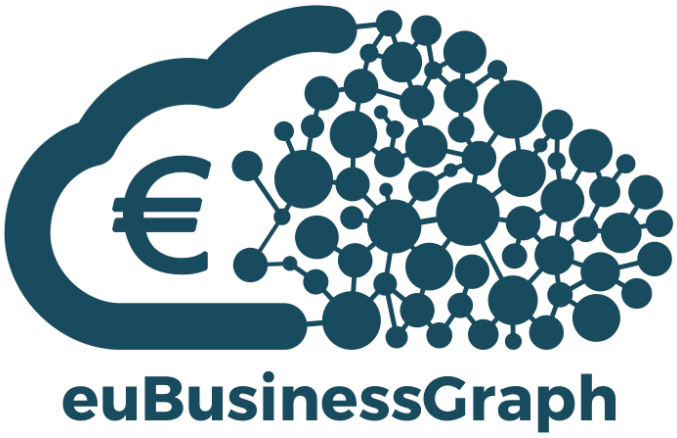
**Innovation Action (IA)**

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

Enabling the European Business Graph for Innovative Data Products and Services




## Deliverable 2.1:

# System of Identifiers, Ontologies and Vocabularies

| | |
|---|---|
| Date | 02.03.2018 |
| Author(s) | Chris Taggart (OCORP), Tatiana Tarasova (SDATI), Javier Paniagua (SDATI), Vladimir Alexiev (ONTO), Fredrik Seehusen (SINTEF), Brian Elvesæter (SINTEF), Dumitru Roman (SINTEF) |
| Dissemination level | Public (PU) |
| Work package | WP2 |
| Version | 1.1 |

# Document metadata

## Quality assurors and contributors

| | |
|---|---|
| Quality assuror(s) | Matjaz Rihtar (JSI) , Inna Novalija (JSI), Lars Tveit (EVRY), Stine Neteland (EVRY), Torsten Utne (EVRY) |
| Contributor(s) | euBusinessGraph Consortium |

## Version history

| Version | Date | Description |
|---|---|---|
| 0.1 | 01.12.2017 | Initial consolidated version. |
| 0.2 | 04.12.2017 | Version for internal review. |
| 0.3 | 11.12.2017 | Version incorporating review comments. |
| 0.4 | 18.12.2017 | Cross references and document styling. |
| 0.5 | 21.12.2017 | Formatting, grammar, and spell-checking. |
| 1.0 | 29.12.2017 | Final formatting and layout. |
| 1.0.5 | 27.02.2017 | Updated version for internal review with extensions to model Identifier Systems |
| 1.1 | 02.03.2017 | Final formatting and layout. |

# Table of contents

# Executive summary

The euBusinessGraph has drawn on the experience of its data providers and technology providers to tackle the complex task of combining company data from multiple sources. This deliverable describes the steps we have taken to solve two key challenges to delivering this technically.

First, we have created a **common semantic model** to represent companies and their attributes in a consistent way, allowing us to combine data on companies from multiple sources, and thus to present richer and more varied data to end users of the platform, as well as allowing them to be classified by different legal forms or economic activity.

These data models are expected to evolve over the lifetime of the project based on feedback from users, from the business cases, and as the technical platform develops (indeed they have already evolved to this point).

Second, we have performed a thorough analysis of **identifiers** in the context of euBusinessGraph. This has provided an iterative path from an MVP implementation towards a more fine-grained approach, dealing with successively more difficult aspects, and adding both more data, and richer meta-data over time.

Among the areas we have addressed:

- Related works for the data model
- Business case requirements for the data model
- Company data model
- Mapping of data to RDF
- Typology of identifiers
- Existing company-related identifiers
- Technical and business requirements for identifiers

From the analysis of the different identifier systems and the requirements of the business cases of the project, we singled out key aspects about identifiers and addressed them in the common semantic model. As such, the Identifier System class includes a series of attributes that model expectations about the identifiers that are issued within a particular system. We modelled validation and cleaning rules and web resources that are used for search, browsing and retrieval of identifier information. Moreover, we distinguished agents that operate to maintain, issue and publish the different identifier systems.

# Namespaces and Abbreviations

This document describes the semantic model used by euBusinessGraph (euBusinessGraph) to represent companies and their attributes, such as those that define companies' online or physical presence, classify them into different legal forms or by economic activity, etc. The reported model is the first iteration and will be revisited and updated in the second period of the project.

The following namespaces are used in the document. Prefixes are used to refer to specific terms of the underlying schema (e.g. **rov:**RegisteredOrganization refers to the term "Registered Organization" defined in the namespace of rov), or to indicate the source of citation (e.g. [**rov:**] at the end of the phrase refers to Registered Organization Vocabulary).

**Table 1: Namespaces used in the document**

| Schema | Prefix | Namespace |
|---|---|---|
| The euBusinessGraph Ontology | ebg | http://data.businessgraph.io/ontology# |
| The Organization Ontology | org | http://www.w3.org/ns/org# |
| Registered Organization Vocabulary | rov | https://www.w3.org/ns/regorg# |
| Core Person Vocabulary | person | http://www.w3.org/ns/person# |
| ISA Programme Location Core Vocabulary | locn | https://www.w3.org/ns/locn# |
| Schema.org | schema | http://schema.org/ |
| XML Schema | xsd | https://www.w3.org/2001/XMLSchema# |
| Global Legal Entity Identifier (GLEI or LEI) | lei | https://www.gleif.org/en/ |
| LEI-CDF Version 2.1 | lei-cdf | https://www.gleif.org/en/about-lei/common-data-file-format/lei-cdf-format/lei-cdf-format-version-2-1 |
| GLEIF Registration Authorities List | lei-ral | https://www.gleif.org/en/about-lei/gleif-registration-authorities-list |
| Entity Legal Form (ELF) Code | lei-elf | https://www.gleif.org/en/about-lei/common-data-file-format/lei-cdf-format |
| DBpedia Ontology | dbo | http://dbpedia.org/ontology/ |
| DC Elements | dc | http://purl.org/dc/elements/1.1/ |
| DC Terms | dct | http://purl.org/dc/terms/ |
| IANA language tag list | iana | https://github.com/euBusinessGraph/eubg-data/blob/master/data/iana-lang-tags.csv |
| Financial Industry Business Ontology | fibo | https://spec.edmcouncil.org/fibo/ |
| Asset Description Metadata Schema | adms | http://www.w3.org/ns/adms# |
| EU NUTS classification as Linked Data | nuts | http://nuts.geovocab.org/ |

Table 2 below explains abbreviations used in the document.

**Table 2: Abbreviations used in the document**

| Abbreviation | Description |
|---|---|
| API | Application Programming Interface |
| ASCII | American Standard Code for Information Interchange |
| ATECO | ATtività ECOnomiche, Italian extension of NACE |
| Atom | Atom, an XML-based web syndication (feed) format |
| BRC | Brønnøysund Register Centre, a Norwegian data provider in euBusinessGraph |

| | |
|---|---|
| BRIS | Business Registers Interconnection System, an EC DG Justice project to integrate register data across the EEA and some other European countries |
| CC | EU Candidate Countries |
| CDF | GLEIF Common Data File format |
| CSV | Comma Separated Values, format for storing tabular data in text form |
| DC | Dublin Core, a small set of metadata terms including author, publisher, date of creation, date of publication, etc. ISO 15836 standard. |
| DG | Directorate General, one of the "ministries" of the EC |
| DW | Deutsche Welle Innovation, an euBusinessGraph business case partner |
| euBusinessGraph | euBusinessGraph, a H2020 project to develop a graph of company and related data |
| EC | European Commission |
| EEA | European Economic Area: EU plus Iceland, Liechtenstein and Norway |
| EFTA | European Free Trade Association, an organization consisting of Iceland, Liechtenstein, Norway, and Switzerland that together with the EU participates in the European single market. |
| ELF | Entity Legal Form Code, an authoritative list of company types. ISO 20275 standard. |
| EU | European Union |
| FIBO | Financial Industry Business Ontology |
| FTS | Full-Text Search |
| GLEI | Global Legal Entity Identifier, same as LEI. A global register of financial and other companies, standard ISO 17442, intended to increase financial transparency and stability. |
| GLEIF | GLEI Foundation, developer of GLEI and related ISO standards |
| GML | Geography Markup Language: XML language for representing vector geometry objects |
| HTML | Hypertext Markup Language |
| HTTP | Hypertext Transfer Protocol |
| IANA | Internet Assigned Numbers Authority. Keeps registers of various entities, including IANA language tags |
| ICAO | International Civil Aviation Organization |
| ID | Identifier |
| INSPIRE | Infrastructure for Spatial Information in the European Community |
| ISA2 | Interoperability solutions for public administrations, businesses and citizens: an EC program for e-Government interoperability |
| ISIC | United Nations International Standard Industrial Classification of economic activities |
| ISO | International Organization for Standardization |
| JSON-LD | JavaScript Object Notation for Linked Data, a web-developer-friendly RDF serialization. euBusinessGraph has selected it as internal data transmission format |
| LAU | Local Administrative Unit, country administrative divisions below the NUTS level 3. Managed by each EU country independently and centralized by Eurostat |
| LEI | Legal Entity Identifier, same as GLEI |
| LOD | Linked Open Data, using semantic web technologies to link open data |
| LOV | Linked Open Vocabularies, a catalogue of ontologies |
| NACE | Nomenclature Statistique des activités économiques, Statistical Classification of Economic Activities in the European Community by Eurostat |
| NACEBEL | Belgian extension of NACE |
| NKID | Национална Класификация на Икономическите Дейности, Bulgarian version of NACE |

| | |
|---|---|
| NUTS | Nomenclature of Territorial Units for Statistics by Eurostat |
| OCORP | OpenCorporates, content and use case partner in euBusinessGraph. Has open data on 130M companies sourced from official registers. |
| ONTO | Ontotext, a semantic technology partner in euBusinessGraph |
| ORM | Object-Role Modelling, a modelling methodology. SINTEF uses it through the tool NORMA[1] that is available as a plugin for Microsoft Visual Studio to model data and to generate an RDF rendition of the euBusinessGraph ontology |
| PDF | Portable Document Format |
| RA, RAL | GLEI Registration Authorities List |
| RAMON | Reference And Management Of Nomenclatures, Eurostat's metadata server |
| RDF | Resource Description Framework, the graph data model of the semantic web |
| RDFS | RDF Schema, a vocabulary for describing subclasses, sub properties and property domains and ranges |
| RSS | Rich Site Summary or Really Simple Syndication, a type of web feed |
| SDATI | SpazioDati, content and business case partner in euBusinessGraph. Has rich data on Italian companies integrated from various sources (Atoka.io) |
| SEMIC | Semantic Interoperability Centre Europe |
| SINTEF | A Norwegian research organization, coordinator of euBusinessGraph |
| SKOS | Simple Knowledge Organization System, the ontology for representing lookup lists: controlled vocabularies, thesauri, classification schemes, taxonomies, subject-heading systems |
| Turtle | Terse RDF Triple Language, a human-readable RDF serialization |
| UN | United Nations |
| URI | Uniform Resource Identifier |
| URL | Uniform Resource Locator, a kind of URI that points to a resource using some internet protocol (HTTP, FTP, mailto, etc.) |
| VAT | Value Added Tax |
| WGS | World Geodetic System: reference frame for the Earth for use in geodesy and navigation |
| XLS | The Microsoft Excel file format |
| XML | Extensible Markup Language |
| XSD | XML Schema Definition, a way to describe XML documents formally |

[1] NORMA - The Software! The ORM Foundation, accessed 6 August 2017

# 1 Introduction

The euBusinessGraph project aims at simplifying cross-border and cross-lingual collection, reconciliation, aggregation and analysis of company-related information from several authoritative and non-authoritative sources.

In this document, we present the first release of a common company model designed to address the challenges mentioned above. This first release is the product of an extensive review of existing approaches for representing company-related information, available sources from the project's data providers and requirements posed by the different business cases in the project.

In its first release, the model focuses on capturing key company information present in official registers such as legal name, registered address and economic classification, and also information coming from online resources related to the company such as company websites, blogs and social media accounts. These aspects are explicitly incorporated into the model and describe company information that is shared across data providers and directly accessible through the graph. Additionally, the model supports advertising other company related information available from data providers directly.

Achieving matching and reconciliation across jurisdictions and registers requires careful modelling of identifier use. This release models the different cases through properties that describe the lifecycle of each identifier issued and by encoding a series of characteristics of the identifier system to which the identifier belongs. We follow a pragmatic approach when describing identifier systems in terms of these characteristics. We model expectations of a particular system that should help determine to which extent an indicator can be used for matching and reconciliation. Additionally, we model web resources that are frequently found for identifier systems such as search endpoints, templates for building identifier URLs through which company information can be reached and other resources that describe the system's rules. Finally, the model supports the representation of the different agents that are in charge of setting and maintaining rules, issuing identifiers and publishing identifier databases.

The rest of the document is structured as follows:

- Section 2 is devoted to the common semantic model for company data. Section 2.1 covers related work about existing company-related models, ontologies and vocabularies, Section 2.2 describes requirements to the semantic model from the business cases in the project and Section 2.3 presents the first iteration of the semantic model.
- Section 3 describes identifier handling in the euBusinessGraph covering the analysis of existing identifier systems and the project's requirements regarding these.
- Section 4 concludes summarising key contributions and describing next steps.

# 2 Data Model, Ontologies and Vocabularies

The euBusinessGraph approach to develop shared company data models is twofold. On one hand, we seek to reuse, align and extend existing schemas/vocabularies/ontologies relevant to modelling and representing business information. We present an overview of the related works in Section 2.1, where Section 2.1.1 provides external references to relevant standardisation organisations, projects and initiatives, and discussions of their results that can be useful to euBusinessGraph.

On the other hand, the development of the shared data model was guided by the needs of the first consumers of data from euBusinessGraph (its business cases), but prioritised towards availability of data in the consortium (its data providers). Table 3 below summarises all data consumers and data providers in the project.

**Table 3: Data Consumers and Data Providers.**

| Organisation | Data Consumer | Data Provider | Business Case |
|---|---|---|---|
| DW | Y | Y | Data Journalism Product (DJP) |
| EVRY | Y | N | EVRY CRM Service (CRM-S) |
| SDATI | Y | Y | Improved content coverage and data quality of Atoka, SaaS B2B lead generation service (ATOKA+) |
| CERVED | Y | N | Tender Discovery Service (TDS) |
| OCORP | N | Y | Corporate Events Data Access Service (CED) |
| BRC | N | Y | Norwegian Registries API Service (BR-S) |
| JSI* | N | Y | - |
| ONTO** | N | Y | - |

*(*) JSI provides business-related news streams annotated with company mentions (**) ONTO provides open access to company data in the Bulgarian Trade Register.*

Section 2.1.2 is dedicated to euBusinessGraph data providers, where we examine data from three partners: SDATI, OCORP and BRC. Analysis of the business cases' requirements is given in Section 2.2. While these requirements guided development of the model in the first year of the project, not all of them were considered in the initial version of the model. Priorities were given to those that could be covered by data that is already available to the consortium through its members.

The initial version of the model is discussed in detail in Section 2.3. It provides an instance diagram of the model, and includes topic-by-topic discussions with informative description of classes and properties, scope notes, examples, data provider rules, as well as suggested RDF bindings. The latter were used to develop the euBusinessGraph ontology http://data.businessgraph.io/ontology.

An Object-Role Modelling (ORM) model was developed. This ORM model was used to generate a RDFS representation of the company data model. The ORM model is described in Section 2.3.3 and Appendix A. Mapping of OpenCorporates and SpazioDati data to RDF is discussed in Section 2.4 and Appendix B.

## 2.1 Related Works

### 2.1.1 External References

In this section, we summarise sources, and which terms and term descriptions we use in defining our model. We select terms according to how well they cover our business case needs.

#### 2.1.1.1 EC Core Vocabularies

The e-Government Core Vocabularies[2] were developed by the ISA2 SEMIC Joinup[3] semantic interoperability initiative and then standardized by the W3C.

Handbook on using the Core Vocabularies[4] was used to adopt the vocabularies to our data model.

#### 2.1.1.2 Schema

Schema.org[5] is a widely used vocabulary spearheaded by the big 4 search engines and developed collaboratively.

It is highly reusable because it makes few ontological commitments, having to cater to a truly global audience of millions of web sites[6]:

- Rather than rdfs:domain and rdfs:range, which are mono-morphic, i.e. admit only a single class and thus force rigid and deep class hierarchies, it uses schema:domainIncludes and schema:rangeIncludes, which are purely advisory.
- While owl:DatatypeProperty and owl:ObjectProperty (or a range being a class) force a property into a dichotomy that you can use, many Schema properties permit either resource or free text (whatever is available), which make them more widely usable.
- Allows free use of external taxonomies/vocabularies (especially in the field schema:additionalType, but not only).

#### 2.1.1.3 Global Legal Entity Identifier

The GLEI Foundation has established a registration structure to issue Legal Entity Identifiers (LEI)[7] to legal entities participating in financial transactions. The LEI structure is standardized as ISO 17442[8]. Accompanying entity data is standardized by GLEIF as the Common Data File[9] (CDF) formats. CDF consists of two levels:

- Level 1[10]: “who is who”
- Level 2[11]: “who owns whom”

We will mostly be concerned with the first level LEI-CDF Version 2.1[12] and two code lists that may be useful to the project.

##### 2.1.1.3.1 Registration Authorities List

Includes 651 national official registers. There are 60 countries with several national registers.

[2] https://joinup.ec.europa.eu/asset/core_vocabularies/asset_release/core-vocabularies-v20#download-links
[3] https://joinup.ec.europa.eu/community/semic/description
[4] https://joinup.ec.europa.eu/site/core_vocabularies/Core_Vocabularies_user_handbook/ISA%20Hanbook%20for%20using%20Core%20Vocabularies.pdf
[5] http://schema.org/
[6] R.V. Guha, Dan Brickley and Steve Macbeth: “Schema.org: Evolution of Structured Data on the Web.” ACM Queue, November-December 2015 http://queue.acm.org/detail.cfm?id=2857276 (Last access: 29/11/2017)
[7] https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei
[8] https://www.gleif.org/en/about-lei/iso-17442-the-lei-code-structure
[9] https://www.gleif.org/en/about-lei/common-data-file-format
[10] https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-1-data-who-is-who
[11] https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom
[12] https://www.gleif.org/en/about-lei/common-data-file-format/lei-cdf-format/lei-cdf-format-version-2-1

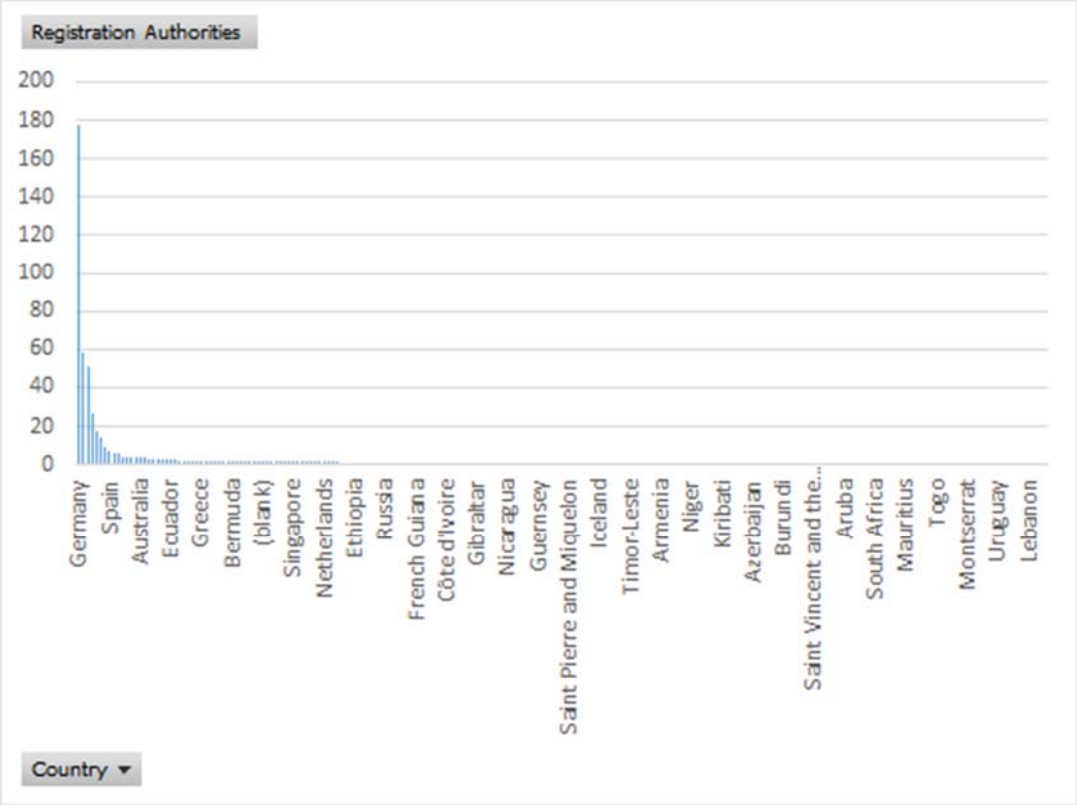



**Figure 1: Registration authorities by country**

- Germany has the most – 177
- United Kingdom's Companies Register by the Companies House is listed 3 times under different jurisdictions (England and Wales, Northern Ireland and Scotland). There is one single register with company IDs unique across these three jurisdictions. From this example, we see that RAL really lists registers, not registration authorities.
- Italy has 4 registers:
    - Business Register (Registro delle Imprese) by Infocamere
    - List of collective investment schemes (UCITS) by Bank of Italy
    - Pension fund supervision commission (Commissione di vigilanza sui fondi pensione)
    - List of investment funds by Italian Companies and Stock Exchange Commission

None of the descriptive fields is guaranteed to be filled. Here are the counts:

| | | |
|---|---|---|
| Count of Registration Authority Code | 653 | 100% |
| Count of International name of Register | 473 | 72% |
| Count of Local name of Register | 406 | 62% |
| Count of International name of organisation responsible for the Register | 603 | 92% |
| Count of Local name of organisation responsible for the Register | 508 | 78% |
| Count of Website | 625 | 96% |

### 2.1.1.3.2 Entity Legal Form Code

Entity Legal Form (ELF) Code[13] will resolve variant names for each valid legal form within a jurisdiction to a single code per legal form.

[13] https://www.gleif.org/en/about-lei/common-data-file-format/lei-cdf-format

On November 30, GLEIF released the first iteration of Entity Legal Forms Code List[14] that lists more than 1,600 entity legal forms across more than 50 jurisdictions, including Italy, Bulgaria and UK. The ELF Code List assigns a unique code to each entity legal form. The ELF code is an alpha-numeric code of four characters from the basic Latin character set, e.g., “ID30” identifies “Limited Partnership” in UK. While the code list does not unify company types across jurisdictions, it provides a single authoritative list of all company types. It is ISO 20275 and is maintained by GLEIF.

#### 2.1.1.4 Wikipedia Business Entities

Wikipedia world-wide List of Business Entities[15] is a useful resource since it includes a translation to English and "approximate equivalents in the company law of English-speaking countries". Compared to the list of legal types defined by euBusinessGraph[16], Wikipedia includes:

- BG: 9 out of 23 euBusinessGraph types
- NO: 18 out of 43 euBusinessGraph types
- UK: 11 out of 25 euBusinessGraph types

#### 2.1.1.5 Business Registers Interconnection System (BRIS)

BRIS is an EU DG-Justice project for exchanging data between registers, thus creating an EU-wide company register. This project interconnects business registers across Europe and provides a single (though limited) search form[17].

There are three data elements potentially interesting for the project:

1. BRIS List of legal forms provides them in the national language, but there is rich explanation in English, e.g.:
    - "società a responsabilità limitata": The limited liability company (società a responsabilità limitata) is a type of limited liability company which has legal personality and meets the company’s obligations only out of its assets. The liability of the members is limited to the contribution of their quotas. It is a company which is usually smaller in size with a more flexible organisational structure than that of a joint stock company. The members are not personally liable for the company’s obligations, even if they have acted in the name and on behalf of the company. Contributions in kind (provision of works/services) are permitted, whereas they are prohibited in joint stock companies.
2. BRIS List of national registers, e.g.:
    - IT RI (Italy’s Registro delle imprese)
    - FI FPRO (Finland’s Patentti-ja Rekisterihallitus)
3. BRIS List of pan-European company identifiers formed from the register ID and company ID, e.g.:
    - ITRI.02866370170 for an Italian company
    - FIFPRO.1670700-0 for a Finnish company

#### 2.1.1.6 Dublin Core

Dublin Core includes common classes and properties for describing resources. Old and venerable, it is reused in many ontologies, including W3C:

- DC Elements[18]
- DC Terms[19]

[14] https://www.gleif.org/en/newsroom/blog/iso-20275-entity-legal-forms-code-list
[15] https://en.wikipedia.org/wiki/List_of_business_entities
[16] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-company-type.xlsx
[17] https://e-justice.europa.eu/content_find_a_company-489-en.do
[18] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#H3

When a property is present in both dc: and dct:, we use dc: for a literal and dct: for a resource.

#### 2.1.1.7 IANA Language Tags

IANA language tag list[20] uses ISO 639-1, 639-2 and 639-3 codes (2 and 3 letter codes) and extends it with additional segments (the main ones are language, script, and region of use). It is the standard to use for XML (xml:lang) and RDF (language tag of an rdf:langString literal). We published the list in the CSV format[21], to have much easier access to the list than the official one. Examples:

- nl: Dutch
- nl-BE: Flemish
- ru: Russian
- ru-Latn: Russian transliterated to Latin characters

It is also possible to construct custom language tags.

#### 2.1.1.8 Financial Industry Business Ontology (FIBO)

FIBO[22] is a very complex set of ontologies for the terms, facts, and relationships associated with financial contracts. FIBO™ covers financial instruments (product reference data), market data pricing, and financial processes financial industry. While it includes entities like companies, directors, control relations, monetary amounts, debts, obligations, contracts, financial instruments, these are primarily in the context of listed companies and other financial market entities. It is available in Turtle, RDF and JSON-LD.

Some simple parts of FIBO are also mapped to Schema[23]. We won't use FIBO directly but refer to it for modelling inspiration for some of the euBusinessGraph entities.

#### 2.1.1.9 EC NUTS and LAU

Eurostat has established a unified hierarchy of regions across the EU, EFTA and Candidate Countries. It consists of:

- Nomenclature of Territorial Units for Statistics[24]: NUTS, 3 levels[25]
- Local Administrative Units[26]: LAU, 2 levels

NUTS is an important geographic resource for euBusinessGraph since:

- Significant Open Data is available, and can support address data mapping (e.g., from postal code to NUTS) and use cases (e.g., hierarchical facets, distance calculations, spatial inclusion)
- The administrative hierarchy varies greatly in different countries, whereas NUTS+LAU provide a uniform hierarchy

The variety and number of regions per country are provided for EU[27] and non-EU[28] (EFTA+CC). euBusinessGraph made a consolidated version[29], and Wikipedia provides a simplified version[30] including only NUTS levels.

---

[19] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#H2
[20] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
[21] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
[22] https://spec.edmcouncil.org/fibo/ontology/master/latest/
[23] http://schema.org/docs/financial.html
[24] https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics
[25] See overview at http://ec.europa.eu/eurostat/web/nuts/overview
[26] https://en.wikipedia.org/wiki/Local_administrative_unit
[27] http://ec.europa.eu/eurostat/web/nuts/national-structures-eu
[28] http://ec.europa.eu/eurostat/web/nuts/correspondence-tables/national-structures-non-eu
[29] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-NUTS-and-LAU-consolidated.xlsx
[30] https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics#Levels

#### 2.1.1.9.1 NUTS RDF

The EC project PlanetData[31] has provided part of NUTS in RDF at http://nuts.geovocab.org/, using the NeoGeo RDF Vocabulary for GeoData[32].

#### 2.1.1.9.2 LAU RDF

euBusinessGraph made RDF representation of LAU, using the consolidated version of EC NUTS and LAU data and following the NUTS RDF representation mentioned above.

LAU RDF is available in the project repository[33] together with the transformation scripts.

### 2.1.1.10 Geonames

Geonames is an important place dataset of 10M places, including:

- Administrative regions and settlements
- Physical places (continents, mountains, oceans, seas, lakes, streams)
- Area features (oil fields, parks)
- Linear features (roads, power lines)
- Spot features (buildings, farms, hotels, oil wells)

Geonames has a 2-level hierarchy of feature types: Class>Code. Figure 2 below shows an example of Administrative Region feature Codes:

| count | Feature Code | Feature name | Feature Description |
|---|---|---|---|
| 10,952,483 | | Total | |
| 337,692 | A | Administrative Boundary Features (country, state, region,...) | |
| 103,896 | A.ADM4 | fourth-order administrative division | a subdivision of a third-order administrative division |
| 86,097 | A.ADM3 | third-order administrative division | a subdivision of a second-order administrative divisio |
| 63,518 | A.ADMD | administrative division | an administrative division of a country, undifferentiate |
| 40,154 | A.ADM2 | second-order administrative division | a subdivision of a first-order administrative division |
| 37,219 | A.ADM5 | fifth-order administrative division | a subdivision of a fourth-order administrative division |
| 3,880 | A.ADM1 | first-order administrative division | a primary administrative division of a country, such as |
| 1,127 | A.ADM4H | historical fourth-order administrative division | a former fourth-order administrative division |
| 423 | A.ADM3H | historical third-order administrative division | a former third-order administrative division |
| 406 | A.ADM2H | historical second-order administrative division | a former second-order administrative division |
| 388 | A.ADM1H | historical first-order administrative division | a former first-order administrative division |
| 193 | A.PCLI | independent political entity | |
| 163 | A.PRSH | parish | an ecclesiastical district |
| 89 | A.ADMDH | historical administrative division | a former administrative division of a political entity, un |
| 44 | A.ZN | zone | |
| 26 | A.PCLD | dependent political entity | |

**Figure 2: Geonames Administrative Region feature codes**

[31] https://www.planet-data.eu/
[32] http://geovocab.org/
[33] https://github.com/euBusinessGraph/eubg-data/tree/master/data/LAU

A rough correspondence between NUTS and the Geonames administrative hierarchy is seen in Table 4 below:

**Table 4: NUTS – Geonames correspondence**

| NUTS | Europe | Geonames | World |
|---|---|---|---|
| NUTS3 | 1493 | A.ADM1 | 3880 |
| LAU1 | 8695 | A.ADM2 | 40154 |
| LAU2 | 162817 | A.ADM3 | 86097 |

The order is not the same and the correspondence is not exact for all countries. Figure 3 illustrates this situation for Italy:


| NUTS3 | Province |
|---|---|
| ITC1 | Piemonte |
| ITC2 | Valle d'Aosta/Vallée d'Aoste |
| ITC3 | Liguria |
| ITC4 | Lombardia |
| ITF1 | Abruzzo |
| ITF2 | Molise |
| ITF3 | Campania |
| ITF4 | Puglia |
| ITF5 | Basilicata |
| ITF6 | Calabria |
| ITG1 | Sicilia |
| ITG2 | Sardegna |
| ITH1 | Provincia Autonoma di Bolzano/Bozen |
| ITH2 | Provincia Autonoma di Trento |
| ITH3 | Veneto |
| ITH4 | Friuli-Venezia Giulia |
| ITH5 | Emilia-Romagna |
| ITI1 | Toscana |
| ITI2 | Umbria |
| ITI3 | Marche |
| ITI4 | Lazio |

**Figure 3: Example of NUTS 3**

The correspondence breaks down at LAU2 because in some countries these correspond to populated places/settlements (P.PPL) and not ADM3 administrative regions.

### 2.1.1.11 Classifications of economic activities

There are many classifications of economic activities available. Some of them come from international organisations, like UN or EC. Others are maintained by national agencies or companies, e.g., OpenCorporates has a list of industry classifications[34]. There are crowd-sourced classifications in Wikipedia, classifications of Products[35] and Industries[36].

Below, we give details about relevant UN and EC classifications, and leave in Appendix B a more comprehensive list of classifications that could be useful to our project.

[34] http://api.opencorporates.com/documentation/API-Reference#industry_codes
[35] https://en.wikipedia.org/wiki/Product_classification
[36] https://en.wikipedia.org/wiki/Industry_classification

#### 2.1.1.11.1 UN classifications

UN Stats provides a list of national classifications[37] mostly for activities and products, and maintains international classifications[38], among which International Standard Industrial Classification of All Economic Activities[39] (ISIC) is the most relevant.

As an alternative to the standard classifications, UN defines alternate structures[40], groupings that delineate certain industries, for example Information economy, Energy, and Financial services. These groupings can be useful for analytical purposes for some customers.

#### 2.1.1.11.2 EC classifications

EC maintains a catalogue of international statistical classifications and nomenclatures[41] and correspondence tables for most EC & UN classifications[42]. Below, we give more details about Statistical Classification of Economic Activities in the European Community (NACE). Other classifications that might be relevant to the project include:

- Products:
    - EC CPA 2.1: Statistical Classification of Products by Activity, Version 2.1); the European version of the CPC
    - EC PRODCOM: PRODCOM List 2016 (List of PRODucts of the European COMmunity).
- EC CN: CN 2017 (Combined Nomenclature), basis of the Common Customs Tariff

##### 2.1.1.11.2.1 NACE

NACE is available from EC RAMON[43], and includes the following fields:

- **Order**: for global sorting
- **Level**: hierarchical level
- **Code**: e.g. A, 01, 01.11
- **Parent**: code of parent concept
- **Description**: label
- **ISIC**: Reference to ISIC Rev. 4 (correlation code)
- **Includes** (we concatenate 3 fields)
    - This item includes (the listed activities)
    - Rulings (more activities included as the result of a ruling)
    - This item also includes (yet more listed activities)
- **Excludes**: this item excludes the listed activities

EC also provides a simple NACE dictionary[44] including just Code and Label.

[37] http://unstats.un.org/unsd/cr/ctryreg/default.asp?Lg=1
[38] https://unstats.un.org/unsd/cr/registry/regct.asp?Lg=1
[39] https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27&Lg=1
[40] https://unstats.un.org/unsd/cr/registry/regat.asp?Lg=1
[41] http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM&StrGroupCode=CLASSIFIC&StrLanguageCode=EN
[42] http://ec.europa.eu/eurostat/ramon/relations/index.cfm?TargetUrl=LST_REL&StrLanguageCode=EN&IntCurrentPage=3
[43] http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=NACE_REV2&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC#
[44] http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&file=dic%2Fen%2Fnace_r2.dic

#### 2.1.1.11.2.2 NACE RDF

RAMON publishes a NACE Ontology[45] that includes some of the master fields, but is not compatible with SKOS.

EstatWrap[46] and Linked Statistics[47] provide dynamic conversion of the simple dictionary to RDF:

- It is SKOS-structured
- Includes sameAs links between URLs of EstatWrap, Linked Statistics, EIONet and RAMON
- Does not include any hierarchy (no skos:broader, the concept scheme has skos:hasTopConcept to every concept)
- Does not have different classes for the different levels and for the additional groupings
- There are only labels, but not detailed descriptions ("This item includes")
- Labels are only in English (so is the RAMON master data)
- Does not specify a base URL

euBusinessGraph made a more complete NACE RDF[48] using most of the data above. Currently, euBusinessGraph NACE RDF has several limitations:

- We include only basic codes, not additional groupings like nace:B-D_X_FOOD
- We provide only English labels and descriptions, and no national extensions (see next section)
- We omit the ISIC correspondence

#### 2.1.1.11.2.3 NACE National Extensions

All EU/EEA countries have translated NACE, and some have provided national extensions:

- IT ATECO: ISTAT Italian Classification of economic activities[49]; derived from NACE rev.2, and extends it as appropriate for the Italian economy
- BE NACEBEL: Belgium NACE[50]; extends NACE with Level 5
- NO SN: Norwegian NACE[51]; extends NACE
- BG KID 2008: Bulgarian NACE[52]; translates NACE to BG
- UK SIC 2007: UK Standard Industrial Classification of Economic Activities[53]; based on NACE 2 and extends it with UK codes

### 2.1.2 Data Providers

Table 5 below summarises data shared with the project by its three main data providers: OCORP, SDATI and BRC. We identified three different modes of data sharing:

- **Shared**: data is fully shared with the Graph's users who can see the value without coming to each partner's site.
- **Matching only**: data value can only be used internally for matching company entities. End users will not see the value unless they go to each partner's graph. The presence of such

[45] http://ec.europa.eu/eurostat/ramon/ontologies/nace.rdf
[46] http://estatwrap.ontologycentral.com/dic/nace_r2#
[47] http://eurostat.linked-statistics.org/dic/nace_r2
[48] https://github.com/euBusinessGraph/eubg-data/tree/master/data
[49] http://www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007
[50] http://statbel.fgov.be/nl/statistieken/gegevensinzameling/nomenclaturen/nacebel/
[51] http://www.ssb.no/en/klass/klassifikasjoner/6
[52] PDF http://www.nsi.bg/sites/default/files/files/publications/KID-2008.pdf and interactive search https://www.kik-info.com/spravochnik/kid-2008.php
[53] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/455263/SIC_codes_V2.pdf

value can be advertised through the graph (see “Data offering” requirements in Section 2.2.1.1)

- **Other**: Data values can have different levels of granularity. When this mode is applied to a data value, coarser granularity is used to share it with the Graph’s users while a more detailed version is used for matching only purposes.

**Table 5: Summary of Data Shared by the Data Providers**

| Field | Description | Partners | Sharing mode |
|---|---|---|---|
| **Company** | | | |
| Legal Name | Official name of the company | OCORP, SDATI, BRC, ONTO | shared |
| Trading As | Informal/popular name of the company (also called Trade Name) | SDATI, ONTO | shared |
| Active? | A flag that identifies whether a company is active or not | SDATI | matching only |
| Startup? | A flag that identifies companies that are startups | SDATI | shared |
| Publicly traded? | A flag indicating whether the company is publicly traded or not | SDATI | shared |
| Languages | Languages used in web resources related to the organisation | SDATI, ONTO | matching only |
| **Company Dates & Status** | | | |
| Incorporation Date | Date entity was created | OCORP, BRC, ONTO | shared |
| Dissolution Date | Date entity was dissolved or removed from register | OCORP, ONTO | Shared |
| Current Status | Current status as reported by company register | OCORP | Shared |
| **Company/Institution Type** | | | |
| Company Type | Type (Legal Form) of entity. String, widely varies per country. May be a hierarchy: it should be reflected in the Types concept hierarchy, but each company has 1 value only | SDATI, OCORP, BRC, ONTO | matching only |
| Public sector organisation? | Flag that identifies organisations belonging to the public sector | SDATI, BRC | shared |
| Public Sector: type of entity | Official ISTAT type of entity (if in public sector) | SDATI | shared |
| **Economic Classification** | | | |
| UKSIC | UK Standard Industrial Classification | SDATI | shared |
| NACE | Statistical classification of economic activities in the European Community | ONTO, BRC | shared |
| ATECO | "Codice ATtività ECOnomica": IT Standard Industrial Classification | SDATI | other |
| **Legal Location** | | | |
| Registered Address | Public Legal Address, where legal papers can be served | OCORP, SDATI, ONTO | shared |

| | | | |
|---|---|---|---|
| Jurisdiction | ISO 3166 code of jurisdiction in which entity is incorporated/registered | OCORP, ONTO | shared |
| Country Code | ISO 3166 code for jurisdiction in which entity is incorporated/registered | SDATI, ONTO | shared |
| **Physical Presence** | | | |
| Locations | Other locations associated with an organisation | SDATI, ONTO | other |
| **Online Presence** | | | |
| Certified Emails | A list of emails that are officially registered and with the same validity as certified (snail) mail | SDATI | matching only |
| Websites | Websites associated with an organisation | SDATI, BRC | matching only |
| Wikipedia URI | Wikipedia URI of the organisation | SDATI | shared |
| Newsfeeds/blogs | RSS/Atom feeds associated with an organisation | SDATI | matching only |
| **Identifiers** | | | |
| Company Number | Identifier issued by national company register | OCORP, BRC, ONTO | shared |
| Companies House Company Number | ID given by Companies House in the UK | SDATI | shared |
| Atoka ID | Internal ID of the company in Atoka | SDATI | shared |
| OpenCorporates URI | Dereferenceable URL of entity on OpenCorporates | OCORP | shared |
| CCIIA Code | CCIIA: Camera di Commercio, Industria, Artigianato e Agricoltura. ID of the chamber of commerce that issues the REA (see next attribute) in that province | SDATI | matching only |
| REA Code | REA stands for "Repertorio Economico Amministrativo"<br>These are codes handed by chambers of commerce to companies upon registration | SDATI | matching only |
| VAT Code | Value added tax identification number | SDATI | matching only |
| IT Tax ID | ID issued by the Italian Revenue Agency (Agenzia delle Entrate) to identify citizens and entities | SDATI | matching only |
| IT IPA Code | Identifier for Italian Public Sector entities | SDATI | shared |

#### 2.1.2.1 Events

While the focus of work in the first period of the project has been on core business information (main objectives of the tasks T2.1 and T2.2), preliminary investigative efforts have been done towards identifying different types of events that will enrich euBusinessGraph company data (tasks T2.3 and T2.4). Below, we present an overview of company-related events that JSI and SDATI work with.

##### 2.1.2.1.1 JSI

The flagship functionality of EventRegistry is to fetch news articles from global news sources and, based on their content and metadata, cluster articles that describe the same world event.

In EventRegistry's top-down terminology, an "Event" is defined as a collection of one or more "Story" instances that report about the same world event. Stories are themselves collections of "Article" instances - enriched news articles - that report about the same world event in the same language.

The article clustering algorithm works by employing the following techniques:

- Article metadata extraction (publish date and time, ...)
- Named entity extraction, website categorization, and text “wikification”
- Active learning, seeded with high-quality annotated data
- Cross-lingual Story instance comparisons using canonical correlation analysis

Some of these techniques enable EventRegistry to enrich Event instances with “Category” and “Concept” attributes. “Source” attribute data is compiled from news source profiles, ranking news sources by their publishing volume over time, the number of times their articles have been “shared” on social media, and by external rankings, such as Alexa’s.

The following subsections provide an overview of the data models that EventRegistry defines. The listed attributes in this document best present each entity model that is being described.

**Event model overview**

| Attribute | Description |
|---|---|
| title | Set of news article titles, one for each language that had Event articles written in it |
| summary | List of news article summaries, one for each language that had Event articles written in it |
| eventDate | Event date |
| location | An event’s approximate geographical location data |
| stories | List of Story instances that were clustered under this Event instance |

**Article model overview**

| Attribute | Description |
|---|---|
| url | Article World Wide Web URL address |
| title | Article title text |
| body | Article body text |
| date | Date of publishing the article |
| time | Time of publishing the article |
| source | Details about the news source that published the article |
| isDuplicate | Whether this article is a duplicate of another found article |
| lang | Detected article language |

**Concept model overview**

| Attribute | Description |
|---|---|
| uri | Concept URI, usually a World Wide Web URL |
| type | Concept type: Person, Location, Organization, or “wiki” |
| description | Textual description of the concept, if available |

**Category model overview**

| Attribute | Description |
|---|---|
| uri | Category URI |
| label | Category label |
| trendingHistory | Aggregate history of articles in which this category had been detected |

**Source model overview**

| Attribute | Description |
|---|---|
| uri | Source URI, usually a URL hostname |
| title | Source title |
| ranking | Collection of rankings for this Source |
| location | A source's approximate geographical location data |

For more exact definitions of Event and Article data models, JSON Schema documents[54] have been produced, which also define the rest of the data models as JSON subschemas. The produced JSON Schema documents can be used for data validation purposes.

EventRegistry users can search Event or Article instances by the attributes described. Additionally, enriched data enables users to arbitrarily categorize these instances to higher-level categories, depending on their needs.

#### 2.1.2.1.2 SDATI

SDATI's core business product, Atoka, is built upon a knowledge graph of company data that, in addition to basic firmographics, contains contextual information about companies, including company relevant events.

Currently, there are two kinds of events in Atoka[55]:

1. Monitoring events: events that reflect changes in basic information about companies, key managers and locations
2. News-derived events: business relevant events that are extracted from news articles

**Monitoring Events**

Changes in basic information about companies, key managers and locations are registered from two kinds of data sources: official data sources (e.g., official provider of company data in Italy, CERVED) and non-official data sources (e.g., corporate websites).

There are four types of events that SDATI monitors:

1. Update (e.g., the legal name of company has been updated)
2. Addition (e.g., a company has added a link to a corporate twitter account on its corporate website)
3. Removal (e.g., a company's inactivity flag has been removed)
4. Sorting (e.g., sorting of companies' phone numbers has been changed)

These events are available through the Monitoring API[56].

**News-derived Events**

SDATI processes news articles to find usage of names of companies and key managers in the news' content. The key technology used is Dandelion API[57]. In addition to this, SDATI annotates news articles into such business events, as "dismissal", "closing", "protests staff shakes", "financial loss crises", "judicial tribunal", "mergers acquisitions", "new products", "change role", etc. These categories were created by domain experts based on their experience and users' demand.

News search API[58] contains documentation of the categories[59].

[54] https://github.com/mihajenko/eubg-data/tree/master/partners/JSI
[55] See https://github.com/euBusinessGraph/eubg-data/tree/master/partners/SDATI for more details.
[56] Atoka Monitoring API is documented at https://developers.atoka.io/v2/monitoring.html#
[57] https://dandelion.eu/
[58] https://developers.atoka.io/v2/news.html#news
[59] Event categories are documented at https://developers.atoka.io/v2/news.html#news_event

The objective of SDATI in the project is to extend and deepen our business information knowledge graph with two more jurisdictions, UK and Norway. In relation to business events, this will imply adaptation of Dandelion company extraction API to the UK and Norwegian languages or usage of other technology that already works with these languages. Besides this, SDATI aims to add other sources of company related events, e.g., registers of public contracts and tenders, trade registers, etc.

**Status progress**

SDATI has worked to implement a notification service in Atoka, to alert users of interesting events. The notification service sources from both monitoring and news-derived event groups.

Since different types of events are important for different user profiles, the notification service matches profile definitions to certain event types. A match indicates that the matched event is interesting to that user profile, and this triggers a notification. There is a third source of events: those that arise from the analysis of the previously mentioned sources. In this analysis, events of the same type are analysed as a time series, and new, higher-level events are detected by interpreting the series.

Additionally, the analysis performed on events of a certain type can be summarised as an indicator. Indicators provide a quick overview of how a company fares regarding the phenomenon they describe (e.g., as a single value ranging from 0 to 100). The following list shows indicators that SDATI is interested in obtaining from event instances generated in Atoka:

- Opening New Facilities
- Generic Executive Change Activity
- Specific Executive Change Activity (e.g., Marketing/Sales Executive)
- Increase in Workforce
- Increase in Reported Revenue
- Increase in Reported Assets

To power its notification service, Atoka is already capable of generating company entity lifecycle events for IT and UK jurisdictions. Additionally, Atoka also generates news-derived events for Italian companies. Presently, these events are being detected using the Dandelion API on Italian news articles. SDATI is collaborating with JSI to extend these capabilities to company entities in the UK.

In the first phase, SDATI is extending annotation of company mentions within news articles for company entities in the UK jurisdiction. The next step will include enhanced detection of events in news in the English language.

### 2.1.2.2 Code Lists, Vocabularies and Classifications

#### 2.1.2.2.1 Company type/ legal form

There is no agreed set of company types that crosses borders [rov:]. After examination, euBusinessGraph concluded there is no chance to "standardize" some shared super-types across jurisdictions. We collected code lists from our data partners: SDATI for Italy, ONTO for Bulgaria and BRC for Norway. The lists can be found in the project repository[60]. Note, OCORP has many jurisdictions and legal forms are not normalised.

#### 2.1.2.2.2 Status

There is no globally accepted list of company states. For Inactive, some providers look at hard evidence, such as when a company has been deregistered its status becomes “inactive; others -- at dissolution date in the past, or an extended period of inactivity (dormant). Because of this, a user cannot assume that Active and Inactive are opposites.

A list and definition of “active” and “inactive” statuses have been accumulated from the data providers and made available in the project repository[61].

[60] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-company-type.xlsx
[61] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-company-status.xlsx

#### 2.1.2.2.3 Economic activity

Data providers use NACE-based national extensions:

- SDATI uses both Italian ATECO and UK SIC, national extensions of NACE rev. 2
- BRC uses NACE rev. 2 with Norwegian NACE national extension

See Section 2.1.1.11.2.3 “NACE National Extensions” for details.

## 2.2 Business Cases’ Requirements to the Model

### 2.2.1 DJP: Data Journalism Product

Deutsche Welle (DW) and similar media organisations that are part of ARD[62] or EBU[63] are non-profit, financed by the public, or direct taxation. Their public remits are based on providing public value via journalism, and media content that informs, educates and entertains. Some open data providers such as OpenCorporates follow a public mission, based on related business models where data is available free for journalists, NGOs and academics with a view to achieve a public benefit. This “public interest” approach to data access is relevant for the viability of Data Journalism Product (DJP), the business case of DW.

DJP is a tool for journalists that helps them gather information for their articles and other news production. DJP is tailored for all journalists involved in company storytelling, those who do factual content production for information/education (public value/interest objective). Such journalists are not familiar with technologies for processing and dealing with data, hence, DJP must be abstracted from these technologies to be useful for the journalists.

DW has the following top-level requirements for euBusinessGraph associated with DJP:

1. Offer a more comprehensive "company profile". This refers to the information that is directly provided by euBusinessGraph following a company name search. It provides significant added value to journalists. It should therefore contain as much shared/free/open data as possible with a view to get an up-to-date “snapshot” overview of a company easily and quickly. Users of the DW Journalism Tool would use this information for research purposes but also to quickly create a “company profile” content item (a template-based profile that can be used in online articles or image galleries). Table 6 below lists attributes relevant for this kind of “company profile”, providing – by default – the most recent data items published (e.g. the most recent turnover figure).

| # | Attributes | Example - Search Term: “Opel” + “Car Company” |
|---|---|---|
| 1 | Name | Adam Opel AG |
| 2 | Company type | Stock Company |
| 3 | Country | Germany |
| 4 | Jurisdiction code | Germany |
| 5 | Address | Bahnhofsplatz, 65423 Rüsselsheim am Main |
| 6 | Key Manager | Ulrich Schumacher (Vorstandssprecher)<br>William F. Bertagni, Aufsichtsrat: Daniel Ammann |
| 7 | Profit/Loss | 2015: -88 Mio |
| 8 | Turnover | 2015: 11.739,1 Mio |
| 9 | Tax Paid | 2015: 96.2 Mio |
| 10 | Number of employees | 18.239 |

[62] Association of German Broadcast Organisations
[63] European Broadcast Organisation

| | | |
|---|---|---|
| 11 | Website | https://www.opel.de/ |
| 12 | Wikipedia URL | https://en.wikipedia.org/wiki/Opel |
| 13 | VAT | DE 111607872 |
| 14 | Certified e-mails | n/a |
| 15 | Other locations | n/a |
| 16 | RSS/Atom feeds | n/a |
| 17 | Web languages | n/a |
| 18 | Publicly traded? | n/a |

**Table 6: Attributes for "company profile" required for DJP**

2. Help journalists (and other users) to overcome ambiguity and complexity related to "company names" (and locations). Unless a user is an expert for the company in question, it is time consuming and painstaking to find out under which names a company is registered (and in which location in case of larger conglomerates). This (reliable/accurate) information is required to find more detailed information from multiple sources and conduct deeper research, if not to avoid confusion and mistakes. This is a key consideration for ordinary, time-short journalists who may only know a company by its common and/or brand name (e.g. "OPEL" in Germany – subject to a company news story in recent months). It is specialist knowledge that in Germany, OPEL is registered under "Adam Opel AG" and "Opel Group GmbH". When working on a story it may also be of interest that the related company in the UK is to be found under "Vauxhall". The naming of legal entities is also subject to change, e.g. it has just been announced in April that there will be a new company in Germany "Adam Opel GmbH".
3. Include detailed information about the data that can be found in other data repositories. Rather than providing a list of data providers to also visit for more information, euBusinessGraph should provide information on the type of additional data that has surfaced for the company of interest in an easy-to-read, explanatory format. The user should be able to link from the data to the respective provider as he/she is likely to be primarily interested in the data (the information), and not who provides it. However, a listing by provider is also useful (e.g. choice of view). Spending time on visiting other providers' websites (including potential registration/payment barriers), the user needs to know whether it is worth it (e.g. it should provide answers to questions such as "what will I get there?", "free or for-pay data?", "registration needed?" "oAuth supported?", "access through API or web?", and "which payment system is used?").
4. Provide easy and smart access to data from other providers. As it is not possible and realistic that all data providers' data is provided on a "fully open" basis, euBusinessGraph must facilitate access to other providers' data as easy, smart and transparent as possible. The options are:
    a. Sufficient information to support decision making on why to go there - see above
    b. Clear indication whether the data is open/closed (free/pay)
    c. Make all or some data open/free for users with a public interest background (see above)
    d. Integrated navigation to the data via underlying business/revenue sharing agreements
    e. Guided navigation to provider site (e.g. a welcome page for referrals with information)
5. Make euBusinessGraph attractive for other data providers to join. The more data providers take part, the more valuable euBusinessGraph becomes, and hence the DJP and other 3rd party applications that consume data and functions via the graph's API.

#### 2.2.1.1 "Data offering" Requirements

To elaborate on requirements #3 and #4 of the DJP business case, we interviewed data providers about business and company data that will not be part of euBusinessGraph, but could be offered through the graph as additional information available via distribution channels of the partners.

The table below contains information about:

- **Jurisdiction**: countries extra information is offer for.
- **Scope**: what extra information will not be freely available in euBusinessGraph, but will be offered through it and can be accessed via data partners' data distribution channels?
- **Coverage/Price variation**: does data coverage or price vary per characteristic, e.g., per region, type (public/private) of organisations or economic activity (better coverage of software companies than agricultural companies), etc.?

| Data Provider | Jurisdiction | Scope | Coverage/ price variation |
|---|---|---|---|
| SDATI | IT | Officers, shareholders, balance sheets (to be confirmed) | none |
| OCORP | BE | Officers, filings | none |
| | GB | Officers, filings, financial reports | |
| JSI | All | Event / Article content | none |
| BRC | NO | Officers (can be shared when we find an acceptable privacy solution), financial reports, bankruptcy filings, more details on special company types | none |

In the table below, data providers elaborated on data access requirements, namely:

- A link to the data provider's home page
- A link to a company page
- Indication of whether extra information is offered for free or not

| Data Provider | Link to provider home page | Link to company page | Free/Paid |
|---|---|---|---|
| SDATI | http://atoka.io | https://atoka.io/it/companies/spaziodati-srl/6da785b3adf2 Registration needed to access the full page | Paid |
| OCORP | https://opencorporates.com | https://opencorporates.com/companies/gb/07444723 No registration needed to access the page | Free |
| BRC | http://brreg.no | http://data.brreg.no/oppslag/enhetsregisteret/enheter.xhtml | Free |
| DW | http://test.screenertool.com/ | http://blogs.dw.com/innovation/ | Free |
| JSI | http://eventregistry.org/ | http://ailab.ijs.si/ | Paid |

### 2.2.2 CRM-S: EVRY CRM Service

EVRY's business case is to extend their current offerings with a CRM service and combine this with the existing InfoTorg[64]. The purpose of the new service will be to offer improved and enriched data quality and coverage of customers' CRMs with up-to-date business and company data sourced from euBusinessGraph.

Based on the needs of their customers, EVRY has identified the following requirements to euBusinessGraph:

1. CRM-S customers are large companies and SMEs in Scandinavia. This determines the focus of CRM-S on Scandinavian data, starting with Norwegian Business Registries.
2. CRM-S customers expect to have consistent services, for which it is important that data is stored locally, and synchronised with euBusinessGraph. EVRY has to store locally a significant subset of euBusinessGraph data, to feed it to the CRM-S recommendation engine that will indicate actual area of business, related business and events that indicate potential or risk. In the initial phase of CRM-S's development, this subset of the data can be determined by the same set of attributes as DJP.
3. If all data that is necessary for building CRM-S cannot be directly accessed through euBusinessGraph, a licensing scheme should be implemented in euBusinessGraph, or as an agreement between the partners.

### 2.2.3 Atoka+

Atoka+ is planned as extension of Atoka, SDATI's lead generation service that currently targets the Italian market. There are two main objectives of Atoka+:

1. Extend data coverage of Atoka to two more jurisdictions; UK and Norway.
2. Deepen and improve the data quality and its coverage in all the three covered countries; Italy, UK and Norway. This specifically concerns information related to business entities found in the news articles and social web, e.g., business events or contacts.

As a consumer of data from euBusinessGraph, Atoka+ introduces the following requirements:

1. Bulk access to UK and Norwegian data is required to initiate new jurisdictions in Atoka.
2. Atoka's main asset is vertical knowledge about companies. One such vertical is contextual information from the news. Currently, Atoka provides news articles associated with companies through mentions of their names or names of key managers in the news' content. euBusinessGraph should provide access to similar information about UK and Norwegian companies.
3. Business relevant events are another important vertical for Atoka+. euBusinessGraph should provide taxonomies of several types of events, considering the following:
    - Categorisation of news into different event types, e.g., whether a news article is about a merger, acquisition or a product launch.
    - Events registered on a corporate website. Examples of these are "a company has a new website", "a company's website has changed its e-commerce technology", and "a new link to corporate social accounts has been created".
    - Events coming from authoritative data sources. Examples of these are "a company has been awarded a new tender", "a company has changed its trading address", and "the information about the shares of a company has changed".
4. All information coming from euBusinessGraph should indicate trustworthiness or fuzziness. Atoka has different business cases that require various degrees of trust. For example, mergers or acquisitions are reflected both in the news and in company registries, but in the news, it usually happens earlier. While information from the news might be less reliable than official information from company registries, some Atoka users prefer it as soon as possible. At

[64] https://www.infotorg.no/

the same time, the credibility of information is important for these users. Hence, they require traceability of the information to its sources.

### 2.2.4 TDS: Tender Discovery Service

Tender Discovery Service (TDS) is the business case of CERVED that aims to create a platform to access data about public contracts enhanced with information from other sources, including business information coming from euBusinessGraph. To enable TDS functionalities, the following company data attributes are required:

- Jurisdiction, country
- Branch locations
- Number of employees
- Linked data keywords
- Company type
- Revenue
- Web languages
- Active/inactive status

## 2.3 Company Data Model

We created an initial company data model considering related works, data available from the partners, and the needs of their business cases. The model covers the following requirements:

- Capture the concept of a company and represent different types of companies.
- Represent company jurisdictions and registration information.
- Capture company contact information, such as the address and other locations.
- Capture social data of companies, such as their websites (together with Web languages), RSS/Atom feeds and Wikipedia URLs.
- Answer if a company is publicly traded or not, if it is state owned or not, and if it is registered in a startup register.
- Support languages: EN, IT, NO.

Figure 4 illustrates the instance diagram of the initial model. All classes and properties are described further in this section.

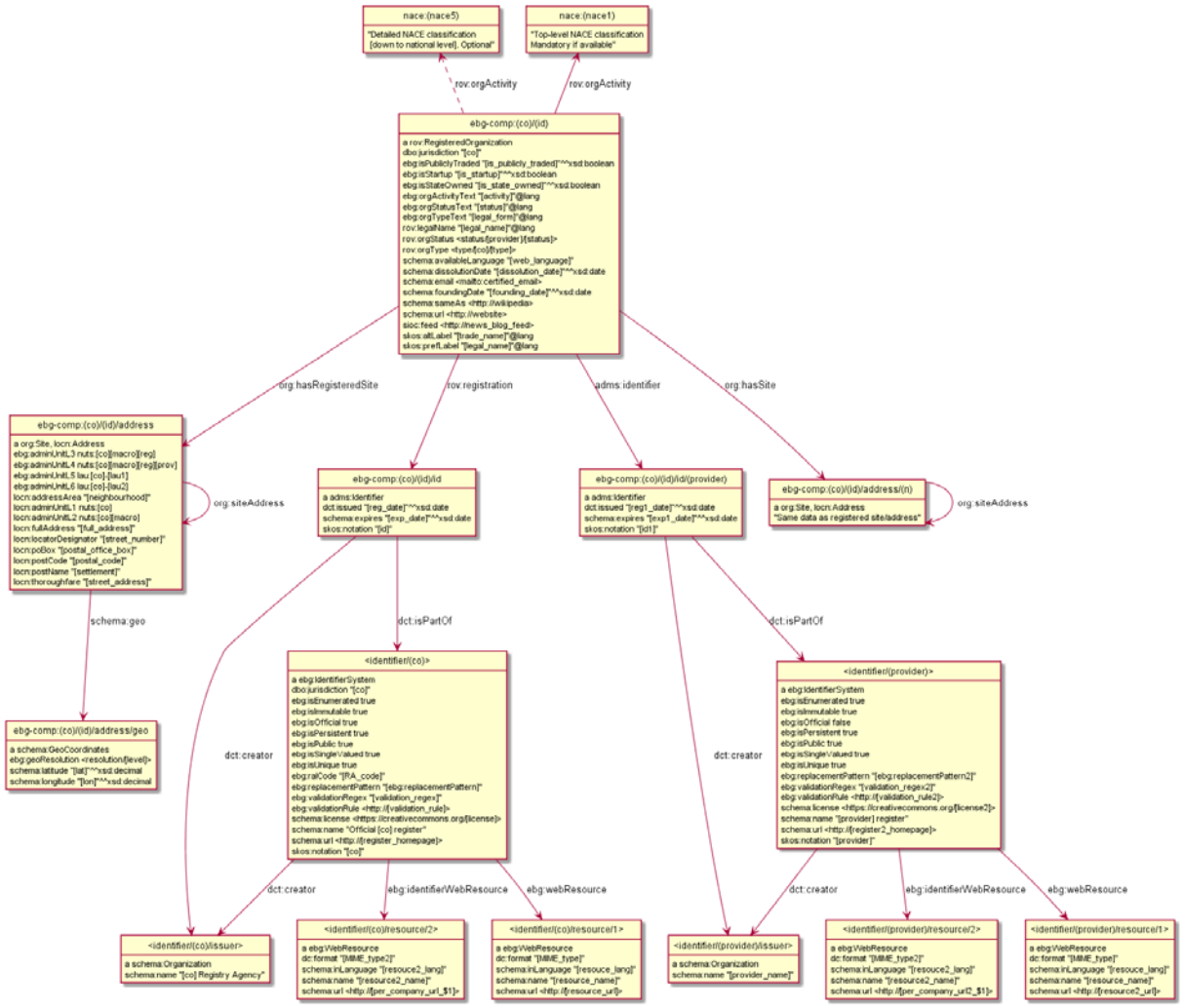



**Figure 4: euBusinessGraph Company Shared Data Model**

Abbreviations used in the diagram:

- (co): country (jurisdiction)
- (macro): macroregion
- (reg): region
- (prov): province
- (lau1): local administrative unit 1
- (lau2): local administrative unit 2
- (prov_code): province code
- (RA_code): Registration Authority List code

### 2.3.1 The Model, Topic by Topic

In this section, we describe the data model divided into several topics:

1. Identifier systems and their web resources (Section 2.3.1.1)
2. Identifiers (Section 2.3.1.2)
3. Company (Section 2.3.1.3)
4. Names (Section 2.3.1.4)

5. Classifications (Section 2.3.1.5)
6. Online presence (Section 2.3.1.6)
7. Other company details (Section 2.3.1.7)
8. Physical presence (Section 2.3.1.8)

For every term, we specify whether it is a Class or a Property (Object, Lookup or Data Property), and include the following information:

| <Term>: <Term Name> | |
|---|---|
| **Description** | A term description, preferably from an established source |
| **Scope Note** | How the term is used, and other clarifications |
| **Examples** | Data rules that providers must conform to |
| **Rules** | A suggested RDF binding |

Additionally, for properties we may specify:

| | |
|---|---|
| **Domain** | Instances of which classes can have this property |
| **Range/ Data Type/ Values** | What kind of value the property should take.<br>For a data property, the default data type is xsd:string and can be omitted. For lookup properties, we give a list of values, or guidance on what to choose from. |
| **Cardinality** | How many values the property can have:<br>• 0 to many<br>• 1 to many<br>• 1 only<br>• 0 or 1 |

### 2.3.1.1 Identifier Systems

A company can have several identifiers issued by different registers and for different purposes. This is modelled by having each company identifier belong to an identifier system. In this way, we can differentiate between “official registration” in official business registers and “alternative registrations” in other kinds of registers. While they have the same nature, only the former can be used to uniquely identify a company in each jurisdiction, and to confirm existence of the company as a legal entity in this jurisdiction. Other registrations may not be unique or persistent. Social accounts are even orthogonal to the existence of the company register.

| Class: Identifier System | |
|---|---|
| **Description** | A system managed by a publisher (e.g., a register or agency) that is used to issue identifiers to companies. |
| **Scope note** | Many registers keep several identifier systems, e.g. http://www.registryagency.bg keeps the systems used by:<br>• The Trade Register http://brra.bg for companies,<br>• The Bulstat Register http://www.bulstat.bg/ for other entities |
| **RDF** | ebg:IdentifierSystem |
| **Data Property: identifier system jurisdiction** | |
| **Description** | Jurisdiction to which the identifier system applies. |
| **Examples** | Consider the following cases:<br>• Website, Twitter, Facebook don't apply to any particular jurisdiction in that they don't have any official status<br>• UK company number applies to jurisdiction "GB"<br>• GLEI applies to the "INT" (world-wide) jurisdiction |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |

| | |
|---|---|
| **RDF** | dbo:jurisdiction |
| **Cardinality** | 0 to many |
| **Data Property: identifier system code** | |
| **Description** | Short mnemonic code for the identifier system, used in its URL. Also used in identifier URLs that are part of the system. |
| **Scope Note** | Issued locally by EBG.<br>For identifier systems published by the sole or preferred official register in a jurisdiction, we use the jurisdiction code (e.g. "BG", "GB"). For others, if the identifier system has no explicit name, we use a short mnemonic code of the publisher: upper-case for company registers (e.g. "OCORP", "SDATI", "RAL", "EU", "BRIS"), mixed-case for social network registers (e.g. "Twitter", "Facebook"). |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |
| **RDF** | skos:notation |
| **Cardinality** | 1 |
| **Data Property: identifier system RAL code** | |
| **Description** | GLEI RAL code for the identifier system. |
| **Scope Note** | Used only for official registers. |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |
| **RDF** | ebg:ralCode |
| **Cardinality** | 0 or 1 |
| **Data Property: identifier system name** | |
| **Description** | The name of the identifier system |
| **Domain** | Identifier System |
| **Data Type** | xsd:string or rdf:langString |
| **RDF** | skos:prefLabel |
| **Cardinality** | 0 or 1 |
| **Data Property: identifier system website** | |
| **Description** | Various websites of the identifier system and/or its associated issuer and register, e.g. home page, search, download. |
| **Scope Note** | To be more specific on what the URL returns, Web Resource should be used (see Section 2.3.1.1.4). Not for per-company or validation rule URLs (see Section 2.3.1.1.3). |
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | schema:url |
| **Cardinality** | 0 to many |
| **Data Property: identifier system license** | |
| **Description** | License that applies to the system |
| **Scope Note** | If possible, use established license URLs, e.g. from https://creativecommons.org/ or http://rightsstatements.org/ |
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | schema:license |
| **Cardinality** | 0 to many |

#### 2.3.1.1.1 Identifier System Agents

In this section we describe 3 agents related to a system:

- Author: in charge of specifying the rules and organization of the system.
- Issuer: issues identifiers and then keeps them in a database (register).
- Publisher: publishes the identifier database (register) in some form

We describe them in the same way:

- URL: use an original web URL if available, else an EBG URL
- Class: schema:Person or schema:Organization
- Property Name: schema:name: xsd:string or rdf:langString

| **Data Property: Identifier System Author** | |
|---|---|
| **Description** | Agent who specified the rules and organisation of the system |
| **Examples** | • W3C specified web URL rules<br>• GLEIF specified GLEI<br>• Dun and Bradstreet specified DUNS numbers |
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | schema:author |
| **Cardinality** | 0 or 1 |
| **Data Property: Identifier System Issuer** | |
| **Description** | Agent that issues identifiers and then keeps them in a database (register) |
| **Scope Note** | Many agencies keep several registries. E.g., http://www.registryagency.bg keeps:<br>• Trade Register http://brra.bg for companies<br>• Bulstat Register http://www.bulstat.bg/ for other entities |
| **Examples** | • Web URLs are not issued by any central agency and there is no register to consult.<br>• Companies House is the issuer for the official UK identifier system.<br>• Dun and Bradstreet issues DUNS numbers.<br>• GLEI Foundation (GLEIF) issues identifiers for the world-wide GLEI register (through its Local Operating Units, LOU). |
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | dct:creator |
| **Cardinality** | 0 or 1 |
| **Data Property: Identifier System Publisher** | |
| **Description** | Agent that publishes the identifier database (register) in some form (i.e., printed form, online) |
| **Scope Note** | Different functions and distributions can be considered (e.g. search or other query, per-company web resources, full download). Often but not always the issuer is also publisher. |
| **Examples** | • GLEIF publishes the GLEI openly, with a full dump that is updated regularly, which has spurned re-publishers such as http://openleis.com/ and http://glei.info.<br>• Companies House and BRC publish the UK and NO registers openly.<br>• DUNS numbers are not openly published. |

| | |
|---|---|
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | schema:publisher |
| **Cardinality** | 0 to many |

#### 2.3.1.1.2 Identifier System Characteristics

Identifier systems have some Boolean characteristics (flags) that represent expectations about their identifiers. Some systems have exceptions, i.e. identifiers that don't satisfy the expectations. Each flag is set to "true" in the desirable (positive) case. We strive to provide all flags for each system, but in some cases the flag could be omitted (e.g., if there is not enough information).

| **Data Property: has unique identifiers** | |
|---|---|
| **Description** | Whether each identifier in the system relates to only one entity |
| **Scope Note** | Similar to:<br>• owl:InverseFunctionalProperty[65]<br>• Wikidata Distinct Values constraint[66] |
| **Examples** | • Company IDs are unique in most national registers<br>• Stock exchange tickers are unique to each exchange (but a company can buy the ticker of another, i.e. the identity changes over time)<br>• Person names are not unique since they are highly ambiguous<br>• Websites are not unique since a website can be shared by several related companies |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isUnique |
| **Cardinality** | 0 or 1 |
| **Data Property: is single-valued** | |
| **Description** | Whether each entity has only one identifier in the system |
| **Scope Note** | Similar to:<br>• owl:FunctionalProperty[67]<br>• Wikidata Single Value constraint[68] |
| **Examples** | • Company IDs are unique in most national registers<br>• Stock exchange tickers are unique to each exchange (but a company can buy the ticker of another, i.e. the identity changes over time)<br>• Person names are not unique since they are highly ambiguous<br>• Websites are not unique since a website can be shared by several related companies |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isSingleValued |
| **Cardinality** | 0 or 1 |
| **Data Property: has persistent identifiers** | |
| **Description** | whether identifiers can be removed from the register |
| **Examples** | Some registers “remove” identifiers when the object they relate to is no longer |

[65] owl:InverseFunctionalProperty as seen in https://www.w3.org/TR/owl-ref/#InverseFunctionalProperty-def
[66] Distinct values constraint: https://www.wikidata.org/wiki/Q21502410
[67] owl:FunctionalProperty as seen in https://www.w3.org/TR/owl-ref/#FunctionalProperty-def
[68] Single value constraint: https://www.wikidata.org/wiki/Q19474404

| | |
|---|---|
| | active(see “Persistent Identifiers” in Section 3.1.2.8). |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isPersistent |
| **Cardinality** | 0 or 1 |
| **Data Property: has immutable identifiers** | |
| **Description** | whether identifiers can change |
| **Examples** | Some registers encode additional information within the identifier (e.g., some identifiers encode the economic sector of the company). In these cases the identifier changes when the extra information that is encoded changes (see “Dumb vs. Intelligent Identifiers” in Section 3.1.2.5). |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isImmutable |
| **Cardinality** | 0 or 1 |
| **Data Property: is public** | |
| **Description** | whether identifiers from the system are available for public use: consulting, search or download. |
| **Scope Note** | Most public systems (but not all) will have some Publishers, and preferably some web resources. |
| **Examples** | • Italian (IT) company identifiers are public. Even though they are not published openly, they are available for a fee, and then can be used freely.<br>• DUNS identifiers are not public, even though there are limited lookup services (e.g., https://www.dnb.com/duns-number/lookup.html). Dun and Bradstreet does not allow their use en-masse. |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isPublic |
| **Cardinality** | 0 or 1 |
| **Data Property: has enumerated identifiers** | |
| **Description** | whether the system has an Issuer and issued identifiers are kept in a database (register) |
| **Examples** | • Every official register is enumerated<br>• Websites are not enumerated |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isEnumerated |
| **Cardinality** | 0 or 1 |
| **Data Property: is official in jurisdiction** | |
| **Description** | Whether the system is considered the official one in all jurisdictions in which it applies. |
| **Examples** | • False for Website, Twitter, Facebook in that they don't have any official status<br>• True for UK company house, the official register for the “GB” |

| | |
|---|---|
| | jurisdiction<br>• False for GLEI, since it is not the official register for any of the world's jurisdictions<br>• False for SDATI identifiers since the SDATI system is not official in Italy. |
| **Domain** | Identifier System |
| **Data Type** | xsd:boolean |
| **RDF** | ebg:isOfficial |
| **Cardinality** | 0 or 1 |

#### 2.3.1.1.3 Identifier Validation Properties

Systems are associated with some properties that can be useful for identifier validation.

| **Data Property: Validation Rule** | |
|---|---|
| **Description** | URL providing human or machine-readable rule(s) for validating identifiers in the system |
| **Scope Note** | Can be in the form of webpage, PDF document, RDF shape, etc. Multiple values about the same validation rule are ok. |
| **Examples** | • http://bsv-bg.com/контролни-цифри-ползвани-в-българия/ describes the BG EIK checksum algorithm in HTML<br>• http://www.nsi.bg/sites/default/files/konkursi/RGP_OPAK_2014_Annex_6.pdf describes the BG EIK checksum algorithm in PDF. It is by an official source, but is less convenient |
| **Domain** | Identifier System |
| **Data Type** | xsd:anyURI |
| **RDF** | ebg:validationRule |
| **Cardinality** | 0 to many |
| **Data Property: Validation Regex** | |
| **Description** | Regular expression for validating identifier values within the system |
| **Examples** | • "([A-Z]{2})/(\d+)" is a regex for the EU Value Added Tax (VIES) register, which consists of a member state code followed by slash and a national numeric identifier<br>• "\d{9}" is a simple regex for validating DUNS numbers<br>• "(\d{2})-?(\d{3})-?(\d{4})" is a regex for validating DUNS numbers that may include optional dashes in the indicated positions (e.g. "36-032-1459") |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |
| **RDF** | ebg:validationRegex |
| **Cardinality** | 0 or 1 |
| **Data Property: Replacement Pattern** | |
| **Description** | Pattern to use together with the Validation Regex to normalise identifier values by removing optional decorations |
| **Examples** | "$1$2$3" can be used together with the validationRegex "(\d{2})-?(\d{3})-?(\d{4})" to extract the pure digits from a DUNS number spelled with optional dashes (e.g. "36-032-1459" -> "360321459") |
| **Domain** | Identifier System |

| | |
|---|---|
| **Data Type** | xsd:string |
| **RDF** | ebg:replacementPattern |
| **Cardinality** | 0 or 1 |

#### 2.3.1.1.4 Identifier System Web resources

A "web resource" is a URL complemented with a MIME type to specify what the URL is about. These web resources are used for Identifier Systems (e.g. to provide the search or download URL) and per-company, as a URL template in which to substitute the identifier value.

There can be several MIME types because some URLs return various resource types using Content Negotiation.

| **Data Property: Identifier System Web Resource** | |
|---|---|
| **Description** | Web resource(s) associated with an identifier system. |
| **Examples** | For BG TR:<br>• https://brra.bg is the homepage.<br>• http://opendata.government.bg/dataset/tbprobckn-pernctbp is the open data download endpoint.<br>For BE Banque-Carrefour des Entreprises, https://kbopub.economie.fgov.be/kbopub/zoeknummerform.html is the search URL. |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |
| **RDF** | ebg:webResource |
| **Cardinality** | 0 to many |
| **Data Property: Identifier Web Resource** | |
| **Description** | Web resource(s) associated with an identifier system that specifies templates to build URLs for all identifiers in the system. |
| **Examples** | For OCORP, the templates:<br>• "https://opencorporates.com/companies/{}",<br>• "https://opencorporates.com/companies/{}.xml",<br>• "https://opencorporates.com/companies/{}.json",<br>• https://opencorporates.com/companies/{}.rdf<br>and the identifier value return company information in "text/html", "application/xml", "application/json" and "application/rdf+xml" respectively. |
| **Domain** | Identifier System |
| **Data Type** | xsd:string |
| **RDF** | ebg:identifierWebResource |
| **Cardinality** | 0 to many |

### 2.3.1.2 Identifiers

Companies are registered using various kinds of identifiers. Some of these identifiers are kept in official registers, others are self-issued and not centralised (e.g. website). Identifiers include but are not limited to:

- the official registration in a trade register. This registration should correspond to the company's jurisdiction and when known, we use it in the EBG company URL.

- Other official government registers (e.g., bank license, insurance company license, register of startup companies, etc.)
- Official international registries (e.g., GLEI)
- EBG data provider or other databases (e.g., OCORP, SDATI, Dun & Bradstreet)
- Various company websites
- Social networks, eg Facebook, Twitter, etc.

| **Class: Identifier** | |
|---|---|
| **Description** | Identifier of a company according to some identifier system. |
| **RDF** | adms:Identifier |
| **Data Property: identifier value** | |
| **Description** | Literal value of the identifier. |
| **Scope Note** | adms:Identifier recommends that this value should be "datatyped with the identifier scheme (including the version number if appropriate)". However, we decided not to do this because:<br>● One cannot search by exact literal value unless one also knows the datatype URL<br>● Currently we do not plan to maintain different identifier system versions |
| **Rules** | Should not include leading, trailing and consecutive spaces, to facilitate exact match. |
| **Domain** | Identifier |
| **Range** | rdfs:Literal |
| **RDF** | skos:notation |
| **Cardinality** | 1 |
| **Object Property: identifier is part of system** | |
| **Description** | The Identifier System this identifier is part of |
| **Domain** | Identifier |
| **Range** | Identifier System |
| **RDF** | dct:isPartOf |
| **Cardinality** | 1 |
| **Object Property: identifier issuer** | |
| **Description** | Agency that issued the identifier, and register that holds it |
| **Scope Note** | In many countries, there is a single registry although in others, such as Spain and Germany, multiple registries exist. [rov:] |
| **Domain** | Identifier |
| **Range** | schema:Person or schema:Organization |
| **RDF** | dct:creator |
| **Cardinality** | 0 or 1 |

| Data Property: identifier has start | |
|---|---|
| **Description** | When was the identifier issued |
| **Scope Note** | This may or may not be the same as the incorporation date of the company. Even for the official registration, the two dates may differ, depending on business rules. |
| **Domain** | Identifier |
| **Data Type** | xsd:date |
| **RDF** | dct:issued |
| **Cardinality** | 0 or 1 |
| **Data Property: identifier has end** | |
| **Description** | Date when the identifier expires |
| **Scope Note** | This may or may not be the same as the dissolution date of the company. Even when it is about the official registration, the two dates may differ, depending on business rules. |
| **Domain** | Identifier |
| **Data Type** | xsd:date |
| **RDF** | schema:expires |
| **Cardinality** | 0 or 1 |

#### 2.3.1.3 Company

| Class: Company | |
|---|---|
| **Description** | An organization that gains legal entity status by the act of registration. Compare to org:FormalOrganization that applies to any legal entity, including those created by other legal means. Registered organizations are distinct from the broader concept of organizations, groups or, in some jurisdictions, sole traders. Many organizations exist that are not legal entities, yet to the outside world they have staff, hierarchies, locations etc. Other organizations exist that are an umbrella for several legal entities (universities are often good examples of this) [rov:]<br>A legal person or structure that is organized under the laws of any jurisdiction [lei-elf:] |
| **Scope note** | Registered organizations are the main entities that euBusinessGraph works with (the project is not concerned with unregistered informal groups). The borderline between public organizations and commercial companies is not always clear-cut (a company may be partially owned by government, and e.g. Public-Private-Partnerships are a prime example), so the project may deal with some organizations that are not companies. But for brevity and convenience we often call them Companies. [ebg:] |
| **Examples** | • All financial intermediaries;<br>• Banks and finance companies;<br>• All entities that issue equity, debt or other securities for other capital structures;<br>• All entities listed on an exchange;<br>• All entities that trade stock or debt, investment vehicles, including mutual funds, pension funds and alternative investment vehicles constituted as corporate entities or collective investment agreements (including umbrella funds as well as funds under an umbrella structure, hedge funds, private equity funds, etc.);<br>• All entities under the purview of a financial regulator and their affiliates, subsidiaries and holding companies;<br>• Counterparties to financial transactions. |
| **RDF** | rov:RegisteredOrganization |

| Data Property: jurisdiction | |
|---|---|
| **Description** | Jurisdiction in which the company is registered. For Europe, this is a country, even for Germany that has 166 separate registers (see [lei-ral:]).<br>Many jurisdictions register foreign companies that are incorporated and fully owned in another country. euBusinessGraph collects all registrations from each register and doesn't attempt to match foreign companies across jurisdictions. |
| **Scope Note** | If we know the official legal ID of the company, the jurisdiction code and legal ID are used to form the company URI |
| **Examples** | "IT", "FR"<br>Values are ISO 3166 country codes |
| **RDF** | dbo:jurisdiction |
| **Data Type** | xsd:string |
| **Cardinality** | 1 |
| **Object Property: registration** | |
| **Description** | An identifier of a company. This includes, but is not limited to:<br>● **Direct**: Identifier in some aggregated dataset, where the official ID is part of the aggregation ID.<br>● **Indirect**: Identifier in some aggregated dataset or for a website, where the aggregation ID bears no resemblance to the official ID<br>● **Multiple**: Identifier that has a many-to-many relation with the company, i.e., a company may have several identifiers of the same kind, and the same identifier can be used by several companies. |
| **Examples** | **Direct**: EU VAT number (in those jurisdictions where it is used as the primary company identifier), BRIS ID, OCORP ID.<br>**Indirect**:<br>● GLEI LEI, SDATI Atoka ID<br>● BG GUID 617f4edf8c154f4296efdf146513de21, which corresponds to official id 204060254 and can be used to reach the official register page for that company[69]<br>**Multiple**: EU VAT number in jurisdictions such as the UK (where they map n:n with companies), bank license, insurance license, license to trade dual-use goods (arms), Twitter account, Facebook account. |
| **RDF** | adms:identifier |
| **Domain** | Company |
| **Range** | Identifier |
| **Cardinality** | 0 to many |
| **Object Property: official registration** | |
| **Description** | Identifier that holds the official company registration in its jurisdiction of registration. It establishes the legal existence of the company. |
| **Scope note** | This property holds redundant information that is captured through the *registration* property. This is done for compatibility with the Registered Organization vocabulary.<br>In some cases, we may not have info about the official registration, though *"it is questionable whether a description of a registered organization without this property and an associated Identifier class will be of any value"* [rov:] |
| **Examples** | Companies House in the United Kingdom<br>Registro Imprese in Italy |
| **RDF** | rov:registration |

[69] https://public.brra.bg/CheckUps/Verifications/ActiveCondition.ra?guid=617f4edf8c154f4296efdf146513de21

| | |
|---|---|
| **Domain** | Company |
| **Range** | Identifier |
| **Cardinality** | 0 or 1 |

### 2.3.1.4 Names

euBusinessGraph adopted two kinds of names used by [lei-cdf:]:

- LegalName: the legal name of the entity. If an entity in a jurisdiction with more than one legal name (e.g., in different languages) this property indicates the primary legal name.
- TRADING_OR_OPERATING_NAME: A "trading as", "brand name" or "operating under" name currently used by this entity in addition to, but not replacing, the (primary) legal, official registered name.

Company names in different languages are captured using other two good practices of lei-cdf:NameType:

- Each company name will include an optional language tag (xml:lang)
- The string is declared Tokenized500Type, which does not allow leading, trailing and two consecutive spaces

| **Data Property: legal name** | |
|---|---|
| **Description** | The legal name of the business, i.e. official name of the company [ebg:] |
| **Scope Note** | A business may have more than one legal name, particularly in jurisdictions with more than one official language (e.g. Belgium). Some registries also treat a transliterated name as official (e.g. Онтотекст vs. Ontotext) [ebg:] |
| **Examples** | "PROGIENE 2.000 I PROFESSIONISTI DELL'IGIENE S.R.L. ENUNCIABILE ANCHE: PROGIENE 2.000 S.R.L."@it<br>"CHRINON LTD" (no lang tag)<br>"Онтотекст"@bg<br>"Ontotext"@en |
| **Rules** | Provide a legal language tag (see [iana:] for a list) if possible. Don't allow leading, trailing and two consecutive spaces |
| **RDF** | rov:legalName<br>Similar to: schema:legalName |
| **Domain** | Company |
| **Data Type** | xsd:string or rdf:langString |
| **Cardinality** | 1 to many |
| **Data Property: trading name** | |
| **Description** | Informal/popular name of the company (also called Trading As) [ebg:] |
| **Scope Note** | Sometimes (e.g., when the legal name is very long) people refer to a company using a different, informal version [ebg:]. Some jurisdictions recognize concepts such as a trading name or alternative forms of a legal entity's name [rov:] |
| **Examples** | "PROGIENE 2.000 I PROFESSIONISTI DELL'IGIENE SRL"@it<br>"OpenCorporates"@en<br>cf. to the first two examples of Legal Name |
| **Rules** | Do not emit such name if it is equal to the legal name [ebg:].<br>Do not use this property to record translations of the primary legal name [rov:].<br>Provide a valid language tag if possible (see [iana:] for a list).<br>Don't allow leading, trailing and two consecutive spaces. |
| **RDF** | skos:altLabel<br>Similar to: schema:alternateName |
| **Domain** | Company |

| | |
|---|---|
| **Data Type** | rdf:langString or xsd:string |
| **Cardinality** | 0 to many |
| **Data Property: preferred name** | |
| **Description** | A single preferred name of a company or register (preferably English). |
| **Scope Note** | Used as a display name of the company or register. For companies, we usually pick one of the Registered Names (English if present, else at random). For registers, we use a commonly used English name. |
| **RDF** | skos:prefLabel |
| **Domain** | Company |
| **Data Type** | rdf:langString or xsd:string |
| **Cardinality** | 1 |

### 2.3.1.5 Classifications

There are three classifications in the first version of the Model: Company Type/ Legal Forms, Company Status, Economic Activity. We modelled them as lookup properties. Values are taken from SKOS concept schemes if available. Alternatively, a free text field is used.

| **Lookup Property: Type** | |
|---|---|
| **Description** | Company Type (Legal Form of the entity) |
| **Scope Note** | Each jurisdiction will have a limited set of recognized company types and these should be expressed in a consistent manner in a SKOS Concept Scheme [rov:] The types may form a hierarchy, but each company can have maximum one value. E.g. for Italy: Società Di Capitale > Società a responsabilità limitata [ebg:] |
| **RDF** | rov:orgType<br>Similar to: lei:EntityLegalFormCode |
| **Domain** | Company |
| **Values** | skos:Concept<br><br>Values are taken from code lists of data providers. Code lists of euBusinessGraph data providers were collected and published in the project repository[70].<br>Alternatively, a free text "Type Text" is used. |
| **Cardinality** | 0 or 1 |
| **Data Property: type text** | |
| **Description** | Company Type (Legal Form of the entity) given in the form of free text. |
| **Scope Note** | Because of the difficulties of standardizing Type lookup, we include a free text field. |
| **Examples** | "Private Limited Company"@en<br>"Дружество с ограничена отговорност"@bg<br>"Società a responsabilità limitata"@it |
| **Rules** | Include valid language tag (see [iana:] for list)<br>If Type is also present, it must express the same value as Type Text. |
| **RDF** | ebg:orgTypeText<br>Same as: lei:OtherLegalForm |
| **Domain** | Company |
| **Data Type** | rdf:langString |
| **Cardinality** | 0 or 1 |

[70] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-company-type.xlsx

| Lookup Property: Status | |
|---|---|
| **Description** | Flag that identifies whether a company is active or not [ebg:]. The operational and/or legal registration status of the entity [lei:]. |
| **Scope Note** | There is no globally accepted list of company states. For Inactive, some providers look at hard evidence (the company has been deregistered), others at dissolution date in the past, or an extended period of inactivity (dormant). Because of this, a user cannot assume that Active and Inactive are opposites. |
| **Rules** | A Best Practice for recording status levels is to use the relevant jurisdiction's terms and to encode these in a SKOS Concept Scheme. [rov:]. |
| **Examples** | • "insolvent", "bankrupt", "in receivership": likely to mean slightly different things with different legal implications in different jurisdictions [rov:]<br>• "normal activity": does appear to have cross-border usefulness and this should be used in preference to terms like "trading" or "operating" [rov:]<br>• "actively trading" vs. "dormant" vs. "closed" could be considered the 3 main divisions of Status |
| **RDF** | rov:orgStatus<br>Similar to: lei:EntityStatus |
| **Domain** | Company |
| **Values** | euBusinessGraph company status concept scheme[71] |
| **Cardinality** | 0 or 1 |
| **Data Property: status text** | |
| **Description** | Company status as it comes from the original register. |
| **Examples** | • "dissolved"@en, "inactive"@en, "revoked"@en<br>• "situation normale"@fr, "en liquidación"@es, "πτώχευση"@el |
| **Rules** | • Include valid language tag (see [iana:] for list)<br>• If both Status and Status Text are present, they must express compatible values (Status Text will be a finer-granularity value) |
| **RDF** | ebg:orgStatusText |
| **Domain** | Company |
| **Data Type** | rdf:langString |
| **Cardinality** | 0 or 1 |
| **Lookup Property: Economic Activity** | |
| **Description** | Economic activity of the organization (NACE code) |
| **Scope Note** | Economic activity is recorded using a controlled vocabulary: EC NACE 2. More detailed national classifications could also be useful (e.g. IT ATECO, UK SIC, BG NKID), but are not supported for now. |
| **Rules** | Each data provider must map codes that are used to established URLs as described below.<br>• Only the top-level classification or a detailed classification may be provided since we will establish a skos:broader hierarchy.<br>• There is no need for separate fields for national classifications. |
| **RDF** | rov:orgActivity<br>Similar to: schema:isicV4, schema:naics |
| **Domain** | Company |

[71] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-company-status.xlsx

| | |
|---|---|
| **Values** | Values are taken from euBusinessGraph NACE RDF concept scheme[72].Template: nace:([A-Z]\|[0-9]{2}\|[0-9]{2}\.[0-9]{1,2}).<br>Rules:<br>● Providers must use exact codes of the specified form (dot separators, no extra spaces).<br>○ Or we could use a more uniform code (e.g. **[A-Z]\d{2,5}**) to make submission easier, but we will need to do some segmenting on output<br>● Providers should drop eventual NACE National Extensions trailing digits (see Section 2.1.1.11.2.3)<br>● If providers find difficulties mapping to NACE, or need national levels, we should discuss this. |
| **Cardinality** | * |
| **Data Property: economic activity text** | |
| **Description** | Economic activity of the organization (free text) |
| **Scope Note** | Such field is widely used in some countries (e.g. BG) and may include much richer info, e.g.:<br>● NACE = 72.19 Scientific research and development in the field of natural, medical, agricultural and technical sciences, without biotechnology<br>● Free text = Development, exploration and production of opto-electronic systems for analogue and digital holographic recording, interferometric measurement of mechanical characteristics and biological objects at micro and macro level, refractometers, light-sensitive materials for holographic recording, blueprints of museum exhibits, artworks and collection surveys, production of holographic illustrations for students, albums, advertising, souvenirs and others, design and manufacture of holographic optical elements with a wide spectrum of applications, marketing and sale of holographic products, consulting, training and promotion of holographic methods and technologies, mediation, brokerage, and any other activity not prohibited by Bulgarian legislation. |
| **RDF** | ebg:orgActivityText |
| **Domain** | Company |
| **Data Type** | rdf:langString |
| **Cardinality** | 0 or 1 |

### 2.3.1.6 Online presence

We represent commonly used electronic resources and channels (website, Wikipedia, email, news feed) as specific properties of companies in this section. We represent social network accounts of the company (e.g. Facebook, Twitter) as Identifiers. Even though these are not official business identifiers, they fulfil a similar role.

| | |
|---|---|
| **Data Property: certified email address** | |
| **Description** | Email that is officially registered and with the same validity as certified (snail) mail[73]. |
| **Scope Note** | euBusinessGraph does not record other company emails, so there is no need to distinguish between different types. |
| **Examples** | mailto:elettorale@pec.comune.trento.it |

[72] https://github.com/euBusinessGraph/eubg-data/blob/master/data/nace.ttl
[73] https://en.wikipedia.org/wiki/Certified_email

| | |
|---|---|
| **RDF** | schema:email |
| **Domain** | Company |
| **Data Type** | rdfs:Literal |
| **Cardinality** | 0 to many |
| **Object Property: Wikipedia page** | |
| **Description** | Wikipedia page pertaining to the company |
| **Examples** | https://it.wikipedia.org/wiki/Trento |
| **RDF** | schema:sameAs |
| **Domain** | Company |
| **Range** | xsd:anyURI |
| **Cardinality** | 0 to many |
| **Object Property: website** | |
| **Description** | Website pertaining to the company |
| **Examples** | http://www.comune.trento.it |
| **RDF** | schema:url |
| **Domain** | Company |
| **Range** | xsd:anyURI |
| **Cardinality** | 0 to many |
| **Object Property: news/blog feed** | |
| **Description** | URL of RSS/Atom feed pertaining to the company |
| **Examples** | http://www.comune.ancona.gov.it/ankonline/anconaentrate/feed/ |
| **RDF** | sioc:feed |
| **Domain** | Company |
| **Range** | xsd:anyURI |
| **Cardinality** | 0 to many |

### 2.3.1.7 Other company details

| | |
|---|---|
| **Data Property: web languages** | |
| **Description** | Languages used in web resources related to the organisation (e.g., corporate websites, twitter accounts)<br>Languages used in web resources related to the organisation (e.g., corporate websites, social network accounts, etc.). |
| **Scope Note** | Currently this data is provided by SDATI, for matching company entities only. |
| **Examples** | Values from ISO 639-1 code list, optionally followed by the country code (as for zh-CN). |
| **RDF** | schema:availableLanguages |
| **Domain** | Company |
| **Data Type** | xsd:string |
| **Cardinality** | 0 to many |
| **Data Property: incorporation date** | |
| **Description** | Date legal entity was created. |
| **Examples** | “2010-11-18” |
| **RDF** | schema:foundingDate |
| **Domain** | Company |
| **Data Type** | xsd:date |
| **Cardinality** | 1 |

| Data Property: dissolution date | |
|---|---|
| **Description** | Date entity was dissolved or removed from register. |
| **Examples** | “2014-03-22” |
| **RDF** | schema:dissolutionDate |
| **Domain** | Company |
| **Data Type** | xsd:date |
| **Cardinality** | 1 |
| **Data Property: is startup** | |
| **Description** | Whether the company is a startup [ebg:] |
| **Examples** | Pitch, sectors and business model of the startup and innovative SMEs registered in the special section of the Business Register http://startup.registroimprese.it/ |
| **RDF** | ebg:isStartup |
| **Domain** | Company |
| **Data Type** | xsd:Boolean |
| **Cardinality** | 0 or 1 |
| **Data Property: is publicly traded** | |
| **Description** | Whether the company is publicly traded (listed at a stock exchange) [ebg:] |
| **RDF** | ebg:isPubliclyTraded |
| **Domain** | Company |
| **Data Type** | xsd:boolean |
| **Cardinality** | 0 or 1 |
| **Data Property: is state owned** | |
| **Description** | Whether this organisation is owned by the government, a government agency, community, city or other public entity. |
| **Scope Note** | In many cases it is not possible to compute this attribute without access to the shareholder register, so it may be missing. |
| **Examples** | • “false” for SpazioDati (a private company)<br>• “true” for Autonomous Province of Trento (local government)<br>• “true” for Statkraft (Norwegian state-owned company) |
| **RDF** | ebg:isStateOwned |
| **Domain** | Company |
| **Data Type** | xsd:boolean |
| **Cardinality** | 0 or 1 |

### 2.3.1.8 Physical presence

Physical Presence of companies is defined via addresses (see class “Address” below). We model addresses in a structured way using a set of attributes such as country, macroregion, province, etc. Addresses may have geographic coordinates specified with a different resolution level. Least precise geographic coordinates are resolved at the level of a country, while most precise are geo points that specify location up to a street and house number.

We also enable data providers to submit full addresses in the form of a free text, which is essentially a string that combines all attributes together into a human-readable format.

To provide RDF binding for the attributes, we considered two ontologies: schema: and locn:. We chose locn: as it has more structured attributes, among which locn:fullAddress that specifies the full address in a free-text form. However, to represent geographic coordinates, schema: was used. It provides a simpler way to model geographic coordinates via two properties (schema:latitude and schema:longitude), as opposed to specifying them via a separate node (cf. locn:Geometry).

In the initial model, we distinguish between registered, and other kinds of addresses. Many jurisdictions have the concept of registered address (i.e., the legal address where summons, subpoenas and other legal documents can be sent). This information is captured via the object property “registered address”.

Note that it is very common for companies to have no physical presence at the registered address. We record physical presence (headquarters, locations of companies’ shops or regional offices) via the object property “address”.

**Address Calculations**

Given some address data, it is possible to derive other data. For each address, we should get at least NUTS3 region, so we can do faceting on these regions:

- From a full address, we can find geo coordinates by using some external georeferencing service[74], record these coordinates and place them in NUTS regions using the NUTS RDF (see Section 2.1.1.9.1).
- From a full address, we can extract the postal code and then use the postal code -> NUTS3 mapping[75].
- We could also try to recognize place names (semantic enrichment) with disambiguation, but that is only possible if there is rich enough context since there are often places with the same name in different regions.
- Given a detailed region (e.g. LAU2), it is trivial to obtain higher-level region (country and NUTS1 to 3) if we have loaded EC NUTS and LAU data (see Section 2.1.1.9).

| **Object Property: registered address** | |
|---|---|
| **Description** | Public legal address where legal papers can be served |
| **Scope Note** | org:hasRegisteredSite leads to a node with types org:Site and locn:Address that has self-link org:siteAddress |
| **RDF** | org:hasRegisteredSite & org:siteAddress |
| **Domain** | Company |
| **Range** | Address |
| **Cardinality** | 0 or 1 |
| **Object Property: address** | |
| **Description** | Other address/location associated with an organisation |
| **Scope Note** | org:hasSite leads to a node with types org:Site & locn:Address that has self-link org:siteAddress. org:hasRegisteredSite is declared a sub-property of org:hasSite, so querying by org:hasSite will obtain all addresses |
| **RDF** | org:hasSite & org:siteAddress |
| **Domain** | Company |
| **Range** | Address |
| **Cardinality** | 0 to many |
| **Class: Address** | |
| **Description** | Mailing or physical address of the company |
| **Scope Note** | When used for company, should be linked as registered address (one) or other address (multiple). |

[74] For example, the withinRegion service “returns the NUTS regions that include a certain point based on its latitude/longitude coordinates”.
[75] http://ec.europa.eu/eurostat/tercet/download.do?file=pc2016_NUTS-2013.zip

| | |
|---|---|
| **Rules** | Every data provider should strive to provide NUTS3 for each address. This can be done in cooperation, using the techniques described in Address Calculations above. |
| **RDF** | org:Site & locn:Address, self-link org:siteAddress.<br><br>For now, we will use a single node with two classes and a self-link since we do not see a need to consider Site and Address as different entities. |
| **Data Property: full address** | |
| **Description** | Full address, free text |
| **Examples** | • Aston House, Cornwall Avenue, London N3 1LF, UK<br>• Via Belenzani 19, 38122, Trento |
| **Rules** | Provide a language tag if possible |
| **RDF** | locn:fullAddress |
| **Domain** | Address |
| **Data Type** | rdf:langString or xsd:string |
| **Cardinality** | 0 or 1 |
| **Lookup Property: Country** | |
| **Description** | Country of the address |
| **Examples** | nuts:IT Italia |
| **RDF** | locn:adminUnitL1 |
| **Domain** | Address |
| **Values** | URLs from [nuts:] |
| **Cardinality** | 1 |
| **Lookup Property: Macroregion** | |
| **Description** | NUTS1 region of the address |
| **Examples** | nuts:ITC NORD-OVEST |
| **RDF** | locn:adminUnitL2 |
| **Domain** | Address |
| **Values** | URLs from [nuts:] |
| **Cardinality** | 0 or 1 (but strongly recommended) |
| **Lookup Property: Administrative Region** | |
| **Description** | NUTS2 region of the address |
| **Examples** | nuts:ITC2 Valle d'Aosta/Vallée d'Aoste |
| **Rules** | Use a valid NUTS region, e.g., “Trentino-Alto Adige/Südtirol”, taken from the Geonames hierarchy (see Section 2.1.1.10), is not a NUTS region and should be resolved to valid NUTS2. |
| **RDF** | ebg:adminUnitL3 |
| **Domain** | Address |
| **Values** | URLs from [nuts:] |
| **Cardinality** | 0 or 1 (but strongly recommended) |

| Lookup Property: Province | |
|---|---|
| **Description** | NUTS3 region of the address |
| **Examples** | nuts:ITC20 Aosta |
| **RDF** | ebg:adminUnitL4 |
| **Domain** | Address |
| **Values** | URLs from [nuts:] |
| **Cardinality** | 0 or 1 (but strongly recommended) |
| **Lookup Property: Municipality/Commune/Settlement** | |
| **Description** | LAU1, LAU2 region of the address |
| **Scope Note** | Some countries (e.g., Bulgaria) use both LAU1 and LAU2 levels. Others (e.g., Italy) use only LAU2 |
| **RDF** | ebg:adminUnitL5, ebg:adminUnitL6 |
| **Domain** | Address |
| **Values** | URLs from [nuts:] |
| **Cardinality** | 0 or 1 |
| **Data Property: locality/city/settlement** | |
| **Description** | Locality/City/Settlement of the address, free text |
| **Rules** | • Provide a language tag if possible<br>• If both ebg:adminUnitL6 and locn:postName are specified, they should correspond to each other |
| **Examples** | "Trento"@it, "Тренто"@bg |
| **RDF** | locn:postName |
| **Domain** | Address |
| **Data Type** | xsd:string or rdf:langString |
| **Cardinality** | 0 or 1 |
| **Data Property: neighbourhood/quarter** | |
| **Description** | Part of a city, village or neighbourhood |
| **Scope Note** | Not likely to appear, but we have Coordinate Resolution corresponding to this field. |
| **RDF** | locn:addressArea |
| **Domain** | Address |
| **Data Type** | xsd:string or rdf:langString |
| **Cardinality** | 0 or 1 |
| **Data Property: street address** | |
| **Description** | Street name (and optionally number) |
| **Scope Note** | It is ok if the street number is also here, because many systems do not enter the number separately. There may be several concatenated lines, separated with newlines (\n) |
| **Examples** | • Via Belenzani 19<br>• Polygraphia Office Center, floor 4 \n 47A Tsarigradsko Shosse |
| **RDF** | locn:thoroughfare |
| **Domain** | Address |
| **Data Type** | xsd:string or rdf:langString |
| **Cardinality** | 0 or 1 |

| Data Property: street number | |
|---|---|
| **Description** | Street number and/or building name |
| **Scope Note** | Not likely to appear separately, but we have Coordinate Resolution corresponding to this field |
| **Examples** | • 47A<br>• Polygraphia Office Center, floor 4 |
| **RDF** | locn:locatorDesignator |
| **Domain** | Address |
| **Data Type** | xsd:string |
| **Cardinality** | 0 or 1 |
| **Data Property: postal code** | |
| **Description** | Postal code of the address |
| **Rules** | If possible, a European postal code should correspond to the fixed values given in NUTS Downloads[76], so we can correlate it to a NUTS region. |
| **RDF** | locn:postCode |
| **Domain** | Address |
| **Data Type** | xsd:string |
| **Cardinality** | 0 or 1 |
| **Data Property: postal office box** | |
| **Description** | Some addresses are associated with a PO box instead of a street address |
| **RDF** | locn:poBox |
| **Domain** | Address |
| **Data Type** | xsd:string |
| **Cardinality** | 0 to many |
| **Object Property: geographic coordinates** | |
| **Description** | Geographic coordinates of a geo-located address |
| **Rules** | If data is integrated from several providers, we may end up with multiple coordinate pairs in that node, which will need to be resolved into one. |
| **RDF** | schema:geo |
| **Domain** | Address |
| **Range** | Geographic Coordinates |
| **Cardinality** | 0 or 1 |
| **Class: Geographic Coordinates** | |
| **Description** | Geographic coordinates |
| **RDF** | schema:GeoCoordinates |
| **Data Properties: latitude, longitude** | |
| **Description** | Latitude and longitude of the coordinates expressed using WGS 84 reference system[77]. |
| **Examples** | (“51.477811", "-0.001475”) |
| **Rules** | • Coordinates must be expressed in the WGS 84 system<br>• If no coordinates available, do not make a schema:GeoCoordinates |

[76] NUTS1..3 available for download as CSV http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=NUTS_2013L&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC# or XLS http://ec.europa.eu/eurostat/ramon/documents/nuts/NUTS_2013.zip

[77] https://en.wikipedia.org/wiki/World_Geodetic_System (Note: some countries may have a different reference system by default)

| | |
|---|---|
| | node |
| **RDF** | schema:latitude, schema:longitude |
| **Domain** | Geographic Coordinates |
| **Data Type** | xsd:decimal (may be omitted) |
| **Cardinality** | 1 |
| **Lookup Property: Geocoordinate Resolution** | |
| **Description** | Resolution of a geographic coordinate location: from continent (least precise) to individual house number (most precise) |
| **Rules** | • If geo coordinates are present, resolution must be present too, even if it is a guess. There is no default value.<br>• Values should be mapped to the <resolution/> ConceptScheme (lookup list). |
| **RDF** | ebg:geoResolution[78] |
| **Domain** | Geographic Coordinates |
| **Values** | Values are taken from <resolution/L0> to <resolution/L10>[79] |
| **Cardinality** | 1 |

## 2.3.2 URI Construction

The euBusinessGraph base URL is http://data.businessgraph.io/ and supports content negotiation (i.e., can serve HTML and several semantic representations for each resource). This URL may still be subject to revision for marketing reasons. Note, http://businessgraph.io/ is used for general project information.

All URIs of individuals, including euBusinessGraph lookup (thesaurus) values are minted in the euBusinessGraph base URL.

Table 7 below summarises URI templates of individuals of different classes and lookup lists.

**Table 7: URI templates of individuals and lookup values**

| Class | URI Template | Example |
|---|---|---|
| Company | URI template to be used for companies with official registration ID | |
| | **company/<jurisdiction>/<ID>** | company/GB/07444723<br><br>Company registered in the GB company register with ID 07444723 |
| | URI template to be used for companies without official registration ID; an ID from a data provider (alternative business register) is used. | |
| | **company/<provider>/<ID>** | company/Atoka/6da785b3adf2<br><br>Company URI in Atoka[80] provided by SDATI. |

[78] Note that the name follows feedback from the LOV community https://plus.google.com/u/1/115593825497938381443/posts/LDxpHf5H1me

[79] https://github.com/euBusinessGraph/eubg-data/blob/master/data/EBG-geo_coordinates_resolution.csv

| | | |
|---|---|---|
| Address | URI of the registered (primary) address is formed by appending "/address" to the end of the company URI. | |
| | **<companyURI>/address** | company/GB/07444723/address[81]<br><br>Primary address of the company registered in GB. |
| | URIs of other addresses include guid or similar internal ID of the address in the provider's system. | |
| | **<companyURI>/address/<guid>** | company/GB/07444723/address/54321<br><br>Other address of the company registered in GB with ID = 54321. |
| | If guid is not provided for other addresses, we append a sequential number to the address. | |
| | **<companyURI>/address/<n>** | company/GB/07444723/address/1<br><br>First other address of the company that does not have ID in the register. |
| Geographic Coordinates | URI of geographic coordinates is formed by appending "/geo" to the corresponding address. | |
| | **<addressURI>/geo** | company/GB/07444723/address/geo<br><br>Geographic coordinates of the primary address of the company registered in GB. |
| Identifier | URI template for the official registration/identifier | |
| | **<companyURI>/id** | company/GB/07444723/id<br><br>Identifier of the company registered in GB with ID 07444723. |
| | URI template for other identifiers | |

[80] http://atoka.io/
[81] Note, we use the template URI of the company that has ID from the official business register. Similarly, URIs are constructed for companies with other kinds of IDs.

| | | |
|---|---|---|
| | **<companyURI>/id/<provider>** | company/GB/123456/id/OCORP<br><br>Identifier of the company not officially registered but provided by OCORP with id 123456.<br><br>If there are several identifiers per company and register, we add a suffix (numeric or some natural key), e.g.,<br>company/GB/123456/id/OCORP/1<br>company/GB/123456/id/OCORP/2 |
| Identifier System | **identifier/<provider>** | |
| **Lookup List** | **URI Template** | **Example** |
| Company Type | **type/<jurisdiction>/<CODE>** | type/BG/OOD<br><br>Company type “Дружество с ограничена отговорност” in Bulgaria. |
| Company Status | **<provider>/<status>** | BRC/inactive<br><br>URI that represents inactive status of companies as defined by BRC. |

### 2.3.3 ORM Specification

A conceptual model has been developed using the Object-Role Modelling (ORM) methodology[82]. The purpose of the ORM model is two-fold:

1. It defines a conceptual model describing the company data model that is more easily understood by non-technical users since it uses a visual modelling language. A set of diagrams showing the representations of the modelled concepts can be found in Appendix A.
2. It has been used to automatically generate an RDFS representation of the company data model. The RDFS generated from the current version of the ORM model is published on GitHub[83].

The ORM specification is work in progress. Currently, it covers the following concepts: Company and company properties, such as company names, incorporation date, dissolution date, language), classifications (company type, legal form, status and economic activity), online presence, jurisdiction, addresses and sites, identifiers, and identifier systems.

## 2.4 Mapping of data into RDF

This section describes the mappings that are used to convert company data (in the form of CSV, JSON, JSONLD or RDF) into RDF conforming to the company data model described above.

[82] http://www.orm.net/
[83] https://github.com/euBusinessGraph/eubg-data/blob/master/model/ebg-ontology-gen.ttl

In the subsections below, we describe the mapping notation and provide a generic example showing how the mapping rules are used. We have so far started to map data from OpenCorporates and SpazioDati. The mapping rules for these data are described in Appendix B.

### 2.4.1 Notation

In the definition of the mappings, we assume that the input data is a set of attribute-value pairs for each company. It is straightforward to convert CSV, JSON, JSONL, or RDF into this format.

For each of our mappings, we provide a table listing the names of all input attributes as well as *parameters* (written in italic), that denote an attribute value for a given company.

The mapping functions are described in tables containing:

- **Name:** the name of the mapping function
- **Definition:** a specification the output of the function for a given set of attribute-value pairs. The definition may contain parameters (written in italic) that will be replaced by their actual values when the mapping is applied to a set of attribute-value pairs. The definition may also contain application of other mapping functions (denoted in bold).
- **Condition (optimal):** Boolean expression which may contain parameters that describe the conditions under which the mapping function can be applied. If no condition is specified, then a default condition requiring all parameter values which are contained in the definition to be non-null is assumed.
- **Comments:** Explanations.

### 2.4.2 Example

Assume we get the following data about two companies as input:

Input =
{{ (id, 1), (registered_id,10), (name, SINTEF), (address_streetname, "Forskningsveien"), (address_streenr, 1)}, { (id, 2), (name, ACME), (address_streetname, "Karl Johan") }}

We first list the attributes and choose parameter names so that we can reference the attribute values in the definition of the mapping functions later.

| Parameter | Attribute |
|---|---|
| Id | Id |
| Rid | registered_id |
| na | Name |
| sn | address_streetname |
| snr | address_streetnr |

Then we defined two mappings named **company** and **address** that together produce the desired conversion when applied to each attribute-value set in the input. Both mappings make use of the following helper functions:

| Name | Definition | Condition | Comments |
|---|---|---|---|
| curi | company/id | rid = null | Company URI |
| | company/rid | rid != null | |
| Cadruri | curi/address | | Company address URI |

Note that an attribute value is assumed to be null if is not defined.

The **company** mapping is defined by the table below:

| Name | Definition | Comments |
|---|---|---|
| company | <curi> rdf:type <rov:RegisteredOrganization> . | Company type |
| | <curi> rov:legalName "na" . | Name |
| | <curi> org:hasRegisteredSite <cadruri> . | Registered address |

The **address** mapping is defined by the table below:

| Name | Definition | Comments |
|---|---|---|
| address | <cadruri> rdf:type <locn:Address> . | Address type |
| | <cadruri> locn:throughfare "sn" . | Street name |
| | <cadruri> locn:locatorDesignator "snr"^^xsd:integer . | Street number |

When the mappings **company** and **address** are applied to the input specified above, we get the following result:

<company/10> rdf:type rov:RegisteredOrganization .
<company/10> rov:legalName “SINTEF” .
<company/10> org:hasRegisteredSite <company/10/address> .
<company/2> rdf:type rov:RegisteredOrganization .
<company/2> rov:legalName “ACME” .
<company/2> org:hasRegisteredSite <company/2/address> .
<company/10/address> rdf:type <locn:Address> .
<company/10/address> locn:throughfare “Forskningsveien” .
<company/10/address> locn:locatorDesignator “1”.
<company/2/address> rdf:type locn:Address .
<company/2/address> locn:throughfare “Karl Johan” .

Note that the mapping rule:

<**cadruri**> locn:locatorDesignator "*snr*"^^xsd:integer .

is not applied to the company ACME because the input data does not contain the address_streetnr attribute for this company.

# 3 Handling identifiers in euBusinessGraph

A key part of the implementation of euBusinessGraph is a coherent understanding of identifiers, and a system for linking them together. This section outlines the conceptual thinking behind the identifier, and how this will be tied together with the user needs to create that implementation.

## 3.1 What is an identifier?

### 3.1.1 Overview

Identifiers are simply names that identify either objects, or classes of objects. Often they are made up of some consistent structure and managed by some sort of authority (in that case they are sometimes referred to as 'identifier codes'). The best of them clearly map to a distinct object in at least the domain in which they live. Others are ambiguous in what they refer to even within a domain, and even more so beyond that. For example, what does the identifier 12345 refer to? What about "John Smith"? Is it one of the tens of thousands of John Smiths in the UK (or another country), the British beer brand[84], a limited company[85], or something else? Because of this, it is important to provide context with an identifier (for example, the identifier system that this code is part of (e.g., 07444723 is a company identifier issued by UK Companies House).

In the context of euBusinessGraph, we are particularly interested in identifiers that are used to identify 'companies', either directly (e.g. company register numbers), or indirectly (stock ticker codes would be one example, but websites, brands, even addresses might be another). Getting to a successful outcome is complicated by several factors, including:

1. The relationship between the identifiers and the company is not always clear – is the identifier directly created for the legal entity (e.g. a company register identifier, or the LEI), or some other thing associated with it (e.g. a security listed on a stock exchange, or a tax registration). If it is the latter, the cardinality is frequently not 1:1, but instead may be 1:n, n:1 or n:n. In some cases, the cardinality will vary even within the same broad class of identifiers, e.g., EU VAT numbers have different cardinalities and coverage depending on the country.
2. Some identifier systems are badly designed, for example, lacking clear business rules (i.e., scope of the identifiers is not well-defined), or allowing identifiers to be reused[86].
3. Identifiers are often used incorrectly by many users, even governments, for example, assuming that they correspond 1:1 with companies when they don't (for example, Tax Identification Numbers in the US; DUNS numbers).

Unpicking the many issues relating to these, and ensuring that there is a consistent, coherent and useful implementation of identifiers in euBusinessGraph requires an in-depth examination of the nature of identifiers, specifically in the context of legal entities, and this is what we attempt to do below.

### 3.1.2 Typology

Note: For this project, a pragmatic analysis of company-related identifiers is important, rather than a formal academic study understanding all aspects of identifiers[87].

#### 3.1.2.1 Formal vs. informal

Many of the things that most people consider to be identifiers are part of a formal identifier system, that is, a system where the identifiers follow a defined set of rules and where some organisation (sometimes called the 'registration authority') administers the system or sets of rules in some way. One such organisation is ISO; others are the company registers, and tax authorities.

---

[84] https://www.johnsmiths.co.uk/
[85] https://opencorporates.com/companies/gb/SC253182
[86] For example, Jersey uses the same company identifiers for different company types, and has reused history numbers. See https://wiki.opencorporates.com/qa/qa_public/migrating_company_numbers/jersey
[87] See also http://www.niso.org/news/events/niso/past/ID-06-wkshp/definitions.html for a list of definitions related to identifiers.

Note that identifiers created for one specific purpose may often be used for another. If this is done carefully, with understanding of the nature of the identifier, it can be extremely useful in linking together datasets – in fact the Global Legal Entity Identifier System, which was set up by the G20[88], is based around the benefits and quality improvements of having a single identifier for legal entities that can be used successfully in a large number of contexts, and by a large number of different actors, particularly in the financial system[89].

Other times, this causes data quality problems, and even security problems, as in the case of using US social security numbers as personal identifiers (see this video for a humorous explanation).

Examples of relevant formal identifier systems include the ISO 3166[90] code for country codes, the SIRENE identifier[91] for French companies, and NACE 2 code for industry classifications[92]. Examples of informal systems are people's names and, in most countries, addresses, although in some countries there is a standardized way of representing addresses.

#### 3.1.2.2 Enumerated (all known identifiers)

Some systems have a list of all known identifiers issued (or sometimes in current use) – for example, the Global LEI System, BIC numbers, ISO 3166. Other systems (often informal ones) have no canonical lists of identifiers, for example the system of UUIDs[93]. In general, addresses are not enumerated (although some governments may maintain an official list of addresses in a country).

#### 3.1.2.3 UID (Unique Identifier) vs. ID

Identifiers can either refer to specific instances of things, or to classes of things. For example, a computer model may have a product identifier (e.g. the EAN/UPC barcode issued by GS1), and each individual computer would in addition have a serial number issued by the manufacturer. These latter, instance-level identifiers are often called UIDs[94]. Neither the classes of things, nor the instance level objects, need to necessarily refer to specific physical items; they can also refer to concepts, or to immaterial things, for example:

- ISO 3166 is a data standard for representing countries and geographic regions as identifier codes (of course "ISO 3166" is itself an identifier)
- EU NACE Rev 2 is an identifier code system for classifying the industry sector
- Identifiers for legal entities (e.g. 391200UUI29C55NSFW45 is the LEI for the German company LOBA GmbH & Co. KG).

#### 3.1.2.4 Validation

While some identifier systems allow any pattern of identifiers to be valid (perhaps with a fixed length), others add restrictions on what may be considered 'valid'. For example:

- A fixed format that includes some sort of restriction on the format (for example the GIIN will never contain the letter "O").
- One or more check digits to act as a way of detecting errors in identifiers, particularly data entry errors. For example, the SIREN company number uses the tenth digit as a checksum, and the LEI has the last two digits (of 20) as check digits.

In addition, if an authority has a complete list of valid identifiers they may offer a lookup service, for example the one provided for EU VAT numbers[95], or by the LEI system.[96]

[88] https://www.gleif.org/en/about/our-vision
[89] https://www.leiroc.org/publications/gls/roc_20120608.pdf
[90] https://www.iso.org/iso-3166-country-codes.html
[91] https://www.sirene.fr/sirene/public/static/documentation
[92] http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2
[93] https://en.wikipedia.org/wiki/Universally_unique_identifier
[94] https://en.wikipedia.org/wiki/Unique_identifier
[95] http://ec.europa.eu/taxation_customs/vies/?locale=en
[96] https://www.gleif.org/en/lei-data/gleif-lei-look-up-api/access-the-api

#### 3.1.2.5 'Dumb' vs. 'intelligent' identifiers

There has been a trend over the past few years to move away from identifiers with embedded information within them[97] to so-called 'dumb' identifiers, where there is no information embedded in the code. To a large extent, this is based on painful experience when such embedded information changes[98].

For example, the company number for CHRINON LTD (the company that publishes OpenCorporates), is 07444723, and will not change when the company changes name, location, corporate structure, industry code, and legal form (e.g., it could change from LTD to PLC and back again[99]). On the other hand, the CIN company identifier used by India includes the industry code, the state of incorporation, and the legal form as part of the identifier (e.g., U63090MH1971PTC015089). When the company changes industry, or becomes a public company, the identifier changes[100], which causes data quality and matching problems (particularly false negatives, resulting in duplicate entries), as well as introducing costs for the company involved.

For dumb identifiers to be truly useful, the identifier and the associated core data need to be open rather than proprietary, otherwise 'lock-in' is likely to occur.

#### 3.1.2.6 Mutable vs. immutable

It is highly desirable to have identifiers be consistent over lifetime and never change, i.e., to be immutable. The most likely reason for identifiers to change, is when the identifiers have information contained within them (see 'dumb vs. intelligent', above). However, change may also occur if the identifier format changes (e.g., because the system runs out of identifiers, the underlying technical platform requires it, or errors are found in the identifier system).

#### 3.1.2.7 Uniqueness

Does the identifier map 1:1 to object it relates to, or could there be multiple identifiers of the same identifier scheme for the same entity? Clearly this is not true where the identifiers have mutable information within them (see 'dumb vs. intelligent', above), but also occurs where duplicate entries are added. Although this might seem an obvious data quality problem that should be prevented, it is surprisingly hard to do so, particularly if the identifiers are private or proprietary, or the mapping between identifiers and objects is not well defined (the proprietary DUNS number suffers from all these issues). The LEI system, for example, has worked hard to ensure that the entities are unique and map 1:1 to legal entities, by having clear, public rules for the mapping, and by using technical tools, data analysis, and challenge system.

#### 3.1.2.8 Persistence

Some issuers of identifiers 'remove' identifiers from the register when the object they relate to is no longer active (e.g., a small number of company registers remove dissolved companies from the register). Of course, if those identifiers are used by third parties, they are rarely actually deleted in the ecosystem, but the canonical source is no longer there, resulting in data quality and provenance problems. If the data or identifier is deleted, it can cause further data quality problems, as the identifier may subsequently be reused to refer to another entity.

#### 3.1.2.9 Public

Some identifiers are not public and only intended to be used within a closed system. As with most issues, there is rarely a binary distinction. For example, personal social security numbers are not

[97] E.g. the G20 when approving the LEI system chose a system where the identifiers were dumb. "The code should be a unique dumb alphanumeric string and not incorporate any intentional embedded intelligence (such as a country reference) which could lead to the code becoming out of date. The code should be persistent, in the sense that the code would never be assigned to another entity."
[98] https://www.capgemini.com/2012/07/why-smart-keys-are-dumb-and-dumb-keys-are-smart/
[99] It could not, however, change to a Scottish Limited company, because such a transition is not allowed, and a new legal entity would have to be formed with the assets transferred to that new entity
[100] There is also a problem when new states are formed, as happened in when the state of Telangana was formed in 2014, with the result that companies within the new area had to change their CIN numbers.

intended to be public, but often leak out of the closed system, either to third-party companies, or for example on the internet.

#### 3.1.2.10 Open vs. proprietary licence

It is debatable whether there can be any IP in just an identifier, at least when it is some sort of code (though there may be in domain names, for example); more often the IP claimed is the combination of the identifier together with some associated data. There are differing levels of licence, from completely open[101] to highly proprietary. The latter is normally enforced through contractual means with users only gaining access if they agree to the limited terms of use (see also the CUSIP identifier for US securities[102]).

#### 3.1.2.11 Existing standards for company-related identifiers

There are many existing identifier standards for companies and related objects. To list them all here would be neither possible, nor beneficial. However, a few of them include:

- **EUID** (European Unique Identifier)[103]: This is a unique identifier for communication between company registers in the EU. We were unable to find a public schema for this, or a consolidated list of how the various parts are made up, e.g., a lookup table for the Register Identifier ("Elements making it possible to identify the domestic register of origin of the company and of the branch respectively"), and we consider this lack of information to be a potential vector for data quality problems in the project.
- **EU VAT numbers**[104]: While the format of VAT numbers vary quite considerably from country to country, they are grouped together under a single EU composite identifier consisting of two letters for the country code, followed by between 8 and 12 characters depending on the country.
- **Global Legal Entity Identifier** (LEI)[105]. The LEI system was set up by the G20 to enable, and was based on recommendations by the Financial Stability Board and the ISO 17442 standard. It is mandated for use by several regulators, including the CFTC in the US and ESMA in Europe, most notably under the MIFID II regulations coming into force in January 2018. Notably the data is made available under a CC0 licence and the identifiers maps 1:1 with legal entities/legal personality.
- **SEC Central Index Key**[106] This is an identifier used by entities making filings to the US Securities & Exchange Commission. However, these do not map 1:1 with the legal entities actually filing them, but actually n:n (as a legal entity may have multiple CIKs and a CIK may be used by different legal entities over time, e.g., when there has been a reverse takeover)
- **Ticker symbols**[107]**.** These are identifiers used by stock markets to identify share listings of publicly listed companies. Although tickers are sometimes used as identifiers for companies, because a company might have several classes of publicly traded shares, each of which may have listings on more than one exchange, there is an n:1 relationship between tickers and the issuing entities.
- **GS1 Global Location Number (GLN)**[108] This is produced by the GS1 organisation to represent locations of companies (rather than the companies themselves), primarily for use in the supply chain data industry. Note that despite this, the GLN has been used for the New Zealand official government Business Number (an identifier that represents all types of business, companies, partnerships, sole traders, etc.[109]).

---

[101] The Global LEI system is an example of an identifier system that is openly licensed, with CC0 licence: https://www.gleif.org/en/meta/lei-data-terms-of-use/
[102] For example, see CUSIP licence FAQ https://www.cusip.com/pdf/CGS%20Website%20FAQ12.4.15%20(ACS).pdf
[103] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R0884
[104] http://ec.europa.eu/taxation_customs/vies/faqvies.do#item_11
[105] https://www.gleif.org
[106] https://www.sec.gov/edgar/searchedgar/cik.htm
[107] https://en.wikipedia.org/wiki/Ticker_symbol
[108] https://www.gs1.org/gln (Standard, as PDF)
[109] http://www.gs1nz.org/standards/identify/

- **Global Intermediary Identification Number** (GIIN)[110]. The GIIN is an identifier controlled by the IRS for use in the FATCA Registration System of foreign financial institutions (FFI), financial institution (FI) branches, direct reporting non-financial foreign entities (NFFE), sponsoring entities, sponsored entities, and sponsored subsidiary branches. Institutions and entities assigned a GIIN can use it to identify themselves to withholding agents and tax administrators for FATCA reporting purposes.
- **DUNS**[111] A proprietary identifier issued by Dun & Bradstreet, and widely used within the US government, although there is now pressure from multiple organisations for the government to move away from them, due to the proprietary licence, data quality issues and the lock-in to D&B systems it provides[112].
- **Business Identifier Code**[113] BIC is fundamentally an international routing ID to facilitate automated processing of information for financial services, identifying banks and bank branches.
- **International Securities Identification Number** (ISIN)[114] An identifier for securities (shares, bonds, etc.) specified by ISO 6166.

The fact that these are standards does not mean that they are without problems. For example, the ISIN contains a flaw in the check characters which mean they will pass validation in certain circumstances when in fact the identifier is invalid. In addition, the use of identifiers changes over time with, for example, the first LEIs issued not conforming to the ISO 17442 standard in the use of the first 4 digits.

## 3.2 Key considerations for identifiers in euBusinessGraph

While all the attributes and behaviours in the above are potentially relevant, there are several key features and considerations that are particularly important in the context of euBusinessGraph, given that we are combining different identifiers from various sources:

1. **Clarity about the identifier.** First and foremost, we must know which identifier we are talking about, and the key attributes of the identifier system (mutability, uniqueness, format, validation, etc.).
2. **Relationship of identifiers to companies.** Does the identifier map directly to a legal entity, or does it map to an object that is one (or more) steps removed from companies?
3. **Confidence of mapping of an identifier to a company.** In many cases (for example, mapping domains to companies, or SEC CIK codes to entities), there is no certainty whether the mapping is correct, and the level of this uncertainty should be understood.
4. **Confidence of mapping companies to company via identifiers.** It is possible that two different data providers might map the same identifier to different entities. This could mean that the different entities:
    - Are actually the same entity
    - Are related in some way (e.g. parent/subsidiary)
    - Are unrelated, for example due to invalid mapping to entities, or to the identifier mapping n:n to entities
5. **Use cases.** Different users will likely have different requirements in terms of accuracy. For example, in general false positives are extremely problematic for many use cases and trust in data. However, for some uses, e.g., sales and marketing, and investigations, some level of false positives would be acceptable.

[110] https://www.irs.gov/businesses/corporations/fatca-online-registration-system-and-ffi-list-giin-composition-information
[111] https://www.dnb.co.uk/duns-number.html
[112] https://www.datacoalition.org/new-gsa-statements-end-in-sight-for-duns-monopoly/
[113] https://www.iso9362.org/
[114] https://www.iso.org/obp/ui/#iso:std:iso:6166:ed-7:v1:en

## 3.3 Requirements (for euBusinessGraph)

### 3.3.1 Why we need identifiers

The euBusinessGraph seeks to identify data related to companies and organisations, such as contracts, corporate events (mergers, acquisitions, etc.), and people involved in companies. The focus is on the design of a cross-border, -domain, -language identification system for company-related data, where the principal task is on the linking and mapping between existing identification systems, including extensions where appropriate (e.g., for events). Operating the system of identifiers requires reliable provisioning of the business graph, including hosting, querying facilities, security and access control, data transformation and onboarding mechanisms, reliability and SLAs.

The euBusinessGraph approach to solving the issues related to systems of identifiers is to collate and define mappings for identifiers related to corporate data and entities. The proposed system of identifiers is meant to support the linking of data between entities in different countries, and across multilingual barriers, and will serve as the core mechanism for linking data for the business cases, and the creation of the business graph.

Core aspects that will be addressed in the system of identifiers are related to understanding/mapping of the different identifiers that are used in Europe, considering three core dimensions: type of identifier, intellectual property (IP), and integrity. The design of such a system will ultimately address and consider aspects such as:

- Consistent, workable and useful data model and schema for identifiers, and their relationship with the things they represent;
- A conceptual methodology for how, when and why to connect identifiers together;
- Sourcing of identifiers;
- A reconciliation system to implement matching using the above-mentioned methodology.

We have decided to tackle this in an iterative way, starting with core business register identifiers and a simplified data model for them.

### 3.3.2 Data model and schema for identifiers

In this section we provide a summary of the main properties and associations in the EBG company model that are relevant for the system of identifiers (see Section 2.3.1 for a detailed description).

Companies are registered using various kinds of identifiers. Some of these identifiers are kept in registers and some others are self-issued and not centralised. The core company model can represent several types of identifiers:

- Official registration in a trade register: this registration should correspond to the company's jurisdiction and it is used to establish the legal existence of the company.
- Other official government registers for specialised purposes (e.g., bank license, insurance company license, register of startup companies)
- Official international registries (e.g., GLEI)
- Registration to non-official company databases (e.g., OCORP, SDATI, Wikidata, Dun & Bradstreet)
- Social networks (e.g., Facebook, Twitter)

To support multiple identifiers, for diverse purposes, the model has the following characteristics:

- the *Company* class can link to several *Identifier* instances
- the *Identifier* class is associated to an *Identifier System*.
- the *Identifier System* class models the different cases in which identifiers are handed to a company.

The *Identifier System* class describes characteristics of an identifier system that reflect whether the system has an identifier database, can be used to uniquely identify a company, has official character

and in which jurisdictions and which are the rules used to determine if identifiers are valid within that system. Moreover, the Identifier System class also encodes information about the agents in charge of creation and maintenance, issuing and publishing of identifiers and information about web resources that can be used to search, browse and retrieve identifier information.

The *Identifier* class holds information about the identifier value and also about its lifecycle (issue date and expiration date, in case the identifiers in the system can expire).

Finally, each *Company* instance can have a redundant *official identifier* link so the model is compatible with the Registered Organisation vocabulary.

The fact that the current version of the model allows multiple identifiers of a company enables its use as a tool for matching and linking data from different data sources about the same company. To provide efficient search, matching and filtering, we envision the business graph to consist of a backbone that will function as an index, and link company data from various data providers to define a (virtual) business graph of company data. This is illustrated in Figure 5 below:

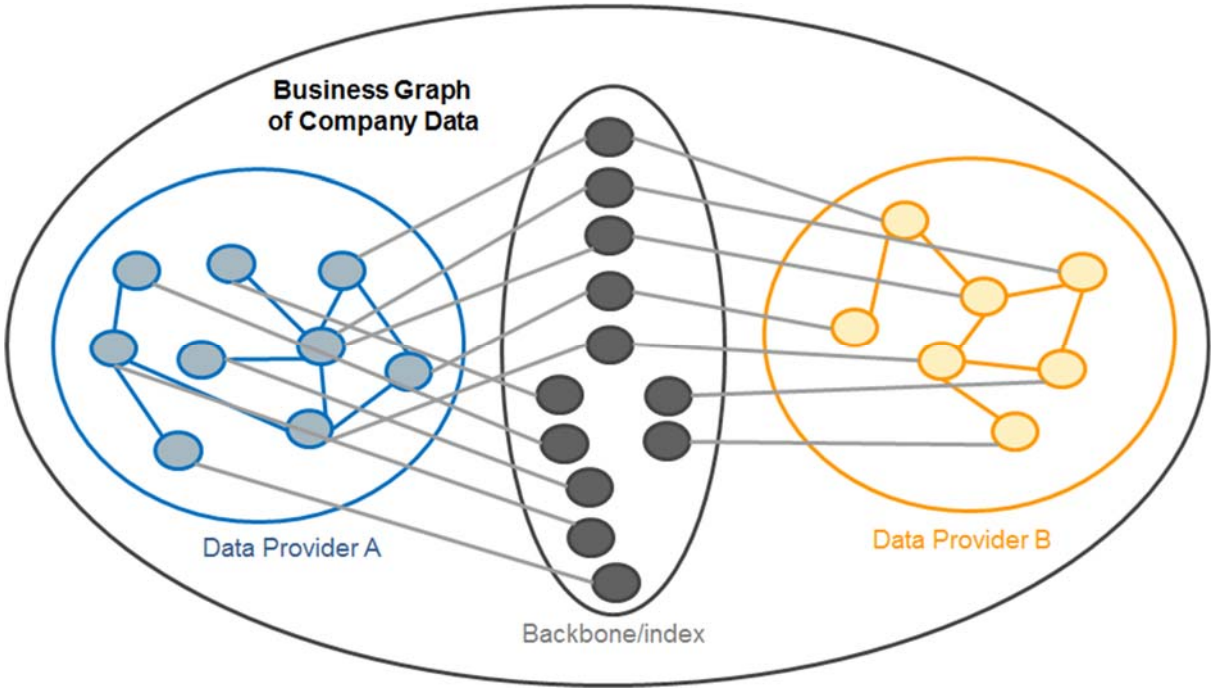



**Figure 5: Connecting identifiers with the euBusinessGraph system**

The backbone/index should contain entries about all companies that are available from the different data providers through the business graph. Each company entry in the backbone/index should contain all identifiers that are supported by the business graph. This enables the aggregation of data about the same company from different data providers, as shown in the figure.

However, not all considerations discussed in Section 3.2 are addressed in the first version of the EBG model. As an example, the model does not cover situations where we have many-to-many mappings between companies and identifiers, and does not resolve ambiguity or integrity.

The EBG model will see further updates during the next period of the project to address these issues.

# 4 Summary and Outlook

## 4.1 Summary

This document reports on the initial iteration of the system of identifiers, and the relevant shared data models developed to support the business cases.

Development of the model was guided by good practices and recommendations for describing the domain of Business Information (BI), including ISO standards, WC3 recommendations, EC and UN vocabularies, classifications and thesaurus. We tried to use terms and term descriptions from well-known established vocabularies and sources.

At the same time, we collected requirements for the model from its first intended users, the business cases of the project: Data Journalism Product (DJP), CRM Service (CRM-S), ATOKA+, and Tender Discovery Service (TDS). The requirements defined the scope of the model. However, in implementation of the initial version of the model, we prioritised those requirements that could be covered by data available in the first period of the project through its members. We surveyed the data from four data providers: OCORP, SDATI, ONTO and BRC. The result of the survey is a set of common fields that each data provider agreed to fully share with the graph, and a set of fields that some data providers agreed to provide for the purpose of matching companies across borders.

The initial version of the model presented in this document covers the concepts of Company, Identifier, Identifier System and Address. Other relevant entities, such as key managers, shareholders and company-related events, are left for future iterations of the model.

The definition of Company is adapted from the Registered Organization Vocabulary and refers to organisations that gain legal entity status by registering in an official business register. This choice was made because registered organisations are the main entities in euBusinessGraph. Registered organisations can have only one official registration number that is given to it by an official business register. However, we identified other kinds of identifiers that companies can obtain by registering with other systems. We captured this information in the model, as it is a requirement for such tasks as company reconciliation and matching across jurisdictions. In the current version of the model, we also keep track of the lifecycle of identifiers from issue to expiration and we encode information about agents in charge of creation, maintenance, issuing and publication of the different identifier systems. Additionally, we are able to represent key properties of several identifier systems studied. Through these properties we are able to determine if a system issues identifiers that univocally identify legal entities, that may change depending on other company information and that are expected to be permanent and thus suitable as foreign keys. Moreover, the model is able to represent information about web resources for tasks such as search, browsing and retrieval.

From the requirements analysis of the project's business cases, we acknowledged the importance of having both legal and trading names of companies. While the former is given during the official registration, these are not necessary the names by which companies are known to public. We included both in the model.

Three company classifications were considered in the first iteration of the model: classification by company type (legal form), status and economic activity. Each jurisdiction has a limited set of company types, and there is no standardised list of types that could be used to normalise company types across jurisdictions. This task is known to be challenging and can be an objective for a separate project by itself. For the purposes of euBusinessGraph, we simply consolidated the types from EU jurisdictions into one concept scheme. Similarly, for company statuses, different data providers understood the meaning of "active" vs. "inactive" differently. We consolidated those definitions per each data provider into a euBusinessGraph company status concept scheme. To specify companies' economic activities, we adapted NACE, as EU jurisdictions use NACE national extensions that are compatible with NACE.

Finally, we distinguished the concepts of legal, physical and online presence of companies. Legal presence is defined as the address a company provides during its registration. It is typically used for mail communication with the company, or for filing legal documents. We use "registered address" to specify companies' legal presence. Addresses of physical locations of companies' offices, stores or other buildings are captured via the "address" property. Online presence of companies is specified through their emails, Wikipedia pages, corporate websites, and news or blog feeds.

Other company attributes captured in the model include incorporation and dissolution dates and languages used in web resources related to the companies. Three flags were introduced to specify whether a company is registered in any startup register, is completely owned by the state or publicly traded.

We concluded the report with realisation of the conceptual model in RDFS using the ORM-to-RDF methodology. We presented the ORM specification of the model and mapping rules. The methodology has been applied to generate RDF representation of data from OCORP and SDATI.

## 4.2 Outlook

The project foresees future development of the model in the light of experience and the work undertaken in the related project work packages and their deliverables.

First, we expect that the business cases' requirements will be refined as the business cases are further developed. At the time of collecting requirements for the current model, only DJP business cases could provide us with details needed to define the scope of the model. We also expect to receive feedback from the technical work package with whom we have been working simultaneously. The model is the foundation for collecting, registering, representing, normalising and distributing data through the marketplace platform. At the time of writing this deliverable, the business case for euBusinessGraph marketplace itself is being finalised. We expect to have the first results during 2018, which will allow us to prioritise data modelling work.

Second, we plan to revise the business cases' requirements that were collected during the first year, but were not included in the current version of the model:

- The model must represent key managers and other officers of companies.
- The model must capture the concept of event, and represent diverse types of events from various sources, such as news (merger, acquisition, etc.); gazettes (corporate events); corporate websites (change on companies' websites) and authoritative data sources (e.g., change of a company's address in the register). WP2 should define the event model that will be used in the project to enrich the graph with company-related events information.
- The model must support trustworthiness/fuzziness of information coming from different data sources (authoritative vs. non-authoritative) -- especially for different events.
- The model should capture key company metrics, such as the number of employees.
- The model should represent company economic indicators, such as profit, loss and taxes paid by companies.
- The model must support data provenance.
- The model must support additional requirements for identifiers, such as confidence in relationship to the matched company.
- The model must represent extra information not available directly through the graph, but offered by the data providers via their repositories (i.e., the data offering requirements). We have started collecting information from the data providers about extra information they'd be willing to "advertise" through euBusinessGraph graph. Depending on the users' needs, this can further be elaborated with the following details:
    - **Pricing strategy**:
        - Usage-based pricing, when charging customers based on data volumes consumed
        - Plan-based pricing, when customers can choose among plans that provide different types of customer support, different scopes of data, different lengths of subscriptions, etc.
        - Other
    - **Free access.** Does your organisation provide free data access to certain organisations, groups or individuals?
    - **Registration**: Is registration required at your own site to access the data?

- **Data access method** defines a standardized procedure for accessing data, e.g., direct download, formats, API access, etc.
- **License** defines proprietary specifications of access rights to the offered data.
- **Unit price** defines how the data is organised into a single paid unit. Possible values are:
  - Field: data provider charges different fields individually
  - Package: data is organised into thematic packages, and each data package has its price
  - Dataset: data is organised into sets, each dataset has its own price
- **Currency**
- **Payment method**

Third, there are different directions in which the model could further be refined. The scope of the model was affected by the availability of the data to the consortium, which resulted in introduction of properties that are only valid to one jurisdiction or data provider. For example, “is startup” information shared by SDATI only. It is taken from a special section of the Italian Business Register[115]. The semantic of the flag is therefore valid for the Italian jurisdiction. It is not clear whether such registers exist in other countries, and if so, whether they have the same semantics. This should be investigated further, and the decision should be made about having a generic “is startup” flag or its specific definition in the case of Italy.

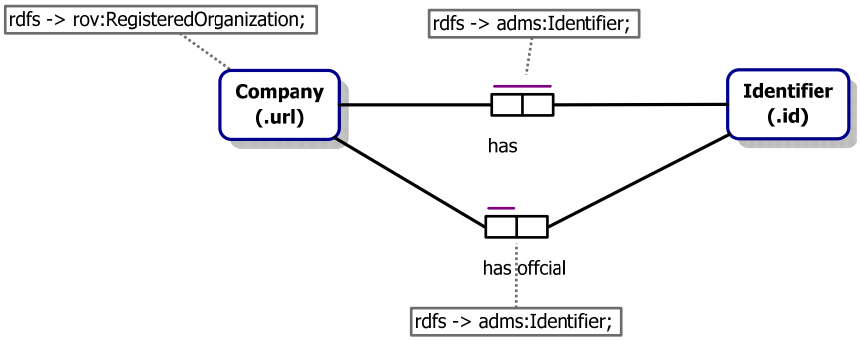
Another issue with the provider-specific properties is that they were designed and implemented for certain business needs. For example, the meaning of the “Wikipedia page” property has been defined as “Wikipedia page pertaining to the company”, as it was defined by its only data provider, SDATI. However, many Wikipedia pages do not relate to legal entities but to human approximation of the company, group or division. For example, https://en.wikipedia.org/wiki/Google could be about Alphabet Inc., or Google LLC or the Group. It appears to be the second of these. However, the history is for Google as a group, and the stock ticker codes are for Alphabet Inc. shares. While SDATI’s customers accept such loose semantics of the property, it must be understood whether there is any value to euBusinessGraph users in having the information in this form.

[115] http://startup.registroimprese.it/

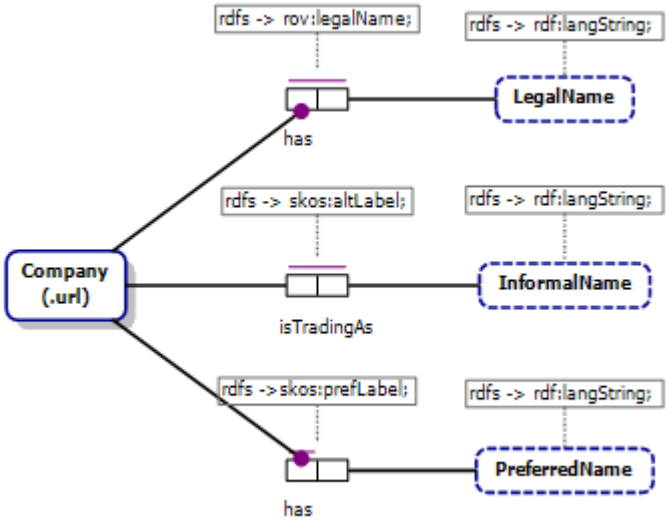
# Appendix A ORM Specification of the Model

This section documents the ORM specification of the semantic data model as well annotations needed to automatically generate RDFS from ORM. The RDFS generated from the current version of the ORM model can be found here. Note that this RDFS only defines classes and domain and range of properties. The other stuff such as skos:definition, skos:scopeNote will be added later.

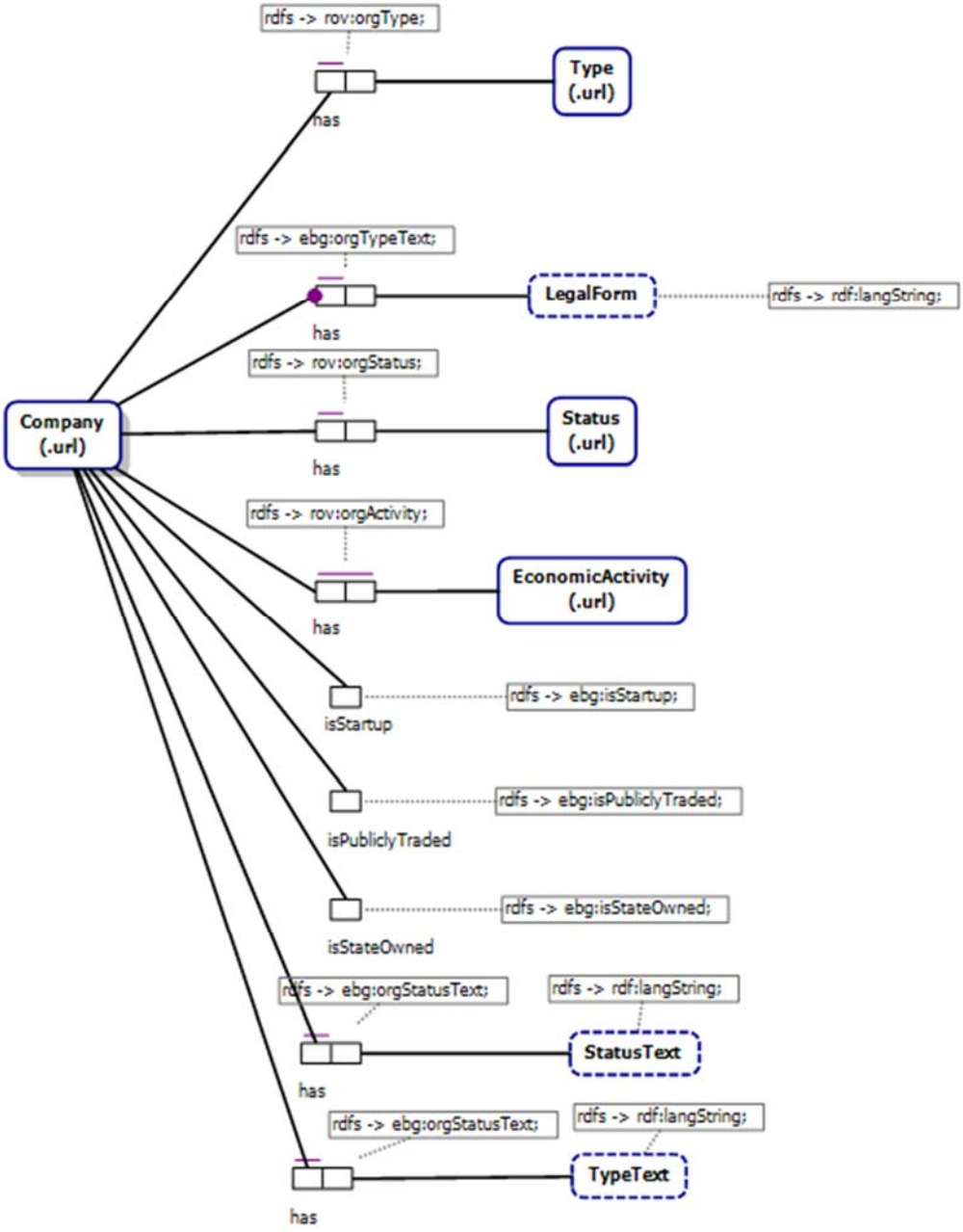
## A.1 Company



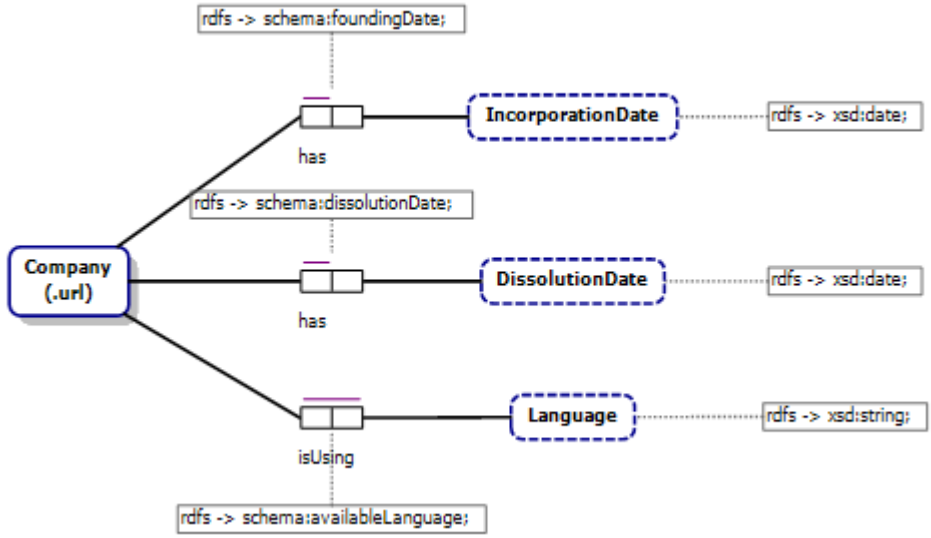

### A.1.1 Names



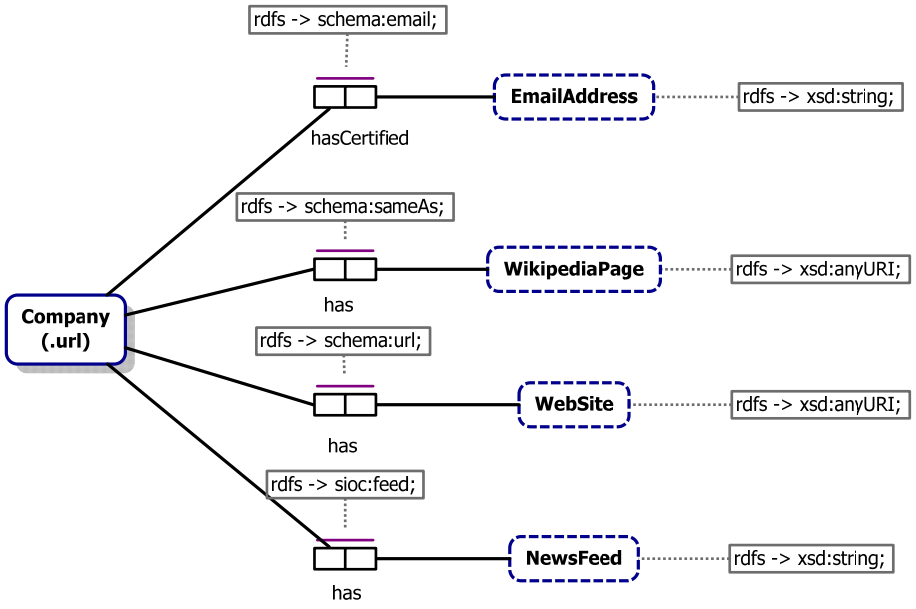
## A.1.2 Classifications



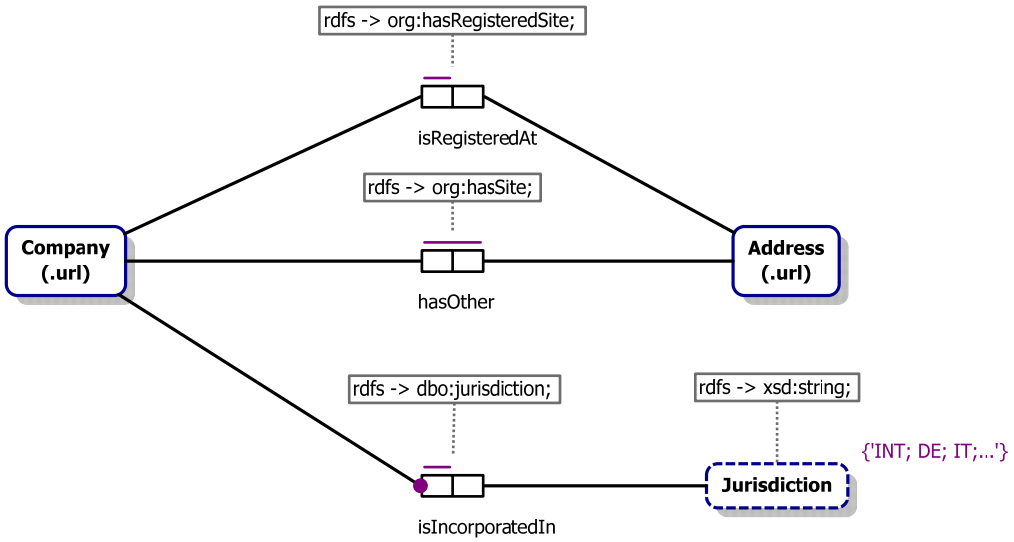
### A.1.3 Other details



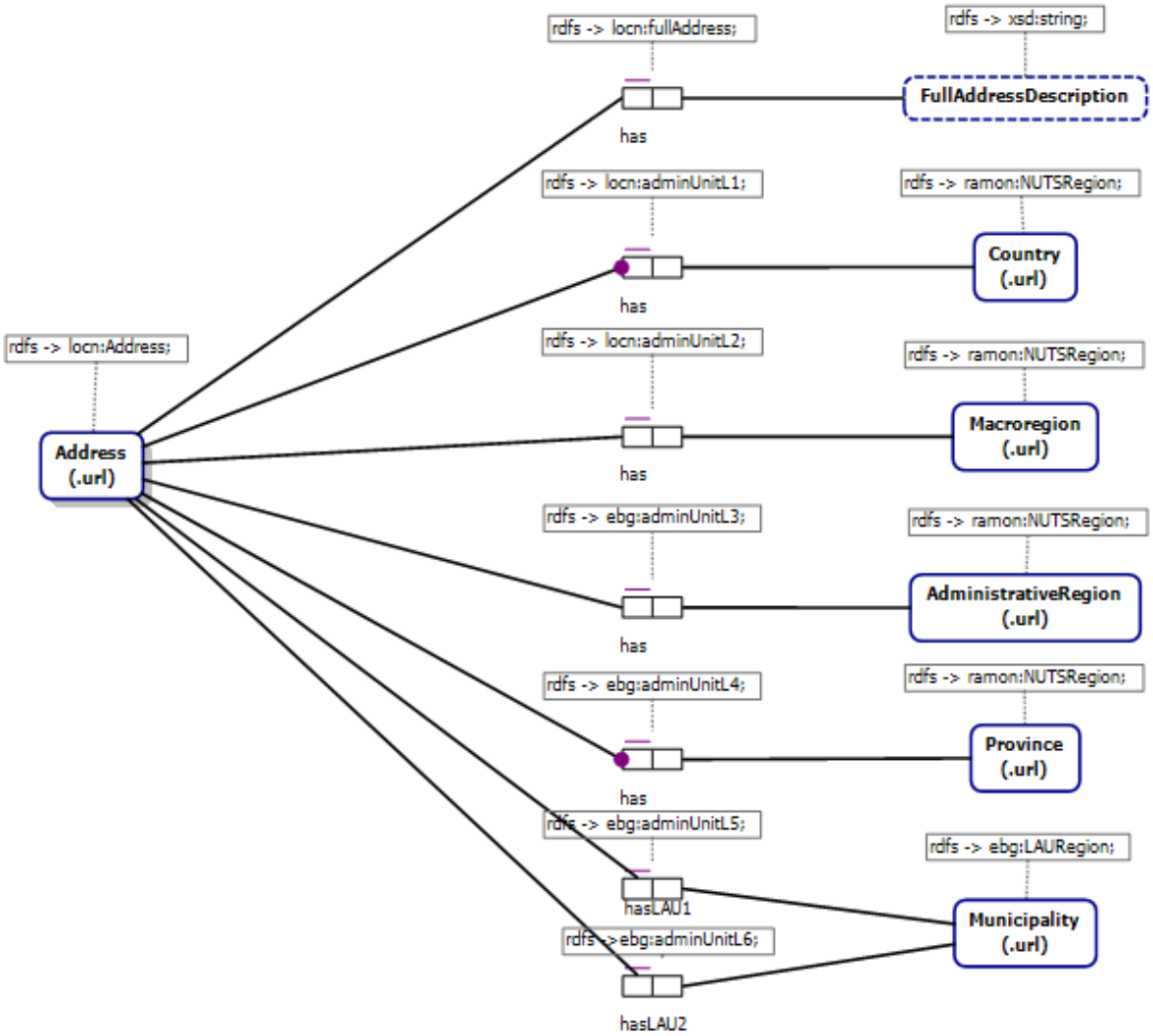

### A.1.4 Online presence



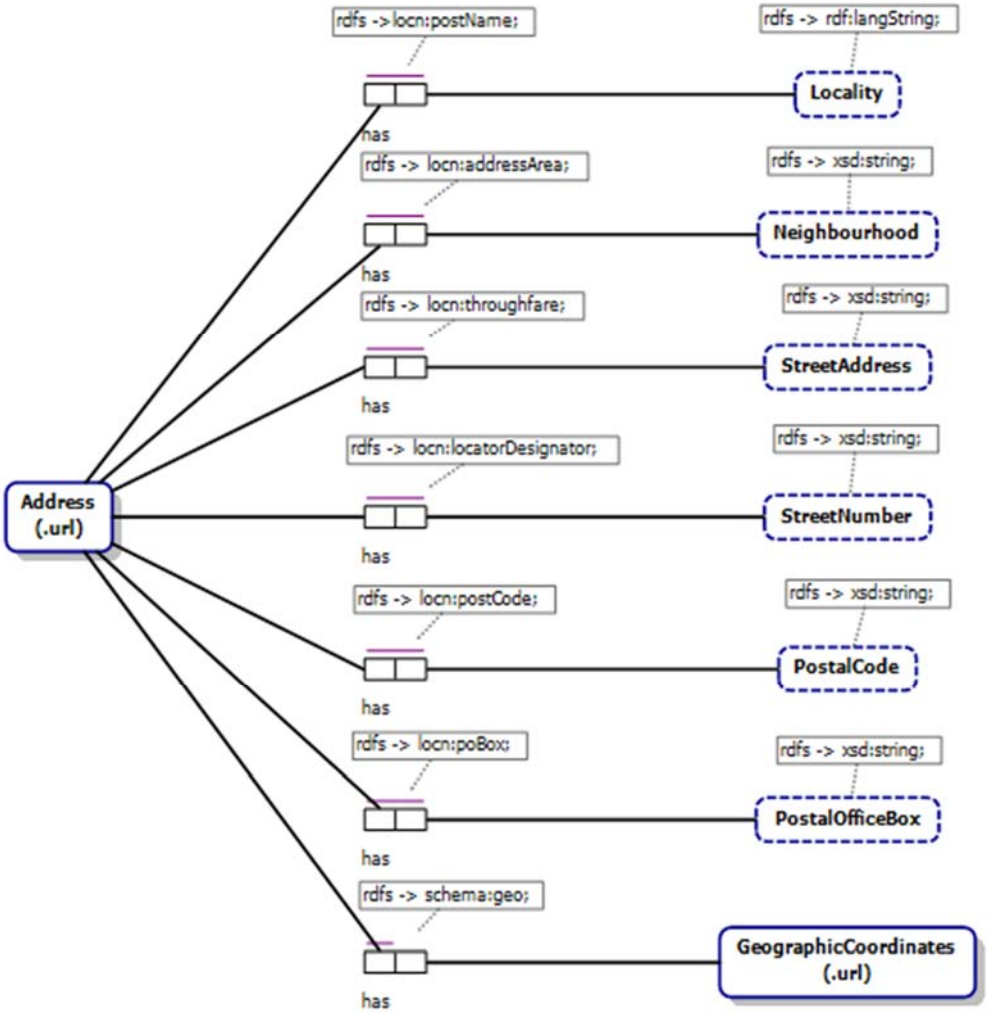
## A.1.5 Physical presence (address and jurisdiction)



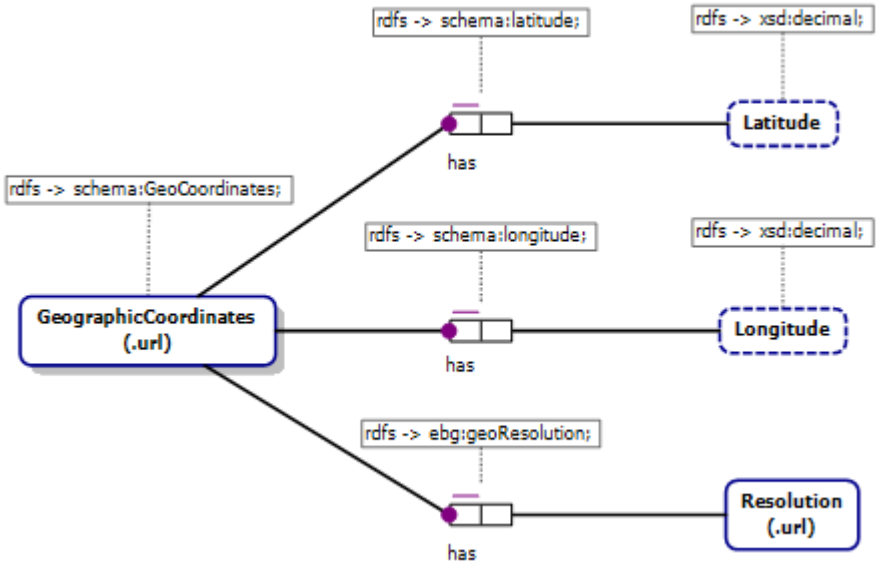

## A.1.6 Address1



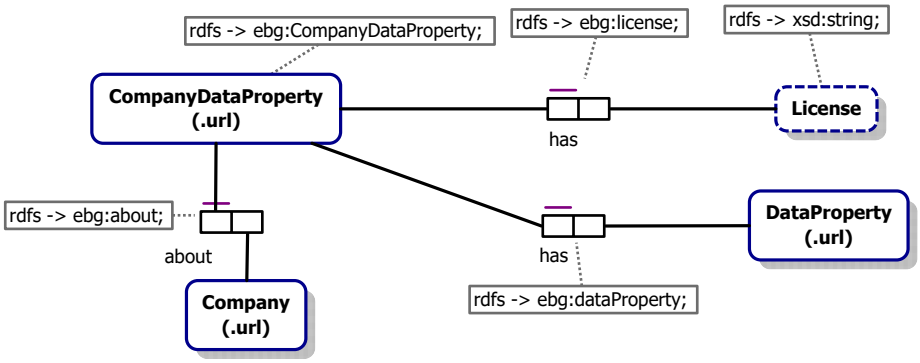
## A.1.7 Address2



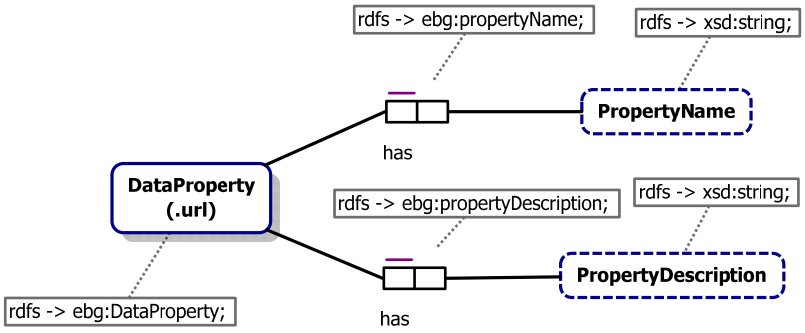

## A.1.8 GeoCoordinate



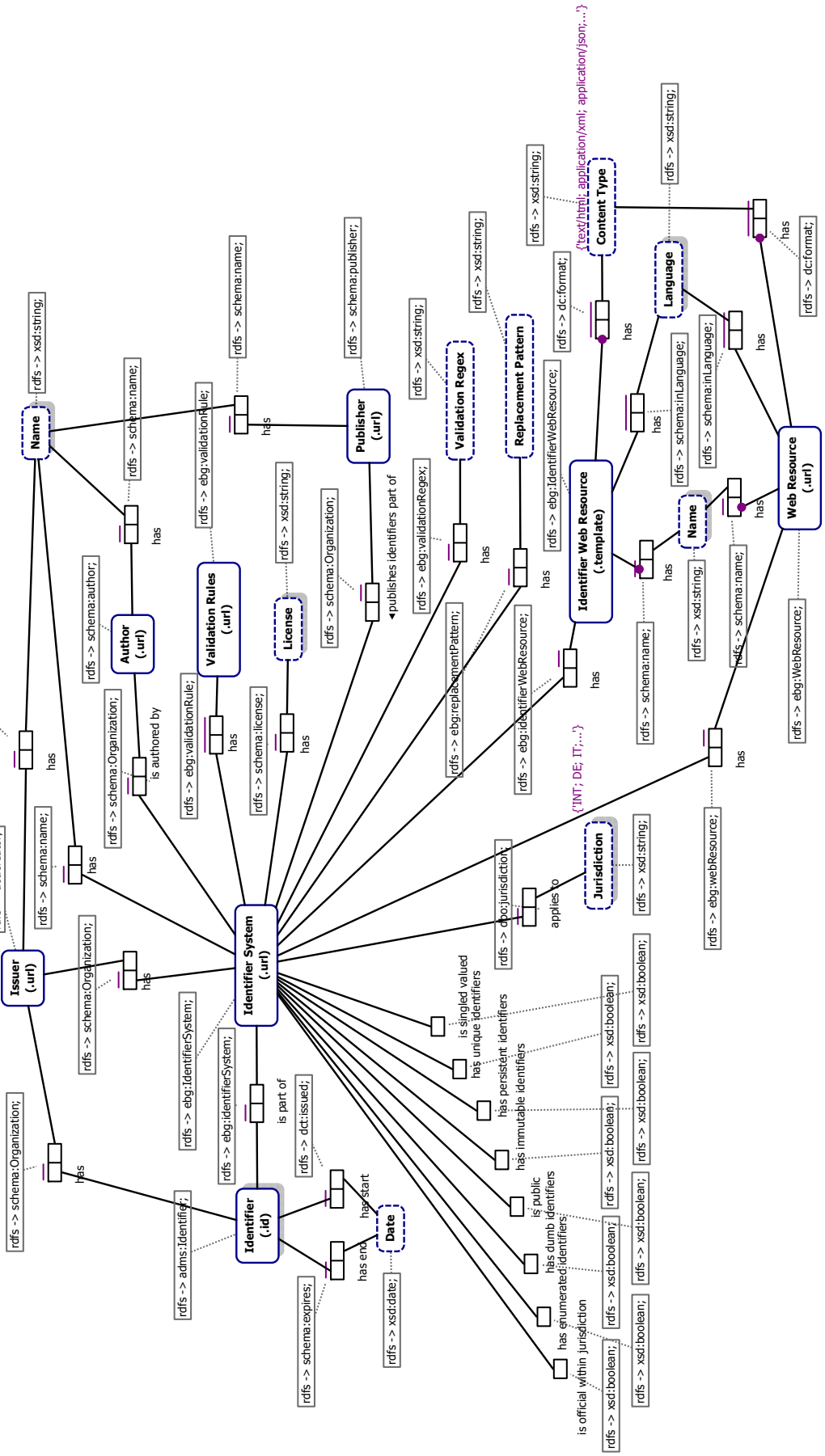
## A.1.9 Data providers




## A.1.10 Data property



# A.2 Identifiers



# Appendix B Data Mapping Rules

## B.1 Mapping OpenCorporates data

### B.1.1 Input

| Parameter | Attribute |
|---|---|
| *Na* | name |
| *jc* | jurisdiction_code |
| *id* | incorporation_data |
| *dd* | dissolution_date |
| *ws* | website |
| *cs* | current_status |
| *ct* | current_type |
| *caln* | current_alternative_legal_name |
| *sa* | registered_address.street_address |
| *lo* | registered_address.locality |
| *re* | registered_address.region |
| *pc* | registered_address.postal_code |
| *co* | registered_address.country |
| *if* | registered_address.in_full |
| *cn* | company_number |

### B.1.2 URI functions

| Name | Definition | Condition | Comments |
|---|---|---|---|
| **curi** | ebg:company/*jc*/*cn* | *jc* != null | Company URI |
| | ebg:company/OCORP/*cn* | *jc* = null | |
| **ciduri** | **curi**/id | | Company identifier URI |
| **cadruri** | **curi**/address | | Company address URI |

| | | | |
|---|---|---|---|
| **ruri** | ebg:register/*jc* | *jc* != null | Company identifier system URI |
| | ebg:register/OCORP | *jc* = null | |

## B.1.3 Mapping rules

### B.1.3.1 Company

| Name | Definition | Comments |
|---|---|---|
| **company** | <**curi**> rdf:type <rov:RegisteredOrganization> . | Company type |
| | <**curi**> rov:registration <**ciduri**> . | Registration identifier |
| | <**curi**> rov:legalName "*na*"**jctolang** . | Legal name |
| | <**curi**> dbo:jurisdiction "*jc*" . | Jurisdiction |
| | <**curi**> schema:foundingDate "*id*"^^xsd:date . | Founding date |
| | <**curi**> schema:dissolutionDate "*dd*"^^xsd:date . | Dissolution date |
| | <**curi**> schema:url "*dd*". | Website |
| | <**curi**> ebg:orgStatusText "*cs*". | Current status |
| | <**curi**> ebg:orgTypeText "*cs*"**jctolang** . | Organization type |
| | <**curi**> org:hasRegisteredSite <**cadruri**> . | Registered address |
| | <**curi**> skos:altLabel "*caln*" . | Alternative name |

### B.1.3.2 Identifier

| Name | Definition | Comments |
|---|---|---|
| **address** | <**ciduri**> rdf:type <adms:Identifier> . | Identifier type |
| | <**ciduri**> skos:notation "*cn*" . | Identifier value |
| | <**cadruri**> dct:creator <**ruri**> . | Register |

### B.1.3.3 Address

| Name | Definition | Comments |
|---|---|---|
| **address** | <**cadruri**> rdf:type <locn:Address> . | Address type |
| | <**cadruri**> rdf:type <org:Site> . | Address type |

| | | |
|---|---|---|
| | <**cadruri**> org:siteAddress <**cadruri**> . | Self reference |
| | <**cadruri**> locn:adminUnitL1 "*co*" . | Country |
| | <**cadruri**> locn:postName "*lo*" . | Post name |
| | <**cadruri**> locn:adminUnitL5 "*re*" . | Region |
| | <**cadruri**> locn:postCode "*pc*" . | Post code |
| | <**cadruri**> locn:throughfare "*sa*" . | Street address |
| | <**cadruri**> locn:fullAddress "*if*" . | Full address |

### B.1.4 Misc. functions

| Name | Definition | Condition | Comments |
|---|---|---|---|
| **jctolang** | @en | *jc* = gb | Jurisdiction code to language mapping. Todo fill in rest of the mapping. |
| | @no | *jc* = no | |
| | ... | | |

## B.2 Mapping SpazioDati data

### B.2.1 Input

| Parameter | Attribute |
|---|---|
| *ac* | base.ateco[1].code |
| *lfna* | base.legalForms[1].name |
| *ln* | base.legalName |
| *pe* | base.pec |
| *rea* | base.rea |
| *cciaa* | base.cciaa |
| *su* | base.startup |
| *co* | country |
| *id* | id |

| | |
|---|---|
| *na* | name |
| *fe* | web.feeds[*] |
| *ws* | web.websites[*] |
| *rafa* | base.registeredAddress.fullAddress |
| *ramu* | base.registeredAddress.municipality |
| *rapc* | base.registeredAddress.postcode |
| *rapr* | base.registeredAddress.province |
| *ramr* | base.registeredAddress.macroregion |
| *rare* | base.registeredAddress.region |
| *rast* | base.registeredAddress.state |
| *rasn* | base.registeredAddress.streetName |
| *rasnr* | base.registeredAddress.streetNumber |
| *rala* | base.registeredAddress.lat |
| *rallp* | base.registeredAddress.latlonPrecision |
| *ralo* | base.registeredAddress.lon |

## B.2.2 URI functions

| Name | Definition | Condition | Comments |
|---|---|---|---|
| **curi** | ebg:company/IT/*cciaa*/*rea* | *rea* != null and *cciaa* != null | Company URI |
| | ebg:company/Atoka/*id* | *rea* = null or *cciaa* = null | |
| **ciduri** | **curi**/id | | Company identifier URI |
| **cadruri** | **curi**/address | | Company address URI |
| **ruri** | ebg:register/IT/*cciaa* | *rea* != null and *cciaa* != null | Company identifier system URI |
| | ebg:register/Atoka | *rea* = null or *cciaa* = null | |
| **guri** | **cadruri**/*geo* | | Geographic coordinate URI |

## B.2.3 Mapping rules

### B.2.3.1 Company

| Name | Definition | Comments |
|---|---|---|
| **company** | <**curi**> rdf:type <rov:RegisteredOrganization> . | Company type |
| | <**curi**> rov:registration <**ciduri**> . | Registration identifier |
| | <**curi**> rov:legalName "*ln*"**cotolang** . | Legal name |
| | <**curi**> dbo:jurisdiction "**cotojc**" . | Jurisdiction |
| | <**curi**> rov:orgActivity "*ac*" . | Activity |
| | <**curi**> ebg:orgTypeText "*lfna*"**cotolang** . | Organization type |
| | <**curi**> schema:url "*ws*" . | Website |
| | <**curi**> schema:email "*pe*" . | Email |
| | <**curi**> ebg:isStartup "*su*"^^xsd:boolean . | Is startup? |
| | <**curi**> skos:altLabel "*na*" . | Alternative name |
| | <**curi**> sioc:feed "*fe*" . | News feeds |
| | <**curi**> org:hasRegisteredSite <**cadruri**> . | Registered address |

### B.2.3.2 Identifier

| Name | Definition | Comments |
|---|---|---|
| **address** | <**ciduri**> rdf:type <adms:Identifier> . | Identifier type |
| | <**ciduri**> skos:notation "**idval**" . | Identifier value |
| | <**cadruri**> dct:creator <**ruri**> . | Register |

### B.2.3.3 Address

| Name | Definition | Comments |
|---|---|---|
| **address** | <**cadruri**> rdf:type <locn:Address> . | Address type |
| | <**cadruri**> rdf:type <org:Site> . | Address type |

| | | |
|---|---|---|
| | <**cadruri**> org:siteAddress <**cadruri**> . | Self reference |
| | <**cadruri**> locn:adminUnitL1 "**cotoaul1**" . | Country |
| | <**cadruri**> locn:fullAddress "*rafa*" . | Full address |
| | <**cadruri**> schema:geao <**guri**> . | Geographic coordinate |
| | <**cadruri**> locn:adminUnitL2 "*ramr*" . | Macro region |
| | <**cadruri**> locn:adminUnitL4 "*ramu*" . | Municipality |
| | <**cadruri**> locn:postCode "*rapc*" . | Post code |
| | <**cadruri**> ebg:adminUnitL5 "*rapr*" . | Province |
| | <**cadruri**> ebg:adminUnitL3 "*rare*" . | Region |
| | <**cadruri**> locn:throughfare "*rasn*" . | Street name |
| | <**cadruri**> locn:locatorDesignator "*rasnr*" . | Street number |

### B.2.3.4 Geographic coordinate

| Name | Definition | Comments |
|---|---|---|
| **geo** | <**ciduri**> rdf:type <schema:GeoCoordinates> . | Type |
| | <**ciduri**> schema:latitute "*rala*" . | Latitude |
| | <**ciduri**> schema:longitude "*ralo*" . | Longitude |
| | <**ciduri**> ebg:geoResolution "*rallp*" . | Resolution |

### B.2.4 Misc. functions

| Name | Definition | Condition | Comments |
|---|---|---|---|
| **idval** | *rea* | *rea* != null and *cciaa* != null | Identifier value |
| | *id* | *rea* = null or *cciaa* = null | |
| **cotolang** | @en | *co* = *uk* | Country to language mapping. |
| | @it | *co* = *it* | |
| | ... | | |
| **cotojc** | GB | *co* = *uk* | Country to jurisdiction code |
| | IT | *co* = *it* | |
| | ,,, | | |
| **cotoaul1** | United Kingdom | *co* = *it* | Country to admin unit L1 |
| | Italia | *co* = *it* | |
| | ... | | |

# Appendix C Classifications

This appendix lists classifications that can be relevant to the project. It contains a more comprehensive overview than the one presented in Section 2.1.1.10:

- **Created by**: which organization created it
- **Name**: classification abbreviation
- **Subject**: the subject of classification
- **Lev**: number of levels (NOT digits)
- **Count**: (+ means leaves only)
- **Updated**: first & last update
- **Freq**: recent frequency of updates
- **Info**: additional resources for details

| Created by | Name | Subject | Lev | Count | Updated | Freq | Info |
|---|---|---|---|---|---|---|---|
| **World** | | | | | | | |
| UN | BEC | trade categories | | 20 | 1971-2003 | n/a | http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=10&Lg=1<br><br>http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_GENINFO_CLASS_5 |
| UN | CPC | product by activity | | | 2005-2015 | 3 | https://en.wikipedia.org/wiki/Central_Product_Classification<br><br>http://unstats.un.org/unsd/cr/registry/regrs.asp?Lg=1 |
| UN | FAOdef | products (agricultural) | | | | | http://www.fao.org/waicent/faoinfo/economic/faodef/faodefe.htm |
| UN | ISIC | industry | 4 | 768 | 1948-2011 | 7 | https://en.wikipedia.org/wiki/International_Standard_Industrial_Classification<br><br>http://unstats.un.org/unsd/cr/registry/isic-4.asp |
| UN | UNSPC | prod/serv | | | | | https://en.wikipedia.org/wiki/UNSPSC |
| UN | SITC | trade categories | | | 2009 | | http://unstats.un.org/unsd/trade/sitcrev4.htm |
| WCO | HS | prod (customs) | 6 | | | | https://en.wikipedia.org/wiki/Harmonized_Schedule_Number<br><br>https://en.wikipedia.org/wiki/Harmonized_Commodity_Description_and_Coding_System |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Supra-national** | | | | | | | |
| EC | CPA | prod by activity | | | | | https://en.wikipedia.org/wiki/Classification_of_Products_by_Activity |
| EC | CPV | prod/serv (procurement) | 8 | 9455 | 2008-2013 | 5 | https://en.wikipedia.org/wiki/Common_Procurement_Vocabulary<br><br>http://simap.ted.europa.eu/cpv |
| EC | NACE | industry | 4 | 615+ | | | https://en.wikipedia.org/wiki/Statistical_Classification_of_Economic_Activities_in_the_European_Community<br><br>http://epp.eurostat.ec.europa.eu/portal/page/portal/nace_rev2/introduction |
| EC | PRODCOM | prod | | | | | PRODCOM List 2016 (List of PRODucts of the European COMmunity) |
| North America | NAICS | industry | 6 | 19745+ | 1885-2012 | 10 | https://en.wikipedia.org/wiki/North_American_Industry_Classification_System |
| **National** | | | | | | | |
| AU | ANZSIC | industry | | | 2006 | | https://en.wikipedia.org/wiki/Australian_and_New_Zealand_Standard_Industrial_Classification<br><br>http://www.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/industrial-classification.aspx |
| BE | NACEBEL | | | | 2003-2008 | 5 | http://statbel.fgov.be/nl/statistieken/gegevensinzameling/nomenclaturen/nacebel/ |
| BG | NKID | industry | | | 2001-2008 | 5 | http://www.nsi.bg/sites/default/files/files/publications/KID-2008.pdf |
| FI | TOL | | | | 2002-2008 | 6 | http://tilastokeskus.fi/meta/luokitukset/toimiala/001-2002/kuvaus_en.html |
| FR | NAF | | | | 2008 | | http://www.insee.fr/fr/methodes/default.asp?page=nomenclatures/naf2008/naf2008.htm |
| GL | GB | | | | 2000-2005 | 3 | http://www.stat.gl/publ/kl/AR/200501/pdf/2005-mi%20suliassaqartitsineq.pdf |
| IN | NIC | industry | | | | | http://www.mospi.gov.in/classification/national-industrial-classification<br><br>http://nicode.su/ |
| IN | MCA | | | | 2004-2009 | | http://www.mca.gov.in/MCA21/dca/efiling/NIC-2004_detail_19jan2009.pdf |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| JP | JSIC | industry | | | >2013 | | |
| LU | NACE LUX | | | | | | [http://www.environnement.public.lu/dechets/informations_pratiques/code_nace.pdf](http://www.environnement.public.lu/dechets/informations_pratiques/code_nace.pdf) |
| ME | KD | | | | 2010 | | http://www.monstat.org/eng/page.php?id=107&pageid=107 |
| NO | SIC | | | | 2007 | | http://stabas.ssb.no/ItemsFrames.asp?ID=8118001&Language=en&VersionLevel=classversion&MenuChoice=Language |
| NZ | BIC | | | | 2006 | | http://www.acc.co.nz/PRD_EXT_CSMP/idcplg?IdcService=GET_FILE&dID=108821&dDocName=WPC133749&allowInterrupt=1 |
| SE | SNI/SSIC | industry | | | | | https://en.wikipedia.org/wiki/Swedish_Standard_Industrial_Classification |
| UK | SIC | industry | | | 1992-2007 | 4 | http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/standard-industrial-classification/index.html |
| US | SIC | industry | 4 | 1004+ | 1937--1987 | n/a | https://en.wikipedia.org/wiki/Standard_Industrial_Classification |
| **Private (created by companies)** | | | | | | | |
| Dun & Bradstreet | SICext | industry | 8 | | current | | |
| FactSet | Revere | line-of-business | | 11000 | | | https://en.wikipedia.org/wiki/FactSet |
| FirstResearch | FS | line-of-business | | | | | http://www.firstresearch.com/ |
| FTSE | ICB | market | 4 | 114 | | | https://en.wikipedia.org/wiki/Industry_Classification_Benchmark |
| IndustryBuildingBlocks | IBB | line-of-business | | 15000 | | | http://industrybuildingblocks.com/ |
| S&P, Morgan Stanley | GICS | market | 4 | 154 | | | https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard |
| Thomson Reuters | TRBC | market | 4 | 124 | | | https://en.wikipedia.org/wiki/Thomson_Reuters_Business_Classification |
| WAND | PST | prod/serv | | | | | http://www.wandinc.com/wand-product-and-service-taxonomy.aspx |