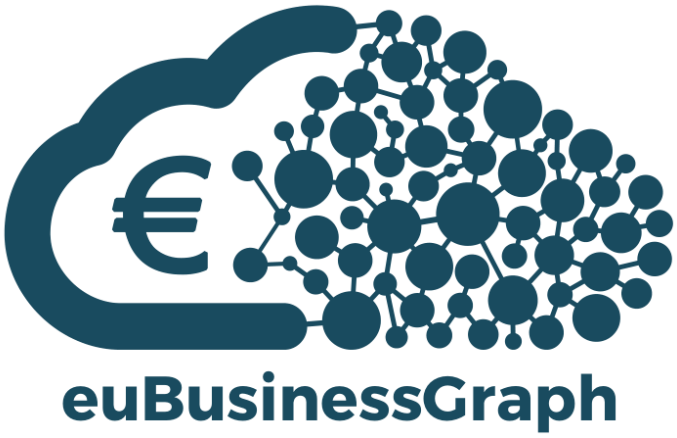
Innovation Action (IA)

ICT-14-2016-2017
H2020-ICT-2016-1

Enabling the European Business Graph for Innovative Data Products and Services



Deliverable D1.1:

Data Gathering, Quality Assessment and Management Plan

Date	02.03.2018
Author(s)	Andrea Maurino (UNIMIB) (Editor)
Dissemination level	Public (PU)
Work package	WP1
Version	1.1

Document metadata

Quality assurors and contributors

Quality assuror(s)	Kay Macquarrie (DW), Vladimir Alexiev (ONTO)
Contributor(s)	euBusinessGraph Consortium

Version history

Version	Date	Description
0.1	10.07.2017	Initial Table of Contents (ToC).
0.2	22.09.2017	Role and responsibility.
0.3	10.09.2017	Insertion of dataset.
0.4	18.10.2017	Frist definition of business rules.
0.5	14.10.2017	Second version of business rule.
0.6	21.11.2017	New rules added.
0.7	07.12.2017	Pre-draft.
0.8	11.12.2017	First draft released.
0.9	20.12.2017	Second draft version, with suggestion of internal reviewers.
0.9.5	27.12.2017	Updated version addressing comments by internal reviewers.
1.0	29.12.2017	Final formatting and layout.
1.0.1	15.01.2018	Update of document.
1.0.2	02.02.2018	New sections added.
1.0.3	21.02.2018	New version released.
1.1	02.03.2018	Final formatting and layout.

Executive summary

The main goal of the euBusinessGraph project is to create the foundations of a European cross-border and cross-lingual business graph through aggregating, linking, and provisioning (open and non-open) high-quality company-related data, thereby demonstrating innovation across sectors where company-related data value chains are relevant. This is achieved by leveraging the power of emerging technologies such as Data-as-a-Service and Linked Data.

This report describes the data gathering and quality strategies applied to collect and process datasets described in the business cases, their data quality requirements and the data management plan for the whole project.

The Data Management Plan (DMP) reports on the data that the euBusinessGraph project will use and generate during its lifetime, from the set-up of the euBusinessGraph platform to the business exploitation of its services. By following Horizon 2020 guidelines, the DMP defines the general approach that will be adopted in the context of the euBusinessGraph project in terms of data management policies. In accordance with these guidelines, this deliverable will include information about the handling of data during and after the end of the project, reserving attention to the methodology and standards to be applied. As a consequence, the DMP describes the approach established in euBusinessGraph to ensure the life-cycle management of the public and proprietary datasets provided by the consortium members to the project as well as other dataset produced by the consortium during the project execution.

In particular, this report describes rules, best practices and standards used with regard to data gathering, to ensure their quality and to make the data findable, accessible, interoperable and reusable (FAIR data) and the process to collect and manage data in compliance with ethical and legal requirements. The deliverable includes a high-level description of the business cases and descriptions of the datasets provided for the euBusinessGraph project. These descriptions aim to detail identification, origin, format, access, security of the data and to take into account legal and ethical requirements. We plan to represent some of them as machine-readable metadata using DCAT and/or VOID.

Table of contents

1 Introduction

This report presents Deliverable D1.1 "Data Gathering, Quality Assessment and Management Plan" of the euBusinessGraph project. This deliverable is developed as part of Work Package 1 (WP1) "Data Gathering, Transformation and Publication".

1.1 Objective

The objective of WP1 is to ensure a process through which the relevant company data for the business cases is collected, quality assured, transformed, and onboarded in the business graph. Specifically, this WP aims to:

- Collect and analyse relevant company related data, both internal/private business data and external Open Data as well as publicly available news and web content.
- Audit datasets from data providers based on agreed business rules to ensure data quality.
- Transform, link and publish data from data providers and business cases as Linked Data to meet the business case requirements.

Having in mind these needs, this report will collect the adopted strategies for collecting data, a set of business rules to assess and ensure the quality of collected data and the general strategies for data management according to best practices in similar EU projects.

1.2 Relationships to other Work Packages and Deliverables

The deliverable D1.1 partially covers the first-year activities related to task T1.1 Data Gathering, task 1.2 Data quality assessment and task T1.4 data management plan. Please note that task T1.1 and task T1.2 end in month 18 and the results of such tasks will be described in deliverable D1.2 Data Transformations and Onboarding.

The definition of the initial version of the data management plan will drive the development of both euBusinessGraph platform (WP3) and development of business cases (WP4). The DMP will be updated if a new data-source is collected during the project.

The list of gathering strategies as well as quality oriented business requirements will be part of the development of business cases (WP4) while the requirement of business case described in D4.1 and the result of data model definition described in D2.1 are other inputs of this deliverable.

1.3 Document structure

The remainder of this report is structured as follows:

- Section 2 describes the data gathering strategies adopted for the project
- Section 3 reports the most important quality assessment rules related to business cases of the euBusinessGraph platform
- Section 3.3.5 describes the Data Management Plan

2 Data gathering requirements

Company related data used in euBusinessGraph comes from various sources. Most of the data is already available at the consortium partners and some of the data can be collected directly online. We will describe the requirements and procedures which are already used by some of the consortium partners and can be used for additional data gathering when there is no data available at the partners.

The term data used in the following chapters means data related to the companies:

- Company firmographics data such as names, identifiers, addresses, etc.
- Official gazettes describing changes about the companies
- News articles about the companies

This section is organised as follows. In Section 2.1 the list of possible external sources are listed. The following section describes the way in which is possible to analyse data sources while Sections 2.3 and 2.4 describe issues related to the data collecting.

2.1 Finding data sources

Quality company related data is very important for making any investigation, analytics or decisions based on this data. A good quality data is typically available for a fee from large providers like Dun & Bradstreet (www.dnb.com) or Bureau van Dijk (https://www.bvdinfo.com), but a lot of this data is also available for free with some pre- and post-processing. For using the free data sources, the following should be observed:

- Sources must be authoritative, like for example:
    - National registers or regional Chambers of Commerce
    - Government agencies (with possibly open data access)
    - Tax authorities
    - Business licensing bodies
    - Official government notices (gazettes)
    - Established/verified news sources
- Checks should be made, in particular:
    - Is the provider the main originator of the data, or a re-publisher?
    - Is the listing complete or just a sub-set?
    - Are unique identifiers present?
    - Which attributes of the data source are available?
    - How easy it is to obtain the data source from a technological point of view?
    - What are the legal terms of data re-use according to the specific licence?

Preferred sources would offer open data, available in bulk in some of the standard formats (such as RDF,) or API without serious limitations like number of accesses or amount of transferred data. If the first found sources are limited in any of these options, other official sources should be searched and possibly combined together to obtain a more complete, up-to-date and precise dataset.

2.1.1 Possible data sources

Some of the possible data sources for company related data could be:

- Wikipedia list of company registers
    - https://en.wikipedia.org/wiki/List_of_company_registers
- OKFN list of company open datasets

  - https://index.okfn.org/dataset/companies/
- EU Commission list of EU company registers
  - https://e-justice.europa.eu/content_business_registers_in_member_states-106-en.do
- EBR European Business Register
  - http://www.ebr.org
- RBA Information Services List
  - http://www.rba.co.uk/sources/registers.htm
- A list of official company registers by country provided by Companies House
  - https://www.gov.uk/government/publications/overseas-registries/overseas-registries
- A list of official company registers around the world maintained by the Commercial Register Office of the Canton St. Gallen, Switzerland
  - http://www.commercial-register.sg.ch/home/worldwide.html

Some of the possible news sources about companies could be:

- International Business Times, http://www.ibtimes.com
- Forbes, https://www.forbes.com
- Economist, https://www.economist.com
- Reuters, https://www.reuters.com
- Business Insider, http://www.businessinsider.com
- BBC, https://www.bbc.com
- Guardian, https://www.theguardian.com
- Wall Street Journal, https://www.wsj.com
- Financial Times, http://www.ft.com
- Bloomberg, https://www.bloomberg.com
- Handelsblatt, https://www.handelsblatt.com

2.2 Analysing the data

Once the data sources are identified and a sample data is obtained, an analysis of the acquired data is needed. Out of possible entity types the scope needs to be determined:

- Which company types will be collected:
  - Legal entities
  - Sole traders
  - Foreign branches
  - State-owned companies
  - Non-profits / foundations
- Which business news events are interesting:
  - Merges
  - Acquisitions
  - Bankruptcy
  - Key person (company officer) changes

To be able to uniquely identify entities each entity needs a unique identifier. These can be identifiers provided with the data if they are universal or a proprietary identifier (like DUNS number from Dun & Bradstreet). In any case, original identifiers and any additional identifiers such as VAT IDs need to be stored for later identification and matching.

All interesting fields from the input data are usually mapped to the internal representation. An example of this mapping would be the dates or the address.

When checking enumerated fields (those that have few possible values, e.g. current status), care should be taken to normalize the input so that these fields can be used for example for searching or filtering.

Company data from most national registers and international business news are obtained in different languages and/or character sets. Some of the input fields (e.g. company name, company type) are mapped to local language representation for faster searching or filtering (e.g. “società a responsabilità limitata” that is company limited by shares in Italian), but the rest of the data is stored in original representation.

2.3 Importing the data

After sample data is analysed, properly mapped and checked for errors/inconsistencies, the bulk amount of data is imported in the internal database. This process typically needs to be repeated, as there are always some lines or fields in the input data, which can't be parsed/processed successfully. In this process the recognition of bad input data improves and the software is made more robust.

These improvements are most notable in news processing, so when new sources are added, news from them can be processed much quicker.

2.3.1 History

The amount of data stored depends on a local policy and national legislation. If new data is added to the already existing set, a history of changes can be preserved.

2.4 Automating data collection

Stored company data needs to be updated from time to time. Sometimes there are requirements that new data should be available almost immediately after official release. In such cases a system of automated data collection, cleaning, mapping and importing must be put in place, as well as data update flows.

If incremental updates are available, this process is easier to implement. If not, a "data difference" mechanism should be implemented in order to understand the modification of data. The new bulk import must then include data from previous imports stored in history records.

When doing automated data collection, the restrictions of the original data or news sources must be respected. These restrictions may range from number of connections to the source per day/month, to the size of transferred data.

When scraping data from the Web, the usage of Web proxy servers is recommended; in fact one of the major issues with using web scrapers is that they request too many pages in too short a period of time from a single IP address, which can be easily traced and blocked by the target website. To limit the chances of getting blocked, it is important to avoid scraping a website with a single IP Address. In such case the use of proxy servers that use different proxy IP addresses whenever the requests are routed over the crawling server, is a typical solution. These can be leased for a small fee typically for a month and are automatically rotated/renewed every month.

2.5 Specific data gathering processes for business cases

2.5.1 Data gathering and updating process for OpenCorporates CED

The data gathering process for the data necessary for OpenCorporates’ Corporates Event Data Product is being implemented in the following steps:

Data discovery

Events data will be inferred primarily from two sources:

- Changes in company attributes, as reported to and by official sources (normally central company registers).
- 'Signals' from filings from other sources, initially Government Gazettes in Europe, and associated countries (e.g. EFTA)

OpenCorporates has now been collecting, collating and standardising company data now for seven years, and has arguably a greater knowledge and understanding of the sources than perhaps any other organisation. While in general, the canonical source for company data is a central company register, this does vary, and it's also not that unusual for the data to be available from multiple different official sources, and we analyse these by a number of measures, as detailed in this blog post: https://blog.opencorporates.com/2017/04/11/from-company-register-to-standardized-open-data-our-processes-explained-part-1-scouting-for-data/

Similar processes have been developed for government gazettes, although these will be refined for this project to understand which gazettes (often a country may have more than one) provide useful signals about company data.

Data selection

The depth and quality of data held at company registers varies considerably, even in Europe, and that made freely available even more so. In addition there is no consistency between company registers about schema, terminology, scope, register structure or even underlying concepts. This means there is a considerable amount of analysis that needs to be done before the data can be collected, understanding the nature of the source, and the information it contains. OpenCorporates has extensive and well-documented procedures for performing this analysis, as detailed in this blog post: https://blog.opencorporates.com/2017/10/23/from-company-register-to-standardized-open-data-our-processes-explained-part-2-analysis/

There are similar processes for government gazettes, but these will be refined for this project to understand which notices map to which event type.

Data collection and importing

Following our extensive analysis of the sources, we collect the data using a data pipeline refined over many years (and still being iterated on). Core company information is collected using the following workflow:

- initial analysis (detailed in 'data selection', above)
- design, and writing of 'bot' using our internal bot framework. Data at the source is made available in a variety of ways:
    - open data dumps, and these may be a single CSV file, multiple CSV files, XML or JSON files, or even a custom format, such as fixed field length files. In addition some sources make only the whole dataset available, some use deltas, and so make only active companies available (meaning inactivity needs to be inferred by the bot)
    - APIs, including REST APIs, SOAP APIs, ElasticSearch APIs, with XML or JSON returned
    - Web pages, which need to be parsed into structured data, with a variety of mechanisms to discover the pages containing the company info, including alpha search, date range search, company number search, company type search
    - Other files e.g. PDFs or Excel files
- QA of data produced by bot - this is a multi-stage, iterative process, initially working on a subset of data, to check that the data does indeed match the initial analysis, and to make any corrections as a result. For example, a schema or data dictionary may list all the company types, yet the actual data may include ones that are not in that list. In many cases the documentation relating to the data is incomplete
- Ingestion of data from the bot, using the OpenCorporates company data pipeline

Gazettes are handled using the same process but using OpenCorporates 'Turbot' bot framework and data pipeline (which is used for non-company data)

Multilingual issues

OpenCorporates has a policy of maintaining the source data in the original language and alphabet, although changing the encoding to UTF-8. This is because its primary goal (and one that end users rely upon) is to reflect the official public record. While it is expected that it will over time develop functionality around transliteration and translation, there are a number of practical and conceptual issues relating to this, and our users have said that they would rather we focus on us increasing the breadth and depth of data on OpenCorporates rather than tackling this issue. However, even though we don't transliterate or translate the data in general (e.g. company names, addresses), for enumerated values (e.g. company type or company status) in non-Latin alphabets we do provide an English translation as well as the original phrase in the original characters.

Data updating

A key part of the Corporate Events Dataset is not just handling updates to the company data that is being ingested into OpenCorporates, but inferring 'events' relating to them from changes contained within those updates. The technical implementation of this is still being finalised, but it is likely to be a three-stage process:

1. The updating of the company data (same process as currently implemented within OpenCorporates)
2. The creation of a 'snapshot' of the company, detailing the attributes after the update has been applied
3. A comparison of the new snapshot with the previous on, triggering the creation of events as necessary

The creation of events can then trigger alerts, be queried by the API, or be concatenated into an events data dump.

Government Gazettes are immutable filings (updates are published as subsequent gazette notices, or amendments rather than the original record being changed), and so do not experience updates from the source. We may in the future investigate the potential for improving gazette parsers, and then updating the parsed records, but that currently isn't in this stage of work.

2.5.2 Data gathering and updating process for TDS

The data gathering process for Tender calls dataset needed for TDS was implemented as in following steps:

- Data discovery, outlining exiting sources of open tender calls in Italy.
- Data selection, based on subject matter expertise of the domain and analysis of the data sources samples including frequency of update, available fields.
    - Mandatory information being existence of at least one of the following:
        - Business name for the public entity publishing the tender call or
        - VAT number of public entity publishing the tender.
    - As outlined in the dataset metadata collection section following data sources were selected for this phase:
        - Italian Municipalities - Albo Pretorio: Milano, Ravenna
        - MePA (Mercato Elettronico della P.A)
        - Calls for public contracts – regional, province and town portals: towns (Rome, Milano, Torino and Napoli); regions/provinces (Alto Adige Lazio Toscana, Emilia-Romagna Molise, Valle d'Aosta); buying centers (Asmel, Consorzio-cev, Empulia, ESTAR-Toscana, RTRT-Toscana, Observatory of Lombardia); companies (Città salute Torino, Atac)
        - Calls for public contracts from SIMOG – ANAC portal

- Calls from national Service for Public Contracts (Ministry of Infrastructure)
- Italian Tender Electronic Daily Public Procurement Notices

Data collection and importing

This phase consists of:

- the web scraping process that was implemented in a Data As A Service (DAAS) scenario including development of scrapers, managing of the process, server maintenance, proxies, and basic quality control on the data. The basic data quality control includes (i.e. retrieving CIG, combine identical CIG from different sources, formatting dates in ISO 8601, alignment of amounts and numbers to 2 decimals, alignment of keys in a unique format when the same fields in defined with different terminologies).
- Initial importing in a bulk format
- Incremental updates with traced history i.e. fields: date of inserting, date of modification, date of obscurement.
- Matching with internal subject in order to couple with a unique internal key that allows enrichment with proprietary information assets in Cerved, and continuous data quality, “cleaning and matching” improvement.
- TDS is focused on Italian market therefore no multilingual issues are present at the data importing level. Any multilingualism at the application level would be handled through relying on euBG graph model.

The Data updating strategy for the Tender calls dataset follows and will follow the business strategy for TDS, that is it consist of periodic reviews of the data being gathered with focus on:

- What aspects covered within the strategic goals of TDS require new data?
- What new markets proposed in the updated TDS business strategy require new data?

2.5.3 Data gathering and updating process for ATOKA+

SDATI enriches company entities using several sources with varying degrees of structuredness of the information to be extracted.

Some sources are available as full database dumps, with several serialisations used. Other sources provide APIs to query and retrieve changes from a particular timestamp. In other cases, the data is published using a format that is not programatically accessible (e.g., public data that requires the resolution of captchas). Finally, some sources are directly published in the web as HTML and require the use of specialised crawlers.

Company-related web content is processed through the Corporate Web Crawler (CWC) looking for company information such as:

- Company names, registration numbers, tax identifiers
- Telephone numbers
- Emails
- Addresses
- Social web accounts
- Positions
- Company locations
- Company-to-company web links

CWC performs the enrichment one jurisdiction at a time. Website knowledge is incrementally built, being enlarged on each run with new sites from several sources. The crawling process is done to cover all known websites for a particular jurisdiction and all page content is crawled.

Afterwards, a series of analyses are performed looking for specific types of information, either data that can be directly used to enrich company information (e.g, telephone numbers and e-mail addresses, social media accounts, e-commerce tools, company description, employees, roles) or features that are ingested by processes that compute metrics or indicators about the company (e.g., company-to-company weblinks that are used to compute the company's web centrality), using different subsets that are sorted differently on each run. The expected relevance of a page for a certain type of analysis is determined looking for specific patterns (e.g., the presence of tokens in the page's URL, the presence of certain elements in the page). These patterns are usually jurisdiction and language-dependent and if found, increase the expected relevance score.

In general, distance from the home page lowers the expected relevance score.

At the end of the analysis all crawled content is annotated and indexed into a Crawled Content Index (CCI).

Cross-lingual issues

Web crawlers set the preferred language to match the primary language of the jurisdiction to be crawled so the page shows content in that language, which is expected to be most up-to-date.

Information such as phone numbers and addresses follow different rules that are bound to the jurisdiction. Detection strategies must vary accordingly.

The preliminary analysis that decides which types of information will be looked for on a specific page vary wildly depending on the language and jurisdiction. As an example, when crawling Italian websites, the presence of permutations or variations of the phrase "chi siamo" (who we are) will increase the expectancy that this page will have text giving an overview of the company.

For some jurisdictions there is more than one transliteration system in place (e.g., Italy uses "SOCIETÁ" and "SOCIETA'"). Before attempting any syntactic or semantic analysis, the transliteration system is annotated if not already present in the metadata.

Updating strategies

Information gathered by the CWC to enrich company information is not kept between runs. The only type of information that is preserved between runs is the list websites that were crawled for a particular jurisdiction. This list is built upon and expanded after each new run.

Company information already present in the database is used to query the CCI. Matching is done using the most relevant query results. Matched content in the CCI is used to update the corresponding information in the company database. Information about the company that come from sources other that the CWC can be used to do further processing and filtering on the candidate information contained in the CCI.

For structured and semi-structured sources (processed outside of CWC), state is preserved for each run. For each item extracted, regardless of the source, the following timestamps are produced:

- Last visit: last time the item was found on the source
- Creation time: the first time the item entered the database
- Last update: last time the retrieved item was different from the version present in the database.
- Deletion time: the moment at which the item is considered to no longer exist

In this way, regardless of the type of source (incremental or full, with explicit or implicit timestamps) it is possible to model the lifecycle of each item.

2.5.4 Data gathering and updating process for CRM-S

EVRY is creating an Analytic Platform that will enrich data in the euBusinessGraph. To be able to do that euBusinessGraph needs to upload historic data that is needed by the machine learning models. The models will use available data from the euBusinessGraph as well as external datasets (that is only used to develop the machine learning models) from other sources. These data will therefore be utilized to create more value to existing euBusinessGraph data. Since external data sources may be proprietary, the pricing model will reflect which data sources that have been used in a model training.

We have started developing models that predicate default risk for companies using the following data:

- Accounting information (BRREG)
- General company information (BRREG)
- Bankruptcy information (BRREG)
- External remarks (external dataset)

The variables used in the models has been selected by domain experts and been examined in statistical analyses to ensure that the most significant variables are chosen. This analysis will also reveal in what degree the variables influence the models result, and if there are any multicollinearity present.

External datasets that is used in combination with euBusinessGraph requires an attribute as a shared identifier. In case of combining Norwegian data from BRREG the key identifier is the organization number. The key is dependent of the business case, and requires the key in euBusinessGraph to be constant.

Cross-lingual issues

Data sources can origin from multiple countries and may therefore vary with a number of unknown factors. The models will therefore be country specific and will not be affected by cross-lingual data.

Updating strategies

euBusinessGraph will upload relevant data to EVRYs Analytic Platform periodically (every month). This data will be persisted and the machine learning models will operate on this data in a sliding window. This means that e.g. for default predictions;

The first run will include data from 01/01-2015-01/01-2018,

The second run will include data from 01/02-2015-01/03-2018,

The third run will include data from 01/03-2015-01/04-2018,

And so on

This ensures that the model training/testing is performed on updated information.

We will use a 36 months duration for the sliding windows. There will be implemented quality assurance mechanisms to ensure that newly added data maintains a high level of quality. When models are trained on new data, they must be manually accepted on the base of Receiver Operating Characteristic (ROC) and the false/positive rate.

2.5.5 Data gathering and updating process for BRC

The process of discovering, gathering data in this business cases is quite simple due to the fact data needed are already available in BRC. In fact the business cases foreseen the publication of existing data of BRC as open data.

Moreover, no multilingual issues are foreseen in the project due to the fact that the jurisdiction is only one.

Updating strategies are still under development, but due to the availability of original data a complete regeneration of that is the most suitable solution.

3 Quality assessment

This section describes the quality dimensions, and set of business rules referring to quality metrics specific to company-related data that are collected in the business cases. Section 3.1 describes commonly adopted quality dimensions in both academia and industry that will be considered in the euBusinessGraph project, while in Section 3.2 for each dataset described in Section 4 the set of most important business rules are reported.

3.1 Data Quality Dimensions relevant for company related data

Quality of data has several possible definitions. Among others, a relevant definition[1] describes data quality as “fitness for (intended) use”. Therefore the definition of what are the facets of quality that a dataset must hold is strictly related to the use of data itself. According to the literature[2,3] there are four categories of data quality rules. They can be classified as:

- Business objects or business entities category
- Data elements or business attributes category
- Types of dependencies between business entities or business attributes category
- Data validity rules category

In the next subsection we briefly summarize such category of rules.

3.1.1 Business Entity Rules

Business entities rules are related to the structure of a dataset and they are subject to three data quality rules: uniqueness, cardinality, and optionality. These rules have the following properties:

- **Uniqueness**—Every instance of a business entity has its own unique identifier. The identifier must always be known.
- **Cardinality**—Cardinality refers to the degree of a relationship. That is the number of times one business entity can be related to another. There are only three types of cardinality possible: One-to-one cardinality, one-to-many (or many-to-one), many-to-many cardinality.
- **Optionality**—Optionality is a type of cardinality. It identifies the minimum number of times they can be related. There are only two options: either two business entities must be related at least once (mandatory relationship) or they do not have to be related (optional relationship). Optionality has a total of five rules; the first three apply to the degree of the relationship: One-to-one, one-to-zero (or zero-to-one), zero-to-zero.

3.1.2 Business Attribute Rules

Business attributes are subject to two data quality rules:

- **Data inheritance**—The inheritance rule applies only to supertypes and subtypes. Business entities can be of a generalized type called a supertype, or they can be of a specialized type called a subtype. For example, FRAMEWORK CONTRACT is a supertype entity, whereas CONTRACT is a subtype of FRAMEWORK COMPANY.
- **Data domains**—Domains refer to a set of allowable values. For structured data, this can be any of the following: list of values, range of values (in RDF called Datatype), and constraints on values (data facets), such as set of allowable characters, a pattern, min-max values, etc.

3.1.3 Data Dependency Rules

The data dependency rules apply to data relationships between two or more business entities as well as to business attributes.

- **Entity dependency**—The three entity-relationship dependency rules are:

[1] Juran on Planning for Quality. The Free Press, New York 1988
[2] Larissa Terpeluk Moss, Majid Abai, Sid Adelma. Data Strategy, Pearson 2005
[3] Batini, Scannapieco, Data Quality: Concepts, Methodologies and Techniques, Springer 2006

- The existence of a data relationship depends on the state (condition) of the other entity that participates in the relationship. For example, Employee numbers cannot be placed for a company whose status is "individual."
- The existence of one data relationship mandates that another data relationship also exists. For example, when an order is placed by a customer, then a sales-person also must be associated with that order.
- The existence of one data relationship prohibits the existence of another data relationship. For example, an employee who is assigned to a project cannot be enrolled in a training program.

- **Attribute dependency**—The four attribute dependency rules are:
  - The value of one business attribute depends on the state (condition) of the entity in which the attribute exists.
  - The correct value of one attribute depends on, or is derived from, the values of two or more other attributes. For example, the value of Pay Amount must equal Hours Worked multiplied by Hourly Pay Rate.
  - The allowable value of one attribute is constrained by the value of one or more other attributes in the same business entity or in a different but related business entity. For example, when Loan Type Code is "ARM4" and the Funding Date is prior to 2015-12-01, then the Ceiling Interest Rate cannot exceed the Floor Interest Rate by more than 6 percent.
  - The existence of one attribute value prohibits the existence of another attribute value in the same business entity or in a different but related business entity. For example, when the Monthly Salary Amount is greater than ZERO, then the Commission Rate must be NULL.
- **Data provenance—** Rules related to this type refers to the assessment of the origin of data value and the process related to data transformation from origin. Example of data provenance rules are:
  - The existence of metadata related to the origin of data
  - The existence of information about the process transforming data, including code source

3.1.4 Data Validity Rules

Data validity rules govern the quality of data values, also known as data domains. There are six validity rules to consider:

- **Data completeness**—This rule specifies that a given set of business attributes must be filled. For example the business attribute CompanyName must be filled.
- **Accuracy**—This rule describes that data values must be correct. For example, values of attribute Company.jurisdiction must be included in the list of recognized nations available at https://en.wikipedia.org/wiki/List_of_sovereign_states.
- **Precision**—This rule specifies that all data values for a business attribute must be as precise as required by the attribute's:
  - Business requirements,
  - Intended meaning,
  - Intended usage,
  - Precision in the real world.
- **Consistency—**This rule specifies that some business attributes must to follow a given pattern. For example the age and date of birth attributes are connected by the following rule age = year (today) – year (dateOfBirth)
- **Time related data—**Rules of this type refer to the temporal dimension of data. They may refer to volatility (the average time between an update of data), timeliness (the average age of a

values) or currency (when a data is entered in the system). An example of such a rule would be “the last modification date of attribute Company. Revenue must be more recent than a year ago”

3.2 RuleSpeak syntax

There are many possible ways to describe a business quality rule. It can range from an algorithmic style such as

- CompanyName EXISTS AND len(trim(CompanyName)) <> 0

to semantic based definition by using the SHACL notation[4]

```
ex:PersonShape
        a sh:NodeShape ;
        sh:targetClass ex:Company ;    # Applies to all companies
        sh:property [
        sh:path ex:CompanyId ;# constrains the values of ex:CompanyId
                sh:maxCount 1 ;
                sh:datatype xsd:string ;
        ] ;
        sh:closed true ;
        sh:ignoredProperties ( rdf:type ) .
```

We chose a more natural and business oriented way to describe rules by means of the RuleSpeak[5] notation. It is an existing business rule notation developed by Business Rule Solutions supporting the definition of business rules with a clear semantics. It is worth nothing that in this phase we are more interested in expressing business oriented quality rules that are independent of specific adopted technologies that will be deployed in WP2, WP3 and WP4. The use of controlled natural language can support both the business manager in understanding the requirements and the technical manager who needs to implement them.

Technology independence imposes the definition of high level business rules. It is possible that in the development of WP4 there will be the need to create new a more technology oriented quality rules. For example, a technology oriented quality rule may require the validation of a set of RDF triples against a SHACL[6] or ShEx[7] specification

The syntax of the model is shown in table Table 1. Notice that r, s, and t, are all parts of the same proposition related to the same business context. In a permissibility formulation (that is in a in a possibility formulation), the ‘only’ is always followed immediately by one of the following:

- an ‘if ’ (yielding ‘only if’)
- a preposition.

Table 1: RuleSpeak syntax

Modal claim type	RuleSpeak keywords
obligation formulation	*r* must *s*
obligation formulation embedding a logical negation	*r* must not *s* *r* may *s* only if *t*
permissibility formulation	*r* may *s* *r* need not *s*
necessity formulation	*r* always *s*
necessity formulation embedding a logical negation	*r* never *s* *r* can *s* only if *t*
possibility formulation	*r* sometimes *s* *r* can *s*

[4] https://www.w3.org/TR/shacl/
[5] http://www.omg.org/spec/SBVR/1.4/Annex-H--The-RuleSpeak-Business-Rule-Notation/PDF
[6] Shape constraint language https://www.w3.org/TR/shacl/
[7] https://www.w3.org/2001/sw/wiki/ShEx

An example of a business rule statement using the "only [preposition]" form follows:

- The official VAT number of a company may be provided only by a national public administration.

An example of a "bidirectional" form "only" follows:

- The value of companySize must be “mediumSize” if and only if the value of Company. NumberOfEmployees is between 15 and 50

3.3 Business rules

This section reports the business quality rules related to the assessment of dataset described in the DMP (see Section 4) according to the quality dimensions reported in Section 3.1. As already described the number business quality rules for each dataset could be very long; moreover in some cases the list is not exhaustive due to the fact business cases are running and new quality rules can be rise. For these reasons we report the top 10 quality rules for each business cases representing the minimum quality level that each dataset must satisfy in order to be used at business level. It is worth to underline that during the implementation of the business cases such quality rules need to be translated according to the specific adopted technology. For example the rule “each company identifier MUST be unique in the dataset” can be translated in different way if data are stored in a relational database of in a graph. In the first case the rule will be “company identifier attribute must be the primary key of company table”, in second case the rule will be “each node type <<company>> must have an not null attribute called <<company identifier>> and values of such attribute must be unique in all instance of node of type <<company>>”. For these reasons we report the top 10 quality rules for each business cases.

More detailed and technological oriented rules are reported in Deliverable 3.3 Requirements Analysis, Architecture and API Specification for the euBusinessGraph Marketplace.

3.3.1 euBusinessGraph data model quality rules

The euBusinessGraph platform will provide an integrated dataset about European level business related data. Each data provider that want to participate to it must satisfy the following minimum quality rules to provide their data. Deliverable D2.1 shows the euBusinessGraph data model; in the definition of the model a number of business requirements are raised and they represented the business rules reported in the following table.

Table 2: Business rule for euBusinessGraph data model

Number	Type	Business Quality Rule
EUBG1	Necessity	The dataset MUST show clearly the licence.
EUBG2	Obligation	euBusinessGraph model MUST support languages: EN, IT, NO.
EUBG3	Obligation	Social data of companies, such as their websites (together with Web languages), RSS/Atom feeds and Wikipedia URLs MUST be included in the data model.
EUBG4	Necessity	Company contact information, such as the address and other locations MUST be included in the model.
EUBG5	Obligation	The model should capture key company metrics, such as the number of employees.
EUBG6	Necessity	The identification of a company MUST be unique.
EUBG7	Obligation	A link to company web site MUST be included.
EUBG8	Necessity	A link to data provider home page MUST be included in the model.
EUBG9	Possibility	Metadata about trustworthiness of source MAY be included.
EUBG10	Necessity	Company jurisdictions and registration information MUST be included.

3.3.2 OCORP business rules

The CED business cases generates the following business rules.

Table 3: Business rule for OCORP

Number	Type	Business Quality Rule
OCORP1	Obligation	A corporate event MUST have a unique, permanent identifier.
OCORP2	Obligation	The type of a corporate event MUST be one of the defined types.
OCORP3	Obligation	A corporate event MUST have a time range.
OCORP4	Obligation	The time range of a corporate event MUST have either a `begin` timestamp or an `end` timestamp.
OCORP5	Obligation	A corporate event MUST have a `description` of the event.
OCORP6	Obligation	A corporate event MUST have a `provenance` detailing the provenance of the event.
OCORP7	Obligation	The provenance of a corporate event MUST have a `created_at` indicating when the event was added to the CED system.
OCORP8	Obligation	The provenance of a corporate event MUST have a `source` indicating where the information for the event originated.

3.3.3 CERVED business rules

With respect to the Tender Discovery Services (TDS) business case, the table below reports the 10 most important business quality rules.

Table 4: Business rule for TDS

Number	Type	Business Quality Rule
CERVED1	Obligation	Business names for the public entity publishing the tender calls coming from field “ragione_sociale_stazione_appaltante” values **must** be coherent with the fiscal codes of this entities in field “codice_fiscale_stazione_appaltante” as in the Cerved’s Public administration (PA) data model.
CERVED2	Permissibility	A publishing entity may not exist in Cerved’s PA datamodel or the association to exiting entity may not be certain, in this case a doubt is created and worked by operators.
CERVED3	Obligation	Fiscal codes of the public entity publishing the tender calls **must** be valid numbers (i.e. using Luhn formula).
CERVED4	Obligation formulation embedding a logical negation	The CIG identifier **may** not be conformant to a sequence of 10 alphanumerical characters **only if** the correct format can be recuperated from one of the fields or the attached pdf blob.
CERVED5	Obligation	At least one of the date fields in the tender call descriptions indicating date of publishing and closing date **must** be a valid date in an adequate range (year).
CERVED6	Obligation	Fiscal codes and official names for Italian business companies participating in and winning tender calls **must** be matched to Cerved’s internal company knowledge graph.
CERVED7	Necessity formulation embedding a logical negation	Fiscal codes and official names for business companies participating in and winning tender calls **can** be matched to euBG graph **only** if the graph API allows matching using relevant jurisdiction and company fiscal codes.
CERVED8	Necessity	The total amount for the tender call excluding VAT **is always a number** greater than zero.
CERVED9	Obligation	At least one of the dates indicated in the description of the open tender call (i.e. fields data_pubblicazione_simog, data_pubblicazione_servizio_contratti_pubblici,

		data_pubblicazione_GUUE etc.) **must** be prior to the closing date of the tender call (i.e. data_termine_offerta).
CERVED10	Possibility	The "id" tender portal identifier and CIG unique identifier in tender call descriptions **can** indicate tender calls that are part of the same lot.
CERVED11	Necessity	If two or more tender call descriptions coming from different sources share tender call ids or CIG ids they refer to the same lot and the same tender call and are **always** uniquely identified as such with "sameAs" id.
CERVED12	Necessity	A tender call description is **always** matched with a normalized location of the public entity publishing a tender call, and location of where the works/service, for which the tender call was published.
CERVED13	Possibility	A tender call description **can** be matched with **a** normalized location of where the works/service, for which the tender call was published, needs to be delivered if this location is valid.

3.3.4 SDATI Business cases

Business rules related to ATOKA+ the business case led by SDATI is shown in Table 5.

Table 5: Business rule for ATOKA+

Number	Level	Business Quality Rule
SDATI1	Obligation	A company in the IT jurisdiction that is marked as belonging to the public sector MUST be present in the PA dataset (the Public Administration Index).
SDATI2	Obligation	A company in the IT jurisdiction that is marked as being a startup MUST be present in the Italian Registry of Startups.
SDATI3	Obligation formulation embedding a logical negation	Latest data about number of employees MUST NOT be older than 3 months Latest data about number of employees MAY be older than 3 months ONLY if it is a one-person company or not actively trading.
SDATI4	Obligation formulation	The preferred contact number of a company MUST be the one with the highest preference score among all the contact numbers of the company.
SDATI5	Necessity formulation	The preference score of a contact number ALWAYS increases whenever a new source contains the contact number.
SDATI6	Obligation formulation	A company MUST have at least one location.
SDATI7	Obligation formulation	A company MUST have at least one of its locations be the headquarters.
SDATI8	Obligation formulation	Revenue tendency MUST be null if latest sample is older than 2 years or the number of consecutive samples up to the latest is less than 3.
SDATI9	Possibility formulation	A company CAN have 1 or more websites.
SDATI10	Obligation formulation	A company with 1 or more websites MUST have a main website.

3.3.5 EVRY Business rules

In Table 6 the business rules supporting the EVRY business cases are swhon

Table 6: Business rule for EVRY

Number	Type	Business Quality Rule
EVRY1	Necessity	A company MUST always have one and only one organization number.
EVRY2	Obligation	euBusinessGraph MUST contain data for the last two years.
EVRY3	Obligation	All financial data for a company MUST be numeric.

EVRY4	Necessity	Persons linked to a company Role MUST be in full name and include a PID.
EVRY5	Possibility	Persons in the dataset MAY be involved in multiple companies.
EVRY6	Obligation	A company MUST belong to a company type.
EVRY7	Obligation	A role MUST belong to a role type.
EVRY8	Obligation	A bankruptcy record MUST contain organization number.
EVRY9	Obligation	A bankruptcy record MUST contain a termination date.
EVRY10	Permissibility	euBusinessGraph MAY support live update of data.

3.3.6 BRC Business rules

Finally the list of business rules supporting the business cases of BRC are shown in Table 7

Table 7: Business rules for BRC

Number	Type	Business Quality Rule
BRC1	Obligation	An organisation number used to refer a legal entity MUST be a registered number to return a result.
BRC2	Obligation	An organisation number used to refer a legal entity MUST have a valid format to return a result.
BRC3	Obligation	An organisation type must exist among registered companies to return a result
BRC4	Obligation	A role MUST be of a valid type.
BRC5	Obligation	Accounts amount values MUST be numbers.
BRC6	Obligation	A legal entity under voluntary or compulsory liquidation or dissolution MUST have exactly one organisation number.
BRC7	Obligation	A deleted legal entity MUST contain a termination date.
BRC8	Obligation	All boolean values must be represented as true or false.
BRC9	Permissibility	Organisation type MAY be specified by code or text.
BRC10	Permissibility	Phrase-based search string MAY be used, as allowed by Elasticsearch.
BRC11	Permissibility	The number of legal entities of a query MUST not exceed 10,000.

3.4 Analysis of business rules

The set of Business rules that are reported in the section 3.3, can be classified according to the quality dimensions described in section 3.1. to a better understanding of what are the most important features that in each business case is assigned to data they use. In Table 8 the rules are annotated to the quality dimensions that considered notice that a business rules can be related to multiple data quality dimensions

Table 8: Business rules and related quality dimensions

Rule Number	Primary quality dimension	Secondary quality dimension
EUBG1	Completeness	
EUBG2	Completeness	
EUBG3	Completeness	
EUBG4	Completeness	
EUBG5	Completeness	
EUBG6	Uniqueness	
EUBG7	Completeness	
EUBG8	Completeness	
EUBG9	Completeness	
EUBG10	Completeness	

OCORP1	Uniqueness	
OCORP2	Accuracy	
OCORP3	Data domain	Completeness
OCORP4	Consistency	
OCORP5	Completeness	
OCORP6	Completeness	
OCORP7	Completeness	
OCORP8	Completeness	
CERVED1	Accuracy	
CERVED2	Entity dependency	
CERVED3	Consistency	
CERVED4	Data domain	Consistency
CERVED5	Data domain	Completeness
CERVED6	Accuracy	
CERVED7	Accuracy	
CERVED8	Consistency	
CERVED9	Consistency	Consistency
CERVED10	Attribute dependency	Consistency
CERVED11	Attribute dependency	
CERVED12	Accuracy	Accuracy
CERVED13	Accuracy	Accuracy
SDATI1	Accuracy	
SDATI2	Accuracy	
SDATI3	Currentless	
SDATI4	Consistency	Attribute dependency
SDATI5	Attribute dependency	
SDATI6	Consistency	
SDATI7	Cardinality	Completeness
SDATI8	Attribute dependency	Consistency
SDATI9	Attribute dependency	
SDATI10	Attribute dependency	
SDATI11	Cardinality	
SDATI12	Attribute dependency	Uniqueness
EVRY1	Uniqueness	
EVRY2	Currentless	
EVRY3	Data domain	

EVRY4	Completeness	
EVRY5	Cardinality	
EVRY6	Accuracy	
EVRY7	Accuracy	
EVRY8	Data domain	Accuracy
EVRY9	Data domain	Accuracy
EVRY10	-	-
BRC1	Consistency	
BRC2	Data domain	
BRC3	Consistency	
BRC4	Data domain	
BRC5	Data domain	
BRC6	Attribute dependency	Uniqueness
BRC7	Attribute dependency	
BRC8	Data domain	
BRC9	Data domain	

It is worth nothing that not all business rules can be mapped to quality dimensions. For example business rules EVRY10 “euBusinessGraph MAY support live update of data” is related to assess the quality of the process of collecting data that have an impact on the curentless of data.

Analysing Table 8 is possible to notice that the top three quality dimensions considered at level of the entire euBusinssGraph project are Completeness, Accuracy and Consistency as shown in Figure 1 and Figure 2 where the distribution of business rules with respect to quality dimensions and business partners are shown. This is quite typical situation as shown also in scientific research[8]

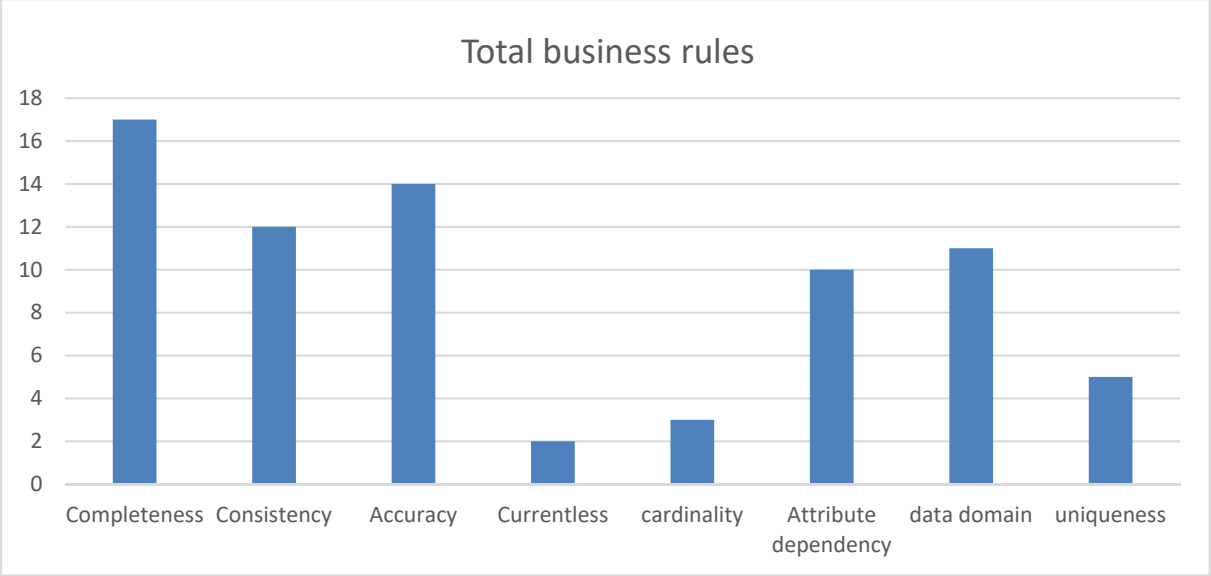


Figure 1: Distribution of business rules

[8] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, Andrea Maurino: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3): 16:1-16:52 (2009)

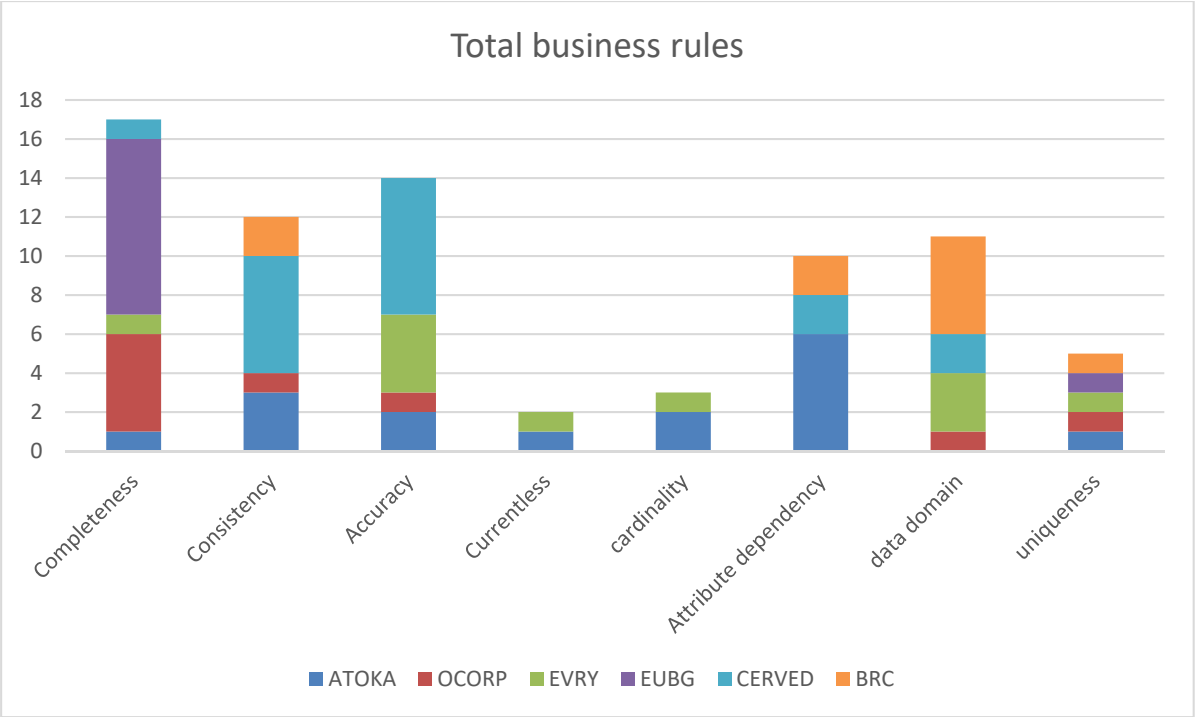


Figure 2: Distribution of business rules with respect to business companies

Considering each single set of business rules provided by business partners a number of insights can be made.

Figure 3 shows the relevance of quality dimensions in the TDS business case. According to Figure 3, data that will be managed need to be accurate and coherent among them. This is reasonable if we consider that the TDS wants to capture data from unstructured sources and the new data must to be integrated with existing data owned by CERVED. The completeness of data is less relevant in this context.

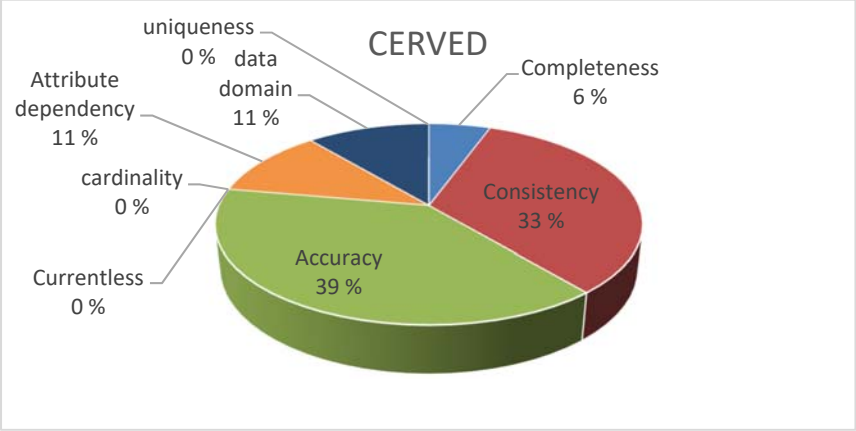


Figure 3: CERVED business rules

A completely different distribution of quality dimensions is reported in where business rules related to the data that will fill in the euBusinessGraph platform are shown in Figure 4. In this case the majority of rules are related to the completeness of data. This can be explained if we consider that such rules are related to the requirements that all data providers must satisfy to provide their data into the platform. As a consequence, business rules are related to the minimum set of data attributes that each companiy must provide.

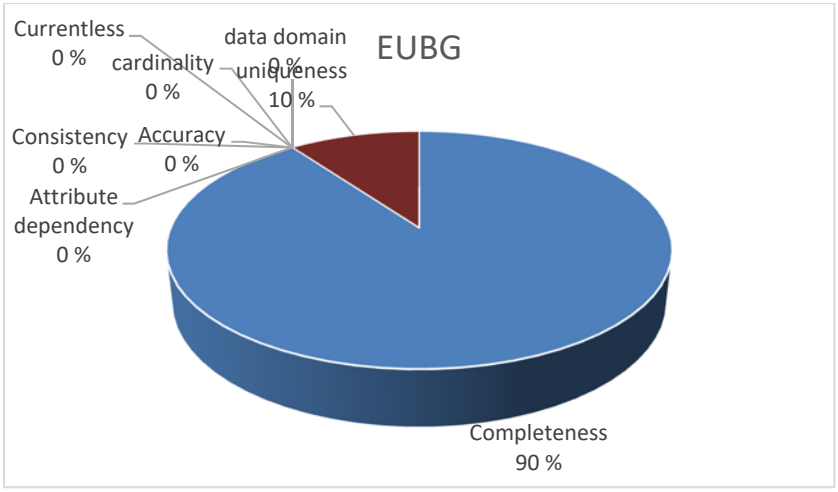


Figure 4: EUBG business rules

The relevance of accuracy is also available in the CRM-s Business cases where data provided by the euBusinessGraph platform will be directly inserted in the EVRY software solutions. In this case not only data must to be accurate, but also there is the need that data must satisfy strictly data domain constraints as show in Figure 5. The relevance of syntax oriented requirements raising by studying the set of business cases provided by EVRY is also demonstrated by considered that 18% of them are related to cardinality and uniqueness quality dimensions

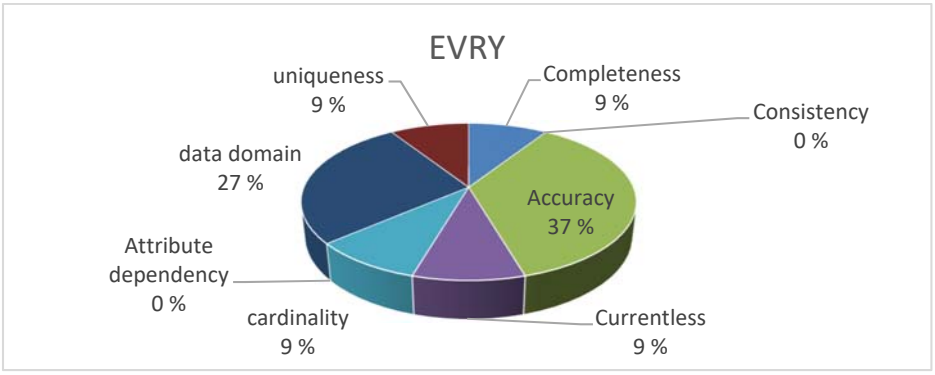


Figure 5: EVRY business rules

The analysis of business rules of OCORP shows some interesting insights (see Figure 6). Conversely to the ones of CERVED, in some sense a competitor of OCORP, in this set of rules, the most important considered quality dimensions is the completeness of data (56%). This can be explained by considered that OCORP collect information about public registries around the world. Thus, they are most interested to have as much possible complete information. The other quality dimensions: accuracy, consistency, data domain and uniqueness are all the same level of interests (11% each)

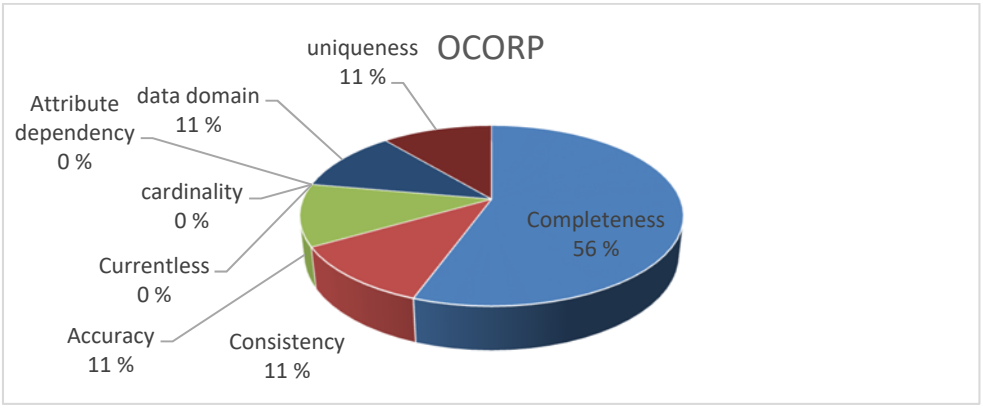


Figure 6: OCORP business rules

The nature of business cases influences also the relevance of quality dimensions considered in the case of ATOKA+ as shown in Figure 7. In this case due to the graph based model of data of ATOKA the most important quality dimensions is the attribute dependency. In fact, new data that will be added to the existing ATOKA graph must be coherent and not in contradiction with existing data. This is also confirmed by considering that the second and third most relevant quality dimensions are consistency and cardinality (with 19% and 13% each). All the three quality dimensions means that data must to be as much as possible coherent.

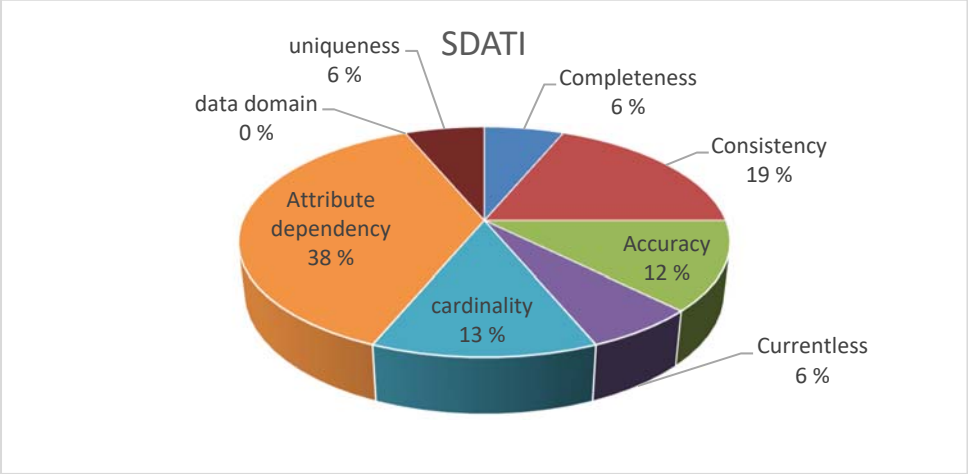


Figure 7: SDATI Business rules

Finally, a different distribution of quality dimensions is reported in Figure 8related to the BRC business case. By remembering that BRC is the official entities that is in charge of publishing business related data for Norway it is possible to understand that the most important quality dimensions is the data domain (50%) and then, with the same percentage consistency and attribute dependency (20% each). In fact, BRC has not the right to modify data provided directly from companies; thus the attention is related to the syntactic level of data it provide.

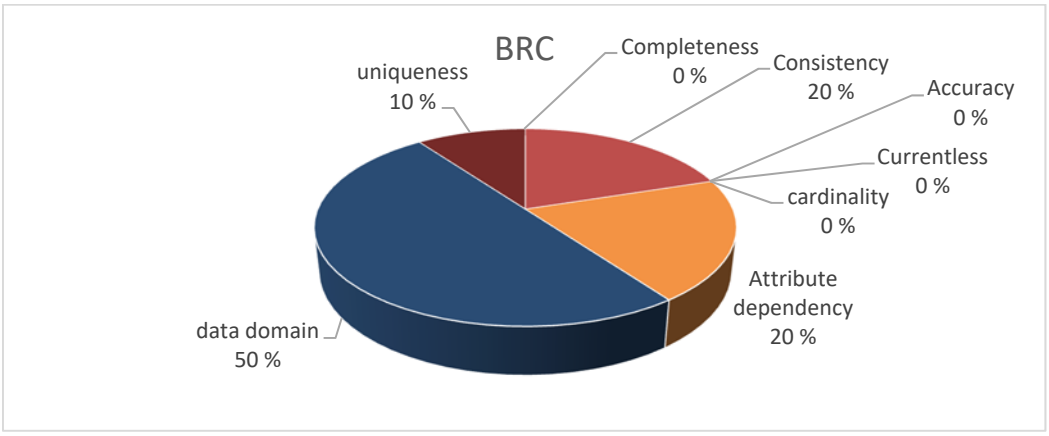


Figure 8: BRC business rules

4 Data Management Plan

This section describes the approach established in euBusinessGraph to ensure the life-cycle management of the public and proprietary datasets provided by the consortium members to the project as well as other dataset produced by the consortium during the project execution, as defined at M12.

The Data Management Plan (DMP) will be in accordance with H2020 Guidelines[9], including information and suggestions about the handling of data during and after the end of the project. It describes what data will be collected, processed and/or generated, which methodology and standards will be applied, whether data will be shared/made open access and how data will be curated and preserved (including after the end of the project).

According to the Guidelines on FAIR Data Management in Horizon 2020, DMP is a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As a consequence, Section 4.1 defines which are the principles underlying euBusinessGraph DMP, and shows the approach followed to generate the DMP. In Section 4.2, the audience and the responsibilities defined around the DMP are described. The next section introduces core concepts and fundamental legal principles as well as outlines an ethical assessment for data owner and, concerning legal requirements, provides detailed guidelines about the obligations that data owners need to comply with. In Section 4.4, relevant information regarding the dataset is explained and the process of collecting all relevant information among data owners is defined. Section 5 shows, for each dataset, all the information required for dataset identification, origin, format, access, security and defines ethical and legal requirements.

4.1 Principles underlying the euBusinessGraph DMP

The euBusinessGraph project aims at deploying and hosting a platform to ease data integration tasks, by embedding shared data models, robust data management techniques and semantic reconciliation methods. This platform will offer a framework for unification of fragmented business data which will support further analysis and services. In general, research data should be 'FAIR', that is Findable, Accessible, Interoperable and Re-usable; in the context of a IA project such as euBusinessGraph such principles must find the right balance with the business goal of industrial partners.

Due to the nature of the project several data providers want to share data, but at the same time some of them want to preserve the added value information in a typical coopetitive environment. The definition of the business model behind the platform is described in deliverable D3.2 euBusinessGraph Marketplace and Services that is released at month 12. As a consequence, several datasets are accessible and reusable under a commercial agreement or a fee-based subscription model

According to recent sentence of the court of justice in March 2017 the right to be forgotten cannot be applied to personal data stored in business data[10]. The court notes first of all that the public nature of company registers is intended to ensure legal certainty in dealings between companies and third parties and to protect, in particular, the interests of third parties in relation to joint stock companies and limited liability companies, since the only safeguards they offer to third parties are their assets. The court further notes that matters requiring the availability of personal data in the companies register may arise for many years after a company has ceased to exist. Anyway, in some jurisdiction personal data cannot be exposed.

In conclusion, personal data can be shown only if the original data stored in official documentation as published in national gazette are available.

The euBusinessGraph DMP was developed by taking into account the DMP template that matches the demands and suggestions of the Guidelines on Data Management in Horizon 2020, and that is available through the DMP online platform[11].

The principal contents indicated in the template are enlisted here below:

[9] European Commission, Directorate-General for Research & Innovation (26 July 2016). Guidelines on FAIR Data Management in Horizon 2020. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[10] https://curia.europa.eu/jcms/upload/docs/application/pdf/2017-03/cp170027en.pdf
[11] https://dmponline.dcc.ac.uk/

- Dataset Description
- Fair data (making data findable, accessible, interoperable and reusable)
- Data security
- Data archiving and preservation
- Ethics and aspects

These principles were utilized as a guide and then the document was customized according to specific study requirements.

The following documents are applicable to the subject discussed in this deliverable, and will be referenced as indicated into round brackets:

1. euBusinessGraph – Grant Agreement ([GA])
2. [GA] Annex 1 – Description of Action ([DoA])
3. euBusinessGraph – Consortium Agreement ([CA])

Short references may be used to refer to project beneficiaries, also frequently referred to as partners. References are listed in the following table:

Table 9: Short references for project partners

Partner Name	Partner Acronym
SINTEF	SINTEF
OpenCorporates	OCORP
Cerved	CERVED
SpazioDati	SDATI
Evry	EVRY
Deutsche Welle	DW
Ontotext	ONTO
Brønnøysund Register Centre	BRC
Jozef Stefan Intitute	JSI
Universit à degli studi di Milano Bicocca	UNIMIB

This D1.1 deliverable will be updated, over the course of the project, whenever significant changes arise, to ensure compliance with Horizon 2020 guidelines. Among these changes it is likely that new datasets will be added, changes in consortium policies or changes in consortium composition will be made and external factors will be added.

4.2 Audience, role and responsibilities

Project data are oriented to:

- The consortium partners;
- All stakeholders involved in the project;
- The European Commission.

Because of the sensitiveness of business data used in the euBusinessGraph innovation action, no commitment to publish datasets provided by business partners as open data is made in [DoA]. For this reason, we do not include external stakeholders in the audience for project data. With external stakeholders we refer to a party that: is not a beneficiary, is not a linked third party in euBusinessGraph, is not the European Commission. Although we do not expect to make all datasets openly accessible to external stakeholders, models and methodologies developed in the project to

support interoperability between different parties will be disseminated to a larger audience of stakeholders.

We describe main roles of beneficiaries in the consortium and their responsibilities with regards to data and services developed in business cases in Table 10. Roles and Responsibilities of Beneficiaries In the table with refer to Business Cases with their number, which are further explained Deliverable 4.1.

In Table 10, we distinguish between three main roles of beneficiaries in the consortium:

Table 10: Roles and Responsibilities of Beneficiaries

Partner	Partner Role		Resp. wrt Business Cases		
	Business	Technology	Provider	Consumer	Facilitator
SINTEF		X			CRM-S, DJP,ATOKA+
OC	X		CED		
CERVED	X		TDS	TDS	
SDATI	X	X	ATOKA+	ATOKA+	
EVRY	X	X			CRM-S
DW	X		DJP	DJP	
ONTO		X			CRM-S, DJP,ATOKA+,CED
BRC	X		BRC-S		
JSI		X			CRM-S, DJP,ATOKA+
UNIMIB		X			CRM-S, DJP,ATOKA+

- **Data provider partners**: partners that develop services within the project, by exploiting the technology developed in the project, i.e., the euBusinessGraph platform, on their own data sets and/or with the help of data sets provided by other partners in the project. These partners will also contribute indirectly to the technology by driving its development with the specification coming from their business cases.
- **Technology partners**: partners whose main role in the project is to develop the technology that will support the euBusinessGraph platform. These partners will also contribute indirectly to the business cases by performing the following activities:
    - Providing or supporting access to project data sets.
    - Supporting the development of pilots and services by helping business partners to integrate the data.
- **Data consumer partners:** partners that will also use the marketplace to enrich their business activities. These partners will also contribute in the definition of the requirements of platform from and end user view point

4.3 Ethics and Legal Compliance

The euBusinessGraph project must comply with all EU laws regarding data protection. The purpose of this section is to explain core principles and concepts of the right of protection of personal data in scientific research[12].

In the 1990s, the European Union started a process of codification of data protection and privacy rights in order to harmonise different national legislation. Directive 95/46/EC[13] ("Data Protection Directive") and Directive 2002/58/EC[14] ("E-Privacy Directive") are the main legal provisions that referred to define the legal framework, considering also the EU Charter of Fundamental Rights[15] and the appropriate national legislation that transposed these EU directives.

This multilevel legal environment is going to change in 2018, when in May a new European Regulation comes into force. Indeed, the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679[16]) was approved, by the EU Parliament, on 14 April 2016. It entered in force 20 days after its publication in the EU Official Journal and will be directly application in all member states two years after this date. It is designed to harmonize data privacy laws across Europe, to protect and empower all EU citizens' data privacy and to reshape the way organizations across the region approach data privacy.

Although the new regulation confirms the main principles of both the above-cited Directives, it will substitute them and all national legislation on data protection and privacy rights.

Generally, every data controller has to notify its national Data Protection Authority (DPA) of its decision to start collection of personal data before starting this process. This notification aims at communicating in advance the creation of a new "database," explaining the reasons for and purposes of this, and the technical and organisational safeguards in place to protect the personal data. Consequently, DPAs are enabled to verify the legal and technical safeguards required by EU legislation. However, the conditions attaching to and the procedures for submitting such a notification differ from EU state to EU state, with the strongest protections in place in Germany and the Netherlands and the least in Ireland and the UK.

The new European regulation will introduce a different way to manage data protection issues, following Privacy by Design principles, however. Each data controller has to carry out an assessment of the impact of processing operations on the protection of personal data before starting the processing itself to evaluate the origin, nature, particularity, and severity of risk attaching to their proposed processing. Such Data Protection/Privacy Impact Assessments (DPIA) can then be utilised to define appropriate measures to assure data protection and compliance with EU legislation.

A DPIA is required in case of:

- Systematic and extensive evaluation of personal aspects in automated processing (e.g. profiling);

[12] According to article 19 Regulation(EU) n. 1291/2013 (Horizon 2020): "all the research and innovation activities carried under Horizon 2020 shall comply with ethical principles and relevant national, Union and international legislation, including the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights and its Supplementar y Protocols. Particular attention shall be paid to the principle of proportionality, the right to privacy, the right to the protection of personal data, the right to the physical and mental integrity of a person, the right to non-discrimination and the need to ensure high levels of human health prot ection."

[13] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[14] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on Privacy and Electronic Communications). Later this Directive was amended with Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009.

[15] Article 8 (Protection of Personal Data) of the EU Charter of Fundamental Rights: "1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3. Compliance with these rules shall be subject to control by an independent authority."

[16] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

- Processing on a large scale of sensitive data or of personal data relating to criminal convictions and offences;
- Systematic monitoring of a publicly accessible area on a large scale.

The main aspects of DPIAs are:

- Systematic description of processing operations and the purposes of the processing;
- Assessment of the necessity and proportionality of the processing operations in relation to the purposes;
- Assessment of the risks to the rights and freedoms of data subjects;
- Measures to deal with the risks, including safeguards, security measures, and mechanisms to ensure data protection and to demonstrate compliance with EU legislation.

In the event that a DPIA indicates a high risk in terms of data protection and privacy rights, the Data Controller must consult the National Data Protection Authority prior to the processing.

The use of datasets within euBusinessGraph project have to comply with applicable international, EU and national law (in particular, EU Directive 95/46/EC).

In order to meet this goal, data owners have been asked to evaluate each of their dataset in order to confirm the nature and sensitivity of data to be used within euBusinessGraph project.

In order to make this evaluation, dataset owners, for each dataset, have to clarify if their own dataset contains Private Data (PD). If the dataset contains PD, they have to provide notification and informed consent for secondary use.

The euBusinessGraph project is implemented considering fundamental ethical standards to ensure the quality and excellence in the process and after the life of the project. In the Horizon 2020 it is specified that Ethical research conduct implies the application of fundamental ethical principles and legislation to scientific research in all possible domains of research. The nature of data managed in the euBusinessGraph project and the role of data distributor of official data of many of the business partners allow us to say that there are no ethical issues that can have an impact on data sharing

In the context of euBusinessGraph project, the IPR ownership is fundamentally regulated by the underlying principles of two main official documents (namely [CA] and [GA]).

Two main concerns on IPR management could impact the current deliverable:

- Existing or developed datasets will be available to the whole Consortium during the project timespan, but any further use in exploitation activities must follow specific limitations and/or conditions (as stated in Article 25.3 of the [GA] and described in its Attachment 1).
- All the identified datasets will be available to all Beneficiaries in order to develop the business cases used to validate the project results, as explicitly mentioned in the description tables contained in “Chapter 6 - Dataset description” (see Dataset ACCESS section).

4.4 euBusinessGraph methodology for DMP

The DMP should address some important points on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced. The DMP, as a key element of good data management, has to describe the life cycle management applied to the data to be collected, processed and/or generated by a Horizon 2020 project.

In order to make data findable, accessible, interoperable and re-usable (FAIR), a DMP should include:

- **Dataset Identification:** specifying what data will be collected, processed and/or generated.
- **Dataset Origin:** specifying if existing data is being re-used (if any), the origin of the data and the expected size of the data (if known).
- **Dataset Format:** describing the structure and type of the data, time and spatial coverage and language and naming conventions.
- **Data Access:** specifying whether data will be shared/made open access. In particular, for:

- **Making data accessible:** specifying if and which data produced and/or used in the project will be made openly available, moreover explaining why certain datasets cannot be shared (or need to be shared under restrictions), separating legal and contractual reasons from voluntary restrictions.
- **Making data interoperable**: specifying if the data produced in the project is interoperable, that is allowing data exchange and re-use. Moreover, specifying what data and metadata vocabularies, standards or methodologies it is meant to follow to make data interoperable.

- **Data Security:** specifying which provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data). Furthermore, specifying Personal Data presence and, in that case, privacy management procedures must be put in practice.

The following sections aim to provide more details, in terms of the class of attributes listed above, and will be used as a guide to describe datasets provided for euBusinessGraph.

4.4.1 Dataset IDENTIFICATION

First of all, it is needed to identify the dataset to be produced and provide dataset details, in terms of description of the data that will be generated or collected.

Following H2020 guidelines, it has been defined a set of relevant information that can help to define the dataset identification:

- Category: Dataset typology (Market, Consumer, Products, Weather, Media).
- Data name: Name of the dataset that should be a self-explaining name.
- Description: Description of the dataset in order to provide more details.
- Provider: Name of the beneficiary providing the dataset (or being in charge of bringing it into the project).
- Contact Person: Name of the person to be contacted for further details about the dataset.
- Business Cases number: BC involved (i.e., BCx)

4.4.2 Dataset ORIGIN

Following H2020 guidelines, it has been defined a set of relevant information that can help to define the dataset origin:

- Available at (M): Project month in which the dataset will be available.
- Core Data (Y|N): Indicate if the dataset is mandatory and will be part of the data shared along the different UCs or if it is discretionary and present only a limited usage.
- Size: A rough order of magnitude (ROM) estimation in terms of MB/GB/TB.
- Growth: A dynamic rough order of magnitude (ROM) estimate by selecting the most appropriate frequency in terms of MB/GB/TB per hour/day/week/months/other.
- Type and format: Dataset format, specifying if it is using, for example, CSV, Excel spreadsheet, XML, JSON, etc.
- Existing data (Y|N): The data already exist or are generated for the project ’s purpose.
- Data origin: How the data in the dataset is being collected/generated (i.e. SQL table, Google API, etc.)

4.4.3 Dataset FORMAT

Following H2020 guidelines, it has been defined a set of relevant information that can help to define the dataset format:

- Dataset structure: description of the structure and type of the data. (i.e. the header columns, the JSON schema, REST response fields, etc.).

- Dataset format: definition of the dataset format (i.e. specifying if it is using CSV, Excel spreadsheet, XML, JSON, GeoJSON, Shapefile, HTTP stream, etc.).
- Time coverage: if the dataset has a time dimension, indication of what period does it cover.
- Spatial coverage: if the dataset relates to a spatial region, indication of what is its coverage.
- Languages: languages of metadata, attributes, code lists, descriptions.
- Identifiability of data: reference to identifiability of data and standard identification mechanism.
- Naming convention: description about how the dataset can be identified if updated or after a versioning task has been performed, if the dataset is not static.
- Versioning: reference to how often is the data updated (i.e. No planned updating, Annually, Quarterly, Monthly, Weekly, Daily, Hourly, Every few minutes, Every few seconds, Real-time) and how the versioning is managed (i.e. if daily, every day a new dataset is generated with the newly created data or every day a new dataset overrides the old one containing all the data generated from the beginning of the collection, …)
- Metadata standards: specification of standards for metadata creation (if any). If there are no standards description of what metadata will be created and how.

4.5 Dataset metadata collection

4.5.1 TDS business case

Metadata item	Description
Dataset name	Tender calls dataset
Contact person + e-mail	Diego.Sanvito@cerved.com
Dataset short description (Original and English language)	Tender calls datasets consist of open and closed tender calls gathered from numerous sources along the following identified lines: • Albo Pretorio, • MePA, • Calls for public contracts on regional, province and town portals, • Calls for public contracts from SIMOG – ANAC portal, • Calls for public contracts from National Service for Public Contracts, • Italian TED. Albo Pretorio data source is a collection of resolutions, ordinances, posters and documents that refer to a tender call for public contracts and are disclosed to the public by Italian regional, province and town administration. MePA data source is a collection of tender calls i.e. bid requests and accompanying documents for public administrations. The MePA (Mercato Elettronico della P.A.) is a digital market operated on the behalf of the Italian Ministry for Economy and Finance. MePA enables public administrations to buy, for values below the market value, goods and services offered by licensed vendors. Calls for public contracts on regional, province and town portals source is a collection of resolutions, ordinances, posters and documents that refer to a tender call for public contract and are published on regional and town portals. Calls for public contracts from SIMOG – ANAC portal source is a collection of documents related to tender calls and from SIMOG portal[17] (Sistema Informativo Monitoraggio Gare). SIMOG is an information system of the Italian national anti-corruption authority ANAC (Autorità nazionale anticorruzione) for monitoring tender calls that allows public

[17] http://www.anticorruzione.it/portal/public/classic/Servizi/ServiziOnline/SistemaSIMOG

	entity publishing a tender call to request an identifying code of the tender call (CIG). The data set contains open tender calls starting September 2017. When published outcomes of tender calls will be added to the database as historical data. Calls for public contracts from National Service for Public Contracts source is a collection of documents related to tender calls for construction contracts on the Service for Public Contracts portal (i.e. national portal of the Ministry of Infrastructure). Italian TED (Tenders Electronic Daily) source is a collection of tender calls i.e. bid requests and accompanying documents from Tenders Electronic Daily relevant to Italy. TED is the 'Supplement to the Official Journal of the EU ("OJ S"), dedicated to European public procurement. The dataset is on CERVED's premises. It includes a step of data cleaning, deduplication and integration as data comes from numerous national, regional, province and town sources (e.g. regional administration as Emilia Romagna and Molise, portals of several major cities as Rome, Milan, Naples, Torino).
Theme / tags	"tender", "public entity", "winning company"
Data collection	Automatically (scrapers)

Dataset structure	Example JSON schemas for a tender call:
	{
	"_id" : ObjectId("59d7cf43c4b41609d68ec816"),
	"codice_fiscale_stazione_appaltante" : "81000250795",
	"cpv" : "55524000-9 - SERVIZI DI RISTORAZIONE SCOLASTICA",
	"data_termine_offerta" : ISODate("2016-10-26T00:00:00.000+0000"),
	"id" : "SIMOG-6527166-6814671C66",
	"id_gara_simog" : "6527166",
	"importo_complessivo_gara" : "392.37000",
	"lotti" : [
	{
	"aggiudicazione" : {
	"aggiudicatari" : [
	{
	"codice_fiscale" : "ND",
	"ragione_sociale" : "--",
	"ruolo" : " "
	}
	],
	"data_aggiudicazione" : " ",
	"importo_aggiudicazione" : " ",
	"numero_offerte_ammesse" : "0",
	"ribasso_aggiudicazione" : "0.0%",
	"tipo_criterio" : "ND"
	},
	"cig" : "6814671C66",
	"importo_base_asta_lotto" : "195.57000",
	"luogo_lavori" : "CROTONE",
	"oggetto" : "SERVIZIO REFEZIONE SCOLASTICA ANNO SCOLASTICO 2016/2017 PER I BAMBINI DELL'ASILO NIDO E PER GLI ALUNNI DELLE SCUOLE DELLE'INFANZIA, PRIMARIE E SECONDARIE DI 1^ GRADO NONCHE' PER GLI INSEGNANTI E PERSONALE ATA - LOTTO A"
	}
	],
	"luogo_stazione_appaltante" : "CROTONE",
	"modalita_realizzazione" : "APPALTO",
	"oggetto" : "SERVIZIO REFEZIONE ANNO SCOLASTICO 2016/2017 PER I BAMBINI DELL'ASILO NIDO E PER GLI ALUNNI

	DELLE SCUOLE DELL'INFANZIA PRIMARIE E SECONDARIE DI 1^ GRADO NONCHE' PER GLI INSEGNANTI E PERSONALE ATA",
	"pubblicazione" : [
	{
	"data" : ISODate("2016-10-17T00:00:00.000+0000"),
	"luogo" : "SITO DELL'AUTORITA' PER LA VIGILANZA SUI CONTRATTI PUBBLICI DI LAVORI SERVIZI E FORNITURE",
	"numero" : "",
	"sito" : "http://portaletrasparenza.avcp.it/microstrategy/asp/download.aspx?id=6527166_213555_1.pdf&check=1",
	"tipo" : "BANDO DI GARA"
	},
	{
	"data" : ISODate("2016-10-17T00:00:00.000+0000"),
	"luogo" : "SITO DELL'AUTORITA' PER LA VIGILANZA SUI CONTRATTI PUBBLICI DI LAVORI SERVIZI E FORNITURE",
	"numero" : "",
	"sito" : "http://portaletrasparenza.avcp.it/microstrategy/asp/download.aspx?id=6527166_213556_2.pdf&check=1",
	"tipo" : "DISCIPLINARE"
	},
	{
	"data" : ISODate("2016-10-04T00:00:00.000+0000"),
	"luogo" : "ALBO PRETORIO",
	"numero" : "",
	"tipo" : "BANDO DI GARA"
	},
	{
	"data" : ISODate("2016-10-04T00:00:00.000+0000"),
	"luogo" : "ALBO PRETORIO",
	"numero" : "",
	"tipo" : "DISCIPLINARE"
	},
	{
	"data" : ISODate("2016-10-04T00:00:00.000+0000"),
	"luogo" : "PUBBLICAZIONE SU SIMOG",
	"numero" : "",
	"tipo" : "DICHIARAZIONE SU SIMOG",
	"data_download" : ISODate("2017-10-06T18:45:20.675+0000")
	}
	],
	"ragione_sociale_stazione_appaltante" : "COMUNE DI CROTONE",
	"settore" : "SETTORE ORDINARIO",
	"tipo_procedura" : "PROCEDURA APERTA",
	"tipologia_intervento" : "SERVIZI",
	"cig" : [
	"6814671C66"
	],
	"data_ultimo_aggiornamento" : ISODate("2017-10-06T20:49:09.596+0000"),
	"fonte" : "bandi-simog"
	}

Standards and metadata	First reduced dataset is available from September 2017 . Currently identified metadata are shown in the previous cell and include: CIG code, where CIG is an identifying code of the tender call defined through the information system for monitoring tender calls by the national authority for public contracts for works, services and supplies.

	CUP code is an Italian government unique identifier that characterizes every public investment project. CPV (Common Procurement Vocabulary, Regulation (EC) 213/2008) code establishes a single classification system for public procurement aimed at standardizing the references used by the contracting authorities and contracting entities to describe the subject of procurement. The date of publication in the official gazette of records for Italy GURI (Gazzetta ufficiale della Repubblica Italiana). The date of publication in the official gazette of records for European Union GUUE (Gazzetta ufficiale dell'Unione Europea).
Dataset owner/publisher/provider name	CERVED
Dataset licence	The enriched dataset is private, as it is cleaned and harmonized with other proprietary sources to enable tool development. The access to the result of the tools being developed in euBusinessGraph will be accessible through an API for a fee.
Data availability	The data will be private and initial versions are at partner premises from September 2017 with limited geographical coverage. The data is stored in institutional repository e.g. NoSQL documental database, and cannot be openly shared for commercial reasons.
Archiving and preservation (including storage and backup)	Original data and enriched dataset will be preserved for future usage.
How data can the dataset be accessed?	Not available by default as it is used in scope of TDS service, outcomes of TDS service can be made available through REST APIs as part of the – data value feedback chain (i.e. consists of the data insights generated by the proposed products and services being fed back into the business graph, therefore enhancing the value and scope of the data in the business graph).
Dataset source URL	Not available
Dataset format (current and target data format)	JSON
Size of the dataset	Tens of GB per month
Update frequency	daily
Time coverage	From April 2017 onwards
Spatial coverage	Italy
Language	Italian
Relation to euBusinessGraph	The database is required for developing the TDS business case.
Data discoverability	The data will we harvested, imported, cleaned, integrated, and linked to non-free and open data and will be a part of the TDS enabling discovery and recommendation of open tender calls that potentially fit well to company's characteristics.
Data identification	As this data is specific to Italy we currently envisage usage of CIG code identifiers and 'sameA's identifiers when a same tender calls is obtained from different sources
Data interoperability	Standard vocabularies and identifiers as described above.
Data privacy	The dataset does not include personally identifiable information and the data will not be anonymized.

4.5.2 BRC business case

Metadata item	Description
Dataset name	Enhetsregisteret - Legal Entities
Contact person + e-mail	Norheim, David <david.norheim@brreg.no>
Dataset short description	Enhetsregisteret (the Central Coordination Register for Legal Entities or

(Original and English language)	in short Entity Register) dataset is a register containing information on all legal entities in Norway – commercial enterprises and governmental agencies. It also includes business sole proprietorships, associations and other economic entities without registration duty that have chosen to join the CCR on a voluntary basis.
Theme / tags	Organization number, Sector, Legal entity, References to branches
Data collection	Mandatory, by law https://lovdata.no/dokument/SF/forskrift/1995-02-09-114/

Dataset structure	See https://confluence.brreg.no/display/DBNPUB/Informasjonsmodell+for+Enhetsregisteret+og+Foretaksregisteret
	
	example
	{
	"organisasjonsnummer": 974760673,
	"navn": "REGISTERENHETEN I BR\u00d8NN\u00d8YSUND",
	"registreringsdatoEnhetsregisteret": "1995-08-09",
	"organisasjonsform": "ORGL",
	"hjemmeside": "www.brreg.no",
	"registrertIFrivillighetsregisteret": "N",
	"registrertIMvaregisteret": "N",
	"registrertIForetaksregisteret": "N",
	"registrertIStiftelsesregisteret": "N",
	"antallAnsatte": 559,
	"institusjonellSektorkode": {
	"kode": "6100",
	"beskrivelse": "Statsforvaltningen"
	},
	"naeringskode1": {
	"kode": "84.110",
	"beskrivelse": "Generell offentlig administrasjon"
	},
	"postadresse": {
	"adresse": "Postboks 900",
	"postnummer": "8910",
	"poststed": "BR\u00d8NN\u00d8YSUND",
	"kommunenummer": "1813",
	"kommune": "BR\u00d8NN\u00d8Y",
	"landkode": "NO",
	"land": "Norge"
	},
	"forretningsadresse": {
	"adresse": "Havnegata 48",
	"postnummer": "8900",
	"poststed": "BR\u00d8NN\u00d8YSUND",
	"kommunenummer": "1813",
	"kommune": "BR\u00d8NN\u00d8Y",
	"landkode": "NO",
	"land": "Norge"
	},
	"konkurs": "N",
	"underAvvikling": "N",
	"underTvangsavviklingEllerTvangsopplosning": "N",
	"overordnetEnhet": 912660680,
	"links": [
	{
	"rel": "self",
	"href": "http:\/\/data.brreg.no\/enhetsregisteret\/enhet\/974760673"
	},
	{

	"rel": "overordnetEnhet", "href": "http:\/\/data.brreg.no\/enhetsregisteret\/enhet\/912660680" }] }
Standards and metadata	Description of concepts are given here https://confluence.brreg.no/display/DBNPUB/Informasjonsmodell+for+Enhetsregisteret+og+Foretaksregisteret
Dataset owner/publisher/provider name	Brønnøysund Register Centre (BRC)
Dataset licence	NLOD, compatible with Creative Commons BY
Data availability	Fully available as open data
Archiving and preservation (including storage and backup)	Fully archived
How data can the dataset be accessed?	REST APIs
Dataset source URL	Not available
Dataset format (current and target data format)	JSON, XML, CSV (in 2017: RDF)
Size of the dataset	Ca 1 mill entities, about
Update frequency	Continuously
Time coverage	From April 2017 onwards
Spatial coverage	Norway
Language	Norwegian
Relation to euBusinessGraph	The database is a source for euBusinesssGraph, but also included in the BRC business case.
Data discoverability	There will be a new version over summer 2017 of the distribution including more details and new APIs. The data will be available in RDF. The updated dataset will also be discoverable by the Norwegian (and EU publication office's) data catalogue using the DCAT standard.
Data identification	There is a national standard for identifiers called "organization numbers". This is mandatory for doing business in Norway. There has also been a URI established being used in the services http://data.brreg.no/enhetsregisteret/enhet/<organization-number>[.<format>]
Data interoperability	Currently there are no identified metadata vocabularies, standards or methodologies to facilitate interoperability. This being said the registry is exported to BRIS – EU company register, which is using the EU core vocabularies. Future APIs will follow these vocabularies.
Data privacy	The dataset currently excludes roles like directors, accountant, and board memberships. However we are working to make this data available as well.

4.5.3 OCORP business case

Metadata item	Description
Dataset name	Business registers from around the world
Contact person + e-mail	Chris Taggart (chris.taggart@opencorporates.com)
Dataset short description (Original and English language)	This is the collection of core company (legal entity) data (on over 130 million entities) collected from more than 120 company registers around the world. The data is sourced only from official public sources, and full provenance (source, and date sourced) is provided. The depth of data varies from jurisdiction to jurisdiction, sometimes including directors and officers, industry codes, even occasionally shareholders and ultimate beneficial owners.

	This dataset is a fundamental basis for understanding of companies in individual jurisdictions and worldwide, and is used by hundreds of thousands of OpenCorporates' users around the world, including banks, law enforcement, journalists, anti-corruption NGOs and business intelligence users. We perform extensive data quality assurance processes, both on ingestion and particularly when understanding and handling the following important issues, among others: • Use of company identifiers (are they unique, reused, consistent, normalisable, etc) • Permissable legal forms • Underlying business rules/legislation, and how that affects the dataset • Language/character set/encoding issues • Treatment of legal status • Profit/non-profit status In order to perform this QA, we have an extensive set of policies, procedures, and workflow processes, and we also publish any known issues with the jurisdiction (see https://blog.opencorporates.com/tag/business-registers/ for examples)
Theme / tags	"Business company", "Official register", "directors", "Legal Entities"
Data collection	The data is collected via a variety of mechanisms, including: • Use of open APIs (e.g. Norway) • Use of APIs under agreement with the company register (e.g. Switzerland, Ireland) • Use of open data dumps in a variety of forms, including XML, CSV and JSON • Use of bulk data supplied under agreement with the company register • Screen-scraping In some jurisdictions, a mixture of the above methods will be used (e.g. in the UK we use a mixture of 1-4).
Dataset structure	The data is mapped to a common schema (see https://github.com/openc/openc-schema/blob/master/schemas/company-schema.json.
Standards and metadata	While some standards do exist for legal entity data, they do not encompass the depth and variety of data in the underlying sources. The core metadata associated with each datapoint in this dataset is the source (including name of the register, URL of source, and date retrieved from the source).
Dataset owner/publisher/provider name	Chrinon Ltd T/A OpenCorporates
Dataset licence	See below
Data availability	All the data held by OpenCorporates is freely available through the OpenCorporates website. The underlying structured data is also made freely available under an open licence for public-benefit uses (particularly to journalists, NGOs and academics) via the OpenCorporates REST API and via bulk dumps. For non-public benefit uses, the data is also available via the API or bulk dumps for a fee, and it is this income that supports OpenCorporates public-benefit work, including the free website, the public benefit data access, and its advocacy work.
Archiving and preservation (including storage and backup)	The data is stored on our own servers, together with periodic offsite backups. Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the

	associated costs are and how these are planned to be covered.
How data can the dataset be accessed?	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data by agreement.
Dataset source URL	N/A
Dataset format (current and target data format)	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data (CSV) by agreement.
Size of the dataset	The dataset size depends on storage mechanism and number of attributes included, but the company register data as stored in a relational database is somewhere under 500G in size
Update frequency	The data is continually being updated automatically with our suite of hundreds of bots
Time coverage	It covers companies incorporated over 140 years ago to the present day
Spatial coverage	N/A
Language	Meta data and attribute names are in English – the underlying data is broadly in the language (and alphabet) of the individual original source
Relation to euBusinessGraph	This one be one of the core datasets provided to the EuBusinessGraph – the attributes supplied are still being agreed, and will depend on both the business model of the EuBusinessGraph platform, and the revenue it is expected to generate
Data discoverability	All the data is available via the website and the API with extensive filtering and queries
Data identification	Wherever possible (and in well over 90% of cases) we use the official company register identifiers, paired with the jurisdiction to identify companies, thus ensuring we do not create any IP in the identifier, and allowing them to be used without restriction. We only vary from that where there are problems with the official identifier (for example, in the Jersey company register they are not unique either across the register, or within company types). We are always transparent about use of identifiers and work to ensure we do not create IP in them.
Data interoperability	See above re identification. We use existing standards where appropriate (e.g ISO 3166 and 3166-2) and have participated in a number of standards processes, including the EU Core Business Vocabulary (and mapped basic attributes to this vocabulary – see https://blog.opencorporates.com/2012/02/22/3-reasons-why-the-eus-new-business-vocabulary-is-so-important/). Our CEO is also on the board of directors of the Global Legal Entity Identifier Foundation, and via that has contributed to the creation of many core standards. However, the depth and detailed nature of the data means that there are many areas where there are no existing standards. To ensure our work can be reused we publish all our schemas under an open licence (https://github.com/openc/openc-schema)
Data privacy	Yes, it does. However, all the data is from public sources and a recent ruling by the ECJ affirmed that there is no right to be forgotten in company register data.

Metadata item	**Description**
Dataset name	Government gazettes from around the world
Contact person + e-mail	Chris Taggart (chris.taggart@opencorporates.com)
Dataset short description (Original and English language)	This is a collection of government gazettes, primarily from Europe. Despite a 300-year-old legacy and being the public record for legal notices, government gazettes (also called official journals in some EU countries) are astonishingly poorly known, and OpenCorporates – with the help of a grant from ODINE, an open data incubator by the European Union – decided to tackle them and make them achieve their true public purpose, by making the information within them properly known to the public. Gazettes are particularly useful when researching or assessing private

	companies, particularly critical corporate events such as liquidation, dissolution, winding-up orders, annual general meetings or director actions. In short, gazettes are an untapped and critical resource, but in their native form are notoriously challenging to work with. They are unstructured, inconsistent and designed for a pre-digital age. OpenCorporates, initially via its OpenGazettes project, and more recently as part of this project is extracting these notices, converting them into data, and matching them to the companies to which they relate. In addition, OpenCorporates will be using Gazette notices as a key input into its Corporate Events Data product, including inferring critical corporate events from gazette notices. We perform extensive data quality assurance processes, both on ingestion and particularly when understanding and handling the following important issues, among others: • Use of identifiers (are they unique, reused, consistent, normalisable, etc) • Language/character set/encoding issues • Classifications of gazette notices (types of notice) • Understanding of nature of entities featured in gazettes (e.g. legal entities only, or mixed with non-legal-entities such as individuals or associations) • Handling of dates, including start and end dates (which may not be explicit) In order to perform this QA, we have an extensive set of policies, procedures, and workflow processes
Theme / tags	"Government Gazettes", "Official Journals", "Company data"
Data collection	The data is collected via a variety of mechanisms, including: • Use of open APIs • Use of open data dumps in a variety of forms, including XML, CSV and JSON • Screen-scraping In some jurisdictions, a mixture of the above methods will be used.
Dataset structure	The data is mapped to a common schema (see https://github.com/openc/openc-schema/blob/master/schemas/gazette-notice-schema.json)
Standards and metadata	No suitable standards exist for this dataset. The core metadata associated with each datapoint in this dataset is the source (including source name, publisher, URL of source, and date retrieved from the source).
Dataset owner/publisher/provider name	Chrinon Ltd T/A OpenCorporates
Dataset licence	As the project is specifically about commercial exploitation of the data contributed by the partners, the data is will be made available under the dual licence system that OpenCorporates currently uses – with free open access for public-benefit uses, and paid access (with still a relatively permissive licence) for non-public benefit and proprietary uses. This applies both to third parties accessing the data via the EUBusinessGraph platform or for other partners in the consortium.
Data availability	All the data held by OpenCorporates is freely available through the OpenCorporates website. The underlying structured data is also made freely available under an open licence for public-benefit uses (particularly to journalists, NGOs and academics) via the OpenCorporates REST API and via bulk dumps. For non-public benefit uses, the data is also available via the API or bulk dumps for a fee, and it is this income that supports OpenCorporates public-benefit work, including the free website, the public benefit data access, and its advocacy work.

Archiving and preservation (including storage and backup)	The data is stored on our own servers, together with periodic offsite backups. Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.
How data can the dataset be accessed?	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data by agreement.
Dataset source URL	N/A
Dataset format (current and target data format)	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data (CSV) by agreement.
Size of the dataset	The dataset size depends on storage mechanism, but the relational database is something over 50G in size
Update frequency	The data is continually being updated automatically with our suite of scores of bots
Time coverage	It primarily covers licences issued in the past 2 years, but also contains some historic information too.
Spatial coverage	N/A
Language	Meta data and attribute names are in English – the underlying data is broadly in the language (and alphabet) of the individual original source
Relation to euBusinessGraph	This dataset will be a key input into the Corporate Events Dataset that OpenCorporates is producing as part of this project – we are also working on extracting the embedded data from the licences and attaching them to the company, enriching the data held on the company
Data discoverability	All the data is available via the website, and via the API
Data identification	While there are identifiers used by some official registers of licences, they not always present, nor when they are there are they always used consistently or using best practices. Where such identifiers are used we make them available as part of the data, but because of these problems cannot use them as primary identifiers, instead using and exposing internal identifiers (which we do not claim any IP in)
Data interoperability	We use existing standards where appropriate (e.g ISO 3166 and 3166-2) and have participated in a number of standards processes, including the EU Core Business Vocabulary (and mapped basic attributes to this vocabulary – see https://blog.opencorporates.com/2012/02/22/3-reasons-why-the-eus-new-business-vocabulary-is-so-important/). Our CEO is also on the board of directors of the Global Legal Entity Identifier Foundation, and via that has contributed to the creation of many core standards. However, the depth and detailed nature of the data means that there are many areas where there are no existing standards, and this is true in the area of business licences, where we had to create our own schema, which is, like all our schemas, published under an open licence (https://github.com/openc/openc-schema)
Data privacy	In some cases it does, for example, individuals in bankruptcy notices, or as shareholders in companies.

Metadata item	**Description**
Dataset name	Business licences from around the world
Contact person + e-mail	Chris Taggart (chris.taggart@opencorporates.com)
Dataset short description (Original and English language)	This is a collection of a variety of business licences from various public registers and regulators around the world. The licences range from banking and other financial licences to gambling licences and (in the US) basic business licences. The data is sourced only from official public sources, and full provenance (source, and date sourced) is provided. The data varies both between types of licences and for the same licence type between jurisdictions. As well as the basic fact of

	whether an entity has a licence to perform certain acts (e.g. to operate as a bank), the licence information also often contains some of the following: • Detailed list of permissions • Headquarters and other addresses • Individuals in senior positions • Website URLs • Telephone numbers • Identifiers (e.g. tax numbers, LEI code, etc) • Balance sheet information • Parent company We perform extensive data quality assurance processes, both on ingestion and particularly when understanding and handling the following important issues, among others: • Use of identifiers (are they unique, reused, consistent, normalisable, etc) • Nature of permissions • Language/character set/encoding issues • Understanding of nature of entities with licences (e.g. legal entities only, or mixed with non-legal-entities such as individuals or associations) • Understanding of source records relationship to licence (one or more record per licence, handling of licence renewals, expiry, etc) • Handling of dates, including start and end dates (which may not be explicit) In order to perform this QA, we have an extensive set of policies, procedures, and workflow processes
Theme / tags	"Business licences", "Bank licences", "Company data"
Data collection	The data is collected via a variety of mechanisms, including: • Use of open APIs • Use of open data dumps in a variety of forms, including XML, CSV and JSON • Screen-scraping In some cases, a mixture of the above methods will be used.
Dataset structure	The data is mapped to a common schema (see https://github.com/openc/openc-schema/blob/master/schemas/licence-schema.json)
Standards and metadata	No suitable standards exist for this dataset. The core metadata associated with each datapoint in this dataset is the source (including source name, publisher, URL of source, and date retrieved from the source).
Dataset owner/publisher/provider name	Chrinon Ltd T/A OpenCorporates
Dataset licence	As the project is specifically about commercial exploitation of the data contributed by the partners, the data is will be made available under the dual licence system that OpenCorporates currently uses – with free open access for public-benefit uses, and paid access (with still a relatively permissive licence) for non-public benefit and proprietary uses. This applies both to third parties accessing the data via the EUBusinessGraph platform or for other partners in the consortium.
Data availability	All the data held by OpenCorporates is freely available through the OpenCorporates website. The underlying structured data is also made freely available under an open licence for public-benefit uses (particularly to journalists, NGOs and academics) via the OpenCorporates REST API and via bulk dumps. For non-public benefit uses, the data is also available via the API or bulk dumps for a fee, and

	it is this income that supports OpenCorporates public-benefit work, including the free website, the public benefit data access, and its advocacy work.
Archiving and preservation (including storage and backup)	The data is stored on our own servers, together with periodic offsite backups. Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.
How data can the dataset be accessed?	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data by agreement.
Dataset source URL	N/A
Dataset format (current and target data format)	Via API (JSON and XML) at http://api.opencorporates.com/ and as bulk data (CSV) by agreement.
Size of the dataset	The dataset size depends on storage mechanism, but the relational database is something over 100G in size
Update frequency	The data is continually being updated automatically with our suite of scores of bots
Time coverage	It primarily covers licences issued in the past 2 years, but also contains some historic information too.
Spatial coverage	N/A
Language	Meta data and attribute names are in English – the underlying data is broadly in the language (and alphabet) of the individual original source
Relation to euBusinessGraph	This dataset will be a key input into the Corporate Events Dataset that OpenCorporates is producing as part of this project – we are also working on extracting the embedded data from the licences and attaching them to the company, enriching the data held on the company
Data discoverability	All the data is available via the website, and via the API
Data identification	While there are identifiers used by some official registers of licences, they not always present, nor when they are there are they always used consistently or using best practices. Where such identifiers are used we make them available as part of the data, but because of these problems cannot use them as primary identifiers, instead using and exposing internal identifiers (which we do not claim any IP in)
Data interoperability	We use existing standards where appropriate (e.g ISO 3166 and 3166-2) and have participated in a number of standards processes, including the EU Core Business Vocabulary (and mapped basic attributes to this vocabulary – see https://blog.opencorporates.com/2012/02/22/3-reasons-why-the-eus-new-business-vocabulary-is-so-important). Our CEO is also on the board of directors of the Global Legal Entity Identifier Foundation, and via that has contributed to the creation of many core standards. However, the depth and detailed nature of the data means that there are many areas where there are no existing standards, and this is true in the area of business licences, where we had to create our own schema, which is, like all our schemas, published under an open licence (https://github.com/openc/openc-schema)
Data privacy	In some cases it does, for example either authorised individuals in the case of financial licences, or business licences that may be issued to both companies and individuals. However, all the data explicitly forms part of the public record, and is made available for a public purpose.

4.5.4 SDATI business case

Metadata item	Description
Dataset name	Aziende Italiane – Italian Legal Entities
Contact person + e-mail	Javier Paniagua paniagua@spaziodati.eu
Dataset short description (Original and English	Dataset contains detailed up-to-date company and contact information on legal entities in Italy. There is basic firmographics about 6 mln business

language)	entities, and information about 13 mln directors and managers. Data comes from both authoritative data sources and social web. Our main source of authoritative data is Cerved, one of the official distributors of the Register of Companies (Registro Imprese) held by the Chambers of Commerce in Italy (Camere di Commercio italiane). Other relevant authoritative source is the National Anti-Corruption Authority (Autorità Nazionale AntiCorruzione, ANAC) that provides data on company contracts for the government. Social web data comes from various online public sources that can be categorised as: (1) news and (2) corporate websites and social media channels The data is updated every week. We maintain data quality procedures, both automatic and manual.
Theme / tags	“Business company”, “firmographics”, “Italy”
Data collection	Official company data is provided by Cerved on a weekly basis. We collect the data, verify it and update our knowledge graph accordingly using automatic procedures. Data from social web and corporate websites is collected via purposefully developed Web crawlers.
Dataset structure	Please, refer to this document for description of the company attributes we will share https://docs.google.com/spreadsheets/d/1ENBasWNfCiK_fh39KavaJQPJn_Shm4RI5si31YUytvA/edit?usp=sharing
Standards and metadata	Standard or officially endorsed: • Industrial classification: o ATECO (Codice ATtività ECOnomica) o UKSIC (UK Standard Industrial Classification) • Legal forms’ classification o **Company Register hierarchical classification** • Public administration type: o **ISTAT type of public administrations** http://www.agid.gov.it/sites/default/files/documentazione/02_amm_adempienti_presenti_in_ipa_per_tipologia_istat_e_per_regione.pdf • Country code: o ISO 3166
Dataset owner/publisher/provider name	SpazioDati
Dataset licence	Various company data will be shared using one of the three sharing modes: o Shared -- full sharing; graph users see the value without coming to each partner's graphs/apps o Matching only -- the value will be used internally for matching company entities. End users will not see the value unless they go to each partner's graphs/apps. The presence of such value can be advertised though. o Other – depending on the shoring mode -- “shared” or “matching only” – the value can be presented in full or at a coarser granularity Data shared with a “shared” mode will be open. “Matching only” will be shared for free but only with the partners of the project, and licensed with the Atoka license for the users of the graph.
Data availability	In general, the data is available via a RESTful API, however, it will take time to implement access to the selected set of company data we will share with the project. Hence, we will start with a CSV dump of the data. The CSV structure will resemble the one described in the Excel document, one row in the CSV will correspond to one company.

	More information is available online in the documentation of the Atoka API https://developers.atoka.io/v2/companies_base.html - companies_packages. However, this documentation contains description of the whole Atoka product. Hence, you will find documentation of all the companies' attributes, but not all of them will be shared with the project.
How data can the dataset be accessed?	Dump
Dataset source URL	N/A
Dataset format (current and target data format)	CSV
Update frequency	Weekly
Time coverage	N/A, up-to-date information
Spatial coverage	Italy
Language	English
Relation to euBusinessGraph	• Source of the business information graph • Used in BCs
Data discoverability	Data is described at the Atoka website https://atoka.io/en/, data sources, data fields and data quality procedures are described at https://atoka.io/en/data-quality/. Detailed documentation of the data fields is provided at https://developers.atoka.io/v2/.
Data identification	• **ATOKA ID** – internal ID of the company in Atoka • **CCIIA** (Camera di Commercio, Industria, Artigianato e Agricoltura) **+ REA** (Repertorio Economico Amministrativo) can be used to identify a company. Companies can have more than one CCIAA - REA, since they can be registered to more than one chamber of commerce: o REA are codes handed by chambers of commerce to companies upon registration o ID of the chamber of commerce that issues the REA (see next attribute) in that province • **Companies House Company Number** • **Wikipedia URL** • **IPA code** governmental identifier of public sector companies http://www.indicepa.gov.it/documentale/n-domande.php - A2
Data interoperability	Besides the standard classifications and identifiers described above, registered addresses of companies are similar to the addresses in the OpenCorporates data.
Data privacy	No personally identifiable information

4.5.5 Bulgarian Trade Register business case

Note: This is a new dataset that ONTO created during the project related to the Bulgarian Trade Register. It will be used in the project similar to other data sources for the euBusinessGraph platform.

Metadata item	Description
Dataset name	Bulgarian Trade Register
Contact person + e-mail	milena.yankova@ontotext.com, vladimir.alexiev@ontotext.com
Dataset short description (Original and English language)	This dataset contains complete document submissions to the Bulgarian Registry Agency and the resulting company data. One XML file is provided for every business day, about 8.1Mb on average. The total size covering 10 years (2008 to Nov 2017) is about 20Gb unzipped. The XML schema is quite comprehensive and complex and comprises the following: • DeedV2.xsd: 20 elements/attributes; 56 nomenclatures with about 60 values. Defines documents submitted by the entity to the Register. Eg registration, pledge (залог), distraint (запор), etc. Lists 282 "fields" in SubDeedType (structures that describe aspects of the company).

	• Envelopev2.xsd: 36 elements/attributes. Message carrying Deeds and/or SearchCriteria • FieldsSchema.rnc: 1439 elements/attributes. Defines the "fields" that can be used in a Deed. For some it's hard to understand what they mean without examining the data or the Law • xmldsig-core-schema.xsd: digitally signed messages, used in Envelopev2
Theme / tags	"Business company", "trade register", "firmographics", "Bulgaria"
Data collection	Provided by the Bulgarian Registry Agency, converted to RDF in the EBG data model by ONTO
Dataset structure	The data about a company is spread across many XMLs: each day when the company made a filing to the Registry Agency. Furthermore, to find out any piece of data (e.g. current address), the complete history of filings by the company must be tracked, and the most recent data must be selected. The problem is exacerbated for list fields (e.g. list of directors) that need to be added to and subtracted from, rather than overwritten, creating a need for consolidation ("data fusion"). ONTO processed BG TR as part of a Datathon held on 24-26 Mar 2017 in Sofia. A Bulgaria Company Data Mapping was created by OCORP and ONTO that describes a mapping from XML to RDF. Technical information and conversions are available in the google folder Datathon. Later this work was extended by adding BG procurement data as part of the Hackathon "10 years Bulgaria in the EU" #BG10xEU on 12-13 May 2017 in Sofia. The RDF data was represented according to the EBG semantic model and loaded in the EBG repository.
Standards and metadata	The data is described in XSD (converted to RelaxNG Compact by ONTO). The following code lists are used: • Trader type (e.g. ЕООД Еднолично Ограничено Отговорно Дружество, EOOD Sole Limited Trader) • NKID Economic activity (Bulgarian extension of Eurostat NACE) • EKATTE territorial classification • Internal codes, e.g. SubDeedType, EuropeanEconomicInterestRepresenterTypes, ForeignAuthority, etc. • Some fields don't have a defined code list, e.g. MandateTypeText The data was linked by ONTO to Eurostat NUTS+LAU, Geonames and Linked Leaks
Dataset owner/publisher/ provider name	XML data: Bulgarian Registry Agency, http://brra.bg. RDF conversion: ONTO
Dataset licence	CC0
Data availability, how data can the dataset be accessed?	The original data is a XML data dump. RDF data is available for querying at the EBG SPARQL endpoint.
Dataset source URL	http://opendata.government.bg/dataset/tbprobckn-pernctbp (BG CKAN)
Dataset format (current and target data format)	XML, converted to RDF in the EBG data model by ONTO
Update frequency	The initial dump covered 8 years (2008-2016). 3 subsequent updates were provided covering 9, 7, and 4 months of data respectively. So the update frequency is about 6 months, but is increasing recently.
Time coverage	2008 to the present. 2008 was the year of trade register reform in Bulgaria, and all companies were asked to confirm their registration to cut outo the large number of inactive companies. As of Feb 2018, the last update was Nov 2017. So for all practical purposes, the data dump covers the complete register scope.
Spatial coverage	Bulgaria (domestic companies, foreign branches, European companies)
Language	Bulgarian (some English company names)
Relation to euBusinessGraph	Source for the business information graph
Data discoverability	The availability of this data was publicized by ONTO in a series of blog

	posts and news articles under the heading "Hacking the Trade Register", e.g. see in these news outlets: manager.bg, banker.bg, kaldata.com.
Data identification	• Official company registration code (EIK), e.g. 200356710 for ONTO • Unofficial GUID that allows access to a per-company HTML page, e.g. for ONTO: https://public.brra.bg/CheckUps/Verifications/ActiveCondition.ra?guid=617f4edf8c154f4296efdf146513de21
Data interoperability	Besides the standard classifications and identifiers described above, company addresses are mapped to Eurostat NUTS+LAU (for Bulgarian addresses only) and matched to Geonames (which is of higher value for foreign addresses).
Data privacy	Director names are included. Person IDs (EGN) are substituted by a cryptographic hash. This allows to identify when the same person participates in several companies, but not to discover his/her EGN, which is considered private data.