within WP4 will primary come from users (and stakeholders) of the system to be developed, both in terms of survation of use, and use itself, in the form of evaluation studies. The data generated as part of WP3 will come from experiments on cross-lingual real-time monitoring and analytics.

1.6 Expected size of the data

The current OpenOpps dataset is 309.2 GB. We estimate this will increase by about 50% with the additional data gathered for TheyBuyForYou, with the final dataset being around 450 GB; while OpenCorporates data includes 140M legal entities resulting in the order of 100s of GB data.

Slovenian use case data is already collected by JSI. Data about Slovenian legal entities (Slovenian business registry, Public posting of company annual reports, transaction account numbers) is about 1.8 GB; data about public procurement for years 2015-2017 is around 30 MB; and compressed (ZIP) public

[6] http://dati.anticorruzione.it/#/l190

procurement contracts (for years 2015-2017) in PDF form is about 44 GB. These are already "full data" and not the experimental one. Cerved's data on award notices in Italy includes more than 12M records from 2013 till 2018.

Regarding WP3 and WP4 experimental data, it is difficult to estimate at this stage. However, very roughly, WP4 data is expected to be less than 50GB depending on type of data collected (textual logs vs full screen capture, etc.).

1.7 Outline the data utility

The generated knowledge graph is expected to be utilised through end-points and tools built upon it by:

- public buyers committed to transparency and accountability, interested in (i) assessing their purchasing decisions; (ii) supporting healthy competition and economic growth through SMEs; and (iii) becoming more effective and productive through better demand and supplier management
- private buyers looking for taking better decisions and become more accountable towards their shareholders;
- procurement service providers and IT developers seeking to innovate in this space;
- SMEs that could not afford participating in public calls for tender because of lack of information or unrealistic bidding costs;
- businesses working with government looking for subcontractors;
- control authorities, data journalists, transparency activists, open data enthusiasts, and
- researchers in various disciplines interested in public spending, procurement processes, and data-driven innovation.

Data collected in WP4 through user studies of visualisation/interaction components will be used by: consortium researchers, for developing and refining novel visualisation/interaction tools and researchers in disciplines such as human-data interaction, interested in reproducing (or building on) experimental results. The data generated in WP3 will be used primarily for providing means to compare and link documents across languages.

2 FAIR Data

2.1 Making data findable, including provisions for metadata

2.1.1 Discoverability of data (metadata provision)

We will obtain, process and maintain the metadata associated to each dataset or data item, as required for the project, creating a Data Catalogue Vocabulary (DCAT) profile for the main datasets. We will also include collecting support data for entity recognition, addresses, identifiers, URLs and domains used by buyers. There is a specific task (T1.5) for this in the project dealing with metadata creation and a deliverable (D1.1) that will provide a catalogue of identified data sources, together with the evaluation of their quality, and the provision of their corresponding metadata.

We will create a well-curated catalogue of such data sources, with metadata descriptions (e.g., on corporate identifiers (linking to Open Corporates), buyer identifiers, post codes, CPV codes, Proclass classification and budgets) according to common metadata standards used in open data catalogues. European and global publication standards (e.g., the Common Procurement Vocabulary (CPV), the Standard Industry Classification (SIC), and the United Nations' Standard Product and Services Code (UNSPSC)) can be practically integrated into practical approaches for the publication of procurement knowledge graphs.

2.1.2 Identifiability of data

We will follow best practices for the publication of data on the Web as linked data (including RDF and JSON-LD formats, and the use of permanent URIs). The resulting data APIs will be thoroughly documented to facilitate their usage outside the context of the project, including examples of how to use them (API usage, SPARQL queries, etc.). Some data dumps (views over our knowledge graph) will be also regularly published in Zenodo, so that they become citable and can be also used for research purposes without the need to access our Linked-Data enabled APIs.

Each dataset made available during the project will be assigned a Digital Object Identifier and annotated using the latest version of the DataCite schema. We will also describe them using the DCAT-AP extension for scientific datasets recently proposed by the European Commission Joint Research Centre[7]. The extension is a close direct map to DataCite, but in RDF format, ensuring the link of the dataset with the Web of Data and improving its discoverability.

2.1.3 Naming conventions

We will use common metadata standards and vocabularies for describing the data as described above.

2.1.4 Search keyword

Search facilities will be provided in the TheyBuyForYou architecture, by means of state-of-the-art tools like ElasticSearch or alike, so as to allow for a better access to the knowledge graph data.

2.1.5 Versioning

Given the fact that public procurement data is constantly being updated in a cumulative manner, there will be no explicit versions of the data sources. The knowledge graph will be constantly updated, increasing its size while maintaining the previous data. There will be no versioning for the experimental data collected/generated in WP3 and WP4.

2.1.6 Standards for metadata creation

We will use common metadata standards, for example DCAT, as described above.

2.2 Making data openly accessible

2.2.1 Openly available data

A significant share of the knowledge graph created and enriched in WPs 1 and 2 will be accessible via open APIs and tools such as the public portal for buyers and suppliers to facilitate easy exploration, and

[7] https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27

for download under an open data license (again, CC-BY or similar). We will promote state-of-the-art data citation practices and technologies to help track data reuse.

OpenOpps data will be available under their standard share-alike attribution Open Database Licence[8]. OpenCorporates core data (company name, jurisdiction, company identifier) will be available under their standard share-alike attribution Open Database Licence[9]. Other data will potentially be available under other as-yet-to-be-decided licences.

Datasets published by WP4 in connection with a scientific publication will be made available under an ODC-ODbL license. The availability of experimental data generated in WP3 will be described later.

2.2.2 Publishing data

The resulting knowledge graph will be made available through data access APIs and query endpoints (i.e., SPARQL). The data from OpenOpps, OpenCorporates and Zaragoza city council are already available through respective APIs and end-points. The first version of the knowledge graph is estimated to be available around end of 2018.

Data linked to scientific publications by WP4 will be published in the institutional repository of the University of Southampton, Pure [10], in accordance with the (publicly available) Research Data Management policy of the University[11]. This will be decided later for WP4's experimental data.

2.2.3 Methods and tools

Data will be made available through the TheyBuyForYou REST API, as well as via SPARQL queries on the SPARQL endpoint that will be provisioned by the project. So far, the following main data sources are being published and serve as the basis of the future API:

- OpenCorporates API
- OpenCorporates Reconciliation API,
- OpenOpps API,
- Zaragoza invoice's API,
- Zaragoza's procurement data in RDF.

The guidelines on how to access such data can be found in deliverable D5.2. Furthermore, data will be regularly published in Zenodo for a complete downloadable and citeable option.

This will be clarified for WP3 and WP4 experimental data later in the project.

2.2.4 Data, metadata, and documentation

Data and their associated metadata will be made available on the developers' portal at the main project website. Data dumps will be also made available in Zenodo.

[8] https://openopps.com/legal/
[9] https://opencorporates.com/info/licence
[10] https://pure.soton.ac.uk
[11] http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html

2.2.5 Access under restriction

The data published by public administrations in the TheyBuyForYou knowledge graph will be open. In any case, state-of-the-art access restrictions will be imposed in order to ensure an appropriate quality of service of the corresponding APIs.

OpenOpps's data is made open so it can be reused by anyone provided they do so openly, reference OpenOpps and link back to OpenOpps. However, if the data is reused for commercial purposes and is not made open then this will have to be negotiated and payment made to OpenOpps. Full details of our data restrictions can be found in Open Opps Legals[12].

OpenCorporates: For OpenCorporates data that is not openly available users will be able to access the data from OpenCorporates directly (by OpenCorporates' website, API or bulk data), subject to the usual OpenCorporates terms and conditions as stated earlier.

Cerved's data coming from open source award notices is subject to intellectual property of the supplier contract, originating in the surrounding consolidation and integration activities performed by the data supplier. Hence it is not openly available, if the need is identified the raw open data can be made available in the project.

2.3 Making data interoperable

2.3.1 Interoperability

Metadata will be published using DCAT, which is widely used by government worldwide to expose the metadata of their open datasets. Data will be organised according to the Open Contracting Data Standard and the upcoming eProcurement Ontology, which is currently under development, for the public contracting data; according to the agreements done in the euBusinessGraph for the publication of data about entities; and according to open and well-documented ontologies for the rest of the data.

To ensure interoperability, datasets will be annotated using the DataCite[13] schema, complemented with the DCAT-AP extension for scientific datasets developed by the Joint Research Centre of the European Commission.

2.3.2 Standard vocabularies

As discussed above:

- Open Contracting Data Standard,
- eProcurement ontology,
- adoptions taken by the euBusinessGraph project,
- Spanish-wide standard ontologies for the rest of economic information,
- DCAT-AP extension for scientific datasets.

[12] https://openopps.com/legal/
[13] https://schema.datacite.org/

2.4 Increase data re-use (through clarifying licenses)

2.4.1 Licensing

The core public procurement data generated under the context of the project will be openly licensed under a combination of CC-BY 4.0 and Open Database License, so that any 3rd party will be able to use it with the only restriction of providing attribution. Some of the existing data sources are already coming with such a license or a compatible one (e.g., Zaragoza's data, OpenCorporates data). Data transformed from existing data sources (curated or not) will have the same license as the original source.

Data collected and generated from the project that is connected to a scientific publication will be made available with an ODC-ODbL license. In the case of data coming from user-studies/crowdsourcing, appropriate anonymisation processes will be applied before release. This is to be decided later for experimental data collected/generated by WP3.

2.4.2 Re-use and sharing

Data will be available under open licenses. No restrictions will be made to use the data that will be generated during the project. Data generated after the project finishes, using the tools generated in the project, may be subject to certain restrictions, which will be identified further in the project, except for the data that has been generated already as open data by public organisations (e.g., Zaragoza's open data), which will remain open for reuse.

OpenCorporates will provide for free basic identifying attributes (e.g., name, number, type – exact list will be defined later in the project) for matched entities, along with an OpenCorporates identifier/URL should the user wish to fetch more information about the entity from OpenCorporates directly. In the latter situation the user will be subject to the usual OpenCorporates terms and conditions.

During the TBFY project, OpenOpps's data will be free to all partners. After the project, it will revert to normal license restrictions. This means any partners using OpenOpps data commercially and not openly will have to pay for access, but the data will remain free to researchers or other non-commercial users under an Open License.

Cerved's consolidated data on award notices in Italy is under supplier license, the original raw data is made available under open license.

2.4.3 Data quality assurance

OpenOpps undertake the following steps to assure the quality of the data gathered:

- Record when the data was gathered and the data source
- Map data to open standards using manual techniques. OpenOpps has mapped millions of documents from Tenders Electronic Daily and other European procurement portals to the Open Contracting Data Standard (OCDS), creating the largest source of OCDS documents in the world.
- Testing for data integrity
- Augment data with CPV codes where none are available, thereby allowing for categorisation of data
- Detect any issues with the data sources through daily monitoring
- Collecting data on the performance of scrapers, such as when they run and how many documents they gather. Upgrading our scraper estate to allow greater classification richer

metadata to be made available, including new information on the quality of the data sourced and links to the licensing data published by each scraper.

SOTON, as part of WP4, will ensure the quality of gathered data by:

- Recording when the data was gathered and basic demographics of any participants
- Describing the quality control methods applied and their assumptions in the metadata
- Describing any software used when performing the quality analysis, including code where practical
- Including in the metadata who did the quality control analysis, when it was done, and what changes were made to the dataset
- Testing for data integrity

3 Allocation of Resources

WP1 partners are responsible for data capture, metadata production, and data quality, while WP2 is responsible for knowledge graph publication. SINTEF will be responsible for hosting the knowledge graph. A significant share of the knowledge graph created and enriched in WP1 and WP2 will be accessible via open APIs and tools during the project. The key cost element will be the hosting infrastructure.

SINTEF is currently considering alternative hosting and infrastructure solutions, including hosted cloud infrastructures such as Microsoft Azure (https://azure.microsoft.com). In addition to accessing the knowledge graph via APIs and tools, the knowledge graph data (used in demonstrators, research papers, etc.) will also be made available for download to researchers under an open data license (e.g. CC-BY or similar) on platforms such as Zenodo[14] and Datahub[15].

SOTON, as part of WP3, will reuse existing infrastructure (see Sections 2.2 & 4) as much as possible. This will be decided later for WP4's experimental data.

4 Data Security

No sensitive personal data will be collected, stored, or transferred.

5 Ethical Aspects

Any and all experiments and studies conducted by SOTON, as part of WP4, that involve human participants and/or personal data will be reviewed by the University's Ethics and Research Governance body[16].

6 Other

Public administrations in the consortium will adhere to the current and future national regulations in terms of public contracting, as well as to Public Sector Information Reuse regulations, as they are currently doing.

[14] https://zenodo.org
[15] https://datahub.io
[16] https://www.southampton.ac.uk/about/governance/policies/ethics.page