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

THEY BUY FOR YOU

Deliverable D1.2

Data gathering, extraction, pre-processing and normalisation components v1

Date	29/06/2018
Author(s)	Ian Makgill, Helen McNally (OO)
Dissemination level	Confidential
Work package	1
Version	Final

Document metadata

Quality assurers and contributors

Quality assurer(s)	Till Lech, Ahmet Soylu, Ben Symonds
Contributor(s)	Oscar Corcho, Helen McNally, Ian Makgill, Ben Symonds, Francesco Yedro

Version history

Date	Version	Description
06/06/2018	1	Initial report submitted for review
13/06/2018	2	Report rewritten and reformatted to reflect feedback and discussion from review
15/06/2018	3	Report revised in light of further review
19/06/2018	4	Report updated following review by assurer
29/06/2018	Final	Link to code in GitHub added

Executive summary

This report describes deliverable D1.2, the first version of data gathering, extraction, pre-processing and normalisation components. It explains the TBFY consortium's approach, which is to use open data on tenders, contracts and companies through OpenOpps' and OpenCorporates' APIs.

It will provide a brief description of these APIs and the Open Source code snippets for accessing the APIs, which will be made available on Github.

It will describe the data provided by OpenOpps and OpenCorporates, explain how we decided on our chosen approach, and an explanation of our approach to extraction and ingestion. It will also explain what the next steps will be and the further data that will be made accessible in the future.

Table of contents

1 INTRODUCTION 5
2 DECIDING ON APPROACH 5
3 DESCRIPTION OF THE DATA 6
3.1 TENDER AND CONTRACTING DATA 6
3.2 COMPANY DATA 6
4 DESCRIPTION OF DATA GATHERING, EXTRACTION, PRE-PROCESSING AND NORMALISATION ACTIVITY 7
4.1 DATA GATHERING 7
4.2 DATA EXTRACTION 7
4.3 DATA PROCESSING 7
4.4 DATA NORMALISATION 7
4.4.1 Categorisation 7
4.4.2 Linking 7
4.4.3 Mapping 8
5 DESCRIPTION OF OPENOPPS API 8
5.1 ACCESSING THE API 9
5.2 CONTINUOUS MONITORING 9
6 DESCRIPTION OF OPENCORPORATES API 10
7 LOOKING FORWARD 11
7.1 UPGRADES TO THE API 11
7.2 NEW DATA SOURCES 11
7.3 NEW TYPES OF DATA 13

1 Introduction

This report describes deliverable D1.2, the first version of data gathering, extraction, pre-processing and normalisation components. It explains the TBFY consortium's approach, which is to use open data on tenders, contracts and companies through OpenOpps' and OpenCorporates' APIs.

2 Deciding on approach

This is the first of three deliverables that will focus on data gathering, extraction, pre-processing and normalisation processes, as part of Work Package 1. We have agreed to align data gathering with the priorities of the project's business cases. There are still budgetary pressures and constraints that affect this work and we will need to secure a mutual agreement on funding if the allocated budget is insufficient for the needs of the business cases.

While OpenCorporates are named as the Lead for this deliverable, Open Opps has taken on this role by mutual agreement as OpenOpps have taken on the bulk of the data gathering and code development work so far and have an infrastructure optimised for gathering this data.

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 2 million documents dating back to 2010, while OpenOpps will maintain its own code for validating, mapping and monitoring the data.

The alpha version of the OpenOpps API has been made available to all partners, giving them access to over 400 sources of data around the EU, including TED and most of the accessible national level portals such as Doffin.no (see below).

A new version of the API, which allows for much greater speed and flexibility in querying the data, is currently being tested by the OpenOpps development team. The team expect to release this product to beta by 1 July 2018.

The API and the underlying data accessed through it, will be fully maintained, during the project. Each document is mapped to the Open Contracting Data Standard (OCDS) and we provide an audit trail for each record gathered, as well as validation, so that each record meets minimum requirements before publication.

The OpenOpps development team will commit to maintaining the API and the scrapers that feed data into the database for the duration of the project, so that partners can rely on a consistent, well-structured and reliable feed of data.

OpenOpps is confident that the upgraded API will provide the most dependable and efficient source of data for the project, allowing partners to benefit from the infrastructure and development investments we have made to gather the data and make it accessible. The use of an API and of the OCDS standard aligns with the technical diagrams created by the project and will allow partners to focus on the core tasks laid out in the project, such as translation, interface development and data linking.

All of the data OpenOpps publishes is openly licensed using the Open Database License.

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. OpenCorporates' primary goal is to make information on companies more usable and more widely available for the public benefit, particularly to tackle the use of companies for criminal or anti-social purposes, for example corruption, money laundering and organised crime.

In collecting this information, and matching up to other data, OpenCorporates has acquired database rights, but we strongly believe this information should be freely reusable, and so make it available (to the extent that we have the rights) under the share-alike attribution Open Database Licence.

OpenCorporates will make data on 140 million legal entities available through its OpenCorporates. OpenCorporates' API gives direct, real-time access to the underlying structured data, with powerful queries and results as JSON or XML, ready to enhance data on demand or power onboarding or investigation workflows.

3 Description of the Data

3.1 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.

The full details of the data fields available in OCDS are available here, but additional data fields can be added should the information be required for TBFY:
http://standard.open-contracting.org/latest/en/schema/reference/

3.2 Company Data

OpenCorporates has the largest open database of core company data (i.e. 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.

OpenCorporates already 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			

A full list of OpenCorporates' registers can be found here: https://opencorporates.com/registers

4 Description of Data Gathering, Extraction, Pre-processing and Normalisation activity

4.1 Data Gathering

OpenOpps is already gathering data from many European sites, including TED as well as large national portals like doffin.no. OpenOpps has made data from 477 European sources accessible to TBFY via its API.

The Data Sources Catalogue (see D1.1) shows where this data is already being gathered. Through the API, partners have access to over 100k open tenders and to contract notices.

OpenCorporates accesses data from national company registers and other sources. 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.

4.2 Data Extraction

OpenOpps extracts data from these sources using scraper scripts.

OpenCorporates uses 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.

4.3 Data Processing

OpenOpps takes the following steps to ensure the data is up to date and high quality:
Constant updates with scrapers run daily

- Mapping data to common standards (see below for more details)
- Data validation and integrity checks
- Data cleansing
- Data matching to registers of suppliers and public sector bodies

Open Corporates' scrapers map the data they access to a common schema internal to OpenCorporates. The data is stored in OpenCorporates database and linked.

4.4 Data Normalisation

4.4.1 Categorisation

OpenOpps currently augments data with CPV codes where this data is not available.

OpenCorporates identifies and categorises inactive companies and sole traders where possible.

4.4.2 Linking

OpenOpps links tender and contract data to records on public sector bodies and companies.

Copyright© TheyBuyForYou Consortium 2018-2020

4.4.3 Mapping

OpenOpps has taken in 8.7m procurement documents to its database and mapped them all to a common format, the internationally recognised and highly usable Open Contracting Data Standard.

> [The Open Contracting Data Standard is]... a global, non-proprietary data standard structured to reflect the complete contracting cycle. The standard enables users and partners around the world to publish shareable, reusable, machine readable data, to join that data with their own information, and to create tools to analyze or share that data.
> https://www.open-contracting.org/data-standard/

This format has several benefits:

- Internationally accepted standard
- Interoperable so it can compare to other datasets
- Can add additional data fields

TBFY partners have reviewed the standard to confirm it meets the needs of the business cases and will suggest where links should be made and additional data should be gathered.

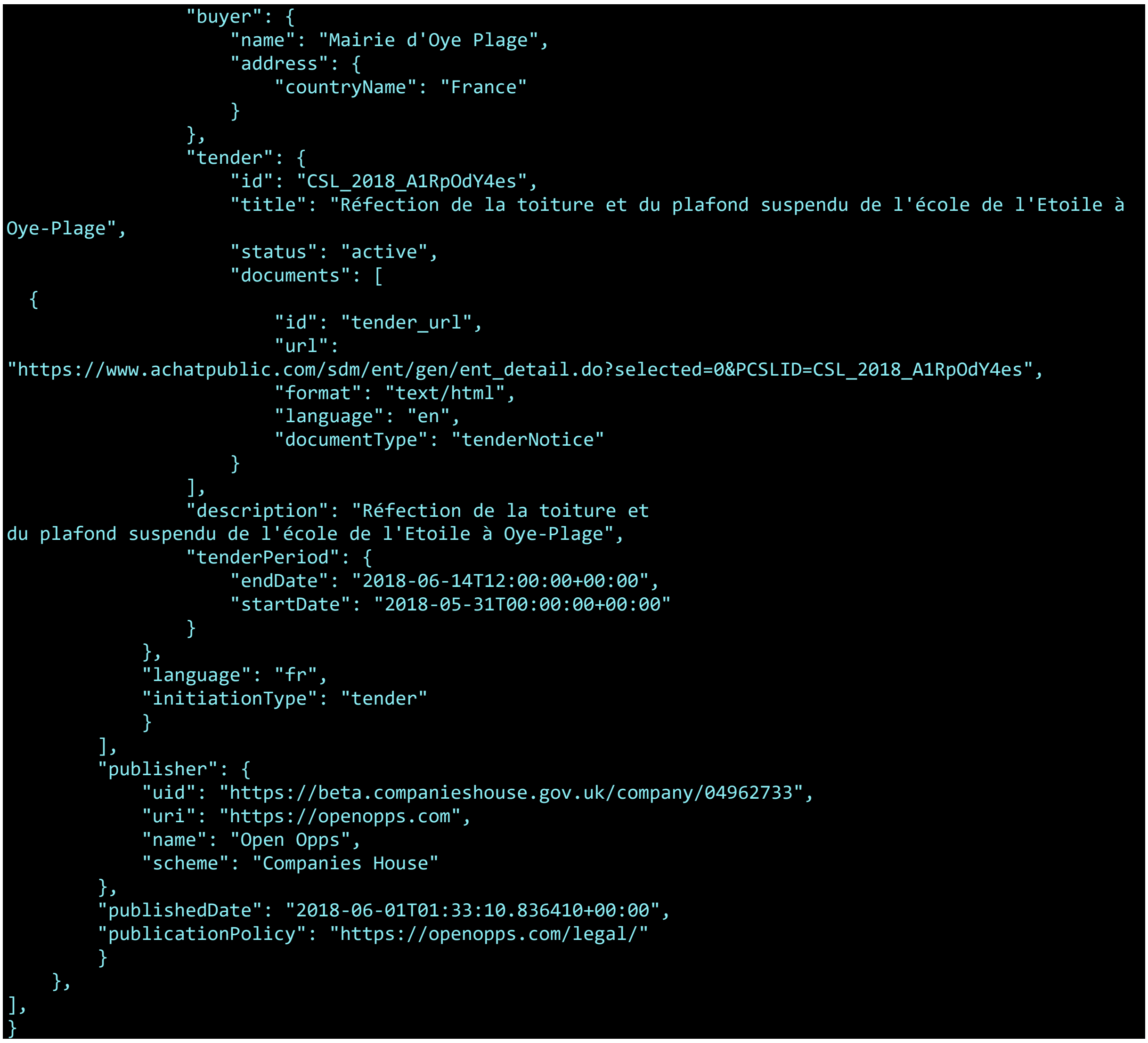
OpenCorporates' company data is mapped to OpenCorporates' own schema which can be seen in its API (see: 6. Description of Open Corporates API).

5 Description of OpenOpps API

OpenOpps has created an API for TBFY partners to access the tender and contract data. Access keys have been made available to all partners. An OpenOpps run training workshop attended by representatives of all the TBFY partners demonstrated how to use the API. This demonstration is included with this deliverable and can be accessed here: https://sintef.sharepoint.com/:v:/r/teams/work-2368/Shared%20Documents/Work/Deliverables/D1.2%20Data%20Gathering,%20Extraction,%20Pre-processing%20and%20Normalisation%20Components%20v1/TBFY%20Open%20Opps%20API%20Demo.mov?csf=1 (accessible only to TBFY team members).

Here is an excerpt of the response to the following API call https://openopps.com/api/tbfy/ocds.

```
{
    "count": 1622,
    "next": "https://openopps.com/api/tbfy/ocds/?page=2",
    "previous": null,
    "results": [
        {
        "ocid": "ocds-0c46vo-0102-CSL_2018_A1RpOdY4es",
        "countryname": "France",
        "language": "fr",
        "source": "td_achatpublic_fr",
        "releasedate": "2018-05-31T00:00:00Z",
        "date_created": "2018-06-01T01:34:54.430619",
        "json": {
            "uri": "https://openopps.com/tenders/ocds-0c46vo-0102-CSL_2018_A1RpOdY4es/?format=json",
            "license": "https://opendatacommons.org/licenses/odbl/",
            "releases": [
     {
                "id": "CSL_2018_A1RpOdY4es",
                "tag": [
              "tender"
                ],
                "date": "2018-05-31T00:00:00+00:00",
                "ocid": "ocds-0c46vo-0102-CSL_2018_A1RpOdY4es",
```



```
                "buyer": {
                    "name": "Mairie d'Oye Plage",
                    "address": {
                        "countryName": "France"
                    }
                },
                "tender": {
                    "id": "CSL_2018_A1RpOdY4es",
                    "title": "Réfection de la toiture et du plafond suspendu de l'école de l'Etoile à 
Oye-Plage",
                    "status": "active",
                    "documents": [
  {
                        "id": "tender_url",
                        "url": 
"https://www.achatpublic.com/sdm/ent/gen/ent_detail.do?selected=0&PCSLID=CSL_2018_A1RpOdY4es",
                        "format": "text/html",
                        "language": "en",
                        "documentType": "tenderNotice"
                    }
                ],
                "description": "Réfection de la toiture et 
du plafond suspendu de l'école de l'Etoile à Oye-Plage",
                "tenderPeriod": {
                    "endDate": "2018-06-14T12:00:00+00:00",
                    "startDate": "2018-05-31T00:00:00+00:00"
                }
            },
            "language": "fr",
            "initiationType": "tender"
            }
        ],
        "publisher": {
            "uid": "https://beta.companieshouse.gov.uk/company/04962733",
            "uri": "https://openopps.com",
            "name": "Open Opps",
            "scheme": "Companies House"
        },
        "publishedDate": "2018-06-01T01:33:10.836410+00:00",
        "publicationPolicy": "https://openopps.com/legal/"
        }
    },
],
}
```

More details are provided in deliverable D5.2. Full API documentation is in the development and testing phase.

Data from the API can be brought into the knowledge graph and linked to other data as part of the Work Package 2 deliverables.

5.1 Accessing the API

TBFY participants have all been granted access to the OpenOpps API. An open source code snippet in Github provides access to the API:
https://github.com/TBFY/WP1-Deliverables/tree/master/D1.2%20Open%20Source%20API%20Access

5.2 Continuous monitoring

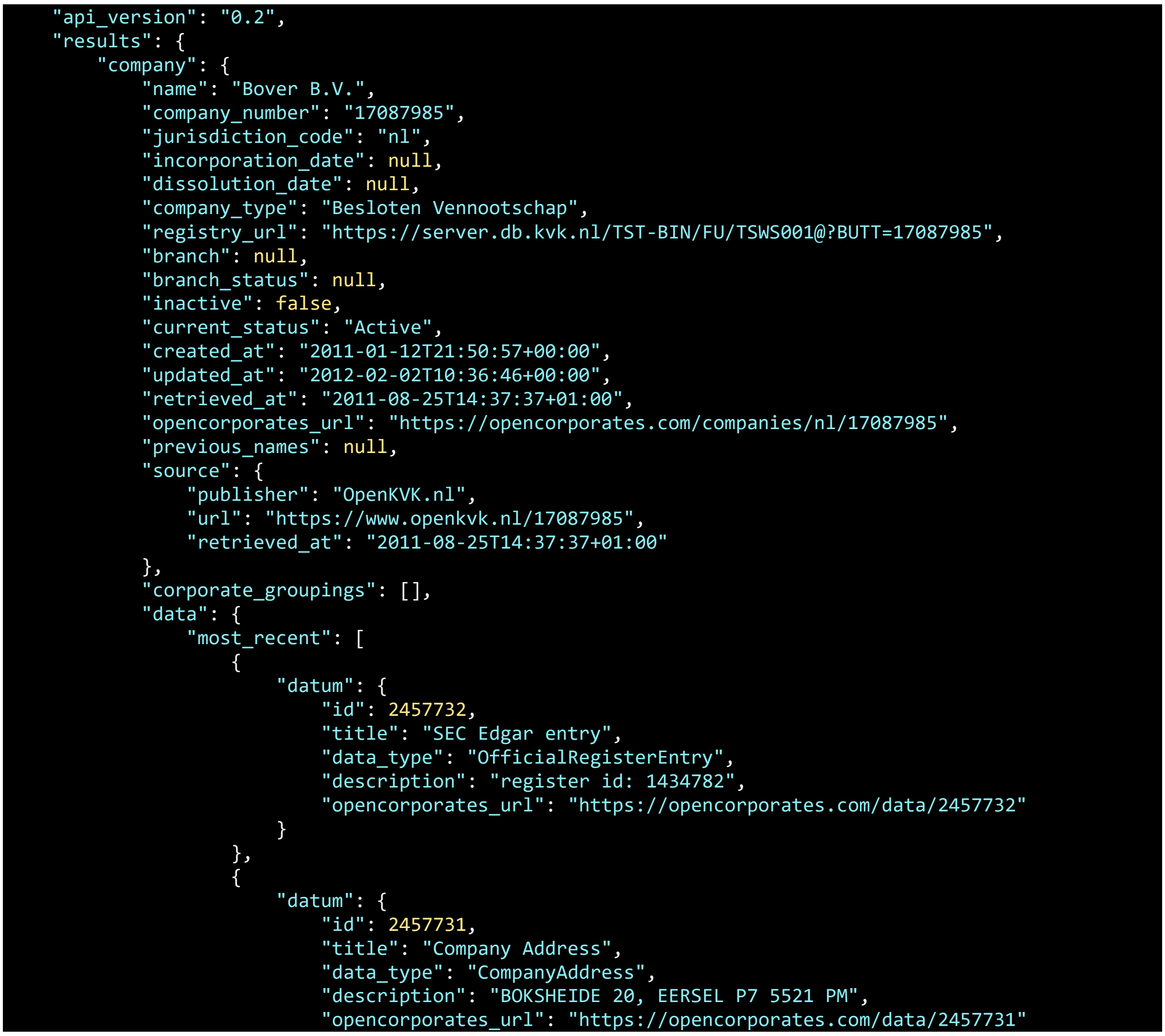
OpenOpps is already collecting data on the performance of scrapers, such as when they run and how many documents they gather. OpenOpps will be upgrading its scraper estate to allow greater classification richer metadata to be made available on our data gathering activity. OpenOpps's new metadata will include new information on the quality of the data sourced and links to the licensing data published by each scraper.

6 Description of OpenCorporates API

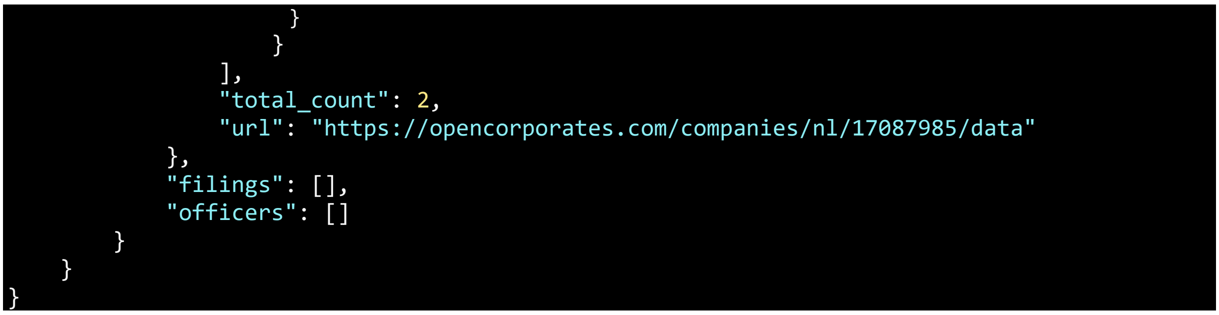
OpenCorporates is a database with information about approximately 100 million companies worldwide, obtained from a wide range of data sources, and which can be accessed via an API. Data are available either as share-alike attribution open data or commercially.

The contents of this section are based on the current online documentation available at the OpenCorporates API. An always up-to-date version is available there. The objective of this section is to provide a general overview of the main data models used by the API as well as the main types of resources and calls that are made available by it.

Among the main data offered by the API, we can find the legal name of the company, the identifier given to the company by the company register, the date the company was incorporated, the previous or alternative names of a company, registered address and so on.
As an example, this is the response to https://api.opencorporates.com/companies/nl/17087985:



```
    "api_version": "0.2",
    "results": {
        "company": {
            "name": "Bover B.V.",
            "company_number": "17087985",
            "jurisdiction_code": "nl",
            "incorporation_date": null,
            "dissolution_date": null,
            "company_type": "Besloten Vennootschap",
            "registry_url": "https://server.db.kvk.nl/TST-BIN/FU/TSWS001@?BUTT=17087985",
            "branch": null,
            "branch_status": null,
            "inactive": false,
            "current_status": "Active",
            "created_at": "2011-01-12T21:50:57+00:00",
            "updated_at": "2012-02-02T10:36:46+00:00",
            "retrieved_at": "2011-08-25T14:37:37+01:00",
            "opencorporates_url": "https://opencorporates.com/companies/nl/17087985",
            "previous_names": null,
            "source": {
                "publisher": "OpenKVK.nl",
                "url": "https://www.openkvk.nl/17087985",
                "retrieved_at": "2011-08-25T14:37:37+01:00"
            },
            "corporate_groupings": [],
            "data": {
                "most_recent": [
                    {
                        "datum": {
                            "id": 2457732,
                            "title": "SEC Edgar entry",
                            "data_type": "OfficialRegisterEntry",
                            "description": "register id: 1434782",
                            "opencorporates_url": "https://opencorporates.com/data/2457732"
                        }
                    },
                    {
                        "datum": {
                            "id": 2457731,
                            "title": "Company Address",
                            "data_type": "CompanyAddress",
                            "description": "BOKSHEIDE 20, EERSEL P7 5521 PM",
                            "opencorporates_url": "https://opencorporates.com/data/2457731"
```



```
                      }
                    }
                ],
                "total_count": 2,
                "url": "https://opencorporates.com/companies/nl/17087985/data"
            },
            "filings": [],
            "officers": []
        }
    }
}
```

More details are provided in deliverable D5.2.
Data from the API can be brought into the knowledge graph and linked to other data as part of the Work Package 2 deliverables.

7 Looking forward

7.1 Upgrades to the API

OpenOpps’ current infrastructure is insufficient for the demands of a complex and responsive API, so we are upgrading the API to provide a more flexible and speedy interface to the data. Backed by an Elasticsearch hosted lucene database, the upgraded API will allow for full text search and significantly increased performance on queries, both in terms of speed and volumes of data returned.

OpenOpps are continuing to develop the API and we anticipate that we will be able to release the new service by the 1st of July 2018.

7.2 New data sources

OpenOpps is establishing a process for adding more data by build scrapers to take in data from other sources, prioritising this work based on the needs of the business cases. The new data sources and prioritisation are all listed in the Data Source Catalogue (D1.1).

OpenOpps has begun with Italian scrapers and have already created scrapers for the following portal sites:

- https://autostrade.bravosolution.com/web/login.html
- https://www.portaleacquisti.rai.it/web/bandi_avvisi/home.shtml
- https://eappalti.regione.fvg.it/web/index.html?_ncp=1523870465457.1073-1
- http://www.carabinieri.it/cittadino/informazioni/gare-appalto/gare-appalto/
- http://www.interno.gov.it/it/amministrazione-trasparente/bandi-gara-e-contratti
- https://sol.regione.molise.it/urbi/progs/urp/ur1ME001.sto?DB_NAME=l1200158&StwEvent=101&OpenTree=1&Archivio
- http://www.sintel.regione.lombardia.it/eprocdata/sintelSearch.xhtml
- https://egas.sanita.fvg.it/it/bandi-e-gare/bandi-attivi/
- https://appalti.aqp.it/web/gare.html
- https://www.acquistionline.trenitalia.it/web/login.html
- http://albo.comunefinaleligure.it/web/trasparenza/papca-ap/-/papca/igrid/314
- http://www.aqp.it/portal/page/portal/MYAQP/FORNITORI/Bandi_gare

Copyright© TheyBuyForYou Consortium 2018-2020

- http://www.comune.venezia.it/it/node/10107
- http://portaletrasparenza.anticorruzione.it/microstrategy/html/index.htm
- http://www.regione.sardegna.it/servizi/cittadino/bandi/
- https://save-procurement.bravosolution.com/web/login.html
- https://start.toscana.it/
- https://trasparenza.escocmvs.it/index.php?id_sezione=876&id_cat=0
- https://atac.i-faber.com/index/index/hideAnnouncements/true
- http://www.comune.latina.it/category/il-cittadino/avvisi-e-bandi-on-line/
- https://inva.i-faber.com/index/index/hideAnnouncements/true
- http://www.labiennale.org/it/bandi-e-gare

Now that TBFY partners have created a prioritised list of data sources, OpenOpps will continue to develop scrapers for them. OpenOpps anticipates that there will be more requests for new sources to be added as business cases develop. OpenOpps is also allowing time to maintain these scrapers.

As well as these new scrapers, OpenOpps will also be gathering data in an XML feed from Slovenia, mapping that to OCDS, and publishing it in the new API service.

Following the review from UPM, JSI and Cerved we will initially target the following data sources:

- www.contrataciondelestado.es
- www.enarocanje.si
- www.contractaciopublica.gencat.cat
- www.contratosdegalicia.gal
- www.contratacion.euskadi.eus
- www.madrid.es
- www.madrid.org
- www.zaragoza.es
- www.prefettura.it
- www.carabinieri.it
- www.juntadeandalucia.es
- www.posteprocurement.it
- www.rfi.it
- www.eproc.ipzs.it
- www.sua-rb.it
- www.comune.latina.it
- www.aimvicenzaspa.it
- www.carm.es
- www.gruppohera.it
- www.plataformadecontractacio.caib.es
- www.gobiernodecanarias.org
- www.cittametropolitanaroma.gov.it
- www.sevilla.org
- www.provincia.vicenza.it
- www.gdf.gov.it
- www.giustizia.it
- www.fnmgroup.it
- www.globalprocurement.enel.com

- www.larioja.org
- www.asturias.es
- www.aragon.es
- www.vitoria-gasteiz.org
- www.regione.piemonte.it
- www.contratacion.gva.es
- www.siab.regione.basilicata.it
- www.contratacion.gobex.es
- www.contratacion.jcyl.es
- www.navarra.es
- www.barcelona.cat
- www.ceuta.es

More sources will be added as the project continues and the business cases develop.
OpenCorporates is adding more company registers from new jurisdictions, which will be made available to the TBFY project:

- Germany
- Russia
- Estonia
- Portugal

7.3 New types of data

OpenOpps has been gathering tender notice documents wherever they have been published, we will be adding references to these documents to the API so that project partners can gather and process these documents, allowing them to use these documents in their long-term analysis, particularly in the area of machine learning and artificial intelligence.

OpenCorpoates will be continue to discover new potential data supplementary sources and to evaluate them based on whether they will assist the matching and linking required for TBFY. Where useful new supplementary data sources are identified, OpenCorporates will work to incorporate them into its database.

Copyright© TheyBuyForYou Consortium 2018-2020