D4.6 Data Value Chain Database v2

Coordinator: Fabrizio Orlandi (Fraunhofer)

With contributions from: Isaiah Mulang' Onando (Fraunhofer), Luis-Daniel Ibáñez (SOTON)

Reviewer: Ryan Goodman (ODI)

Deliverable nature	Other (O)
Dissemination level	Public (PU)
Contractual delivery date	31th July 2017
Actual delivery date	1st August 2017
Version	04
Total Number of Pages	17
Keywords	Data value chain, database, open data

Table of contents

Executive Summary

This deliverable describes the second and final version of the data value chain database. The goal of the database is to collect information about open data datasets and the value chains which are generated around these datasets. An additional aim of this work is to create the basis for deeper studies on open datasets' consumption as performed by companies that applied to the ODINE incubator. This set of companies can be a valuable representative sample for understanding the entire European open data ecosystem.

The first version of the data value chain database (v1, described in D4.3) has been developed in order to ensure that the ODINE project maximizes its reach and keeps tracks of the available open data datasets and shareholders. With the previous version, in D4.3, we provided an essential tool for monitoring open data reuse and evolution. In particular, a tool capable of monitoring users' demand/supply for open datasets was presented. In this deliverable, we describe the final version of the ODINE data value chain database (v2) that includes real data extracted from all the proposals submitted to the ODINE call (1173 proposals).

In this report we describe the software architecture of the extraction pipeline implemented to extract datasets mentioned by the ODINE applicants. By parsing the ODINE proposals received we populate a database of open datasets which are being used by hundreds of companies in Europe (707 different SMEs). These datasets can be correlated with information related to the countries and domains of the companies using them. This allows for interesting analysis of the open data ecosystem that can be performed on top of this data value chain database. Part of this analysis is presented at the end of the report.

1 Introduction

The value chain concept traditionally involves identifying the various activities and roles in manufacturing a product. The value chain is also used as a tool that enables the analysis of interactions between the different activities in order to identify the sources for competitive advantage. However, this concept is not representative of the activities and roles in a value chain where the product is a non-tangible data product. In this context, the data value chain represents all the activities and roles specifically required when the product is a data product.

In D4.3 we defined the data value chain in detail and proposed a "demand and supply" distribution model and a system with the aim of providing insight on how an entity can participate in the global data market by producing a data product. We described our ODINE Data Value Chain Database v1 first as a cloud service prototype called "Demand and Supply as a Service". The service follows the concept of a demand and supply distribution model. This was the first attempt at creating a database of open datasets and tracking their usage by involving users and the community in expressing their demands and offer for open data.

With this deliverable we improve on this approach by populating a large database of open datasets automatically, hence not leaving this effort to the community only. We extract information about open datasets and their usage by programmatically analysing the text of the ODINE submitted proposals. In these proposals information about the datasets used by the SMEs is provided explicitly. Therefore, we created a system to process more than 1100 proposals (from 707 different companies) and populate a large database of datasets, their URLs and their usage frequency, together with information about the companies using them. Companies' information could be retrieved from the ODINE proposals metadata available on the internal Easychair submission management system. As the database contains sensitive information about companies submitting to ODINE, it will not be made available publicly in its entirety. The database will be used internally for some analysis of the open data ecosystem and the outcome of this analysis will be published on the ODINE website. A subset of the database will be publicly available and it will contain only information about the datasets, their URLs, their original publishers and the frequency of usage. Information about the ODINE applicants and their submissions will not be included nor published. Finally, this work complements the study performed by IDC[1] and included in their report on the ODINE program. In that case, the study on the open datasets used by SMEs was performed on a limited sample of 57 companies that applied to ODINE: 47 funded by ODINE and 10 not funded.

The document is structured as follows: in Section 2 the work done previously for D4.3 is briefly reviewed; in Section 3 the architecture of the data extraction pipeline and the data model of the data value chain v2 database are described; in Section 3.4 an analysis of the database is provided before concluding the deliverable (Section 4).

[1] https://www.idc.com/

2 Definitions and data value chain database v1

Due to the potential of data to be used over and over (until it remains relevant), the economic impact of adding value to it and using it as a product differs when compared to traditional product manufacturing. First and foremost this is evident in the reuse of data in another context, or domain, than it was originally envisaged for. For example, e-commerce businesses, use historical purchase data to identify patterns and suggest items to users. Moreover, the data can be processed repetitively in order to make it more usable for a specific use case, for example, by changing its format, removing irrelevant data, or linking it with other data. Data can also be interpreted and made human-readable by extracting knowledge from it. For example, in the case of government data, this data processing would enable all citizens to exploit the data, and potentially even give their feedback. In turn, this feedback could offer added value that the governmental entity can exploit. Figure 1 depicts possible actions that can be performed on a data product, hence contributing in building data value chains[2].

As described in D4.3, we define a ***Data Value Chain*** *(or Data Value Network)* to be:

A set of linked activities having the aim of adding value to data in order to exploit it as a product where different ***actors*** *can participate by executing one or more* ***activities****, and each activity can consist of a number of* ***actions****. In turn, each action can be broken down in one or more* ***data value chains*** [1].

As a first version of the ODINE data value chain database, we created a cloud service in the form of a portal. The aim was to provide an entry point to the ODINE value chain and the Economic Data Ecosystem. The service is currently available online at http://butterbur22.iai.uni-bonn.de/dsaas/. The portal caters for two discrete roles, reflecting the Demand and Supply Distribution Model, namely data producers (Supply) and data consumers (Demand).

The Demand and Supply as a Service provides two different ways for consuming data:

1. A faceted browser enables data consumers (users) to browse the Data Supply/Demand Knowledge Base of existing data that they can consume. In case a consumer needs specific data that is not available yet, we provide an online form that consumers can fill, providing details about the data that is required.
2. The second way of consuming data is through a *RESTful API*. This API enables automated access to the Knowledge Base. This enables third parties to provide their own applications based on the available data.

[2] Attard, Judie, Fabrizio Orlandi, and Sören Auer. "Data driven governments: creating value through open government data." *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVII*. Springer Berlin Heidelberg, 2016. 84-110.

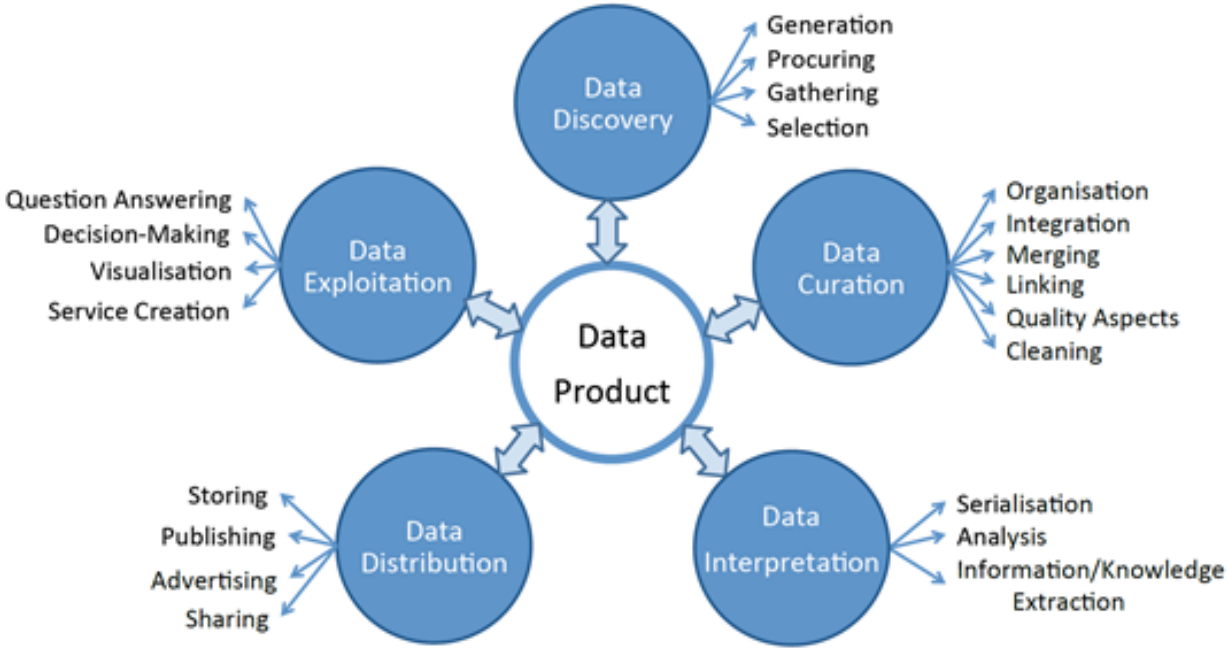

Figure 1: Activities of a data value chain

3 Data value chain database v2

While the first version of the Data Value Chain Database (DVCD) relies on users' input for the curation of the database, the second version of the database follows a different approach: information about the datasets used by different companies is programmatically extracted from the ODINE submitted proposals. Considering the ODINE call terminated after the delivery of DVCD v1, we decided to leverage such a large source of information - the ODINE set of submissions - in order to automatically populate a database of datasets and their usage by many different EU SMEs. Instead of investing effort in dissemination and community building for the DVCD v1, the consortium decided to invest the effort into building an automatic processing pipeline for ODINE submissions and building a new kind of database for the DVCD v2.

First, this solution would complement the independent study performed by IDC[3] on ODINE and included in their report on the ODINE program and its companies. In that case, the study on the open datasets used by SMEs was performed on a limited sample of 57 companies that applied to ODINE (47 funded by ODINE and 10 not funded). The DVCD v2 would follow instead a different approach and be based on a much larger sample of companies. Second, building an entire sustainable community around the aforementioned tool (DVCD v1) revealed to require much more effort than expected. Especially compared to the effort required for building an automatic extraction software pipeline. Moreover, in the context of a time limited 30-months EU project, the time available for building a sustainable community would have not been sufficient. On the other hand, the analysis of the DVCD v2 allows for an interesting study of the European open data ecosystem as it is based on a large representative sample of European open data SMEs. A larger sample of companies (707 distinct SMEs) compared to the one we could have obtained with an online community-driven tool.

The processing of the ODINE proposals is done using state of the art NLP (natural language processing) techniques. The architecture and implementation of the processing pipeline is described in the following subsections. The DVCD resulting data model is described in Section 3.3, while its analysis and application is reported in Section 3.4.

3.1 Functional Architecture

In this section we describe the architecture of the data value chain database v2 (as depicted in Figure 2). In order to collect information about the open datasets used by the ODINE applicants we need to process the PDF files of all the 1105 submitted proposals. In each proposal a specific section is dedicated to the description of the open datasets used by the applicant. It is then

[3] https://www.idc.com/

necessary to parse the text for this section only to extract the datasets mentioned by each SME (see top right box in Figure 2). We considered the possibility that companies could later on move on to using new/different datasets. We sampled a few proposals from those who got accepted/ incubated, checked this eventuality and identified only a very few minor cases in our sample. This showed the validity of the source information we are using. This tool provides a snapshot of the open data ecosystem in a precise point of time. An automated analysis of the evolution of open data usage would not be possible with the ODINE proposals data.

Each dataset mention spotted in the text can be validated and looked up on popular open dataset portals using CKAN such as DataHub[4]. DataHub provides a large catalogue of publicly available open datasets and by using its API it is possible to check if a dataset name is existing, its Web URL and what are related categories/tags assigned by DataHub users. We adopt this solution to disambiguate datasets mentions in order to extract only relevant existing datasets. Moreover, the extracted mentioned datasets can be linked to the information about the company which is available in our internal (already structured) Easychair database for the management of the submissions (top left box in Figure 2). Relevant information that can be extracted from Easychair about each company is related to information such as country, domain, ODINE acceptance status, company name, etc. However, as this is sensitive information that is not supposed to be published online, we will not publish this part of the database and we will use it only for aggregated analysis (e.g. to check what are the most popular datasets and in which countries/domains).

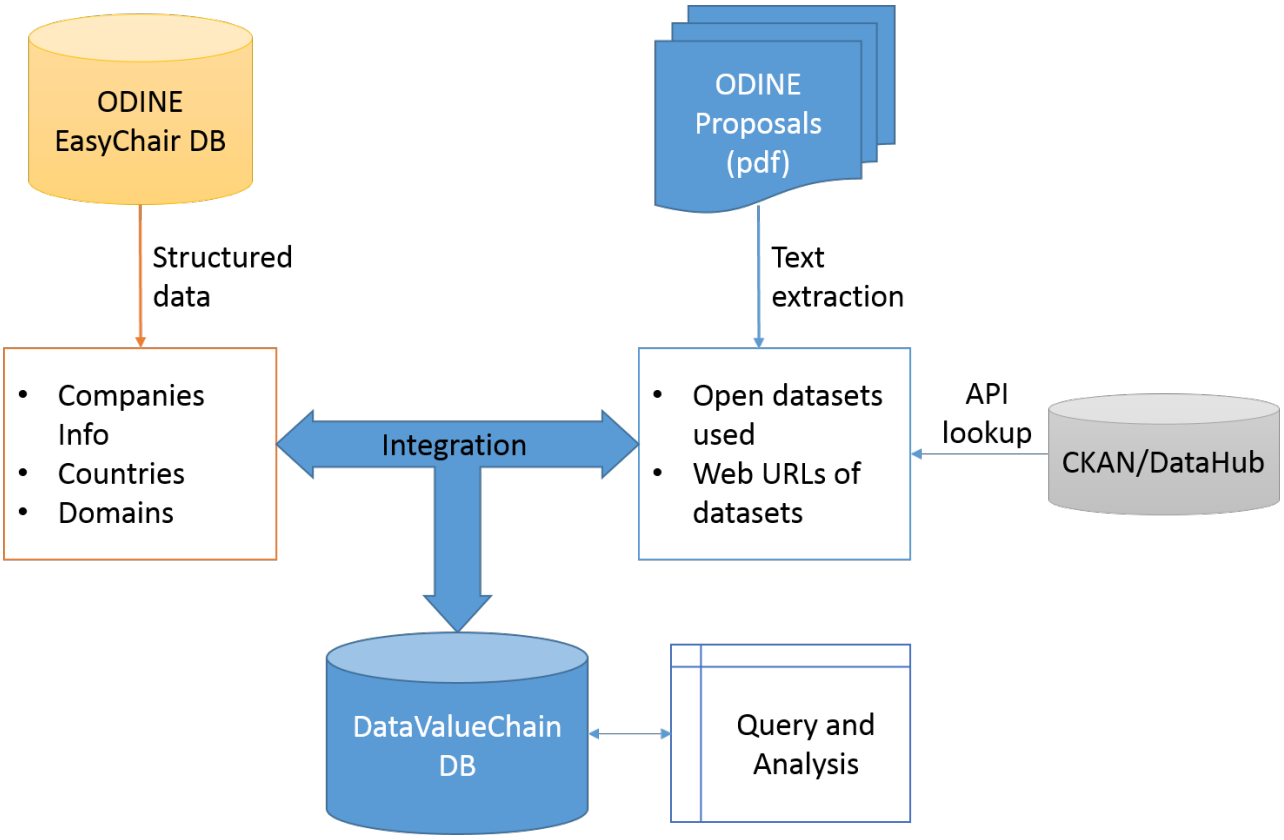

Figure 2: Architecture of the ODINE DVCD extraction pipeline

[4] https://datahub.io

Once all the aforementioned data has been collected it is integrated and stored in a relational database such as SQLite. This makes it easy to query and export it for further analysis. This database is used internally for performing analysis on the open data ecosystem and derive data stories that can be published by the ODINE project on the website. More details about the data model of the database and all its attributes is provided in Section 3.3, while examples of its application and analysis are provided in Section 3.4.

3.2 ODINE proposals text extraction

To derive meaning from the data within the proposals from the consumers of open data, we carry out text processing via the GATE[5] (General Architecture for Text Engineering) text processing tool by converting each proposal into plain text and carrying out iterative text analysis. The initial task involves identification of the sections within the SME proposals that describe data usage. This section of the proposals contains mentions of open and proprietary datasets used by the SMEs as well as possible mention of their URLs. Since our goal in this task is to identify and extract datasets that are utilised by the companies and annotate them with several attributes as shown in Section 3.3, the next iteration involves passing this text through a Natural Language Processing (NLP) pipeline. GATE provides an optimized information extraction pipeline called ANNIE (A Nearly New Information Extraction Pipeline) which we employ in our processing together with customized JAPE rules and Gazetteers built from APIs of open dataset portals (CKAN[6] and European Data Portal[7]). The dataset extraction architecture is shown in Figure 3.

The architecture, mainly consists of the ODINE GATE text extraction pipeline which we have customized for our processing needs. The inputs for this pipeline is plain natural language text from the ODINE proposals, and a list of datasets obtained from the open data portals (CKAN and EU data portal). The pipeline is made up of a linguistic processing stage, an Open Data Gazetteer, and the ODINE transducer.

[5] https://gate.ac.uk/overview.html
[6] http://docs.ckan.org/en/ckan-2.2.3/api.html
[7] https://www.europeandataportal.eu/

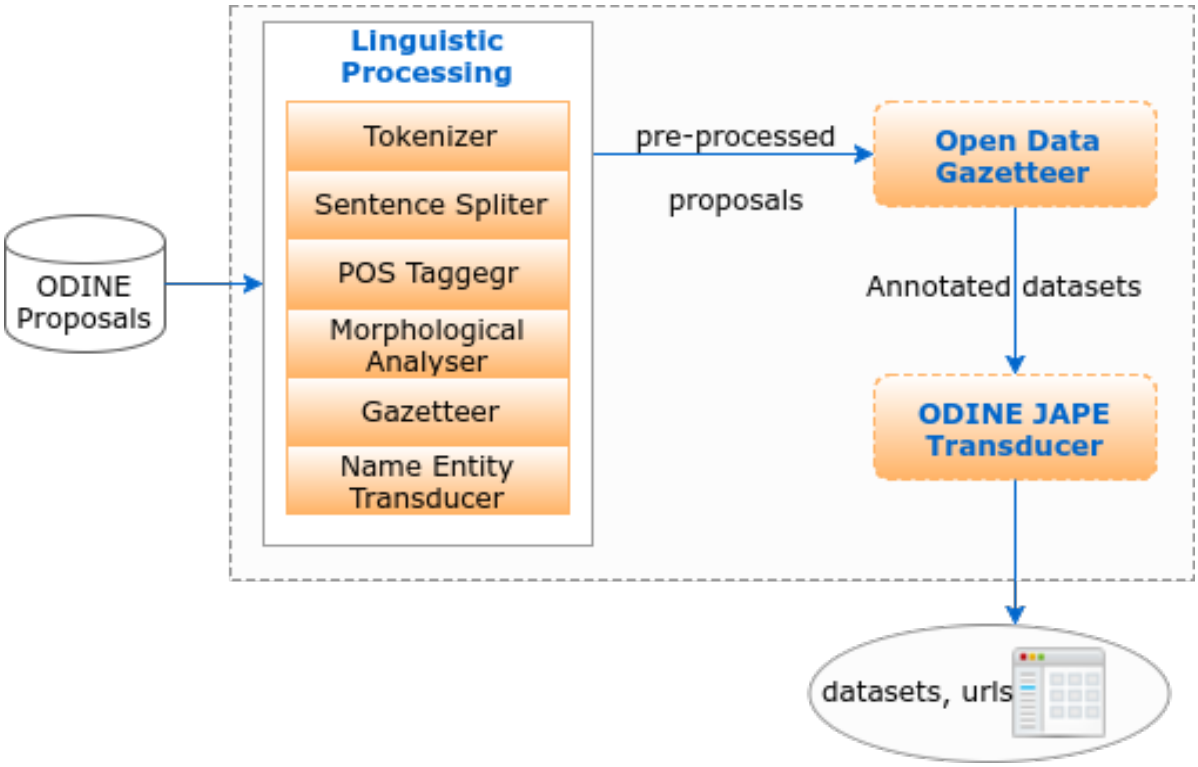

Figure 3: ODINE datasets extraction pipeline

- **Linguistic Processing**

The processing resources (PRs) present in GATE are used in our pipeline as follows:

Tokenizer: It adds two annotation sets: Tokens and space-Tokens.

The Sentence Splitter: Creates an annotation set sentence based on the sentence termination tokens from the tokenizer PR. Following this is the POS Tagger.

POS Tagger: It performs Part-of-Speech tagging and includes an annotation feature called "category" to each Token annotation, which categorizes the different tokens according to their part of speech tags.

The Morphological Analyser: It carries out a deeper analysis of the tokens, adding attributes lemma and affix for each token, these attributes are necessary during our JAPE transduction.

ANNIE Gazetteer: It reuses existing relevant lists (e.g., countries) but we add additional lists covering terms like the city names to help define rules that eliminate these from datasets names.

Named Entity Transducer: identifies named entities and adds categorical annotations such as location, person, organization etc. This information would assist during the JAPE transduction stage to identify datasets.

- **Open Data Gazetteer**

We build a Gazetteer by fetching dataset names from two popular open data portal CKAN APIs,

namely DataHub and the EU Open Data Portal[8]. With this as input, the processing resource adds a new annotation type called ‘dataset’ to the ‘Lookup’ annotation set for all matches to instances of datasets from the open data portals.

- **ODINE JAPE Transducer**

We implement the transducer that works upon the outputs from the previous processing steps and executes the final required annotations, within 4 phases. We have implemented 7 grammar rules to generate the final datasets, URLs and open dataset annotation sets. The final result is two lists, one containing the named datasets identified within the proposals and the other are the URLs directly quoted within the text. An extra annotation is added on the list of datasets to indicate the ones that were directly matched to the open data gazetteer.

- **Generation of Dataset Statistics**

With the lists obtained from the text processing, we seek to obtain the remaining attributes of the data model for each dataset by querying the CKAN API for information such as organization, tags, publisher etc. of the given dataset. We note that such information may only be available for datasets which were matched to the gazetteer entries. For the URLs, we attempt to complete any broken or mistyped URLs, followed by a Web scraping approach to obtain the titles of the sites and companies that they belong to.

We observe that our extraction was able to return all datasets within the proposals (achieving complete recall of the datasets) but the list of datasets returned is quite noisy with some items incorrectly annotated as datasets (this reduces our precision considerably). To try and circumvent this problem, we only represent those datasets that were found within open data portals to contain a valid source URL within the database. With this final constrainment of the datasets, we provide a cleaner and more precise representation of datasets within our database.

3.3 Data Model

The structure of the Data Value Chain Database (DVCD) is described, on a higher logical level, in Figure 4. The diagram represents the elements that could be captured from the ODINE proposals and their logical interconnection.

[8] http://data.europa.eu/euodp/

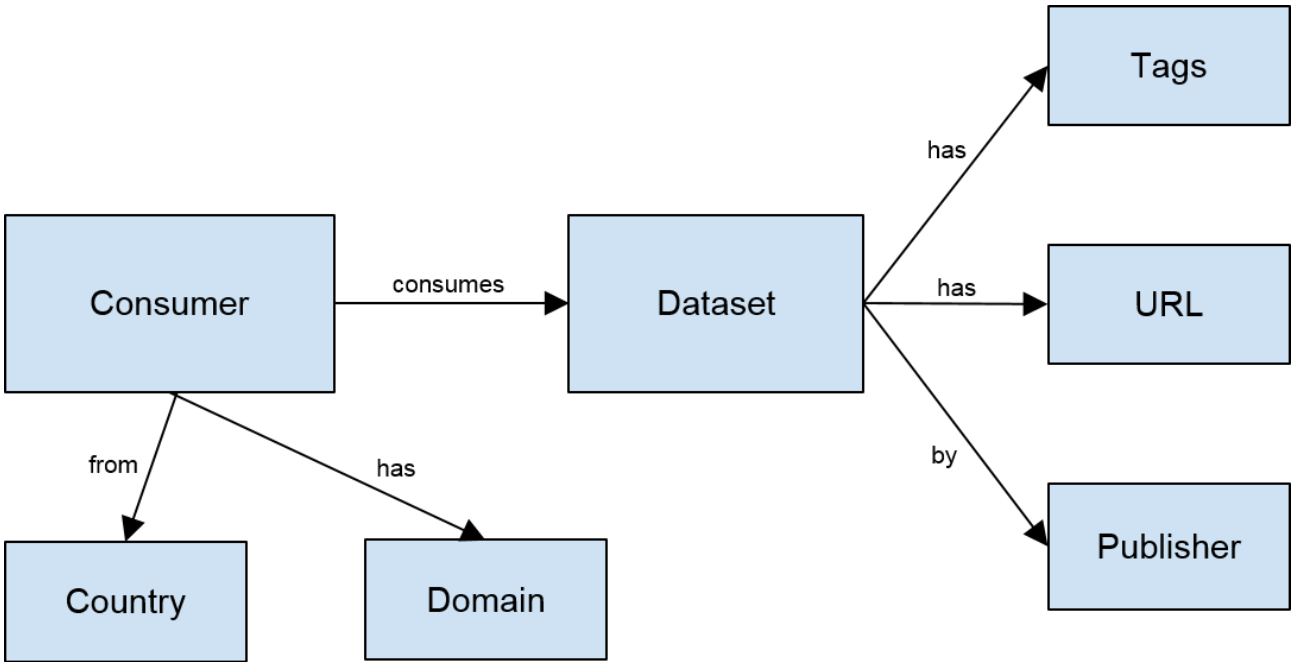

Figure 4: DVCD Logical Data Model

As we can see in Figure 4, each dataset can be used by different consumers (in our case the SMEs) who are associated with a country and a domain of activity (such as healthcare, agriculture, etc.). At the same time a dataset has been produced by a publisher (who, in a data value chain, can also assume the role of a consumer at another stage), it has some tags identifying its content and nature and also has a Web URL.

More precisely, after the text extraction pipeline described in the previous subsection, the DVCD presents the database schema illustrated in Figure 5. Three main tables represent the datasets, the ODINE proposals and the URLs respectively. Two 'JOIN' tables have been generated representing the relations between the proposals and the datasets, as well as the proposals and the URLs.

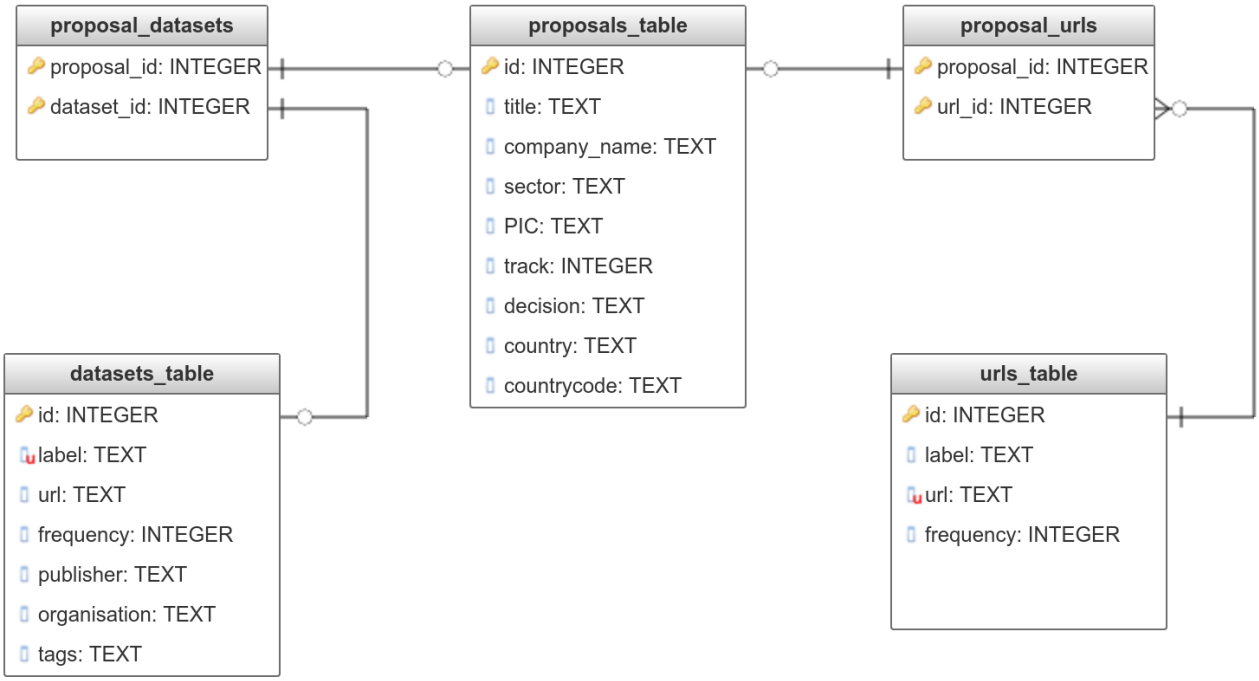

Figure 5: Database Schema

3.4 Application and analysis

The main aim of the DVCD is to enable an analysis of the open data ecosystem by looking at how SMEs utilise open data. The following are some examples of analysis that we could perform on this database.

- *What are the most popular datasets used by ODINE SMEs?*

The following figure displays the Top 15 results obtained which represent the most frequently mentioned datasets, ordered by their frequency of mention. From the table it is clear that geographical datasets are the most popular (i.e. OpenStreetMap and Geonames) followed by encyclopaedic data (e.g. DBpedia and Wikipedia) and data about cities. In the latter case, data about “cities” has been aggregated in the table in one entry only (see second row in Figure 6). This was part of some post-processing of the results that we performed when we realised that e.g. “cities” data was indeed frequently mentioned in the proposals but with our automated extraction process it was always erroneously pointing to the same data source. Instead, it would have been correct to link each mention of cities data to each corresponding EU region or country implied by the applicants and their country of business. However, this was not possible with the current natural language processing pipeline and therefore we decided to mark the entry about “cities” to *multiple data sources*. The same condition applies here for “housing” data and “hotels” data which span across different datasets as well.

#	label	url	publisher	organisation	frequency
1	osm	http://www.openstreetmap.org/	Community maintained	Open Data Day	208
2	cities	*multiple data sources*	-	-	59
3	eurostat	http://ec.europa.eu/eurostat	Eurostat	Linking Open Data	52
4	wikipedia	http://www.wikipedia.org/	Wikimedia Foundation	Wikimedia	50
5	dbpedia	http://dbpedia.org/	DBpedia Team - http://wiki.dbpedia.org	Linking Open Data	45
6	geonames	http://www.geonames.org/export/	Geonames	Open Geospatial Data	36
7	ckan	http://www.ckan.net/	OKF	Global	29
8	opencorporates	http://opencorporates.com/	OpenCorporates (Chris Taggart)	Linking Open Data Cloud	25
9	housing	*multiple data sources*	-	-	18
10	nasa	http://data.nasa.gov/	Global	Global	16
11	eea	http://www.eea.europa.eu/	-	Linking Open Data Cloud	14
12	datahub	http://datahub.io	-	LODCloud2014	10
13	hotels	*multiple data sources*	-	-	10
14	istat	http://www.istat.it	ISTAT	it-ckan-net	8
15	gutenberg	http://www.gutenberg.org/	Project Gutenberg	Bibliographic Data	7

Figure 6: Top 15 most frequently mentioned datasets

- *How many companies are using these datasets?*

In the database we have 707 different companies and 88 datasets. As in the last column in Figure 6 we can see how many companies mentioned to use each of the Top 15 datasets. In the case of OpenStreetMap (the most popular dataset) we can see that 208 companies mention it in the proposals. This shows that approximately 29% of the open data SMEs are using (or interested in using) OpenStreetMap. The reason for this is that not only open geographical information is very interesting for businesses, but also that geographical data can easily be combined with other datasets from different domains to provide new services.

- *From which countries are these datasets used?*

If we take for example the top dataset, OpenStreetMap, we can analyse which countries are using this data most. The result of this query is in Figure 7 below. In this case, Italy and Germany are leading users for that particular dataset, but similar analysis can be performed for all the other datasets in order to study the EU data ecosystem and the data usage distribution across Europe.

	country	count
1	Italy	21
2	Germany	17
3	Spain	11
4	United Kingdom	11
5	France	4
6	Greece	4
7	Romania	3
8	Austria	2
9	Belgium	2
10	Czech Republic	2
11	Hungary	2
12	Bulgaria	1
13	Switzerland	1
14	Denmark	1
15	Israel	1

Figure 7: Top countries of the companies using the OpenStreetMap dataset

This kind of analysis can be performed for many datasets and companies and for different dimensions of the data. More details with interesting results of this analysis will be publicly available on the ODINE website.

4 Conclusions

Data is a commodity in our information society and the “dataification” of products has led to the need for a change in existing value chains. Therefore our aim is to project our vision of generating a new Economic Data Ecosystem based on the concept of data value chains. With this deliverable we developed a database collecting information about 707 different companies who applied to ODINE and are working with open data. We analysed their proposals and derived some interesting analysis over their datasets usage. For example, the most popular datasets are geographical datasets followed by encyclopaedic ones. The most mentioned dataset is OpenStreetMap and it is largely used by ODINE companies based in Italy and Germany. Further analysis of the database will be included in the ODINE website, here we proposed only some of the main examples of analysis that can be performed on the proposed data value chain database. A public version of this data will also be publicly available on the ODINE website for the community to perform their own analysis.

5 References

[1] Attard, J., Orlandi, F., and Auer, S. (2016). Value Creation on Open Government Data. In *Proceedings of the 49th Hawaii International Conferences on System Sciences (HICSS 2016),* Kauai, HI: IEEE.