eventeuBusinessGraph Semantic Data Model

Authors:

* euBusinessGraph Consortium

Version: 1.3

Last updated: 30 Sept 2019

# 1 Introduction

This document describes the semantic model used by euBusinessGraph (EBG) to represent companies, their attributes, addresses, directors and CEOs, datasets by the different providers, provenance. It includes an informative description of classes and properties, gives examples and data provider rules, provides schema and instance diagrams.

It also provides RDF bindings, i.e. classes and properties to be used for representation of company data. The github project [euBusinessGraph/eubg-data/model](https://github.com/euBusinessGraph/eubg-data/tree/master/model) provides a generated ontology and RDF shapes intended to validate submitted data (these are not yet complete).

## 1.1 Contents

Use the "Table of Contents" addon (Addons> Get addons) to regularly refresh the hierarchical section numbering (use its Refresh button in top right). Then refresh the Table of Contents below by clicking the Refresh button below.

[1 Introduction](#_Toc20845948)

[1.1 Contents](#_Toc20845949)

[1.2 Revision History](#_Toc20845950)

[1.3 Abbreviations](#_Toc20845951)

[1.4 Background Documents](#_Toc20845952)

[1.5 Descriptive Info](#_Toc20845953)

[1.6 euBusinessGraph URLs](#_Toc20845954)

[1.7 External References](#_Toc20845955)

[1.7.1 EC Core Vocabularies](#_Toc20845956)

[1.7.2 Schema](#_Toc20845957)

[1.7.3 Global Legal Entity Identifier](#_Toc20845958)

[1.7.3.1 Registration Authorities List](#_Toc20845959)

[1.7.3.2 Entity Legal Form Code](#_Toc20845960)

[1.7.4 Wikipedia Business Entities](#_Toc20845961)

[1.7.5 Wikidata External Identifiers](#_Toc20845962)

[1.7.6 Business Registers Interconnection System](#_Toc20845963)

[1.7.7 Dublin Core](#_Toc20845964)

[1.7.8 IANA Language Tags](#_Toc20845965)

[1.7.9 Financial Industry Business Ontology](#_Toc20845966)

[1.7.10 EC NUTS and LAU](#_Toc20845967)

[1.7.10.1 NUTS Downloads](#_Toc20845968)

[1.7.10.2 NUTS RDF](#_Toc20845969)

[1.7.10.3 LAU RDF](#_Toc20845970)

[1.7.11 NACE](#_Toc20845971)

[1.7.11.1 NACE RDF](#_Toc20845972)

[1.7.11.2 NACE Display](#_Toc20845973)

[1.7.11.3 NACE National Extensions](#_Toc20845974)

[1.7.12 Geonames](#_Toc20845975)

[1.8 Dataset Ontologies](#_Toc20845976)

[2 Business Data](#_Toc20845977)

[2.1 Company](#_Toc20845978)

[2.1.1 Names](#_Toc20845979)

[2.1.1.1 Name Language Discussion](#_Toc20845980)

[2.1.1.2 Legal Name](#_Toc20845981)

[2.1.1.3 Trade Name](#_Toc20845982)

[2.1.1.4 Preferred Name](#_Toc20845983)

[2.1.2 Classifications](#_Toc20845984)

[2.1.2.1 Type](#_Toc20845985)

[2.1.2.2 Type Text](#_Toc20845986)

[2.1.2.3 Is Startup](#_Toc20845987)

[2.1.2.4 Is State Owned](#_Toc20845988)

[2.1.2.5 Is Publicly Traded](#_Toc20845989)

[2.1.2.6 Status](#_Toc20845990)

[2.1.2.7 Status Text](#_Toc20845991)

[2.1.2.8 Economic Activity](#_Toc20845992)

[2.1.2.9 Economic Activity Text](#_Toc20845993)

[2.1.3 Other Details](#_Toc20845994)

[2.1.3.1 Incorporation Date](#_Toc20845995)

[2.1.3.2 Dissolution Date](#_Toc20845996)

[2.1.3.3 Language](#_Toc20845997)

[2.1.4 Online Presence](#_Toc20845998)

[2.1.4.1 Website](#_Toc20845999)

[2.1.4.2 Wikipedia Page](#_Toc20846000)

[2.1.4.3 Certified Email](#_Toc20846001)

[2.1.4.4 News/Blog Feed](#_Toc20846002)

[2.2 Jurisdiction/Addresses/Sites](#_Toc20846003)

[2.2.1 Jurisdiction](#_Toc20846004)

[2.2.2 Address Discussion](#_Toc20846005)

[2.2.2.1 Address Data](#_Toc20846006)

[2.2.2.2 Schema Address Props](#_Toc20846007)

[2.2.2.3 locn Address Props](#_Toc20846008)

[2.2.2.4 locn/schema Geographic Coordinates](#_Toc20846009)

[2.2.2.5 Org Site Class](#_Toc20846010)

[2.2.2.6 Address Calculations](#_Toc20846011)

[2.2.3 Address](#_Toc20846012)

[2.2.3.1 Full Address](#_Toc20846013)

[2.2.3.2 Address Country](#_Toc20846014)

[2.2.3.3 Macroregion](#_Toc20846015)

[2.2.3.4 Administrative Region](#_Toc20846016)

[2.2.3.5 Province](#_Toc20846017)

[2.2.3.6 Municipality/Commune/Settlement](#_Toc20846018)

[2.2.3.7 Locality/City/Settlement](#_Toc20846019)

[2.2.3.8 Neighbourhood/Quarter](#_Toc20846020)

[2.2.3.9 Street Address](#_Toc20846021)

[2.2.3.10 Street Number](#_Toc20846022)

[2.2.3.11 Postal Code](#_Toc20846023)

[2.2.3.12 Postal Office box](#_Toc20846024)

[2.2.3.13 Geographic Coordinates](#_Toc20846025)

[2.2.3.13.1 Latitude, Longitude](#_Toc20846026)

[2.2.3.13.2 Geo Coordinate Resolution](#_Toc20846027)

[2.2.4 Registered Address](#_Toc20846028)

[2.2.5 Other Address](#_Toc20846029)

[2.2.6 Event](#_Toc20846030)

[2.2.6.1 EventValue](#_Toc20846031)

[2.2.6.2 EventType](#_Toc20846032)

[2.3 Company Model Diagrams](#_Toc20846033)

[2.3.1 Company Instance Diagram](#_Toc20846034)

[2.3.2 Company JSON-LD](#_Toc20846035)

[2.3.3 Company RDF Shapes](#_Toc20846036)

[2.3.4 Company ORM Schema](#_Toc20846037)

[2.3.4.1 Company ORM](#_Toc20846038)

[2.3.4.2 Names ORM](#_Toc20846039)

[2.3.4.3 Classifications ORM](#_Toc20846040)

[2.3.4.4 Other Details ORM](#_Toc20846041)

[2.3.4.5 Online Presence ORM](#_Toc20846042)

[2.3.4.6 Jurisdiction/Addresses/Sites ORM](#_Toc20846043)

[3 Identifiers and Identifier Systems](#_Toc20846044)

[3.1 Identifiers ORM Diagram](#_Toc20846045)

[3.2 Identifier Systems](#_Toc20846046)

[3.2.1 Identifier System Agents](#_Toc20846047)

[3.2.1.1 Identifier System Author](#_Toc20846048)

[3.2.1.2 Identifier System Issuer](#_Toc20846049)

[3.2.1.3 Identifier System Publisher](#_Toc20846050)

[3.2.2 Identifier System Properties](#_Toc20846051)

[3.2.2.1 Identifier System Jurisdiction](#_Toc20846052)

[3.2.2.2 Identifier System Code](#_Toc20846053)

[3.2.2.3 Identifier System RAL Code](#_Toc20846054)

[3.2.2.4 Identifier System Name](#_Toc20846055)

[3.2.2.5 Identifier System Website](#_Toc20846056)

[3.2.2.6 Identifier System License](#_Toc20846057)

[3.2.3 Identifier System Characteristics](#_Toc20846058)

[3.2.3.1 Has Unique Identifiers](#_Toc20846059)

[3.2.3.2 Is Single-Valued](#_Toc20846060)

[3.2.3.3 Has Persistent Identifiers](#_Toc20846061)

[3.2.3.4 Has Immutable Identifiers](#_Toc20846062)

[3.2.3.5 Is Public](#_Toc20846063)

[3.2.3.6 Has Dumb Identifiers](#_Toc20846064)

[3.2.3.7 Has Enumerated Identifiers](#_Toc20846065)

[3.2.3.8 Identifier System Official in Jurisdiction](#_Toc20846066)

[3.2.4 Identifier Validation Properties](#_Toc20846067)

[3.2.4.1 Validation Rule](#_Toc20846068)

[3.2.4.2 Validation Regex](#_Toc20846069)

[3.2.4.3 Replacement Pattern](#_Toc20846070)

[3.2.5 Identifier Kinds Discussion](#_Toc20846071)

[3.3 Web Resources](#_Toc20846072)

[3.3.1 Web Resource](#_Toc20846073)

[3.3.1.1 Identifier System Web Resource](#_Toc20846074)

[3.3.1.2 Web Resource URL](#_Toc20846075)

[3.3.1.3 Web Resource Name](#_Toc20846076)

[3.3.1.4 Web Resource MIME Type](#_Toc20846077)

[3.3.1.5 Web Resource Language](#_Toc20846078)

[3.3.2 Identifier Web Resource](#_Toc20846079)

[3.3.2.1 Identifier Web Resource](#_Toc20846080)

[3.3.2.2 Web Resource URL Template](#_Toc20846081)

[3.4 Company Identifier](#_Toc20846082)

[3.4.1 Identifier Value](#_Toc20846083)

[3.4.2 Identifier Part Of System](#_Toc20846084)

[3.4.3 Identifier Issuer](#_Toc20846085)

[3.4.4 Identifier Issued On](#_Toc20846086)

[3.4.5 Identifier Expires On](#_Toc20846087)

[3.4.6 Official Identifier](#_Toc20846088)

[3.4.7 Company Identifier](#_Toc20846089)

[4 Datasets](#_Toc20846090)

[4.1 Dataset Hierarchy](#_Toc20846091)

[4.2 Dataset Metadata](#_Toc20846092)

[4.3 Dataset Sources](#_Toc20846093)

[4.4 Company to Dataset Links](#_Toc20846094)

[4.5 Dataset Statistics and Partitions](#_Toc20846095)

[5 Officers and Membership](#_Toc20846096)

[5.1 Officer](#_Toc20846097)

[5.1.1 Officer Name](#_Toc20846098)

[5.1.2 Officer Address](#_Toc20846099)

[5.2 Membership](#_Toc20846100)

[5.2.1 Membership Company](#_Toc20846101)

[5.2.2 Member Person](#_Toc20846102)

[5.2.3 Membership Interval](#_Toc20846103)

[5.2.4 Membership Role](#_Toc20846104)

[6 SCRATCH](#_Toc20846105)

[6.1 Descriptive Template](#_Toc20846106)

## 1.2 Revision History

19 July 2017

* Initial version

27 Jul 2017

* Move [NACE](#_ns56oyboj30d) lookup info to its own section with 2 subsections
* Describe chance that [Economic Activity](#_ejdfl7tx154d) may be described in a more complex way

31 Jul 2017

* Added examples in [NACE RDF](#_a5yh4l8jgdb0)
* Added section [NACE Display](#_ywk8o1a6pvao)

6 Aug 2017

* Added section [Status Text](#_zgzb95gdgnv5)
* Completed [Abbreviations](#_bm47yrjr9vqw)
* Moved [Geo Coordinate Resolution](#_fvvq191h4sy7) values to a google sheet in preparation for generating RDF
* Many editorial fixes (eg there were two sections "Revision History")

7 Aug 2017

* Clarified how to represent Registration Date, as distinct from [Incorporation Date](#_46pgy1vmzn6q)

8 Sep 2017

* Added [Economic Activity Text](#_53fiy7z04ap5)

1.0, 28 Feb 2018

* Added [Identifiers and Identifier Systems](#_hofh07qhoz6m)
* Added [Dataset](#_bffa3ligs8w6) descriptions

1.1, 26 Oct 2018

* Revised [Dataset](#_bffa3ligs8w6) descriptions to remove the free/paid distinction

1.2, 15 Mar 2019

* Added void:inDataset links for each company

1.3, 21 Mar 2019

* Use literal in dct:spatial instead of URL: "GB" instead of nuts:GB (which does not exist, it's nuts:UK), see [4.5 Dataset Statistics and Partitions](https://docs.google.com/document/d/1dhMOTlIOC6dOK_jksJRX0CB-GIRoiYY6fWtCnZArUhU/edit#heading=h.l8elmhubgcu6)

## 1.3 Abbreviations

The list below excludes common country and currency abbreviations, as well as local abbreviations used only once and whose meaning is clear from the context.

|  |  |
| --- | --- |
| **Abbr** | **Description** |
| API | [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) |
| ASCII | [American Standard Code for Information Interchange](https://en.wikipedia.org/wiki/ASCII) |
| ATECO | [ATtività ECOnomiche](https://it.wikipedia.org/wiki/ATECO), Italian extension of NACE |
| Atom | [Atom](https://en.wikipedia.org/wiki/Atom_(standard)), an XML-based web syndication (feed) format |
| BRC | [Brønnøysund Register Centre](https://www.brreg.no/home/), a Norwegian data provider in EBG |
| BRIS | [Business Registers Interconnection System](https://e-justice.europa.eu/content_business_registers_at_european_level-105-en.do), an EC DG Justice project to integrate register data across the EEA and some other European countries |
| CC | [EU Candidate Countries](https://en.wikipedia.org/wiki/Future_enlargement_of_the_European_Union) |
| CDF | GLEIF [Common Data File format](https://www.gleif.org/en/about-lei/common-data-file-format) |
| CSV | [Comma Separated Values](https://en.wikipedia.org/wiki/Comma-separated_values), format for storing tabular data in text form |
| CTO | [Chief Technology Officer](https://en.wikipedia.org/wiki/Chief_technology_officer) |
| DC | [Dublin Core](https://en.wikipedia.org/wiki/Dublin_Core), a small set of metadata terms including author, publisher, date created, date published, etc. ISO 15836 standard. |
| DG | [Directorate General](https://en.wikipedia.org/wiki/Directorate-General), one of the "ministries" of the EC |
| DW | [Deutsche Welle Innovation](http://blogs.dw.com/innovation/), an EBG business case partner |
| EBG | [euBusinessGraph](http://eubusinessgraph.eu/), a H2020 project to develop a graph of company and related data |
| EC | European Commission |
| EEA | [European Economic Area](https://en.wikipedia.org/wiki/European_Economic_Area): EU plus Iceland, Liechtenstein and Norway |
| EFTA | [European Free Trade Association](https://en.wikipedia.org/wiki/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](https://www.iso.org/standard/67462.html) Code, an authoritative list of company types. ISO 20275 standard. |
| EU | [European Union](https://en.wikipedia.org/wiki/European_Union) |
| FIBO | [Financial Industry Business Ontology](https://www.edmcouncil.org/financialbusiness) |
| FTS | [Full-Text Search](https://en.wikipedia.org/wiki/Full-text_search) |
| GLEI | [Global Legal Entity Identifier](https://www.gleif.org/), same as LEI. A global register of financial and other companies, standard ISO 17442, intended to increase financial transparency and stability. |
| GLEIF | [GLEI Foundation](https://www.gleif.org/), developer of GLEI and related ISO standards |
| GML | [Geography Markup Language](https://en.wikipedia.org/wiki/Geography_Markup_Language): XML language for representing vector geometry objects |
| HTML | [Hypertext Markup Language](https://en.wikipedia.org/wiki/HTML) |
| HTTP | [Hypertext Transfer Protocol](https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol) |
| IANA | [Internet Assigned Numbers Authority](https://en.wikipedia.org/wiki/Internet_Assigned_Numbers_Authority). Keeps registers of various entities, including [IANA language tags](https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry) |
| ICAO | [International Civil Aviation Organization](https://en.wikipedia.org/wiki/International_Civil_Aviation_Organization) |
| ID | Identifier |
| IMHO | In my humble opinion |
| INSPIRE | [Infrastructure for Spatial Information in the European Community](https://en.wikipedia.org/wiki/Infrastructure_for_Spatial_Information_in_the_European_Community) |
| ISA2 | [Interoperability solutions for public administrations, businesses and citizens](https://ec.europa.eu/isa2/home_en): an EC program for eGovernment interoperability |
| ISIC | United Nations [International Standard Industrial Classification](https://en.wikipedia.org/wiki/International_Standard_Industrial_Classification) of economic activities |
| ISO | [International Organization for Standardization](https://en.wikipedia.org/wiki/International_Organization_for_Standardization) |
| JSON-LD | [JavaScript Object Notation for Linked Data](https://en.wikipedia.org/wiki/JSON-LD), a web-developer-friendly RDF serialization. EBG has selected it as internal data transmission format |
| LAU | [Local Administrative Unit](http://ec.europa.eu/eurostat/web/nuts/local-administrative-units), 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](https://en.wikipedia.org/wiki/Linked_Open_Data), using semantic web technologies to link open data |
| LOV | [Linked Open Vocabularies](http://lov.okfn.org), a catalog of ontologies |
| NACE | [Nomenclature Statistique des activités économiques](https://en.wikipedia.org/wiki/Statistical_Classification_of_Economic_Activities_in_the_European_Community), Statistical Classification of Economic Activities in the European Community by Eurostat |
| NACEBEL | [Belgian extension](http://statbel.fgov.be/nl/statistieken/gegevensinzameling/nomenclaturen/nacebel/) of NACE |
| NKID | [Национална Класификация на Икономическите Дейности](https://www.mi.government.bg/files/useruploads/files/sme/kid_2008.pdf), Bulgarian version of NACE |
| NUTS | [Nomenclature of Territorial Units for Statistics](https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics) by Eurostat |
| OCORP | [OpenCorporates](https://opencorporates.com/), content and use case partner in EBG. Has open data on 130M companies crawled from official registers. |
| ONTO | [Ontotext](https://ontotext.com/), a semantic technology partner in EBG |
| ORGS | [The Organization Ontology](https://www.w3.org/TR/vocab-org/) a semantic ontology to describe organizations |
| ORM | [Object-Role Modeling](https://en.wikipedia.org/wiki/Object-role_modeling), a modeling methodology. SINTEF uses it through the tool NORMA[[1]](#footnote-1) that is available as a plugin for Microsoft Visual Studio to model data and to generate an RDF rendition of the EBG ontology |
| PDF | [Portable Document Format](https://en.wikipedia.org/wiki/Portable_Document_Format) |
| PP | [Is proper part of](http://geovocab.org/spatial#PP), a relation in the [NeoGeo Spatial Ontology](http://geovocab.org/spatial) |
| RA, RAL | GLEI [Registration Authorities List](https://www.gleif.org/en/about-lei/gleif-registration-authorities-list) |
| RAMON | [Reference And Management Of Nomenclatures](http://ec.europa.eu/eurostat/ramon), Eurostat's metadata server |
| RDF | [Resource Description Framework](https://en.wikipedia.org/wiki/Resource_Description_Framework), the graph data model of the semantic web |
| RDFS | [RDF Schema](https://en.wikipedia.org/wiki/RDF_Schema), a vocabulary for describing subclasses, subproperties and property domains and ranges |
| RSS | [Rich Site Summary](https://en.wikipedia.org/wiki/RSS) or Really Simple Syndication, a type of web feed |
| SDATI | [SpazioDati](http://www.spaziodati.eu/), content and business case partner in EBG. Has rich data on Italian companies integrated from various sources ([Atoka.io](https://atoka.io/)) |
| SEM | [The Simple Event Model Ontology](https://semanticweb.cs.vu.nl/2009/11/sem/) an ontology to describe simple events |
| SEMIC | [Semantic Interoperability Centre Europe](https://en.wikipedia.org/wiki/Semantic_Interoperability_Centre_Europe) |
| SHACL | [Shapes Constraint Language](https://en.wikipedia.org/wiki/SHACL), a language for expressing RDF Shapes. A W3C specification |
| SINTEF | [A Norwegian research organization](https://www.sintef.no/en/), coordinator of EBG |
| SKOS | [Simple Knowledge Organization System](https://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System), the ontology for representing lookup lists: controlled vocabularies, thesauri, classification schemes, taxonomies, subject-heading systems |
| SPARQL | [SPARQL Protocol and RDF Query Language](https://en.wikipedia.org/wiki/SPARQL), the query language of the semantic web |
| ShEx | [Shape expressions](https://en.wikipedia.org/wiki/Shape_expression), a language for expressing RDF Shapes. Specification by a W3C Community Group |
| TERCET | Eurostat [Territorial classification and typology](http://ec.europa.eu/eurostat/tercet/): provides search for NUTS and LAU and additional data about postal codes, distances, etc |
| TOC | Table of Contents |
| TODO | Tag indicating some incomplete task (to do) |
| Turtle | [Terse RDF Triple Language](https://en.wikipedia.org/wiki/Turtle_(syntax)), a human-readable RDF serialization |
| UN | [United Nations](https://en.wikipedia.org/wiki/United_Nations) |
| UN/CEFACT | United Nations [Centre for Trade Facilitation and Electronic Business](https://en.wikipedia.org/wiki/UN/CEFACT) |
| URI | [Uniform Resource Identifier](https://en.wikipedia.org/wiki/Uniform_Resource_Identifier) |
| URL | [Uniform Resource Locator](https://en.wikipedia.org/wiki/URL), 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](https://en.wikipedia.org/wiki/World_Geodetic_System): reference frame for the Earth for use in geodesy and navigation |
| WKT | [Well-known text](https://en.wikipedia.org/wiki/Well-known_text): text language for representing vector geometry objects |
| XLS | The [Microsoft Excel file format](https://en.wikipedia.org/wiki/Microsoft_Excel_file_format) |
| XML | [Extensible Markup Language](https://en.wikipedia.org/wiki/XML) |
| XSD | [XML Schema Definition](https://en.wikipedia.org/wiki/XML_Schema_(W3C)), a way to describe XML documents formally |
| rdfpuml | RDF visualization tool by ONTO [[2]](#footnote-2) |
|  |  |
|  |  |

## 1.4 Background Documents

[Google folder](https://drive.google.com/drive/u/1/folders/0B3niP30AbzzOalNTeFY1bHppdnM)

* [EBG model fields](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit): data collection spreadsheet. Includes fields, descriptions, and extra tabs for some lookup fields
* [EBG model notes](https://docs.google.com/document/d/1mrARYqxs0cSSj-G-zAfdJDQnxD049LV44CYOmXuEiNs/edit): initial working document with discussion on specific fields and some modeling approaches
* [Conceptual modelling requirements](https://docs.google.com/document/d/1yHmJBmCGFJFB6Cr3HBKN4s4Uz23qbO7YoVSrMl1laww/edit): data requirements from different use case partners. Some of the cases pose data requirements that may not be covered (e.g. DW would like to have NetWorth, Profit/Loss, number of employees)
* [EBG model team](https://docs.google.com/document/d/1SIWXwccIsQ-xtWP4fuCgMC5QY3_POZsUFWYsxocTynM/edit): please use these emails to post tasks to collaborators. Write your comments in the document body, not in comments (unless you **want** them to disappear when resolved)

[DropBox T2.2](https://www.dropbox.com/home/H2020-euBusinessGraph/work/wp/WP2/T2.2)

* vocabulary/: ORM and generated Turtle drafts by SINTEF
* EU-core-vocabs/: Org, RegOrg, Person and a few others (see [EC Core Vocabularies](#_j9cc8980vbtx))
  + uses/: press releases about their use in BRIS and by OCORP. Sample OCORP data with diagram

## 1.5 Descriptive Info

The data model consists of terms and the connections between them. Terms include:

* Class (entity),
* Property (field)
  + Object property (relation)
  + Lookup property (relation to a lookup value, typically skos:Concept)
  + Data property (literal)
* Value (fixed lookup values, also called concepts).

Each term is described using the following pieces of info. We give in parentheses the RDF representation of each piece

* Class / Property / Value (rdfs:label): name
* Range / Data Type (schema:rangeIncludes): what kind of value the property should take
  + The kinds of values could be multiple. That's why we don't use rdfs:range but the more permissive schema:rangeIncludes
  + 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 what to choose from
* Cardinality: how many values the property can have: \* any, + many, 1 one, ? zero or one
* RDF (prefixed URL): suggested RDF binding (class or property)
* Description (skos:definition): term description, preferably from an established source (see below)
* Scope Note (skos:scopeNote): how the term is used or other clarifications
* Rules (skos:scopeNote): data rules that providers must conform to
* Examples (skos:example): examples

For classes we also include

* URL: description how to form URLs for entities of that class

For properties we also include:

* OCORP, BRC, SDATI, etc: source fields in data provider systems
* OCORP example, BRC example, SDATI example, etc: example values from data provider systems

For lookup values, we may include:

* Parent (skos:broader): broader concept (hypernym)

In addition, major groupings of terms (e.g. a class with its pertaining properties) may include the following sections:

* Schema Diagram: data layout modeled in ORM
* Instance Diagram: precise example of RDF representation
* Queries: queries to find and get information. For each class of entity, we first formulate some target queries in English, then implement them in SPARQL.
* Discussion: sometimes long discussions that do not (yet) have an impact on the data model. If resolved positively, the corresponding terms and descriptions will be incorporated in the body of the data model.

## 1.6 euBusinessGraph URLs

Experience shows that due consideration should be given to forming URLs that are sensible, permanent, and comply to strict rules. Leaving URL formation to the imagination of each data provider is a bad practice. There are several documented best practices that we follow [[3]](#footnote-3) [[4]](#footnote-4) [[5]](#footnote-5) [[6]](#footnote-6).

* The EBG base URL is <http://data.businessgraph.io/> and supports content negotiation (i.e. can serve HTML and several semantic representations for each resource).
  + May still be subject to revision for marketing reasons
  + We leave <http://businessgraph.io/> available for general project information, that's why we use the "data" hostname
* Individuals, including companies, addresses, lookup (thesaurus) values are in this namespace, with an appropriate word in singular. They are defined further in the document, e.g.
  + company/GB/07444723 (OpenCorporates), company/US\_DE/4888154 (Ontotext USA), company/Atoka/6da785b3adf2
  + type/GB/PLC
  + status/OCORP/active
* The EBG ontology is at [http://data.businessgraph.io/ontology#](http://data.businessgraph.io/ontology) and is used with prefix ebg:
* We maintain the namespaces of all EBG and external (referenced) RDF vocabularies in the file **prefixes.ttl** in this folder. There is no need to repeat them here.

## 1.7 External References

We try to use terms and term descriptions from the following sources, approximately in the given order. We select terms according to how well they cover our business need.

Please reference the source of each term or description in brackets, and provide a hyperlink to a textual source whenever possible (eg [[schema:PostalAddress](http://schema.org/PostalAddress)]).

### 1.7.1 EC Core Vocabularies

The [e-Government Core Vocabularies](https://joinup.ec.europa.eu/asset/core_vocabularies/asset_release/core-vocabularies-v20#download-links) were developed by the [ISA2 SEMIC Joinup](https://joinup.ec.europa.eu/community/semic/description) semantic interoperability initiative and then standardized by the W3C. They include basic data about the following classes (amongst others). The first link is to a descriptive document, and the second to the namespace document:

* [Organization](http://www.w3.org/TR/vocab-org/) ([org:](http://www.w3.org/ns/org)): formal or informal organizations, classification, related people, events.
* [Registered Organization](http://www.w3.org/TR/vocab-regorg/) ([rov:](https://www.w3.org/ns/regorg)): organizations that are registered in some register.
* [Person](https://joinup.ec.europa.eu/asset/core_person/asset_release/core-person-vocabulary#download-links) ([person:](http://www.w3.org/ns/person)). Does not have a descriptive doc at W3C ([vocab-people](https://www.w3.org/TR/vocab-people/) is a different thing), so the link is to the EC descriptive PDF.
* [Location](https://www.w3.org/ns/locn) ([locn:](https://www.w3.org/ns/locn)): addresses and geographic locations. The namespace is also a descriptive doc at W3C.

Also see [Handbook on using the Core Vocabularies](https://joinup.ec.europa.eu/site/core_vocabularies/Core_Vocabularies_user_handbook/ISA%20Hanbook%20for%20using%20Core%20Vocabularies.pdf).

### 1.7.2 Schema

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

* [Schema.org](http://schema.org/) ([schema:](http://schema.org/))

It is highly reusable because it makes few ontological commitments, having to cater to a truly global audience of millions of web sites. See [[7]](#footnote-7) for details:

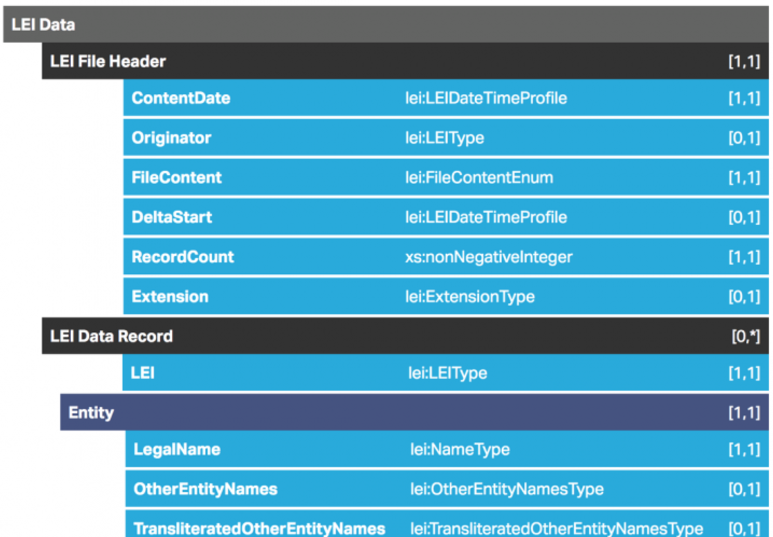
* Rather than rdfs:domain and rdfs:range, which are mono-morphic i.e. admit only a single class and thus force rigid and deep class hierarchies, it uses schema:domainIncludes and schema:rangeIncludes, which are purely advisory.
* While owl:DatatypeProperty and owl:ObjectProperty (or a range being a class)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

### 1.7.3 Global Legal Entity Identifier

The GLEI Foundation has established a registration structure to issue [Legal Entity Identifiers (LEI)](https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei) to legal entities participating in financial transactions. The LEI structure is standardized as [ISO 17442](https://www.gleif.org/en/about-lei/iso-17442-the-lei-code-structure). Accompanying entity data is standardized by GLEIF as the [Common Data File](https://www.gleif.org/en/about-lei/common-data-file-format) (CDF) formats. CDF consists of two levels:

* [Level 1](https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-1-data-who-is-who): "who is who"
* [Level 2](https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom): "who owns whom". This consists of:
  + Relationship Records (ownership when the parent has a LEI, with qualifying and quantifying attributes),
  + Parent Reference Data (ownership relations when the parent doesn't have a LEI therefore more of its descriptive data is included)
  + Reporting Exceptions (why owners are not reported).
  + It was introduced in May 2017 and as of June 2017, there are only 9k ownership relations (and 68k reporting exceptions)

We will mostly be concerned with [LEI-CDF Version 2.1](https://www.gleif.org/en/about-lei/common-data-file-format/lei-cdf-format/lei-cdf-format-version-2-1) (May 2017). An informative data layout diagram is shown at that page (a small part is reproduced below).



* [[lei-cdf:](https://www.gleif.org/content/2-about-lei/5-common-data-file-format/1-lei-cdf-format/2-lei-cdf-format-version-2-1/2017-03-21_lei-cdf-v2-1.pdf)] v2.1: [PDF description](https://www.gleif.org/content/2-about-lei/5-common-data-file-format/1-lei-cdf-format/2-lei-cdf-format-version-2-1/2017-03-21_lei-cdf-v2-1.pdf), [XML XSD Schema](https://www.gleif.org/content/2-about-lei/5-common-data-file-format/1-lei-cdf-format/2-lei-cdf-format-version-2-1/2017-03-21_lei-cdf-v2-1.xsd), [RelaxNG Compact Schema](https://github.com/Ontotext-AD/GLEI/blob/master/schema/20170321_lei-cdf-v2-1-nodoc.rnc) (ONTO)

ONTO has made a draft mapping and conversion of GLEI to RDF using the FIBO ontology, see <https://github.com/Ontotext-AD/GLEI>.

LEI also includes two codelists that may be useful to EBG, see next.

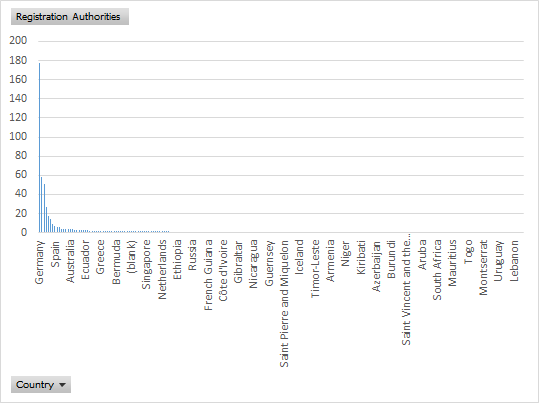
#### 1.7.3.1 Registration Authorities List

* [[lei-ral:](https://www.gleif.org/en/about-lei/gleif-registration-authorities-list)] (PDF, CSV, XLS).
* [RAL google sheet](https://docs.google.com/spreadsheets/d/1u3-rSC95FdoR7_nac6kvAca9axWOqWit6vXQs5CRKHQ/edit)

Includes 651 national official registers. The descriptive columns are:

* Registration Authority Code
* Country
* Country Code
* Jurisdiction (country or region)
* International name of Register
* Local name of Register
* International name of organisation responsible for the Register
* Local name of organisation responsible for the Register
* Website

There are 60 countries with several national registers.



* Germany has the most: 177:
  + A separate Commercial Register (Handelsregister) for every local court,
  + A central Company Register (Unternehmensregister), with the note: IMPORTANT: when this source is used, the name of the local court must be added to the field BUSINESS REGISTER ENTITY ID
  + Federal Financial Supervisory Authority
* United Kingdom's Companies Register by the Companies House is listed 3 times under different jurisdictions.
  + England and Wales
  + Northern Ireland
  + Scotland

This is quite an interesting view since there is one single register: company IDs are unique across these 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% |

E.g. Commissione di vigilanza sui fondi pensione has only "Local name of organisation responsible for the Register": its register doesn't have its own name, and the commission's name is not yet translated to English.

Only two codes lack all four "name" fields: RA888888 and RA999999, and they indicate "no data".

As of 20170608, the Consolidated LEI data file has:

* 287981 entities with specific RegistrationAuthority
* 200179 entities with RA999999 no registration authority available for this legal entity
* 17287 entities with RA888888 temporary code; new registration authority code requested

Because of this last item, we can expect the RA list to grow somewhat in the near future.

#### 1.7.3.2 Entity Legal Form Code

* [lei-elf:] [ISO 20275: Entity Legal Forms Code List](https://www.gleif.org/en/about-lei/iso-20275-entity-legal-forms-code-list)

ELF will resolve variant names for each valid legal form within a jurisdiction to a single code per legal form. While it will not unify company types across jurisdictions, it will provide a single authoritative list of all company types. It is ISO 20275 and is maintained by GLEIF.

* GLEI states "There will be a period of transition during which the LOUs will implement the ELF code list (likely until 1 year). As of 2017-06-08, the Consolidated LEI file has no real ELF codes:
* 501120 entities have 8888 "temporary code", the other
* 4327 entities have 9999 "not available for this entity"
* ELF was released in Nov 2017. It uses "randomly" generated codes that have nothing in common with established national codes. Furthermore, there are some omissions and mistakes in the list (see discussion in FIBO issue [BE-141](https://jira.edmcouncil.org/browse/BE-141) Legal Form), and we contacted GLEI with feedback but received no reply. So we preferred to work with codes generated within the project. For example:
  + Uses "random" codes (e.g. “946C”), which is not a good idea because they are impossible to remember. Established national abbreviations should have been used (e.g. “BG-ET”).
  + ELF does not provide an English translation, so its value for a non-English speaking user is diminished
  + BG TR has 23 types, ELF has only 13 codes for BG
  + Uses some obsolete abbreviations, eg “S-ie” is an obsolete form from 50-70 years ago, now only “SD” is used
  + Misspellings in two types: Командиртно дружество с акции, Командиртно дружество.
  + Some lack capitalization: европейско дружество, европейско кооперативно дружество

### 1.7.4 Wikipedia Business Entities

* [wp-lbe:] Wikipedia world-wide [List of Business Entities](https://en.wikipedia.org/wiki/List_of_business_entities).

It's a useful resource since it includes a translation to English and "approximate equivalents in the company law of English-speaking countries". It doesn't describe all types: comparing to the [EBG list of legal types](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit#gid=1134528076), Wikipedia includes:

* BG: 9 of 23
* NO: 18 of 43

### 1.7.5 Wikidata External Identifiers

Wikidata has a datatype "external-id" used to link Wikidata entities to external databases. As of Feb 2018, there are [2496 external-id properties](https://www.wikidata.org/w/index.php?title=Special:ListProperties/external-id&limit=500&offset=2000) on Wikidata; of which 1000 are covered in the [Mix-n-Match](https://tools.wmflabs.org/mix-n-match/#/) crowd-sourced coreference development tool. Wikidata properties are defined in a comprehensive way and there is an elaborate constraint checking system, some of which has provided inspiration for our description of Identifier Systems. For example, take property [P245 ULAN ID](https://www.wikidata.org/wiki/Property:P245) that links to the Getty Union List of Artist Names (an Authority file for artists, sitters and other people):

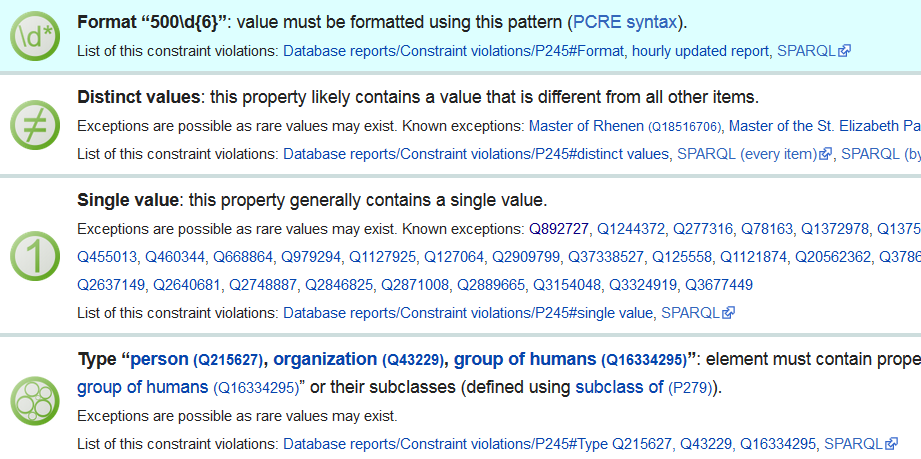
* "Issued by: Getty Research Institute" shows the author of that identifier system, and issuer of identifiers that are part of it.
* [formatter URL](https://www.wikidata.org/wiki/Property:P1630) "http://vocab.getty.edu/page/ulan/$1" and [formatter URI of the RDF resources](https://www.wikidata.org/wiki/Property:P1921) "http://vocab.getty.edu/ulan/$1" provide URL templates for web resources related to every entity using the identifier (HTML and RDF respectively). "$1" indicates where the identifier value should be inserted
* [format as a regular expression](https://www.wikidata.org/wiki/Property:P1793) "500\d{6}" provides a simple validation mechanism, and the link to [https://regex101.com/](https://regex101.com/?regex=500%5Cd%7B6%7D) is a tool where values can be tested against that regex.
* [stability of property value](https://www.wikidata.org/wiki/Property:P2668) "sometimes changes" indicates whether the identifier is expected to be permanent (Getty records are sometimes merged, and the recessive ID is kept as Obsolete for a couple of years).

[Property constraint](https://www.wikidata.org/wiki/Property:P2302) captures a number of constraint checking mechanisms:

* [distinct values constraint](https://www.wikidata.org/wiki/Q21502410) indicates that ID values are expected to identify unique entities.
  + An [exception to constraint](https://www.wikidata.org/wiki/Property:P2303) is recorded for [Master of Rhenen](https://www.wikidata.org/wiki/Q18516706) and [Master of the St. Elizabeth Panels](https://www.wikidata.org/wiki/Q18516719). Wikidata has two different entities, but art researchers contributing to ULAN have decided that this unknown master is in fact the same (the "hand" of an artist can be identified through art style research).
* [single value constraint](https://www.wikidata.org/wiki/Q19474404) indicates that an entity is expected to have only one ID in that system.
  + Again some [exception to constraint](https://www.wikidata.org/wiki/Property:P2303) are recorded. E.g. the [Bonnefanten Museum](https://www.wikidata.org/wiki/Q892727#P245) has two ULAN entries; examination shows they should be merged by the Getty: one is Maastricht Bonnefantenmuseum, the other is described as Bonnefantenmuseum (Dutch repository, Maastricht, contemporary)

The [property discussion page](https://www.wikidata.org/wiki/Property_talk:P245) includes further details, eg:

* [Current uses](https://query.wikidata.org/#SELECT%20%3Fitem%20%3FitemLabel%20%3Fvalue%20%3FvalueLabel%0A%7B%0A%09%3Fitem%20wdt%3AP245%20%3Fvalue%20.%0A%09SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22en%2Cen%22%20%20%7D%20%20%20%20%0A%7D%0ALIMIT%201000) – [Statistics by class](https://query.wikidata.org/#%23top%2050%20for%20P31%20and%20P279%20of%20items%20using%20P245%0ASELECT%20%3Fclass%20%3FclassLabel%20%3Fcount%20%3Fuse_as_Label%0A%7B%09%7B%09SELECT%20%3Fclass%20%28COUNT%28*%29%20AS%20%3Fcount%29%20%28wd%3AP31%20as%20%3Fuse_as_%29%0A%09%09%7B%09%3Fa%09wdt%3AP245%09%3Fp%09%3B%20wdt%3AP31%09%3Fclass%7D%0A%20%20%20%20%20%20%20%20GROUP%20BY%20%3Fclass%20ORDER%20BY%20DESC%28%3Fcount%29%20LIMIT%2050%0A%20%20%09%7D%0A%20%20%09UNION%0A%09%7B%09SELECT%20%3Fclass%20%28COUNT%28*%29%20AS%20%3Fcount%29%20%28wd%3AP279%20as%20%3Fuse_as_%29%0A%09%09%7B%09%3Fa%09wdt%3AP245%09%3Fp%09%3B%20wdt%3AP279%09%3Fclass%7D%0A%20%20%20%20%20%20%20%20GROUP%20BY%20%3Fclass%20ORDER%20BY%20DESC%28%3Fcount%29%20LIMIT%2050%0A%20%20%09%7D%0A%20%20%09SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22en%2Cen%22%20%7D%0A%7D%0AORDER%20BY%20DESC%28%3Fcount%29%20%3Fclass) – [String length](https://query.wikidata.org/#%23String%20lenghts%20by%20frequency%20with%20sample%20value%0A%23by%20Jura1%2C%202016-01-24%2C%202017-07-28%0A%0ASELECT%20%3Fstringlength%20%28COUNT%28DISTINCT%28%3Fstringvalue%29%29%20AS%20%3Fcount%29%20%28SAMPLE%28%3Fstringvalue%29%20as%20%3Fsamplestringvalue%29%20%20%28SAMPLE%28%3Fitem%29%20as%20%3Fsampleitem%29%0A%28URI%28CONCAT%28%22https%3A%2F%2Fquery.wikidata.org%2F%23SELECT%2520%253Fitem%2520%253FitemLabel%2520%253FitemDescription%2520%253Fstringvalue%2520%253Fstringlength%250A%22%2C%0A%20%20%20%20%20%20%20%20%20%20%20%20%22WHERE%2520%250A%257B%250A%2509%253Fitem%2520wdt%253A%22%2C%22P245%22%2C%22%2520%253Fstringvalue%2520.%250A%2520%2520%2520%2520BIND%2528strlen%2528%253Fstringvalue%2529%2520as%2520%253F%22%2C%0A%20%20%20%20%20%20%20%20%20%20%20%20%22stringlength%2529%250A%2520%2520%2520%2520FILTER%2528%253Fstringlength%2520%253D%2520%22%2Cstr%28%3Fstringlength%29%2C%22%2529%250A%2520%2520%2520%2520SERVICE%2520wikibase%253Alabel%2520%257B%2520bd%253A%22%2C%0A%20%20%20%20%20%20%20%20%20%20%20%20%22serviceParam%2520wikibase%253Alanguage%2520%2522%255BAUTO_LANGUAGE%255D%252Cen%2522.%2520%257D%250A%257D%250ALIMIT%2520100%250A%22%29%29%20as%20%3Flist%29%0AWHERE%0A%7B%0A%09%3Fitem%20wdt%3AP245%20%3Fstringvalue%20.%0A%20%20%20%20BIND%28strlen%28%3Fstringvalue%29%20as%20%3Fstringlength%29%0A%7D%0AGROUP%20BY%20%3Fstringlength%0AORDER%20BY%20DESC%28%3Fcount%29%0ALIMIT%2010) (for ULAN should be fixed to 9) –[Count](about:blank)
* Links to Database Constraint violation reports: [Format](https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P245#Format) (regex), [distinct values](https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P245#distinct_values), [single value](https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P245#single_value)



* Additional constraints that are not mandatory but likely, eg   
  **Probably death:** People with [ULAN ID (P245)](https://www.wikidata.org/wiki/Property:P245) born before the year 1900, but no [date of death (P570)](https://www.wikidata.org/wiki/Property:P570)
* A template that groups this with other Authority control properties that are used often



* A table of submitted [Error reports](https://www.wikidata.org/wiki/Property_talk:P245#Error_reports), together with columns to track the resulting data flow

All these tools show the value of identifiers in working out coreferencing and data flows between different datasets. We have used several of these features when designing our model for identifiers.

### 1.7.6 Business Registers Interconnection System

This project interconnects business registers across Europe and provides a single (though limited) [search form](https://e-justice.europa.eu/content_find_a_company-489-en.do). See [EBG confluence](https://confluence.ontotext.com/display/EBG/BRIS+EBR) for details. We should liaise and attempt to obtain two data elements:

* [bris-lf:] List of legal forms. They are provided in the national language, but there is rich explanation in English, e.g.:
  + "societa’ a responsabilità limitata": The limited liability company (societa’ 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.
* [bris-ra:] List of national registers, e.g.
  + IT RI (Italy's Registro delle imprese)
  + FI FPRO (Finland's Patentti-ja Rekisterihallitus)
* [bris-euid:] A pan-European company identifier formed of the register ID and company ID, e.g.
  + ITRI.02866370170 for an Italian company
  + FIFPRO.1670700-0 for a Finnish company

### 1.7.7 Dublin Core

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

* [DC Elements](http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#H3) ([dc:](http://purl.org/dc/elements/1.1/))
* [DC Terms](http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#H2) ([dct:](http://purl.org/dc/terms/))

When a property is present in both dc: and dct: we use dc: for a literal and dct: for a resource. An example of a class we could need is e.g. [dct:Jurisdiction](http://purl.org/dc/terms/Jurisdiction).

### 1.7.8 IANA Language Tags

* [[iana:](https://docs.google.com/spreadsheets/d/1M1yv9aBUmc-NyCJX69vOLUmH2uIglSwmDwgRgByI1AI/edit)] [IANA language tag](https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry) list

The IANA language code registry 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 a rdf:langString literal). The first link is a [google sheet](https://docs.google.com/spreadsheets/d/1M1yv9aBUmc-NyCJX69vOLUmH2uIglSwmDwgRgByI1AI/edit) that provides 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, but hopefully we won't need this.

### 1.7.9 Financial Industry Business Ontology

* [[fibo:](https://spec.edmcouncil.org/fibo/)] [latest ontology](https://spec.edmcouncil.org/fibo/ontology/master/latest/)

FIBO is a very complex set of ontologies for the financial industry that formalizes entities like companies, directors, ownership and control relations, business registers, monetary amounts, debts, obligations, contracts, financial instruments, etc etc. It is available in Turtle, RDF and JSON-LD.

* Consists of 30 domains and 448 Turtle files: 361 ontology, 59 "about" and 28 "test"
* Uses 505 prefixes: 404 internal ontologies, 61 "about", 25 "test", 15 external ontologies
* Defines 7215 ObjectProperties and 301 DatatypeProperties

Some simple parts of FIBO are also [mapped to Schema](http://schema.org/docs/financial.html).

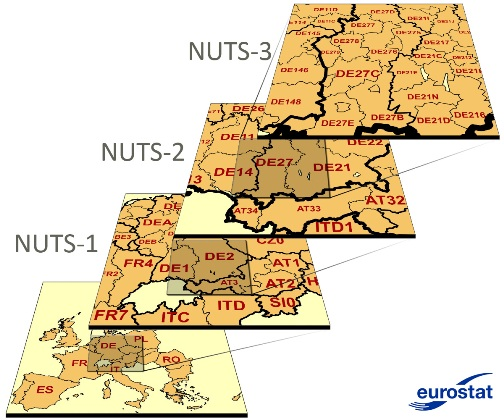
We don't use it directly but refer to it for modeling inspiration for some of the EBG entities.

### 1.7.10 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](https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics): NUTS, 3 levels
* [Local Administrative Units](https://en.wikipedia.org/wiki/Local_administrative_unit): LAU, 2 levels

The image below is from [NUTS Overview](http://ec.europa.eu/eurostat/web/nuts/overview):



NUTS is an important geographic resource for EBG since:

* Significant Open Data is available that 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](http://ec.europa.eu/eurostat/web/nuts/national-structures-eu) and [non-EU](http://ec.europa.eu/eurostat/web/nuts/correspondence-tables/national-structures-non-eu) (EFTA+CC), also see a [cleaned up consolidation](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit?pli=1#gid=1485440757) by EBG or a [simplified version](https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics#Levels) including only NUTS levels by wikipedia.

* Total number of regions:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Country group** | **NUTS1** | **NUTS2** | **NUTS3** | **LAU1** | **LAU2** |
| EU-28 | 98 | 276 | 1343 |  | 118495 |
| EFTA-4 (IS, LI, NO, CH) | 4 | 16 | 48 | 265 | 3242 |
| Candidate Countries-5 (ME, MK, AL, RS, TR) | 15 | 31 | 102 | 1028 | 41080 |
| **Total** | **117** | **323** | **1493** | **8695** | **162817** |

* A sampling of region types per country:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Country** | **NUTS1** | **NUTS2** | **NUTS3** | **LAU1** | **LAU2** |
| AT | Gruppen von Bundesländern | Bundesländer | Gruppen von Gemeinden | - | Gemeinden |
| BG | Зони (Zoni) | Райони (Rajoni) | Области  (Oblasti) | Общини  (Obshtini) | Населени места  (Naseleni mesta) |
| DE | Länder | Regierungsbezirke | Kreise | Verwaltungs-  gemeinschaften | Gemeinden |
| IT | Gruppi di regioni | Regioni | Provincie | - | Comuni |
| SI | - | Kohezijske regije | Statistične regije | Upravne enote | Občine |
| NO | - | Landsdeler | Fylker | Økonomiske  regioner | Kommuner |
| UK | Government Office Regions; Country | Counties (some grouped); Inner and Outer London; Groups of unitary  authorities | Upper tier authorities or groups of lower tier authorities (unitary  authorities or districts) | Lower tier authorities (districts) or individual unitary authorities;  Individual unitary authorities or LECs (or parts thereof); Districts | Wards (or parts thereof) |

Smaller countries have few NUTS1 regions (e.g. BG has BG3, BG4) or even a singleton (eg CZ0, LI0, NO0).

#### 1.7.10.1 NUTS Downloads

* NUTS1..3 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), including:  
  NUTS code, label (in national language and transliterated to Latin), level, country, country order, global sorting order
* [NUTS1..2 translation](http://publications.europa.eu/code/en/en-5001000.htm) to English, e.g.   
  BG41 Югозападен (Yugozapaden) South-West  
  CZ08 Moravskoslezsko Moravian Silesia
* [Detailed LAU info](http://ec.europa.eu/eurostat/web/nuts/local-administrative-units) (we use [Correspondence table LAU 2 – NUTS 2013, EU-28](http://ec.europa.eu/eurostat/documents/345175/501971/EU-28_LAU_2016+%28zip%29/6131e000-381a-42a1-b661-4edddd47cf40)), including:
  + NUTS3, LAU1, LAU2 (these are local codes), name1 in national language, name2 Latinized, population, area.
  + Doesn't have info for EFTA including NO. Could be extracted from [List\_of\_municipalities\_of\_Norway](https://en.wikipedia.org/wiki/List_of_municipalities_of_Norway)
  + Doesn't have LAU1 names for CY, FR, IE, LT. We got FR Canton names from Wikidata, but it's a bit of a mess. The other countries have fewer LAU1, so it's manageable
  + Name2 Latinized is:
    - Transliteration for non-Latin alphabets e.g. BG Абланица-Ablanitsa, EL Δημοτική Κοινότητα Κομοτηνής-Municipal Commune of Komotini, CY
    - Second official name for countries that have such, e.g. BE Antwerpen-Anvers, MT Valletta-Citta' Umilissima
    - A preposition if that's an acceptable variant, e.g. FR Boisse - (La)
    - Missing for LU, NL
    - Same as name1 for all other countries
  + Historic info for 2010-2016, flag whether there is a change
* [NUTS3 typology](http://ec.europa.eu/eurostat/documents/345175/6807882/Ttypologies+and+local+information+corresponding+to+NUTS3.xls) (urban, rural, coastal region) and local info (main city, ports)

[NUTS TERCET flat files](http://ec.europa.eu/eurostat/tercet/flatfiles.do) downloads (also see [NUTS TERCET search](http://ec.europa.eu/eurostat/tercet/) engine):

* Postal codes mapped to NUTS3.
  + Includes NO
  + "Due to provisions on copyright protection post codes are not available for ES, FR, PL, PT, SE. For post code - NUTS links in these countries please contact the relevant national post office."
* Airports: ICAO code, name, country, NUTS3
* Road distances between NUTS3, including distance to travel in each intervening country
* Air distance from each airport to each NUTS3

#### 1.7.10.2 NUTS RDF

The EC project PlanetData has provided part of NUTS in RDF at <http://nuts.geovocab.org/>, using the [NeoGeo RDF Vocabulary for GeoData](http://geovocab.org/) (see [datahub description](https://datahub.io/dataset/nuts-geovocab)).

The data includes:

* Administrative region boundaries (coordinates)
* GeoSPARQL spatial relations (e.g. country includes regions, country has neighbors, etc)
* GeoSPARQL services based on latitude/longitude: NUTS regions that include a certain point, and NUTS regions that are contained within a specified bounding box.
* Correspondences to a number of other datasets. E.g. http://nuts.geovocab.org/id/BG is declared owl:sameAs the following resources:   
  <http://estatwrap.ontologycentral.com/dic/geo#BG>,   
  <http://nuts.psi.enakting.org/id/BG>, <http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/Bulgaria>, <http://aims.fao.org/aos/geopolitical.owl#Bulgaria>,   
  <http://rod.eionet.europa.eu/spatial/7>,   
  <http://dbpedia.org/resource/Bulgaria>,   
  <http://sws.geonames.org/732800/>,   
  <http://ec.europa.eu/eurostat/ramon/rdfdata/countries/BG>, <http://rdfdata.eionet.europa.eu/iso3166/countries/BG>, <http://rdfdata.eionet.europa.eu/ramon/nuts2008/BG>, <http://rdfdata.eionet.europa.eu/ramon/nuts2003/BG>
  + You can check for yourself how many of these return data…

E.g. see the data for Sofia, BG: <http://nuts.geovocab.org/id/BG412.html>

TODO: check whether this data corresponds to NUTS 2013, or an older version

* nuts: <http://nuts.geovocab.org/id/>

#### 1.7.10.3 LAU RDF

EBG should make LAU RDF, using [EC NUTS and LAU](#_fn3icsx21uas) data and following the [NUTS RDF](#_wezm1c5p6t7g) representation.

* We should nest LAU1,2 regions under the respective NUTS3 regions.
* We should record population and area, which can support interesting queries like "give me companies in a large city"

Consider these examples (indentation shows the administrative hierarchy):

* nuts:ITH20 "Trento" province=NUTS3
  + <lau/IT-022006> "Arco" commune=LAU2
* nuts:BG311 "Vidin" oblast=NUTS3
  + <lau/BG-VID09> "Vidin" obshtina=LAU1
    - <lau/BG-10971> "Vidin" city=LAU2
  + <lau/BG-VID16> "Dimovo" obshtina=LAU1
    - <lau/BG-00672> "Archar" village=LAU2

We represent the nodes shown in blue like this. You can see that BG uses level LAU1,2 while IT uses only LAU2. Population and area are available only at the lowest level (settlement).

lau:IT-022006 a ebg:LAURegion;

void:inDataset lau: ;

rdfs:label "IT-022006 Arco";

ramon:code "IT-022006";

ramon:name "Arco"@en, "Arco"@it;

ramon:level 5; spatial:PP nuts:ITH20;

dbo:populationTotal 17526; dbo:areaTotal 63218400.

lau:BG-VID09 a ebg:LAURegion;

void:inDataset lau: ;

rdfs:label "BG-VID09 Vidin";

ramon:code "BG-VID09";

ramon:name "Vidin"@en, "Видин"@bg;

ramon:level 4; spatial:PP nuts:BG311.

lau:BG-10971 a ebg:LAURegion;

void:inDataset lau: ;

rdfs:label "BG-10971 Vidin";

ramon:code "BG-10971";

ramon:name "Vidin"@en, "Видин"@bg;

ramon:level 5; spatial:PP lau:BG-VID09;

dbo:populationTotal 43595; dbo:areaTotal 63218000.

This representation uses the following terms.

* ebg:LAURegion: Local Administrative Unit. We use [specific LAU type names](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit?pli=1#gid=1485440757) as examples, see ebg-ontology-part.ttl
* [spatial:PP](http://geovocab.org/spatial#PP): is proper part of
* [ramon:code](http://rdfdata.eionet.europa.eu/page/ramon/ontology/code): NUTS/LAU Code of an administrative region
* [ramon:level](http://rdfdata.eionet.europa.eu/page/ramon/ontology/level): Level of an administrative region: NUTS1..3=1..3, LAU1..2=4..5
* [ramon:name](http://rdfdata.eionet.europa.eu/page/ramon/ontology/name): Name of an administrative region or settlement
* [dbo:populationTotal](http://dbpedia.org/ontology/populationTotal): Number of people inhabiting the administrative region or settlement
* [dbo:areaTotal](http://dbpedia.org/ontology/areaTotal): Area of the administrative region or settlement, in square meters

### 1.7.11 NACE

A number of classification vocabularies exist. See [EBG Confluence](https://confluence.ontotext.com/display/EBG/Economic+Classifications) for a detailed discussion.

* [UN ISIC](https://unstats.un.org/unsd/cr/registry/isic-4.asp) 4 ([detailed info](http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27)) is the international classification that many others are modeled after.
* [EC NACE](http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=NACE_REV2&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC) 2 is the preferred choice for European interoperability. EU National classifications are required to map to NACE since that is how they report to Eurostat

NACE data:http:%2F%2Flinkedlifedata.com%2Fresource%2Fchebi&role=context

* The [master NACE data as HTML or CSV](http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=NACE_REV2&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC) is available from EC RAMON. It includes the following fields. See eg [NACE 01.11 at unstat](https://unstats.un.org/unsd//cr/registry/regcssm.asp?Cl=242&Lg=1&Co=01.11) for an example of these 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](http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&file=dic%2Fen%2Fnace_r2.dic) including just Code and Label.
  + Does not include hierarchy
  + Includes additional groupings, e.g. nace:B-D\_X\_FOOD is classes B..D except food, i.e. "Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply (except food, beverages and tobacco)"

#### 1.7.11.1 NACE RDF

* RAMON includes a [NACE Ontology](http://ec.europa.eu/eurostat/ramon/ontologies/nace.rdf) that includes some of the master fields, but is not compatible with SKOS
* [EstatWrap](http://estatwrap.ontologycentral.com/dic/nace_r2) and [Linked Statistics](http://eurostat.linked-statistics.org/dic/nace_r2) 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
  + Doesn't include any hierarchy (no skos:broader, the concept scheme has skos:hasTopConcept to every concept)
  + Doesn't have different classes for the different levels (which is ok) and for the additional groupings (which is not ok).
  + There are only labels, but not detailed descriptions ("This item includes").
  + Labels are only in English (but the RAMON master data is also only in English)
  + Doesn't specify a base URL so it's unclear whether the base is [http://estatwrap.ontologycentral.com/dic/nace\_r2#](http://estatwrap.ontologycentral.com/dic/nace_r2) or [http://ontologycentral.com/2009/01/eurostat/ns/nace\_r2#](http://ontologycentral.com/2009/01/eurostat/ns/nace_r2) or a locally saved file
* [EIONet](http://rdfdata.eionet.europa.eu/eurostatdic/nace_r2) and RAMON had NACE data in the past, but not anymore.

EBG made a more complete NACE RDF using most of the data above. 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

Mapping:

nace:(Code) a skos:Concept; skos:inScheme nace:;

skos:notation "(Code)"; skos:broader nace:(Parent);

ebg:order "(Order)"^^xsd:integer; ebg:level "(Level)"^^xsd:integer;

skos:prefLabel "(Description)"@en; skos:scopeNote "(Includes)"@en;

ebg:exclusionNote "(Excludes)"@en.

Example:

nace:F a skos:Concept; skos:inScheme nace: ; skos:notation "F"; ebg:order 398951; ebg:level 1;

skos:prefLabel "CONSTRUCTION"@en; skos:note "This section includes… also includes …"@en;

ebg:exclusionNote "If these activities are carried out not for later sale…"@en.

nace:42 a skos:Concept; skos:inScheme nace: ; skos:notation "42"; ebg:order 398957; ebg:level 2;

skos:broader nace:F;

skos:prefLabel "Civil engineering"@en; skos:note "This division includes… ".

nace:42.1 a skos:Concept; skos:inScheme nace: ; skos:notation "42.1"; ebg:order 398958; ebg:level 3;

skos:broader nace:42;

skos:prefLabel "Construction of roads and railways"@en.

nace:42.11 a skos:Concept; skos:inScheme nace: ; skos:notation "42.11"; ebg:order 398959; ebg:level 4;

skos:broader nace:42.1;

skos:prefLabel "Construction of roads and motorways"@en; skos:scopeNote "This class includes: …";

ebg:exclusionNote "This class excludes: …".

#### 1.7.11.2 NACE Display

We could use the NACE RDF text fields to create a popup with rich info about every NACE node. Please note they include newlines that should be preserved on display. Alternatively, we can use this link at UN Stats: <https://unstats.un.org/unsd/cr/registry/regcssm.asp?Cl=242&Lg=1&Co=> e.g. [01](https://unstats.un.org/unsd/cr/registry/regcssm.asp?Cl=242&Lg=1&Co=01), [01.11](https://unstats.un.org/unsd/cr/registry/regcssm.asp?Cl=242&Lg=1&Co=01.11):

|  |
| --- |
| **NACE Rev.2**  **Class 01.11 - Growing of cereals (except rice), leguminous crops and oil seeds**  This class includes all forms of growing of cereals, leguminous crops and oil seeds in open fields. The growing of these crops is often combined within agricultural units.  This class includes:  - growing of cereals such as:  · wheat  ...  · sesame seed  · sunflower seed  · other oil seeds  *This class excludes:*  *- growing of rice, see 01.12*  *- growing of sweet corn, see 01.13*  *- growing of maize for fodder, see 01.19*  *- growing of oleaginous fruits, see 01.26* |

If we make our own NACE display, we should seek to discover NACE codes and turn them into hyperlinks.

#### 1.7.11.3 NACE National Extensions

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

* [BE NACEBEL](https://unstats.un.org/unsd/cr/ctryreg/ctrydetail.asp?id=1029) 2008 ([homepage](http://statbel.fgov.be/nl/statistieken/gegevensinzameling/nomenclaturen/nacebel/)) extends NACE with Level 5, which includes 943 Belgian subclasses, eg see extensions of class [43.32](https://kbopub.economie.fgov.be/kbopub/naceToelichting.html?nace.code=43320), eg
  + F43 Specialised construction activities
    - F43.3 Building completion and finishing
      * F43.32 Joinery installation (Travaux de menuiserie): **extended with**
        + F43.32.001 Montage de menuiseries extérieures et intérieures: portes, fenêtres, escaliers, placards de cuisines équipées, équipements pour magasins, dormants de portes et fenêtres, etc.
        + F43.32.002 Montage de cloisons mobiles; revêtement de murs, de plafonds, etc.
* [BG KID](https://unstats.un.org/unsd/cr/ctryreg/ctrydetail.asp?id=1040) 2008 ([BG PDF](http://www.nsi.bg/sites/default/files/files/publications/KID-2008.pdf), [interactive search](https://www.kik-info.com/spravochnik/kid-2008.php)) translates NACE to BG and doesn't seem to extend it
* [IT ATECO](https://unstats.un.org/unsd/cr/ctryreg/ctrydetail.asp?id=1147) 2007 ([PDF](https://www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007)) extends NACE as appropriate for the Italian economy, eg
  + C14 Manufacture of wearing apparel
    - C14.1 Manufacture of wearing apparel, except fur apparelC14.19 Manufacture of other wearing apparel and accessorie
    - s: **extended with**
      * + C14.19.2: sportswear and special clothing

C14.19.29: production of clothing for infants, track suits, ski suits, swimwear and the like

* [UK SIC](https://unstats.un.org/unsd/cr/ctryreg/ctrydetail.asp?id=1263) 2007 ([EN PDF](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/455263/SIC_codes_V2.pdf)) is based on NACE 2 and extends it with UK codes

For now, we'll use only standard NACE in English, not national translations or extensions.

* This means that providers should national extension trailing digits
* If some providers want to provide national extensions or translation, please give us the data so we can convert it to RDF

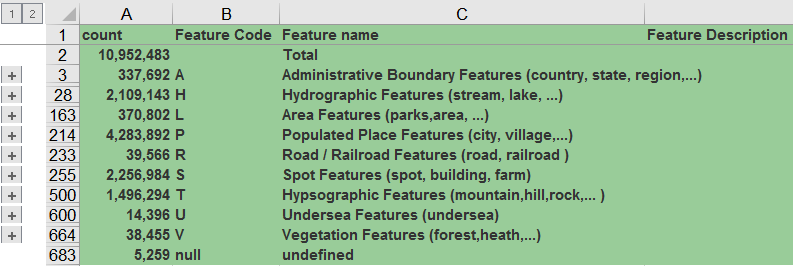
### 1.7.12 Geonames

[Geonames](http://www.geonames.org/) is an important place dataset of 11M places, including:

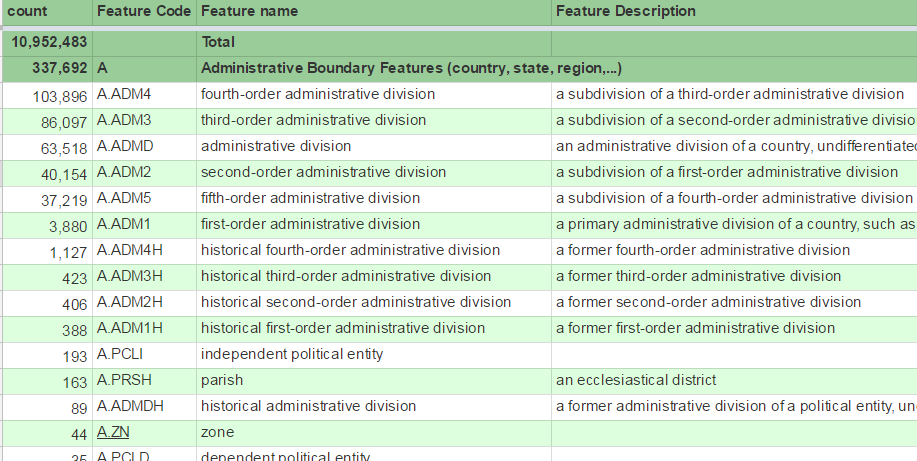
* Administrative regions and settlements
* Physical places (continents, mountains, oceans, seas, lakes, streams)
* Area features (oil fields, parks)
* Linear features (roads, power lines)
* Spot features (buildings, farms, hotels, oil wells)

Geonames has a 2-level hierarchy of feature types: Class>Code.

* Number of feature Classes as of Mar 2017:



* Example of Administrative Region feature Codes:



<http://geotree.geonames.org/> shows the administrative hierarchy consisting of A.PCLI (countries) and A.ADM1..3, e.g.:



A **rough** correspondence between NUTS and the Geonames administrative hierarchy:

|  |  |  |  |
| --- | --- | --- | --- |
| **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, e.g. for Italy:



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

## 1.8 Dataset Ontologies

A main function of euBusinessGraph is to aggregate free company datasets from various providers, to advertise paid dataset extensions, and to describe adequately their internal structure and content (e.g. number of companies per jurisdiction).

There are several ontologies dedicated to describinig datasets:

* The [ADMS ontology](https://www.w3.org/TR/vocab-adms/) describes various interoperability assets, including xml schemats, generic data models, code lists, taxonomies, dictionaries, vocabularies. It was developed by EC SEMIC and standardized by the W3C. It is not dedicated to datasets and has no means for describing the internal structure of a dataset.
* The [DCAT ontology](https://www.w3.org/TR/vocab-dcat/) describes datasets, their distributions, and dataset catalogs. DCAT and its various profiles (DCAT-AP, DCAT-Stat, etc) are the foundation of open data description, especially in Europe. It was developed by EC SEMIC and standardized by the W3C. It has rich descriptive metadata (eg spatial and temporal coverage, keywords and themes, authors and publishers, distributions and their timestamps and formats, etc). However, it doesn't describe the internal structure of a dataset: there's only one property dcat:byteSize describing the size of a distribution, which is not useful in understanding dataset content or value.
* The W3C [Data Exchange Working Group](https://www.w3.org/2017/dxwg/) (DXWG) is working on a ["next-generation" DCAT](https://w3c.github.io/dxwg/dcat/), but at this time there is still a lot of work remaining. There is a [Use Cases and Requirements](https://w3c.github.io/dxwg/ucr/) (UCR) document in active development, and more [cases and issues](https://github.com/w3c/dxwg/issues) are being collected on github.
  + In particular, requirement [ID33 Summarization/Characterization of datasets](https://w3c.github.io/dxwg/ucr/#ID33) fits the euBusinessGraph needs quite closely "Summary/descriptive statistics that characterize a dataset are important elements to have a high-level overview of the dataset.
  + This is particularly important for datasets that are not publicly accessible, but whose access could be requested under certain conditions". ID7, RDSAT, RSS and [issue#161 Dataset size characteristics](https://github.com/w3c/dxwg/issues/161) raised by euBusinessGraph "Data consumers often need to know how many of what sort of entities are included in a dataset. In an aggregation scenario, different subsets (parts of a dataset) need to be expressed, eg because they come from different data providers. We need the ability to express the fine-grained content of a dataset:
    - Ability to express subsets of a dataset.
    - Describe subsets by kind of entity (e.g. Companies vs Events) and/or entity characteristics (e.g. Italian companies, Startups, Startups in Italy)
    - The kinds and characteristics should be expressed by URLs
    - Express the count of entities in a dataset or subset
    - Optionally, express other dataset size characteristics. E.g. in RDF context, that's number of triples and nodes".
* The [VOID ontology](https://www.w3.org/TR/void/) is widely used for describing RDF datasets and includes means for describing the internal structure (property and class partitions, entity and triples counts, etc). VOID makes no assertions regarding values (individuals in the object position of triples), e.g. you can't state that a dataset includes companies only from a certain jurisdiction (dbo:jurisdiction), economic classification (org:orgActivity) or province (ebg:adminUnitL3). One could use DCAT's "geographic coverage" (dct:spatialCoverage), but cannot be specific about which property is being restricted.
* [VOID-ext](http://ldf.fi/void-ext) extends the VoID vocabulary with a total of 18 statistics and 14 partitions. It can capture various interesting characteristics of an RDF dataset, eg average literal size, number of literals with a certain language tag (vext:languagePartition), total number of languages used in a dataset (vext:languages), etc. There are a number of partitions characterizing the object of triples: object class partition, object datatype partition, object IRI length partition, object language partition, object namespace partition, object partition. The last one allows to characterize a dataset "that contains **only triples** with a certain object", which nearly matches euBusinessGraph needs, but not quite.
* [SEVOD](http://www.w3.org/2015/03/sevod) (Semagow Vocabulary of interlinked Datasets) [[8]](#footnote-8) extends VOID with the ability to describe a hierarchy of multi-dimensional histogram buckets, each providing statistics about triple patterns or sets of joined triple patterns. It does it by describing when certain subsets are a complete and disjoint partition of a parent dataset, the joint statistics (selectivity) of non-star triple patterns. This is useful for federated querying frameworks that can use such statistics in query planning. SEVOD allows to describe an object using svd:objectRegexPattern (restriction on object literal or URL) or svd:objectVocabulary (the authority component, i.e. host name, of object URLs). Unfortunately svd:objectRegexPattern is not a less natural and economical way to describe an object (rather than using the value itself), and the published ontology is missing some properties described in the paper.

ADMS, DCAT and VOID are fairly compatible, and here are a few examples where they are used together:

* The [DataId ontology](https://github.com/dbpedia/DataId-Ontology/) [[9]](#footnote-9) [[10]](#footnote-10) uses VOID, DCAT and PROV to describe datasets and their physical instances more precisely, as well as their relation to agents like persons or organizations in regard to their rights and responsibilities. It is the foundation of recent DBpedia dump catalogs (eg the [2016-10\_dataid\_catalog.ttl](http://dataid.dbpedia.org/2016-10/2016-10_dataid_catalog.ttl) of DBpedia is 595k), has been applied to NLP metadata (CLARIN MetaShare and CMDI), and has extensions for linked data (linksets), data management plans, preservation.
* The [Descriptive Information](http://vocab.getty.edu/doc/#Descriptive_Information) of the Getty Vocabularies Linked Open Data is captured using all 3 ontologies.
* [Dataset Descriptions: HCLS Community Profile](https://www.w3.org/TR/hcls-dataset/) is a profile that combines VOID and DCAT for describing datasets in the Health Care and Life Science community.

A recent survey [[11]](#footnote-11) describes various tools for creating RDF dataset descriptions and collecting RDF statistics (called dataset characterization or dataset profiles), and a few ontologies for capturing this data. Such tools include SWOOGLE, RDFStats, LODStats, ProLod++, ExpLod, STRHist, SchemEx, voiDge.

After researching numerous dataset description ontologies, we settled on using VOID and eventually some of its extensions.

# 2 Business Data

In this section we describe "business payload" data, i.e. data that customers may want to obtain. This includes companies, classifications, addresses, directors, business registries, identifiers, events, etc

## 2.1 Company

* Class: RegisteredOrganization
* RDF: rov:RegisteredOrganization or org:OrganizationalUnit, self-link org:hasUnit or ebg:hasHQUnit & self-link org:unitOf
* URL:
  + If we have the official registration ID of the company, we use that  
    **company/<jurisdiction>/<id>**, e.g.   
    company company/GB/07444723 (OpenCorporates)  
    company/US\_DE/4888154 (Ontotext USA)
  + Otherwise we use an ID from a data provider (alternative business register)  
    **company/<provider>/<id>,** e.g.  
    company/Atoka/6da785b3adf2
* 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:]  
  A legal person or structure that is organized under the laws of any jurisdiction [lei-elf:]

A Registered Organization can be also an organizational unit which refers to a company. The organizational unit has almost all the characteristic of the Organization, for those characteristics which are not applicable for the Organizational Unit that information will be added in the description.

* Scope note: Registered organizations are the main entities that EBG 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 include, without limitation [lei-elf:]:
  + all financial intermediaries;
  + banks and finance companies;
  + all entities that issue equity, debt or other securities for other capital structures;
  + all entities listed on an exchange;
  + 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.);
  + all entities under the purview of a financial regulator and their affiliates, subsidiaries and holding companies;
  + counterparties to financial transactions.
* BRC: this is a complex area where we are also struggling ourselves. Our registries contain any organization that needs to be registered, hence also foundations and sole ownerships. Based on the definitions we can extract which organizational form that fits the definition.

### 2.1.1 Names

[[lei-cdf:](https://github.com/Ontotext-AD/GLEI/blob/master/schema/20170321_lei-cdf-v2-1-nodoc.rnc)NameType] uses two good practices that we also adopt:

* Includes an optional language tag (xml:lang) for each name
* The string is declared Tokenized500Type, which does not allow leading, trailing and two consecutive spaces

Discussion: [lei-cdf:] allows a number of different names to be provided. Only the **bold** ones have been adopted by EBG:

* **LegalName**
* ALTERNATIVE\_LANGUAGE\_LEGAL\_NAME
* **PREVIOUS\_LEGAL\_NAME**
* **TRADING\_OR\_OPERATING\_NAME**
* PREFERRED\_ASCII\_TRANSLITERATED\_LEGAL\_NAME (human-transliterated)
* AUTO\_ASCII\_TRANSLITERATED\_LEGAL\_NAME (automatically transliterated)

#### 2.1.1.1 Name Language Discussion

* Chris: In many cases the language is not known, so IMHO it will not be possible to emit xml:lang. We always store in the native characters in the register, and it's usually not clear what language this is. So in general, language won't be included from OCORP.
* Vlado: you could guess per country, e.g. if the company comes from the Italian register then the language is probably @it
* David: Vlado, this will not work for several Norwegian companies that operates e.g. in the oil industry, where most company names are english. E.g. is “Statoil” a Norwegian company name?

But BE has several official languages.

* Vlado: I notice at <https://opencorporates.com/companies/be/0820966428> you don't mark the language while at <https://kbopub.economie.fgov.be/kbopub/toonondernemingps.html?lang=fr&ondernemingsnummer=0820966428> they say "Dénomination en néerlandais". Could you crawl that too?

The benefit of knowing the language is that one could then use language-specific rules for removing "parasitic words" like company type ("Limited"), "advanced", "international", etc

#### 2.1.1.2 Legal Name

* Data Property: Legal Name
* Data Type: xsd:string or rdf:langString
* Cardinality: +
* RDF: rov:legalName
* Similar to: [schema:legalName](http://schema.org/legalName)
* 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:]
* Rules:
  + Provide a legal language tag (see [iana:] for a list) if possible. Cons: if a user wants to search by exact name, they also need to provide the lang tag. But they'll usually perform FTS (won't know the full exact name).
  + Don't allow leading, trailing and two consecutive spaces
* Examples:
  + "PROGIENE 2.000 I PROFESSIONISTI DELL'IGIENE S.R.L. ENUNCIABILE ANCHE: PROGIENE 2.000 S.R.L."@it
  + "CHRINON LTD" (no lang tag)
  + "Онтотекст"@bg
  + "Ontotext"@en
* OCORP: Legal name
* BRC: navn
* SDATI: base.legalName

#### 2.1.1.3 Trade Name

* Data Property: Trade Name
* Data Type: rdf:langString or xsd:string
* Cardinality: \*
* RDF: skos:altLabel
* Similar to: [schema:alternateName](http://schema.org/alternateName)
* Description: Informal/popular name of the company (also called Trading As) [ebg:], or former name.
* Scope Notes: 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:]. We also use this for former names.
* Rules:
  + Do not emit such name if it's equal to the legal name [ebg:].
  + Do not use this property to record translations of the primary legal name [rov:].
  + Provide a valid language tag if possible (see [iana:] for a list)
  + Don't allow leading, trailing and two consecutive spaces
* Examples: (parallelling the first two above)
  + "PROGIENE 2.000 I PROFESSIONISTI DELL'IGIENE SRL"@it
  + "OpenCorporates"@en
* SDATI: name
* BRC: we do not have this and it is unlikely that we will until our register is modernized in 2020-2021. However the issue has been raised recently due to very formal names for public agencies, e.g. ourselves which is called “Registerenheten i Brønnøysund”, but everyone knows us as “Brønnøysundregistrene” (at least somewhat shorter :). We will look into ways of collecting that information.

#### 2.1.1.4 Preferred Name

* Data Property: Name
* Data Type: rdf:langString or xsd:string
* Cardinality: 1
* RDF: skos:prefLabel
* 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.
* BRC: we do not have this. Vladimir: duplicate the legal name, or pick one of them if there are several.

### 2.1.2 Classifications

[org:] includes a classification field that is further split in [rov:] into three: type, status, activity.

#### 2.1.2.1 Type

* Lookup property: Type
* Cardinality: ?
* RDF: rov:orgType
* Similar to: lei:EntityLegalFormCode
* Description: Company Type (Legal Form of the entity), not suitable for Organizational Unit
* Scope Notes: There is no agreed set of company types that crosses borders [rov:]. After examination, EBG concluded there is no chance to "standardize" some shared super-types across jurisdictions. This said, 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. Eg for Italy: Società Di Capitale > Società a responsabilità limitata [ebg:]
* OCORP: Company Type
* SDATI: base.legalForms[] and base.govType
* BRC: orgform (code + description). In the future we will provide a link to a SKOS representation of all organisational forms.

Values:

* [[ebg-lf:](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit?pli=1#gid=1134528076)] EBG draft list of types per jurisdiction.
  + Each data partner should complete it for the countries it will provide data for. 2017-08-06: We have many jurisdictions from OCORP (not normalized), BG from ONTO, NO from BRC. Awaiting IT from SDATI.
  + I think we should strive to provide English translations to as many types as possible. They won't be legally precise but still will provide value to our users
* [lei-elf:] will provide a global list for all jurisdictions, but that is not yet available.
* [wp-lbe:] doesn't describe all types but is a useful resource nevertheless
* [bris-le:] includes cross-Europe company types, we may be able to obtain this list (see [EBG confluence](https://confluence.ontotext.com/display/EBG/BRIS+EBR#BRISEBR-BRISSearch) for more details)
* [[it-istat:](http://www.agid.gov.it/sites/default/files/documentazione/02_amm_adempienti_presenti_in_ipa_per_tipologia_istat_e_per_regione.pdf)] defines Italian public sector entity types
* The URLs below are in the [euBusinessGraph URLs](#_8y8dut2o2ton) namespace

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **URL** | **src** | **label (national)** | **description (English)** | **broader** |
| type/BG/OOD | BG | "Дружество с ограничена отговорност"@bg | "Limited Liability Company"@en | Ubbi Dubbi |
| type/BG/AD | BG | "Акционерно дружество"@bg | "Joint-Stock Company"@en |  |

#### 2.1.2.2 Type Text

* Data property: Type Text (Legal Form)
* Data type: rdf:langString
* Cardinality: ? (but strongly recommended)
* RDF: ebg:orgTypeText
* Same as: lei:OtherLegalForm
* Description: Current entity status as reported by a national company register
* Scope Note: Because of the difficulties of standardizing Type lookup, we include a free text field
* Examples:
  + "Private Limited Company"@en
  + "Дружество с ограничена отговорност"@bg
  + "Società a responsabilità limitata"@it
* Rules:
  + Include valid language tag (see [iana:] for list)
  + If Type is also present, it must express the same value as Type Text
* BRC: part of orgform

#### 2.1.2.3 Is Startup

* Data property: Is Startup?
* Data type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isStartup
* Description: Whether the company is a startup [ebg:]
* SDATI: base.startup
* Discussion: need to define more strictly. Source: <http://startup.registroimprese.it/>

#### 2.1.2.4 Is State Owned

* Data property: Is State Owned Organization?
* Data type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isStateOwned
* Description: Whether this organisation is owned by the government, a government agency, community, city or other public entity.
* Scope Notes: In many cases it's not possible to compute this attribute without access to the shareholder register, so it may be missing.
* Examples:
  + false for Spaziodati (a private company)
  + true for Autonomous Province of Trento (local government)
  + true for Statkraft (Norwegian state-owned company)
* SDATI: public
* BRC: not directly but can be deduced. Several values for orgform indicates various forms of public sector organizations (the state itself, state agency, municipality, regional, commercial company fully owned by one of the above). We also have institusjonellSektorkode= "Statsforvaltningen" (institutional sector code) which is a more direct approach but does not include state owned commercial companies.

Discussion on difficulty defining and computing this field:

* David: this flag is somewhat difficult. First need a clearer definition (e.g. is a business owned by the government, but fully acting as a commercial company like the Norwegian company Statkraft relevant?
* Vladimir: I think that both government branches, and commercial companies owned by the government, are relevant.
* David: And why only State and not local governments?
* Vladimir: Local governments are included, see Description. Previously was called "Publicly Owned" but Chris thought this may be confused with isPubliclyTraded (see below), so now it's isStateOwned
* David: there is a lot of gray areas. Some examples:

1. State government is major shareholder, in a publicly traded company: E.g. Statoil, Telenor
2. State government is a blocking minority shareholder, publicly traded: E.g. Norsk Hydro, Yara
3. State government is only owner/shareholder and company is a regular limited (norway: AS and ASA): E.g. Avinor AS, Flytoget AS, Entra ASA
4. State government is only owner and company is governed by a special law (Norway: SF). E.g. Statkraft, Statskog
5. State government is a minority shareholder: E.g. SAS AB
6. Local government is a major shareholder, in a publically traded company
7. Local government is a minor shareholder
8. Several local governments together own a company
9. Local government is only owner and company is governed by a special law (Norway: KF)
10. Several local governments together own a company governed by a special law (Norway: IKS)

* David: some of these cases (1, 2, 3, 5, 6, 7, 8) are currently not possible to decide from our registry, as you need the shareholder registry. We can only state it for clear-cut cases of ownership (4, 9, 10). And this is for the Norwegian jurisdiction only: others will have different rules. So at least we cannot say "no" to this attribute because we do not have enough information (open world).
* Vladimir: Of course: just because it’s Boolean doesn’t mean it's mandatory. If you can’t provide it in some cases (or even in a majority of cases), that is fine.

Discussion on need/motivation:

* David: what is the purpose of this attribute. It would be helpful to know why this is important information, and what quality you will expect.
* Vladimir: Cannot answer, will ask the mailing list… (SDATI proposed the field since they can provide it)
* Vladimir: I didn't propose this field so I don’t know the exact use case. But one purpose is fairly obvious. The majority of tenders are run by Government entities, and there are some suppliers that go only after Government entities.
* Chris: I fully agree with David. This is a complex area, and there's no clear accepted definition. There's also not a clear use case, and I think the driving force of the euBusinessGraph should be user need.
* Ben: If there is not a clear use case/requirement at this stage, I would say we should just drop it for now. Sounds like it would take a lot of effort to get it correct (and even then it probably wouldn't be 100% correct). It's easier to **add** later (when it's required) than have a suboptimal version now which we have to **update** later, possibly making breaking changes to the model. (This is a general point which applies to all properties in the model - if something is (a) hard to get right and (b) isn't required for now, why spend time doing it?)
* Vladimir: It's intuitively clear what "State owned" means, just because we find difficulties defining it precisely is not enough reason to kill it. (Which is the case for any "synthesized" info we'd like to include to aid the user, eg type, status). I hope nobody disputes that SOME data is better than no data, and that perfect is the enemy of the good.

not suitable for Organizational Unit

#### 2.1.2.5 Is Publicly Traded

* Data property: Is Publicly Traded Company?
* Data type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isPubliclyTraded
* Description: Whether the company is publicly traded (listed at a stock exchange) [ebg:]

not suitable for Organizational Unit

* SDATI: economics.public
* BRC: no such info

#### 2.1.2.6 Status

* Lookup property: Status
* Values: concept scheme <status/>, see below
* Cardinality: ?
* RDF: rov:orgStatus
* Similar to: lei:EntityStatus
* Description: flag that identifies whether a company is active or not [ebg:]. The operational and/or legal registration status of the entity [lei:]
* Scope Notes: 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 a long 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:]
  + 'Normal Activity': does appear to have cross-border usefulness and this should be used in preference to terms like 'trading' or 'operating' [rov:]
  + 'Actively trading' vs 'Dormant' vs 'Unregistered' could be considered the 3 main divisions of Status
* OCORP: Current Status
* SDATI: base.active
* BRC: konkurs=’J’. We are adding in our next release a response for lookups on companies that for various reasons have folded (on request from OCORP)

See [Status gsheet](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit#gid=1169560552) for values

#### 2.1.2.7 Status Text

* Data property: Status Text
* Data type: rdf:langString
* Cardinality: ?
* RDF: ebg:orgStatusText
* Description: Company status as it comes from the original register.
* Examples:
  + "Dissolved"@en, "Inactive"@en, "Revoked"@en
  + “Situation Normale”@fr, “En Liquidacion”@es, “Πτώχευση”@el
* Rules:
  + Include valid language tag (see [iana:] for list)
  + If both Status and Status Text are present, they must express compatible values (Status Text will be a finer-granularity value)

#### 2.1.2.8 Economic Activity

* Lookup property: Economic Activity
* Cardinality: \*
* RDF: rov:orgActivity
* Similar to: schema:isicV4, schema:naics
* Description: Economic activity of the organization (NACE code)
* Scope Notes: recorded using a controlled vocabulary: EC NACE 2. More detailed national classifications could also useful (e.g. IT ATECO, UK SIC, BG NKID) but are not supported for now
* Rules: each data provider must map codes that it uses to established URLs described below.
  + It may provide only the top-level classification or a detailed classification since we'll establish a skos:broader hierarchy.
  + There is no need for separate fields for
* SDATI: base.sicUk (UK SIC) and base.ateco[] (IT ATECO)
* BRC: we have NACE 2 with NO NACE national extension (extra digit)

Values: **nace:([A-Z]|[0-9]{2}|[0-9]{2}\.[0-9]{1,2})**, see [NACE RDF](#_a5yh4l8jgdb0). Rules:

* Providers must use exact codes of the specified form (dot separators, no extra spaces).
  + Or we could use a more uniform code (eg **[A-Z]\d{2,5}**) to make submission easier, but will need to do some segmenting on output
* Providers should drop eventual [NACE National Extensions](#_x04cr66uwejw) trailing digits
* If providers find difficulties mapping to NACE or need national levels, we should discuss this.

Discussion: possible extensions/complications:

* Distinction between Primary activity and Other activities

#### 2.1.2.9 Economic Activity Text

* Data property: Economic Activity
* Datatype: rdf:langString
* Cardinality: ?
* RDF: ebg:orgActivityText
* Description: Economic activity of the organization (free text)
* Scope Notes: Such field is widely used in some countries (eg BG) and may include much richer info, eg
  + NACE = 72.19 Scientific research and development in the field of natural, medical, agricultural and technical sciences, without biotechnology
  + 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, designing and manufacture of holographic optical elements with a wide spectrum of application, 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

### 2.1.3 Other Details

#### 2.1.3.1 Incorporation Date

* Data property: Incorporation Date
* Data type: xsd:date or xsd:gYearMonth or xsd:gYear
* Cardinality: ?
* RDF: [schema:foundingDate](http://schema.org/foundingDate)
* Description: Date when the organization was created
* Examples: "2010-11-18"^^xsd:date, "2010-11"^^xsd:gYearMonth, "2010"^^xsd:gYear
* BRC: brc:stiftelsesdato (founding)

Note: registration date should be represented as [Identifier Issued On](#_4p3g538cxon0) of the main company identifier.

#### 2.1.3.2 Dissolution Date

* Data property: Dissolution Date
* Data type: xsd:date or xsd:gYearMonth or xsd:gYear
* Cardinality: ?
* RDF: [schema:dissolutionDate](http://schema.org/dissolutionDate)
* Similar to: lei-cdf:EntityExpirationDate
* Description: Date when the organization was dissolved or removed from the primary register
* Examples: "2010-11-18"^^xsd:date, "2010-11"^^xsd:gYearMonth, "2010"^^xsd:gYear
* BRC: no

#### 2.1.3.3 Language

* Data property: Language
* Data type: xsd:string
* Values: valid [iana:] language tags
* RDF: [schema:availableLanguage](http://schema.org/availableLanguage)
* Cardinality: \*
* Description: Languages used by the organization
* Scope Note: Deduced from languages appearing in web resources related to the organisation, so it may be imprecise
* Examples: "it", "en"
* SDATI: web.languages[]
* BRC: we register “målform” (language form) either one of Bokmål or Nynorsk to indicate what language form the company wants to communicate in written. Both forms of Norwegian.

### 2.1.4 Online Presence

We represent commonly used electronic resources and channels (website, wikipedia, email, news feed) as specific properties in this section.

We represent social network accounts of the company (eg Facebook, Twitter) as [Identifiers](#_gzr1h05zqb77). Even though these are not official business identifiers, they fulfill a similar role.

#### 2.1.4.1 Website

* Object property: Website
* RDF: [schema:url](http://schema.org/url)
* Cardinality: \*
* Description: Website pertaining to the company
* Example: <<http://www.comune.trento.it>>
* Discussion: I also considered [schema:mainEntityOfPage](http://schema.org/mainEntityOfPage), but [schema:url](http://schema.org/url) is used a lot more and is more natural
* BRC: hjemmeside

#### 2.1.4.2 Wikipedia Page

* Object property: Wikipedia page
* RDF: [schema:sameAs](http://schema.org/sameAs) "... E.g. the URL of the item's Wikipedia page"
* Cardinality: \*
* Description: Wikipedia page pertaining to the company
* Example: <<https://it.wikipedia.org/wiki/Trento>>

#### 2.1.4.3 Certified Email

* Object property: Certified Email
* RDF: [schema:email](http://schema.org/email)
* Cardinality: \*
* Description: Email that is officially registered and with the same validity as [certified (snail) mail](https://en.wikipedia.org/wiki/Certified_email)
* Scope note: EBG does not record other company emails, so there is no need to distinguish between different kinds. The value should be a mailto: URL
* Example: <<mailto:elettorale@pec.comune.trento.it>>
* BRC: no

#### 2.1.4.4 News/Blog Feed

* Object property: News/Blog Feed
* RDF: [sioc:feed](http://rdfs.org/sioc/spec/#term_feed)
* Cardinality: \*
* Description: URL of RSS/Atom feed pertaining to the company
* BRC: no

## 2.2 Jurisdiction/Addresses/Sites

### 2.2.1 Jurisdiction

* Data property: Jurisdiction
* RDF: [dbo:jurisdiction](http://dbpedia.org/ontology/jurisdiction)
* Cardinality: 1
* Description: jurisdiction in which the company is registered.
  + For Europe, this is a country, even for Germany that has 166 separate registers (see [Registration Authorities List](#_53ymmfdsusiu)).
  + Many jurisdictions register foreign companies that are incorporated and fully owned in another country. EBG collects all registrations from each register and doesn't attempt to match foreign companies across jurisdictions.
* Examples of foreign companies:
  + NO has type NUF "Norskregistrert utenlandsk foretak"
  + BG has many types, including a number of European companies

|  |  |  |
| --- | --- | --- |
| **code** | **Name (BG)** | **Name (EN)** |
| KCHT | Клон на чуждестранен търговец | Branch of a Foreign Trader |
| EUIE | Европейско обединение по икономически интереси ЕОИИ | European Economic Interest Grouping (EEIG) |
| DEUIE | Поделение на ЕОИИ | Division of European Economic Interest Grouping (EEIG) |
| ED | Европейско дружество | European Company |
| EKD | Европейско кооперативно дружество | European Cooperative Company |
| LEKD | Европейско кооперативно дружество с ограничена отговорност | European Limited Cooperative Company |
| CHD | Дружество регистрирано в юрисдикция с преференциален данъчен режим | Company registered in a jurisdiction with preferential tax regime |
| CHDF | Чуждестранно дружество - физическо лице | Foreign sole trader |
| CHDU | Чуждестранно дружество - юридическо лице | Foreign trader |

* Scope note: if we know the official legal ID of the company, the jurisdiction code and legal ID are used in the company URL
* Values: country code, eg "IT", "DE", "NO"
* OCORP: Jurisdiction code
* BRC: Always "NO"

### 2.2.2 Address Discussion

Companies may have a variety of addresses:

* By "importance": legal/registered, headquarters, other locations
* By "mode of delivery": mailing, physical

EBG has info about registered vs other, and does not yet have info about mailing vs physical. This could be represented

#### 2.2.2.1 Address Data

An address may be represented as a free text (full address), e.g.

* Aston House, Cornwall Avenue, London N3 1LF, UK
* Via Belenzani 19, 38122, Trento

An address may also have structured attributes. Part of these form a natural hierarchy (see [EC NUTS and LAU](#_fn3icsx21uas) and [Geonames](#_adddl8oep3pq) above)

* Country (state): country code; English or national name; resource; e.g. IT; Italia; [dbr:Italy](http://dbedia.org/resource/Italy), [nuts:IT](http://nuts.geovocab.org/id/IT)
* Macroregion, e.g. IT Nord-est, which is NUTS=ITC NORD-OVEST
* Region, e.g. Trentino-Alto Adige/Sudtirol (note: this is **not** a NUTS region)
* Province, e.g. Trento
* Locality/city/settlement/municipality, e.g. Trento
* Street address, e.g. Via Belenzani 19
* Postal code
* Latitude, Longitude
* Coordinate resolution (from 10 continent to 90 civic number, i.e. down to individual house)

#### 2.2.2.2 Schema Address Props

Discussion: we have considered two ontologies that define addresses: schema: and locn: (see next section).

[schema:PostalAddress](http://schema.org/PostalAddress): mailing or physical address. It has the following structured attributes that are quite obvious and need no explanation

* [schema:addressCountry](http://schema.org/addressCountry)
* [schema:addressRegion](http://schema.org/addressRegion) (state/province)
* [schema:addressLocality](http://schema.org/addressLocality) (city/settlement)
* [schema:streetAddress](http://schema.org/streetAddress)
* [schema:postOfficeBoxNumber](http://schema.org/postOfficeBoxNumber)
* [schema:postalCode](http://schema.org/postalCode)

Schema doesn't have a notion of "full address", but [schema:address](http://schema.org/address) can be either schema:PostalAddress or free text, which covers this case

#### 2.2.2.3 locn Address Props

[locn:Address](https://www.w3.org/ns/locn#locn:Address): An "address representation" as conceptually defined by the [INSPIRE Address Representation data type](http://inspire-twg.jrc.ec.europa.eu/data-model/approved/r937/fc/#_C2538). It has more structured attributes that however are described somewhat abstractly (according to the INSPIRE model) and there are no examples in the spec:

* [locn:adminUnitL1](https://www.w3.org/ns/locn#locn:adminUnitL1): country (or other uppermost administrative unit)
* [locn:adminUnitL2](https://www.w3.org/ns/locn#locn:adminUnitL2): region/county/province (typically encompasses several localities)
  + What to use for deeper administrative regions (eg municipality)? One option is to repeat locn:adminUnitL2. But that's likely to be confusing, so we'll use EBG props: ebg:adminUnitL3, ebg:adminUnitL4
* [locn:postName](https://www.w3.org/ns/locn#locn:postName): city or other key postal division of the address
* [locn:addressArea](https://www.w3.org/ns/locn#locn:addressArea): part of a city, a neighbourhood or village
* [locn:thoroughfare](https://www.w3.org/ns/locn#locn:thoroughfare): street address
* [locn:locatorName](https://www.w3.org/ns/locn#locn:locatorName): name of property, complex, (part of) building, or room
* [locn:locatorDesignator](https://www.w3.org/ns/locn#locn:locatorDesignator): "number or a sequence of characters that uniquely identifies the locator within the relevant scope": I think this means "house number" but I'm not sure
* [locn:poBox](https://www.w3.org/ns/locn#locn:poBox): Postal Office box
* [locn:postCode](https://www.w3.org/ns/locn#locn:postCode): postal code, zip code
* [locn:addressId](https://www.w3.org/ns/locn#locn:addressId): globally unique identifier for each instance of an address (almost never used)
* [locn:fullAddress](https://www.w3.org/ns/locn#locn:fullAddress): free-text full address

#### 2.2.2.4 locn/schema Geographic Coordinates

locn: supports geographic latitude/longitude:

* The resource (eg org:Site) uses [locn:geometry](https://www.w3.org/ns/locn#locn:geometry) to link to a [locn:Geometry](https://www.w3.org/ns/locn#locn:Geometry) node
* Coordinates should be put in that [loc n:Geometry](https://www.w3.org/ns/locn#locn:Geometry) node.
* locn: doesn't include specific properties for latitude/longitude, but [locn:geometry](https://www.w3.org/ns/locn#locn:geometry) includes a number of examples how this can be expressed (we use the first one):
  + schema: eg  
    a schema:GeoCoordinates; schema:latitude "51.477811"; schema:longitude "-0.001475"
  + the simple [wgs](http://www.w3.org/2003/01/geo/wgs84_pos): ontology, eg  
    wgs:Point; wgs:lat "51.477811"; wgs:long "-0.001475"
  + GeoSPARQL literal (WKT or GML), eg  
    a sf:Point; gsp:asWKT "Point(-0.001475 51.477811)"^^geo:wktLiteral
  + geo: URL, eg  
    <geo:51.477811,-0.001475>
  + geohash URL

#### 2.2.2.5 Org Site Class

[org:Site](https://www.w3.org/TR/vocab-org/#org:Site): An office or other premise at which the organization is located.

* "Many organizations are spread across multiple sites and many sites will host multiple *locations*" (the italicized word is probably a mistake, should be "organizations")
* It's not obvious to me that the distinction between site and address is useful. Maybe the Org ontology means that a site is per-company, while an address can be shared between companies? But the mistake above makes this unclear.

So why bother with Site?

* Neither schema: nor locn: have a notion of "registered address". So if we want to use existing properties, we must use [org:hasRegisteredSite](https://www.w3.org/TR/vocab-org/#org:hasRegisteredSite) and an extra class [org:Site](https://www.w3.org/TR/vocab-org/#org:Site) (Company --hasSite-> Site --siteAddress-> Address)
* If we want to use schema:address for both full address (literal text) and structured (schema:Address), this intermediate node Site will tie the two addresses together.

#### 2.2.2.6 Address Calculations

Given some address data, it is possible to derive other data. The idea is that 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. Then we can record these coordinates, and place them in NUTS regions using the [NUTS RDF](#_wezm1c5p6t7g) **withinRegion** service "Returns the NUTS regions that include a certain point based on its latitude/longitude coordinates". Note: we'll have to re-host this service since the example <http://nuts.geovocab.org/services/withinRegion?lat=49&long=8.4#point> returns "Failed to connect to PostgreSQL"
* From a full address, we can extract the postal code, then use [postal code -> NUTS3](http://ec.europa.eu/eurostat/tercet/download.do?file=pc2016_NUTS-2013.zip) mapping (see [NUTS Downloads](#_bzgahz6i936c))
  + One option is to use per-country regexps, but because of the concerns below, it's better to use a full whitelist of all postal codes.
  + Spaces seem to be optional, so we'll need to account for that
  + If there are several numbers in the full address (e.g. "1234 Some street, Sofia 1234") we should take the last occurrence
  + I have extracted the different [postal code patterns per country](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit?pli=1#gid=1537849788) (A=capital, a=lowercase, d=digit) and there are some special cases:
  + FR includes a number of postal codes ending with "CED\*", "AIR" or "SP \*" (unclear)
  + DE includes some codes (eg 37956, 39052) that are not NUTS3 but probably LAU1
  + UK comes with the biggest number of variations:

|  |  |
| --- | --- |
| AAd | 39 |
| AAd dAA | 678783 |
| AAdAdAA | 11382 |
| AAdd | 378 |
| AAdddAA | 788508 |
| Ad dAA | 44349 |
| AdA dAA | 9867 |
| Add dAA | 156263 |

* 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. Eg Bulgaria has 9 villages named "Bania"
* Given a detailed region (eg LAU2) it is trivial to obtain higher-level region (country and NUTS1..3), since we have loaded [EC NUTS and LAU](#_fn3icsx21uas) data

TODO: decide which of these it's worth implementing, and consequently what data to convert and integrate.

* Chris: If this is implemented, it should be done at the platform level, not by the individual data suppliers
* Vladimir: agreed, but we should figure out the complete data flow:
  + Some providers may not want such enrichment because they have better structured data
  + Conversely, other providers may want to get such enrichments back

### 2.2.3 Address

* Class: Address
* RDF: [org:Site](https://www.w3.org/TR/vocab-org/#org:Site) & [locn:Address](https://www.w3.org/ns/locn#locn:Address), self-link [org:siteAddress](https://www.w3.org/TR/vocab-org/#org:siteAddress).
  + For now we will use a single node with two classes and a self-link since we don't see a need to consider Site and Address as different entities.
* URL: we form the address URL by appending to the end of the company URL
  + For the registered (primary) address use:   
    **<company>/address**
  + For other addresses, if there is a guid or similar internal ID in the provider system, append that:   
    **<company>/address/<guid>**
  + Otherwise use a sequential number (this may lead to confusion if data is integrated from several providers):  
    **<company>/address/<n>**
* Description: mailing or physical address
* Scope note: when used for company, should be linked as registered address (one) or other address (multiple)
* Rules: every data provider should strive to provide NUTS3 for each address. This can be done in cooperation, eg using the techniques described in [Address Calculations](#_aodfm12bfotf)

#### 2.2.3.1 Full Address

* Data property: full address
* Data type: rdf:langString or xsd:string
* RDF: [locn:fullAddress](https://www.w3.org/ns/locn#locn:fullAddress)
* Cardinality: ?
* Description: full address, free text
* Rules: Provide a language tag if possible
* Examples:
  + Aston House, Cornwall Avenue, London N3 1LF, UK
  + Via Belenzani 19, 38122, Trento

#### 2.2.3.2 Address Country

* Lookup property: Country
* RDF: [locn:adminUnitL1](https://www.w3.org/ns/locn#locn:adminUnitL1)
* Cardinality: 1
* Description: Country of the address
* Values: [NUTS RDF](#_wezm1c5p6t7g) URLs
* Example: [nuts:IT](http://nuts.geovocab.org/id/IT) (Italy)

#### 2.2.3.3 Macroregion

* Lookup property: Macroregion
* RDF: [locn:adminUnitL2](https://www.w3.org/ns/locn#locn:adminUnitL2)
* Cardinality: ? (but strongly recommended)
* Description: NUTS1 region of the address
* Values: [NUTS RDF](#_wezm1c5p6t7g) URLs
* Example: [nuts:ITC](http://nuts.geovocab.org/id/ITC) NORD-OVEST (Nord-est)

#### 2.2.3.4 Administrative Region

* Lookup property: Administrative Region
* RDF: ebg:adminUnitL3
* Cardinality: ? (but strongly recommended)
* Description: NUTS2 region of the address
* Values: [NUTS RDF](#_wezm1c5p6t7g) URLs
* Example: [nuts:ITC2](http://nuts.geovocab.org/id/ITC2) Valle d'Aosta/Vallée d'Aoste, [nuts:ITH2](http://nuts.geovocab.org/id/ITH2) Provincia Autonoma di Trento
* Rules: use a valid NUTS region. Eg Trentino-Alto Adige/Sudtirol (taken from the [Geonames](#_adddl8oep3pq) hierarchy above) is **not** a NUTS region

#### 2.2.3.5 Province

* Lookup property: Province
* RDF: ebg:adminUnitL4
* Cardinality: ? (but strongly recommended)
* Description: NUTS3 region of the address
* Values: [NUTS RDF](#_wezm1c5p6t7g) URLs
* Example:, nuts:ITH31 Verona

#### 2.2.3.6 Municipality/Commune/Settlement

* Lookup property: Municipality/Commune/Settlement
* RDF: ebg:adminUnitL5, ebg:adminUnitL6
* Cardinality: ?
* Description: LAU1, LAU2 region of the address
* Scope notes: some countries (eg Bulgaria) use both LAU1 and LAU2 levels. Others (eg Italy) use only LAU2
* Values: EBG [LAU RDF](#_8b4005dqv5ya) URLs

#### 2.2.3.7 Locality/City/Settlement

* Data property: Locality/City/Settlement
* Data type: xsd:string or rdf:langString
* RDF: [locn:postName](https://www.w3.org/ns/locn#locn:postName)
* Cardinality: ?
* Description: Locality/City/Settlement of the address, free text
* Rules:
  + Provide a language tag if possible
  + If both ebg:adminUnitL6 and locn:postName are specified, they should correspond to each other
* Example: "Trento"@it, "Тренто"@bg

#### 2.2.3.8 Neighbourhood/Quarter

* Data property: Neighbourhood/Quarter
* Data type: xsd:string or rdf:langString
* RDF: [locn:addressArea](https://www.w3.org/ns/locn#locn:addressArea)
* Cardinality: ?
* Description: part of a city, village or neighborhood
* Scope note: not likely to appear, but we have Coordinate resolution corresponding to this field

#### 2.2.3.9 Street Address

* Data property: Street Address
* Data type: xsd:string or rdf:langString
* RDF: [locn:thoroughfare](https://www.w3.org/ns/locn#locn:thoroughfare)
* Cardinality: ?
* 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)
* Example:
  + Via Belenzani 19
  + Polygraphia Office Center, floor 4 \n 47A Tsarigradsko Shosse

#### 2.2.3.10 Street Number

* Data property: Street Number
* Data type: xsd:string
* RDF: [locn:locatorDesignator](https://www.w3.org/ns/locn#locn:locatorDesignator)
* Cardinality: ?
* Description: Street number and/or building name
* Scope note: not likely to appear separately, but we have Coordinate resolution corresponding to this field
* Example:
  + 47A
  + Polygraphia Office Center, floor 4

#### 2.2.3.11 Postal Code

* Data property: Postal Code
* Data type: xsd:string
* RDF: [locn:postCode](https://www.w3.org/ns/locn#locn:postCode)
* Cardinality: ?
* Description: Postal code of the address
* Rules: If possible, a European postal code should correspond to the fixed values given in [NUTS Downloads](#_bzgahz6i936c), so we can correlate it to a NUTS region
* Discussion: should we remove spaces to make matching easier?

#### 2.2.3.12 Postal Office box

* Data property: Postal Office box
* Data type: xsd:string
* RDF: [locn:poBox](https://www.w3.org/ns/locn#locn:poBox)
* Cardinality: \*
* Description: Some addresses are associated with a PO box instead of a street address

#### 2.2.3.13 Geographic Coordinates

* Property & Class: Geographic Coordinates
* RDF: [schema:geo](http://schema.org/geo) leading to class [schema:GeoCoordinates](http://schema.org/GeoCoordinates)
* Description: Geographic coordinates, eg of a geo-located address
* URL: we append "/geo" to the corresponding address URL, eg  
  **<company>/address/geo** for the registered address  
  **<company>/address/<n>/geo** for other addresses
* Cardinality: ?. If data is integrated from several providers, we may end up with multiple coordinate pairs in that node, which is too bad
* Example:  
  <company/BG/123456/address> schema:geo <company/BG/123456/geo>.  
  <company/BG/123456/address/geo> a schema:GeoCoordinates;  
   schema:latitude "51.477811"; schema:longitude "-0.001475";  
   ebg:geoResolution <resolution/L6>.

##### 2.2.3.13.1 Latitude, Longitude

* Data properties: Latitude, Longitude
* Data type: xsd:decimal (may be omitted)
* Domain: Geographic Coordinates
* RDF: schema:latitude, schema:longitude
* Cardinality: 1
* Rules:
  + Coordinates must be expressed in the [WGS 84](https://en.wikipedia.org/wiki/World_Geodetic_System) system
  + If you don't have coordinates, do not make a [schema:GeoCoordinates](http://schema.org/GeoCoordinates) node

##### 2.2.3.13.2 Geo Coordinate Resolution

* Lookup property: Geo Coordinate Resolution
* Domain: [schema:GeoCoordinates](http://schema.org/GeoCoordinates)
* Description: Resolution of a geographic coordinate location: from continent (least precise) to individual house number (most precise)
* RDF: ebg:geoResolution  
  Note: the name follows feedback of the [LOV community](https://plus.google.com/u/1/115593825497938381443/posts/LDxpHf5H1me)
* See also: [datex:locationresolution](http://vocab.datex.org/terms/#locationPrecision), but that's a data property with unclear values
* Cardlinality: 1
* SDATI: latlonresolution. Uses numeric values in the range [10..100] that are mapped to column "notation" in the lookup table. Added 65 because Italy doesn't use LAU1 (doesn't make a difference between municipality and settlement)
* Rules:
  + if geo coordinates are present, resolution must be present too, even if it's a guess. There is no default value
  + Values should be mapped to the <resolution/> ConceptScheme (lookup list)

Values: from <resolution/L0> to <resolution/L10>, see [this sheet](https://docs.google.com/spreadsheets/d/1GjDUDQ5LMyZwFBI8mpuUtk_XGWuaO_E9EFmF9T8FmoY/edit?pli=1#gid=212809552).

### 2.2.4 Registered Address

* Object property: Registered Address
* RDF: [org:hasRegisteredSite](https://www.w3.org/TR/vocab-org/#org:hasRegisteredSite) & [org:siteAddress](https://www.w3.org/TR/vocab-org/#org:siteAddress)
* Cardinality: ?
* Description: Public legal address where legal papers can be served
* Scope note: [org:hasRegisteredSite](https://www.w3.org/TR/vocab-org/#org:hasRegisteredSite) leads to a node with types org:Site and locn:Address that has self-link [org:siteAddress](https://www.w3.org/TR/vocab-org/#org:siteAddress)
* BRC: brc:postadresse (postal address)
* OCORP: Registered Address
* SDATI: base.registeredAddress

### 2.2.5 Other Address

* Object property: Address
* RDF: [org:hasSite](https://www.w3.org/TR/vocab-org/#org:hasSite) & [org:siteAddress](https://www.w3.org/TR/vocab-org/#org:siteAddress)
* Cardinality: \*
* Description: Other address/location associated to an organisation
* Scope note: [org:hasSite](https://www.w3.org/TR/vocab-org/#org:hasSite) leads to a node with types org:Site & locn:Address that has self-link [org:siteAddress](https://www.w3.org/TR/vocab-org/#org:siteAddress). [org:hasRegisteredSite](https://www.w3.org/TR/vocab-org/#org:hasRegisteredSite) is declared a sub-property of [org:hasSite](https://www.w3.org/TR/vocab-org/#org:hasSite), so querying by [org:hasSite](https://www.w3.org/TR/vocab-org/#org:hasSite) will obtain all addresses
* BRC: brc:forretningsadresse (business address, visiting address of the main office)

### 2.2.6 Event

* Class: Event
* RDF: sem:Event & ebg-event:hasEvent
* URL: company/<jurisdiction>/<co>/<id>

example url (FR/44098064700/SIREN/event/2002-02-01Bought\_by00555002500018)

Or

site/<jurisdiction>/<co>/<id>

* Description: Events are things that happen. This comprises everything from historical events to web site sessions and mythical journeys. In the case of Ebg model Events are things that happen to a company or a site

##### 2.2.6.1 EventValue

* Object property: Event Value
* Data type: rdf:langString or xsd:string or org:site
* RDF: ebg-event: eventValue
* Cardinality: 1
* Description: is the value linked to an eventType that happens to a company or a site

(example: The company A changes its principal actitity to “agricolture”

Company: A

Event type: change\_principal\_activity

Value: “agricolture”)

##### 2.2.6.2 EventType

* Class: EventType
* Values: concept scheme <Type/>, see below
* Cardinality: 1
* RDF: sem:eventType
* url: /event/type/<eventType>
* Description: types of pre-established events that can happen to a company or a site  
  below are listed the pre-established events

Possible EventType values:

change\_administrative\_state

change\_ensign

change\_usual\_denomination

change\_principal\_activity

change\_type\_employeer

change\_administrative\_state

change\_lastname

change\_used\_lastname

change\_denomination

change\_denomination\_used

change\_juridical\_category

change\_principal\_activity

change\_NIC\_HQ

change\_in\_social\_solidarity\_economy

change\_type\_employeer

sold\_by

bought\_by

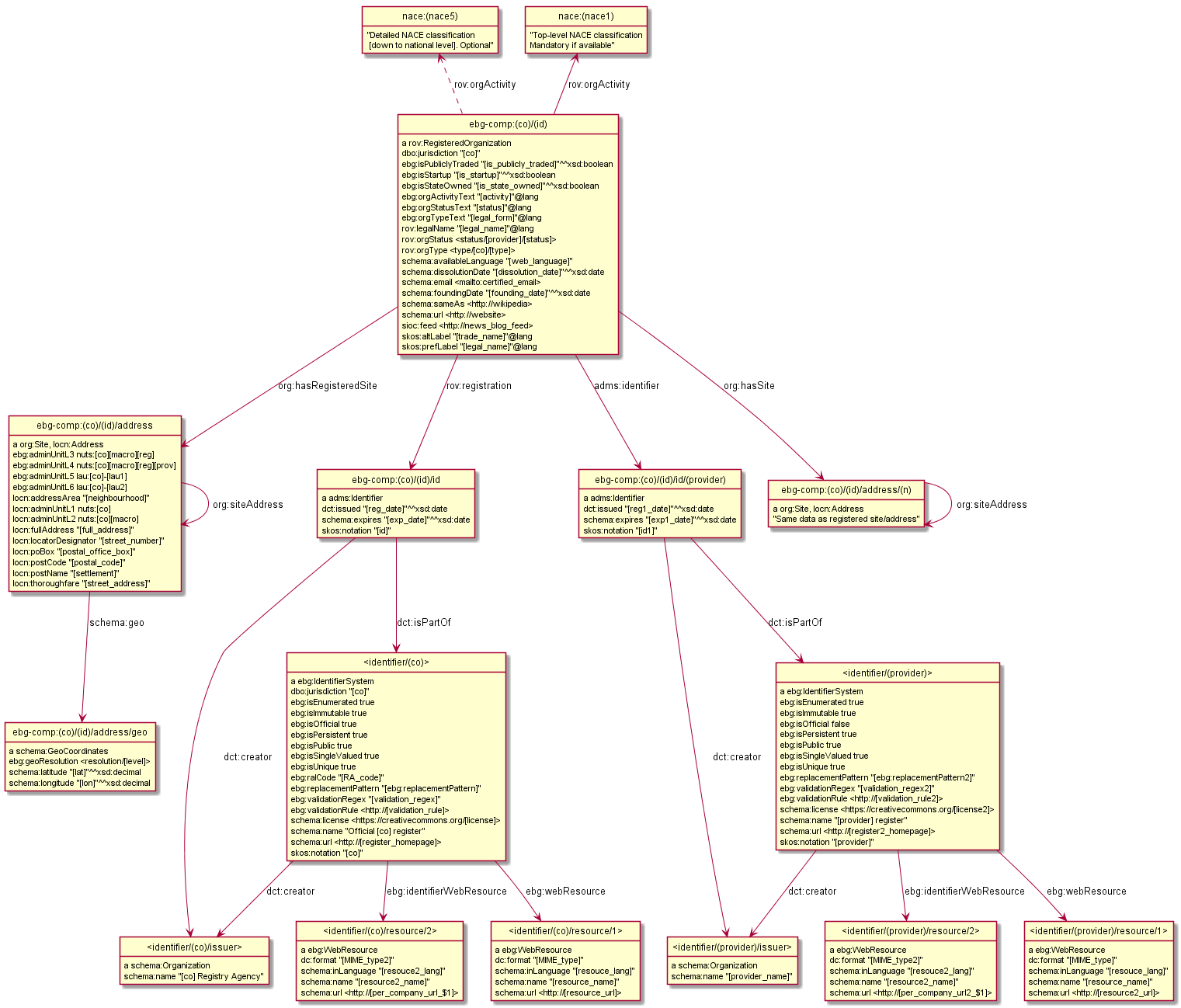
## 2.3 Company Model Diagrams

This section collects various diagrams illustrating the model: instance diagrams made with rdfpuml, and diagrams made with ORM (similar to UML class diagrams).

### 2.3.1 Company Instance Diagram

All classes and properties described in section 2 are illustrated on the following diagram ([company.png](https://github.com/euBusinessGraph/eubg-data/raw/master/model/company.png), [company.ttl](https://github.com/euBusinessGraph/eubg-data/raw/master/model/company.ttl)), this is the old diagram, before the new additions. With the extensions the model will cover: 1) events, from the company URI

2) self-reference to link the Establishments and the Legal Unit starting from the company URI



Abbreviations used:

* (co): country (jurisdiction)
* (macro): macroregion
* (reg): region
* (prov): province
* (lau1): local administrative unit 1
* (lau2): local administrative unit 2
* (prov\_code):
* (RA\_code): Registration Authority List code

Note: the <register/RAL> identifier of the official country register <register/(co)> (bottom node) is not part of the Company data model, it's explained in section [Business Register](#_jg8rjjwyxqos).

### 2.3.2 Company JSON-LD

TODO: create JSON-LD Context, and present the example above as JSON-LD, so data partners can use it as a template to follow.

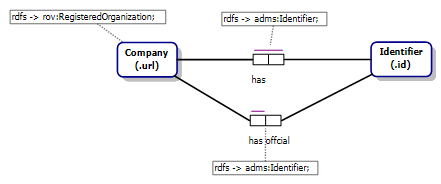
### 2.3.3 Company RDF Shapes

TODO: make ShEx and/or SHACL Compact shape definitions of the information above.

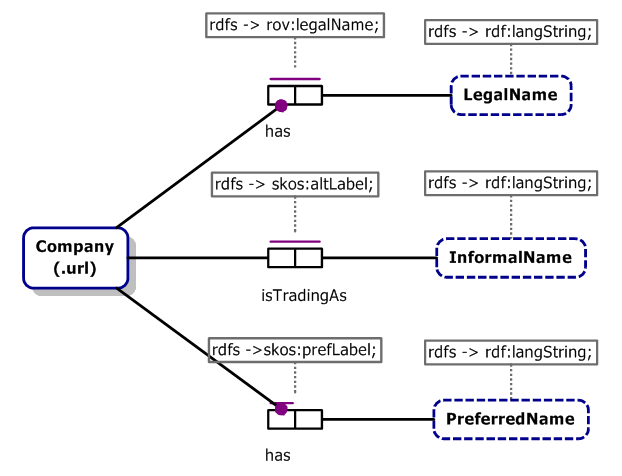
### 2.3.4 Company ORM Schema

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](https://github.com/euBusinessGraph/eubg-data/blob/master/model/ebg-ontology-gen.ttl). Note that this RDFS only defines classes and domain and range of properties.

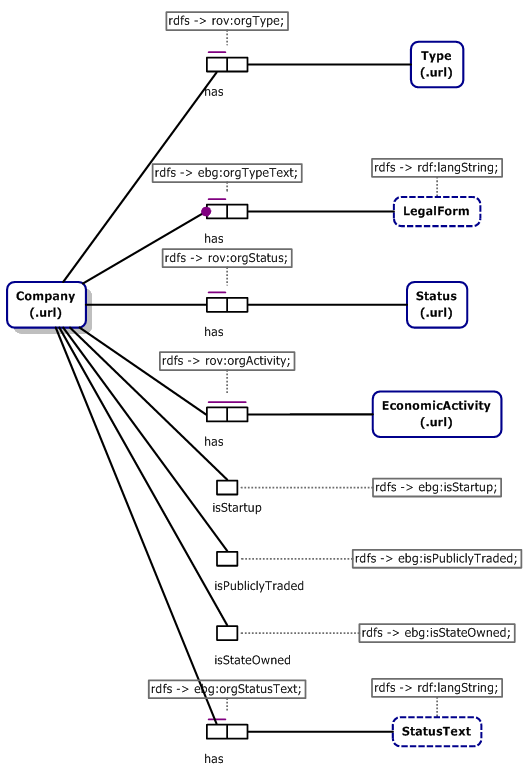
#### 2.3.4.1 Company ORM



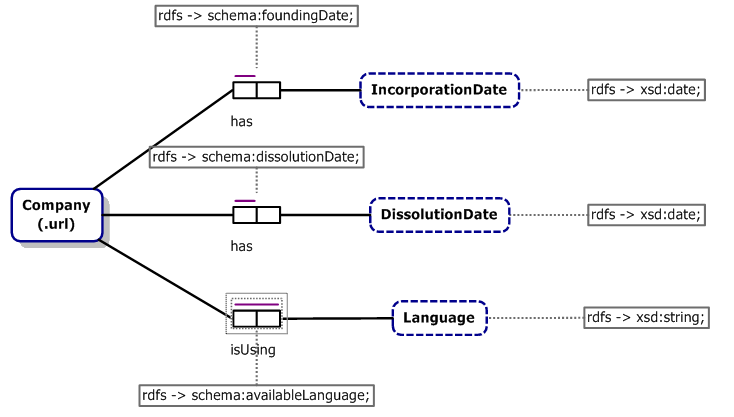
#### 2.3.4.2 Names ORM



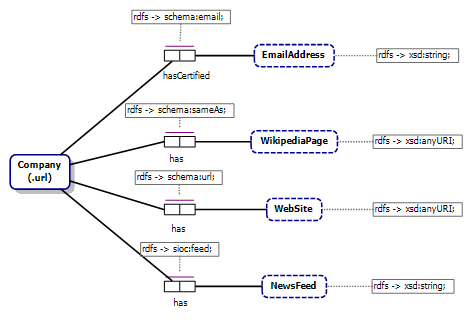
#### 2.3.4.3 Classifications ORM



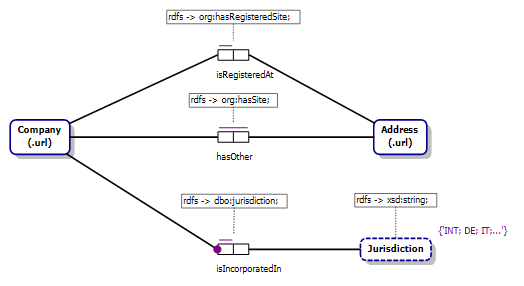
#### 2.3.4.4 Other Details ORM

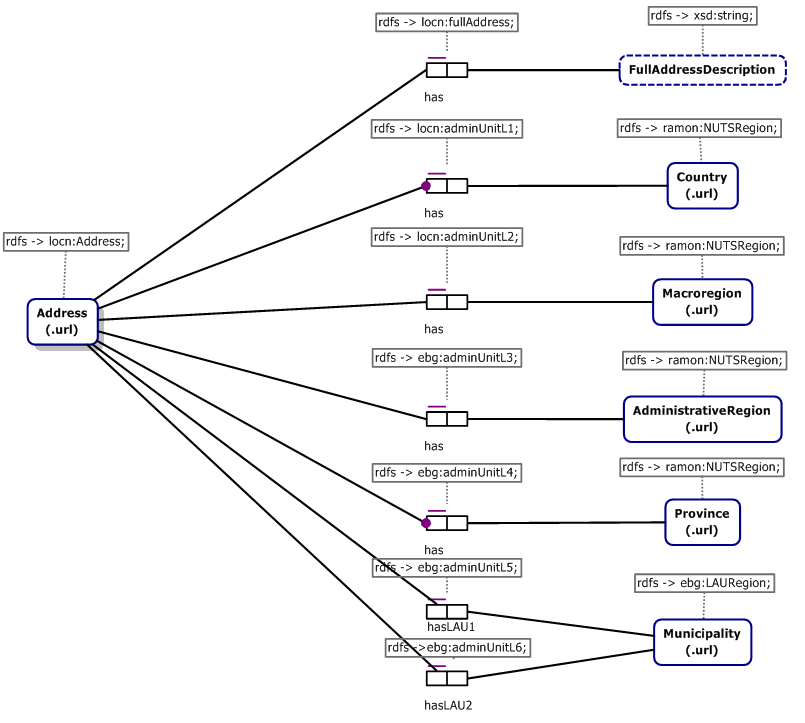


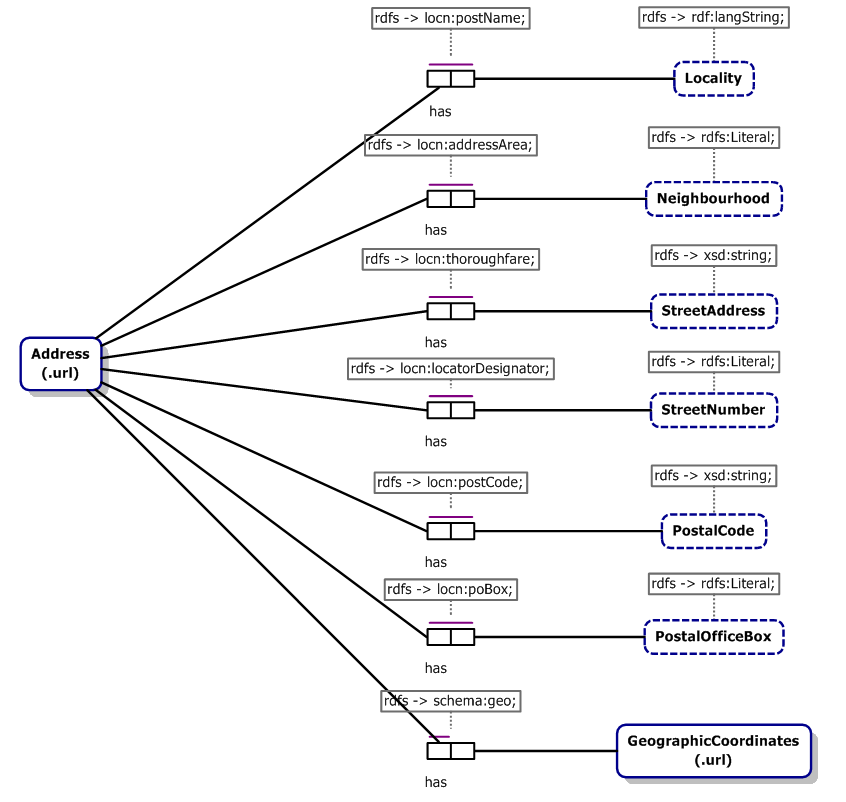
#### 2.3.4.5 Online Presence ORM

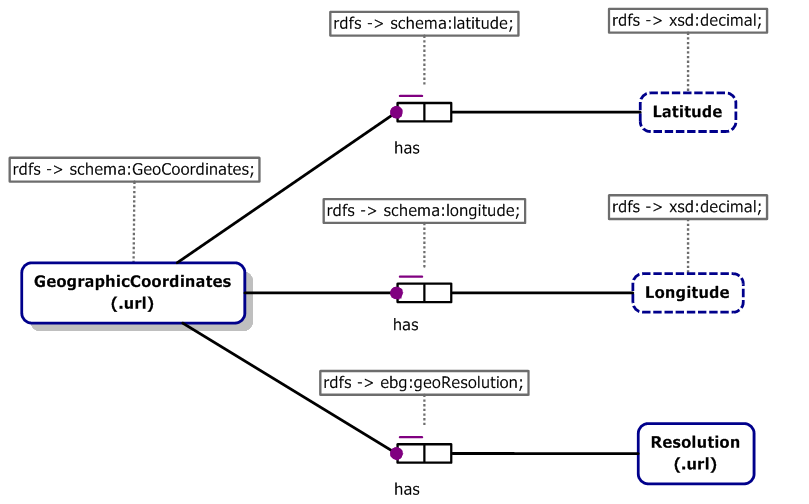


#### 2.3.4.6 Jurisdiction/Addresses/Sites ORM









# 3 Identifiers and Identifier Systems

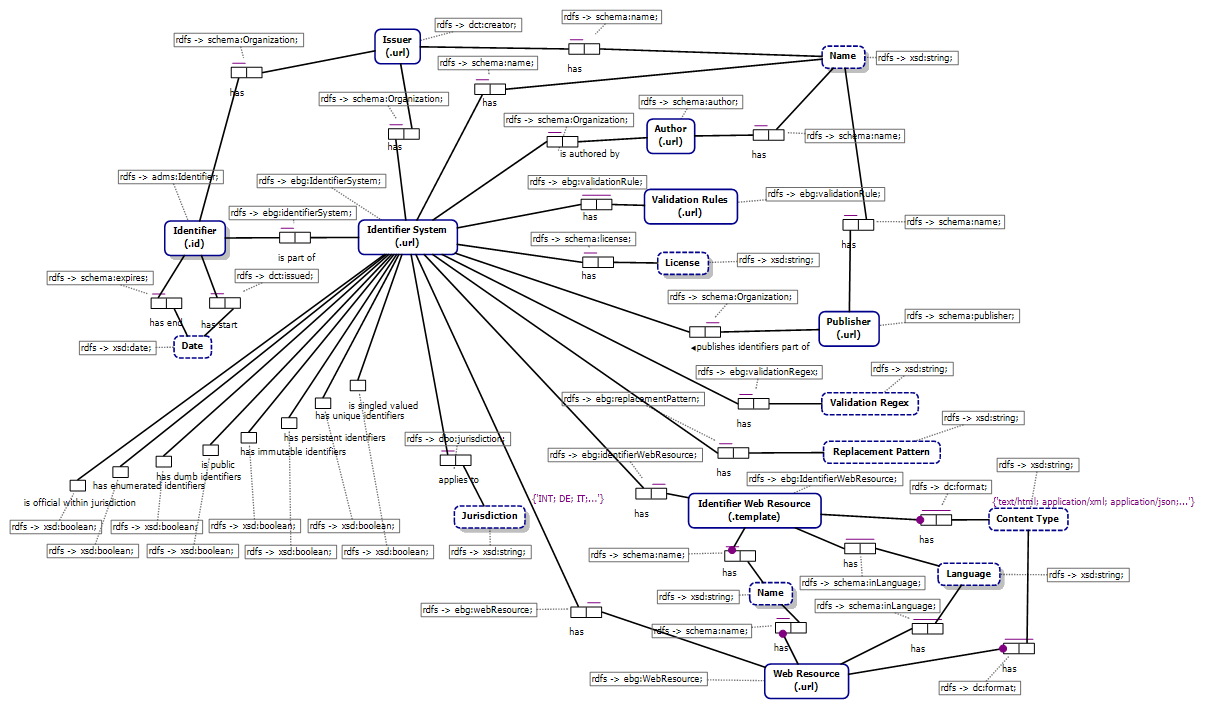
Companies are registered by various kinds of identifiers. Some of them are kept in official registers, others are self-issued and not centralized (e.g. website). Identifiers include but are not limited to:

* Official registration in a trade register, which is singled out as [rov:registration](https://www.w3.org/TR/vocab-regorg/#ref_rov_registration). This registration should correspond to the company's [Jurisdiction](#_wudfhlvh78es) and when known, we use it in the EBG company URL.
* Other official government registers, eg bank license, insurance company license, register of startup companies, etc
* Official international registries, eg [Global Legal Entity Identifier](#_itf84se7o12s), etc
* EBG data provider or other databases, eg OCORP, SDATI, Wikidata, Dun & Bradstreet
* Various company websites
* Social networks, eg Facebook, Twitter, etc

Currently we use two different ways:

* [adms:Identifier](https://www.w3.org/TR/vocab-adms/#identifier) records the identifier value and identifier system (see below). This allows to search or group companies by system and to record additional info about the system (website, URL patterns for per-company resources, etc).
* Section [Online Presence](#_9qy4ycuhh70d) uses 4 Schema properties to provide direct company website URLs: website, blog, Wikipedia page ([schema:sameAs](http://schema.org/sameAs) is used only for a reference website that unambiguously identifies the company), and email. TODO: Should we rework to cast them as Identifiers? Blog presents a problem, since it's the same system as Website, only the purpose is more specific. I am against this.

## 3.1 Identifiers ORM Diagram



## 3.2 Identifier Systems

We consider identifiers that share the same specification, characteristics and origin to be organized in an Identifier System.

* Class: Identifier System
* URL: identifier/<code>. E.g. <identifier/GB> is the Companies House system (register)
* RDF: ebg:IdentifierSystem

TODO: make RDF from the [List of registers](https://docs.google.com/spreadsheets/d/1ENBasWNfCiK_fh39KavaJQPJn_Shm4RI5si31YUytvA/edit) (identifier kinds) used by partners.

### 3.2.1 Identifier System Agents

There are 3 agents related to a system:

* **Author**: specified 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, with property
* schema:name: xsd:string or rdf:langString

#### 3.2.1.1 Identifier System Author

* Object Property: Author
* Range: schema:Person or schema:Organization
* Cardinality: ?
* RDF: schema:author
* Description: Agent who specified the rules and organization of the identifier system
* Examples:
  + W3C or TimBL specified web URLs.
  + GLEIF specified GLEI
  + Dun and Bradstreet specified DUNS numbers

#### 3.2.1.2 Identifier System Issuer

* Object Property: Issuer
* Range: schema:Person or schema:Organization
* Cardinality: ?
* RDF: dct:creator
* Description: Agent that issues identifiers and then keeps them in a database (register)
* Examples:
  + Web URLs are not issued by any central agency and there is no register to consult (web crawlers work in a different way).
  + Companies House does that for the official UK identifier system (register)
  + GLEI Foundation (GLEIF) does it for the world-wide GLEI register (through its Local Operating Units, LOU).
  + Dun and Bradstreet issus DUNS numbers
* Scope Note: Many agencies keep several registries, eg <http://www.registryagency.bg> keeps:
* Trade Register <http://brra.bg> for companies,
* Bulstat Register <http://www.bulstat.bg/> for other entities

#### 3.2.1.3 Identifier System Publisher

* Object Property: Publisher
* Range: schema:Person or schema:Organization
* Cardinality: \*
* RDF: schema:publisher
* Description: Any agent that publishes the identifier database (register) in some form (e.g. paper, CD, 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>.
  + Companies House and BRC publish the UK and NO registers openly.
  + DUNS numbers are not openly published

### 3.2.2 Identifier System Properties

Identifier systems have some basic properties (jurisdiction, name, URL, codes).

#### 3.2.2.1 Identifier System Jurisdiction

* Data property: Jurisdiction
* RDF: [dbo:jurisdiction](http://dbpedia.org/ontology/jurisdiction)
* Cardinality: \*
* Description: jurisdiction to which the identifier system applies.
* Examples:
  + Website, Twitter, Facebook don't apply to any particular jurisdiction in that they don't have any official status
  + UK company number applies to jurisdiction "GB"
  + GLEI applies to the "INT" (world-wide) jurisdiction

#### 3.2.2.2 Identifier System Code

* Data Property: Code
* Cardinality: 1
* RDF: skos:notation
* Description: A short mnemonic code for the identifier system, used in its URL. Also used in identifier URLs that are part of that system.
* Scope note: Issued locally by EBG.
* Rules:
  + For the "preferred" company register of a jurisdiction, we use the jurisdiction code (e.g. "BG", "GB")
  + For other company registers, we use upper-case (e.g. "OCORP", "ATOKA", "EU", "BRIS")
  + For web and social network identifier, we use mixed case (e.g. "Twitter", "Facebook", "Website", "Blog")

#### 3.2.2.3 Identifier System RAL Code

See [Registration Authorities List](#_53ymmfdsusiu).

* Data Property: RAL Code
* Cardinality: ?
* RDF: ebg:ralCode
* Description: GLEI RAL code for the identifier system
* Scope note: used only for official registers

#### 3.2.2.4 Identifier System Name

* Data Property: Name
* Data Type: xsd:string or rdf:langString
* Cardinality: +
* RDF: schema:name
* Rules: it is desirable to provide national and English name of the identifier system (same as in the RAL list) and use language tag to distinguish them

#### 3.2.2.5 Identifier System Website

* Object Property: Website
* Cardinality: \*
* RDF: schema:url
* 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 what the URL returns, use [Web Resource](#_quxcjrrzwaic). Don't provide per-company or validation rule URLs here

#### 3.2.2.6 Identifier System License

* Object Property: License
* Cardinality: \*
* RDF: schema: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/>

### 3.2.3 Identifier System Characteristics

Systems have some Boolean characteristics (flags) that represent expectations about the identifiers in the system. 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 (if there is not enough information)

#### 3.2.3.1 Has Unique Identifiers

* Data Property: Has Unique Identifiers
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isUnique
* Description: Whether each identifier in the system relates to only one entity
* Similar to:
  + [owl:InverseFunctionalProperty](https://www.w3.org/TR/owl-ref/#InverseFunctionalProperty-def)
  + Wikidata Unique Values constraint (see [Wikidata External Identifiers](#_b1uhtsrkgdlr))
* Examples:
  + Company IDs are unique in most national registers
  + Stock exchange tickers are unique to each exchange (but a company can buy the ticker of another, i.e. the identity changes over time)
  + Stock exchange tickers are not unique across exchanges (e.g. MET identifies [http://www.metlife.com](http://www.metlife.com/) on the New York exchange, but [http://www.metropolitan.co.za](http://www.metropolitan.co.za/) on the Johannesburg exchange).
  + Person names are not unique since they are highly ambiguous
  + Websites are not unique since a website can be shared by several related companies

#### 3.2.3.2 Is Single-Valued

* Data Property: Is Single-Valued
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isSingleValued
* Description: Whether each entity has only one identifier in the system
* Similar to:
  + [owl:FunctionalProperty](https://www.w3.org/TR/owl-ref/#FunctionalProperty-def)
  + Wikidata Single Values constraint (see [Wikidata External Identifiers](#_b1uhtsrkgdlr))
* Examples:
  + Company IDs are single-valued in most national registers. If two companies merge, the recessive company ID is usually deprecated.
  + Stock exchange tickers are not single-valued, since a company may have several tickers for different kinds of its stock.
  + Official person names are single-valued, since each person has one official name (but that can change over time)
  + Websites are not single-valued since a company may have several websites

#### 3.2.3.3 Has Persistent Identifiers

* Data Property: Has Persistent Identifiers
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isPersistent
* Description: whether identifiers can be removed from the register (e.g., when a company is dissolved). Some registers “remove” identifiers when the object they relate to is no longer active, whereas the best practice is to make such identifiers Obsolete.
* Examples: TODO

#### 3.2.3.4 Has Immutable Identifiers

* Data Property: Has Immutable Identifiers
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isImmutable
* Description: TODO
* Examples: TODO

#### 3.2.3.5 Is Public

* Data Property: Is Public
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isPublic
* 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.
  + 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.

#### 3.2.3.6 Has Dumb Identifiers

* Data Property: Has Dumb Identifiers
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isDumb
* Description: "Intelligent" or "Smart" identifiers contain built-in "‘intelligence" (semantic information) embedded in the identifier. This is increasingly considered bad practice[[12]](#footnote-12) since when the attributes change, the identifier must also change, making it unreliable, particularly as a foreign key. "Dumb" identifiers on the other hand contain no intelligence and will not change
* Examples:
  + Indian (IN) company number is an "intelligent" identifier, containing company type, state of incorporation and industry code, meaning that it changes each time one of these transient attributes changes
  + The Global LEI is a "dumb" identifier, and does not change when the legal entity changes, nor when the legal entity moves from one registration unit to another.

#### 3.2.3.7 Has Enumerated Identifiers

* Data Property: Has Enumerated Identifiers
* Data Type: xsd:boolean
* Cardinality: ?
* RDF: ebg:isEnumerated
* Description: Whether the system has an Issuer, and issued identifiers are kept in a database (register)
* Examples:
  + Every official register is enumerated
  + Websites are not enumerated

#### 3.2.3.8 Identifier System Official in Jurisdiction

* Data property: Jurisdiction
* Data type: xsd:boolean
* RDF: ebg:isOfficial
* Cardinality: ?
* Description: whether the system is considered the official one in all jurisdictions in which it applies (see [3.4.6 Official Identifier)](#_l5xd4lubuppo)
* Examples:
  + False for Website, Twitter, Facebook: they don't apply to any particular jurisdiction and don't have any official status
  + True for UK Company House, the official register for the “GB” jurisdiction
  + False for GLEI, since it is not the official register for any of the world’s jurisdictions. (We list it in "INT" jurisdiction, and there are no official registers for it).
  + False for SDATI identifiers since that system is not official in Italy.

### 3.2.4 Identifier Validation Properties

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

#### 3.2.4.1 Validation Rule

* Object Property: Validation Rule
* Cardinality: \*
* RDF: ebg:validationRule
* Description: URL providing human or machine-readable rule(s) for validating identifiers in the system
* Scope Note: Can be in the form of web page, PDF document, RDF shape, etc. Multiple values about the same validation rule are ok.
* Examples:
  + <[http://bsv-bg.com/контролни-цифри-ползвани-в-българия/](http://bsv-bg.com/%D0%BA%D0%BE%D0%BD%D1%82%D1%80%D0%BE%D0%BB%D0%BD%D0%B8-%D1%86%D0%B8%D1%84%D1%80%D0%B8-%D0%BF%D0%BE%D0%BB%D0%B7%D0%B2%D0%B0%D0%BD%D0%B8-%D0%B2-%D0%B1%D1%8A%D0%BB%D0%B3%D0%B0%D1%80%D0%B8%D1%8F/)> describes the BG EIK checksum algorithm in HTML
  + <<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

#### 3.2.4.2 Validation Regex

* Data Property: Validation Regex
* Cardinality: ?
* RDF: ebg:validationRegex
* Description: Regular expression for validating identifier values of that system
* Rules: Don't forget that you need to escape back-slashes in Turtle strings by doubling them
* 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
  + "\\d{9}" is a simple regex for validating DUNS numbers
  + "(\\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")

#### 3.2.4.3 Replacement Pattern

* Data Property: Replacement Pattern
* Cardinality: ?
* RDF: ebg:replacementPattern
* Description: Pattern to use together with the Validation Regex to normalize 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")

### 3.2.5 Identifier Kinds Discussion

TODO: do we need to distinguish between different identifier kinds listed above? Would we use a government register number differently from a Twitter account?

Chris: These are different, sometimes subtly, sometimes substantively. Official entry in the company register is a registration of the existence of the company.

* Vladimir: The single official registration is already singled out: it uses a different property (rov:registration), is single (cardinality), is nearly mandatory (if we can get hold of it), and is used in the company URL.

Chris: The LEI register is in some ways a proxy for the official (existence) register.

* Vladimir: agree, and there are other examples:
* "Straight proxies" where the official id is part of the proxy id:
  + EU VAT number
  + BRIS id (the one you said they botched)
  + OCORP id
* "Remote proxies": datasets that use their own id and sometimes even can't/won't provide mapping to the official id
  + SDATI Atoka id

Chris: An entry in a bank register is an record of the entity's bank licence (it may have more than one of these, may lost it/regain it), which is independent of the company register. Twitter accounts, websites etc are orthogonal to existence, and in general there's a many to many relationship between these and legal entities.

* Vladimir: so banking/insurance/twitter have the same nature

The question is whether the system should treat them somehow differently. So far I don't see such need.

## 3.3 Web Resources

A "web resource" is a URL complemented with a MIME type to specify what the URL is about.

* We use that 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

### 3.3.1 Web Resource

* Class: Web Resource
* RDF: ebg:WebResource
* URL: blank node
* Description: URL complemented with name, language and MIME type(s) to specify what the URL is about

#### 3.3.1.1 Identifier System Web Resource

* Object Property: web resource
* Domain: Identifier System
* Range: Web Resource
* Cardinality: \*
* RDF: ebg:webResource
* Description: Web resource(s) associated with an identifier system

#### 3.3.1.2 Web Resource URL

* Object Property: web resource URL
* Cardinality: 1
* RDF: schema:url
* Description: URL of the web resource
* Examples
  + For BG TR, <<https://brra.bg>> with language "bg" and MIME type "text/html" is the homepage
  + For BG TR, <<http://opendata.government.bg/dataset/tbprobckn-pernctbp>> with language "bg" and MIME type "application/xml" is open data download in XML
  + For BE Banque-Carrefour des Entreprises, <<https://kbopub.economie.fgov.be/kbopub/zoeknummerform.html>> with MIME type "application/x-search" is the search URL

#### 3.3.1.3 Web Resource Name

* Data Property: Web Resource Name
* Data type: xsd:string or rdf:langString
* Cardinality: +
* RDF: schema:name
* Description: Name or short (generic) description of the resource
* Rules: Provide this since it's useful to the user. Can provide it in several languages, in which case the language tag should be specified
* Examples: "Homepage", "Search", "Download", "Company data", etc

#### 3.3.1.4 Web Resource MIME Type

* Data Property: MIME Type
* Cardinality: \*
* RDF: dc:format
* Description: MIME type(s) of the resource. If several are provided, the server must provide all these resource types using Content Negotiation
* Examples:
  + "text/html", "application/xml", "application/rdf+xml", "text/turtle", "text/csv" ("text/comma-separated-values" could also be used, but the official type is text/csv), "text/tab-separated-values"
  + For a search page, use the custom type "application/x-search"
  + For a [Description document supporting the OpenSearch](http://www.opensearch.org/Specifications/OpenSearch/1.1) standard, use <<http://a9.com/-/spec/opensearch/1.1/>>

#### 3.3.1.5 Web Resource Language

* Data Property: Language
* Cardinality: \*
* RDF: schema:inLanguage
* Description: Language of the web resource
* Rules: use a valid [iana:] language code from the [IETF BCP 47 standard](http://tools.ietf.org/html/bcp47)

### 3.3.2 Identifier Web Resource

* Class: Identifier Web Resource
* RDF: ebg:IdentifierWebResource
* URL: blank node
* Description: URL template complemented with name, language and MIME type(s) to specify what the URL is about. Substitute the identifier value into the template

This class has:

* a specific property "web resource URL template",
* the same general properties as Web Resource: name, MIME type, language (not repeated below)

#### 3.3.2.1 Identifier Web Resource

* Object Property: identifier web resource
* Domain: Identifier System
* Range: Identifier Web Resource
* Cardinality: \*
* RDF: ebg:identifierWebResource
* Description: Web resource(s) associated with an identifier system
* Scope note: Specifies a template that can be used uniformly to build URLs for all identifiers in the system.

#### 3.3.2.2 Web Resource URL Template

* Data Property: web resource URL template
* Cardinality: 1
* RDF: ebg:urlTemplate
* Description: URL template of the web resource.
  + If it has a placeholder {}, substitute the identifier value there
  + If it has placeholders like $1, $2… substitute the groups extracted by the validationRegexp
* Examples
  + For BE Banque-Carrefour des Entreprises, "https://kbopub.economie.fgov.be/kbopub/toonondernemingps.html?ondernemingsnummer={}" with MIME type "text/html" shows a web page for the company.
  + For the EU Value Added Tax (VIES) register, "http://ec.europa.eu/taxation\_customs/vies/vatResponse.html?memberStateCode=$1&number=$2" with MIME type "text/html" shows a web page for an identifier that can be parsed with validationRegex "([A-Z]{2})/(\d+)". Here $1 is the member state code, and $2 is the national company id
  + For OCORP, the templates   
    "https://opencorporates.com/companies/{}", "https://opencorporates.com/companies/{}.xml", "https://opencorporates.com/companies/{}.json",  
    "https://opencorporates.com/companies/{}.rdf" return  
    "text/html", "application/xml", "application/json" and "application/rdf+xml" respectively
  + For BR, "http://data.brreg.no/enhetsregisteret/enhet/{}" with either  
    "application/xml" or "application/json" returns the respective MIME type using Content Negotiation

## 3.4 Company Identifier

We follow the [RegOrg model](https://www.w3.org/TR/vocab-regorg/#overview-of-the-vocabulary), but complement with [Identifier System](#_h8y86haj7zkl) as described above

* Class: Identifier
* RDF: [adms:Identifier](https://www.w3.org/TR/vocab-adms/#identifier) (based on the UN/CEFACT Identifier class)
* Description: identifier of a company according to some identifier system
* URL:
  + For the official identifier:  
    **<company>/id**, eg
    - <company/GB/123456/id> with value "123456"
  + For other identifiers:  
    **<company>/id/<system>**, eg
    - <company/GB/123456/id/OCORP> with value "gb/123456"
    - <company/GB/123456/id/Twitter> with value eg "mycorp"
  + If there are several identifiers per company and system, we add a suffix (numeric or some natural key)

### 3.4.1 Identifier Value

* Data property: Identifier Value
* Domain: Identifier
* RDF: skos:notation
* Cardinality: 1
* Description: Literal value of the identifier.
* Rules: should not include leading, trailing and consecutive spaces, to facilitate exact match. Should be validated against validationRegex for that system
* Discussion: [adms:Identifier](https://www.w3.org/TR/vocab-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:
  + One cannot search by exact literal value unless one also knows the datatype URL
  + There is no property defined for the identifier scheme (as opposed to the agency)
  + Currently we don't plan to maintain different identifier scheme versions
* BRC: organisasjonsnummer

### 3.4.2 Identifier Part Of System

* Object property: Identifier System
* Domain: Identifier
* Range: Identifier System
* RDF: dct:isPartOf
* Cardinality: 1
* Description: System this identifier is a part of

### 3.4.3 Identifier Issuer

* Object property: Issuer
* Domain: Identifier
* Range: schema:Person or schema:Organization
* RDF: dct:creator
* Cardinality: ?
* Description: agency that issued the identifier
* Scope note:
  + In many countries there is a single registry although in others, such as Spain and Germany, multiple registries exist. [rov:]
  + If the system has an issuer, in most cases the identifier issuer will coincide with that issuer. So in a sense this is redundant, but we keep it because the [RegOrg model](https://www.w3.org/TR/vocab-regorg/#overview-of-the-vocabulary) demands it.
  + Furthermore, the original agency that issued the identifier may later transfer issance responsibility to a new agency. Thus the system issuer will change, but the identifier issuer stays fixed (this is often printed on a physical rendition of the identifier, such as company certificate or personal ID card).

### 3.4.4 Identifier Issued On

* Data property: Issued On (Start Date)
* Domain: Identifier
* Data type: xsd:date
* RDF: dct:issued
* Cardinality: ?
* Description: when was the identifier issued
* Scope note: this may or may not be the same as the [Incorporation Date](#_46pgy1vmzn6q). Even for the official registration, the two dates may differ, depending on business rules
* BRC: registered, registreringsdatoEnhetsregisteret

### 3.4.5 Identifier Expires On

* Data property: Expires On (End Date)
* Domain: Identifier
* Data type: xsd:date
* RDF: schema:expires
* Cardinality: ?
* 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.

### 3.4.6 Official Identifier

* Object property: Official Registration
* Domain: Company
* Range: Identifier
* RDF: [rov:registration](https://www.w3.org/TR/vocab-regorg/#ref_rov_registration)
* Cardinality: ?
* Description: identifier that holds the official company registration in its jurisdiction of registration. It establishes the legal existence of the company.
* Scope note: 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:]
* Note: this is an RDF link from Company to Identifier. The ID itself is in Identifier Value

### 3.4.7 Company Identifier

* Object property: Company Identifier
* Domain: Company
* Range: Identifier
* RDF: [adms:identifier](https://www.w3.org/TR/vocab-adms/#adms-identifier)
* Cardinality: \*
* Description: Some other identifier of a company (not the official identifier). This includes but is not limited to:
  + **Direct**: Identifier in some aggregated dataset, where the official id is part of the aggregation id.
  + **Indirect**: Identifier in some aggregated dataset or for a website, where the aggregation id bears no resemblance to the official id.
  + **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 (but the VAT number is not the official registration id in some jurisdictions), BRIS id, OCORP id
  + **Indirect**: GLEI LEI, SDATI Atoka id
  + **Indirect**: BG GUID 617f4edf8c154f4296efdf146513de21, which corresponds to official id 204060254 and can be used to reach the official register page for that company: <https://public.brra.bg/CheckUps/Verifications/ActiveCondition.ra?guid=617f4edf8c154f4296efdf146513de21>
  + **Multiple**: bank license, insurance license, license to trade dual-use goods (arms), Twitter account, Facebook account.
* Note: this is an RDF link from Company to Identifier. The registration date is "Identifier Issued On"

# 4 Datasets

Key functions of euBusinessGraph include:

* Aggregating company datasets from various providers
* Advertising dataset partitions: by jurisdiction, covered fields, etc
* Describing the different subsets that come from different data providers
* Describing internal dataset structure and content (e.g. number of companies per jurisdiction)
* Describing rich dataset metadata, eg source, publisher, dates, license, etc

euBusinessGraph data consumers need to know how many companies are included in a dataset, from which jurisdictions, and what depth of data is included (eg which properties, addresses with what geo resolution, etc). So we need to express both metadata about the dataset itself, and fine-grained statistics about the content of a dataset, eg:

* Publisher, source, last modified, license, home page, download distribution, etc
* Describe subsets by provider
* Describe subsets by kind of entity (e.g. Companies vs Addresses), field coverage (which fields are included in which subsets), entity characteristics (e.g. Italian companies, Startups, Startups in Italy)
* Count of entities in a dataset or subset

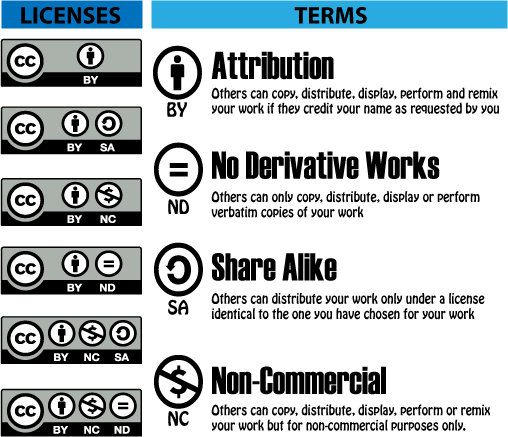
After researching numerous dataset description ontologies (see section [1.8 Dataset Ontologies](#_egkos2e9ec0y)), we settled on using VOID with some extensions. VOID describes RDF datasets (in terms of entities, property/class partitions, number of triples, etc). The original datasets by most providers are not in RDF, so how is VOID applicable?

* The parts of datasets aggregated in euBusinessGraph are in RDF
* We find it useful to describe the kinds of entities and available characteristics (properties) using URLs rather than strings. In this way we harmonize properties across providers, can include extra linked information about them, reuse definitions from other ontologies, and are prepared to capture them in RDF when the opportunity arises.

## 4.1 Dataset Hierarchy

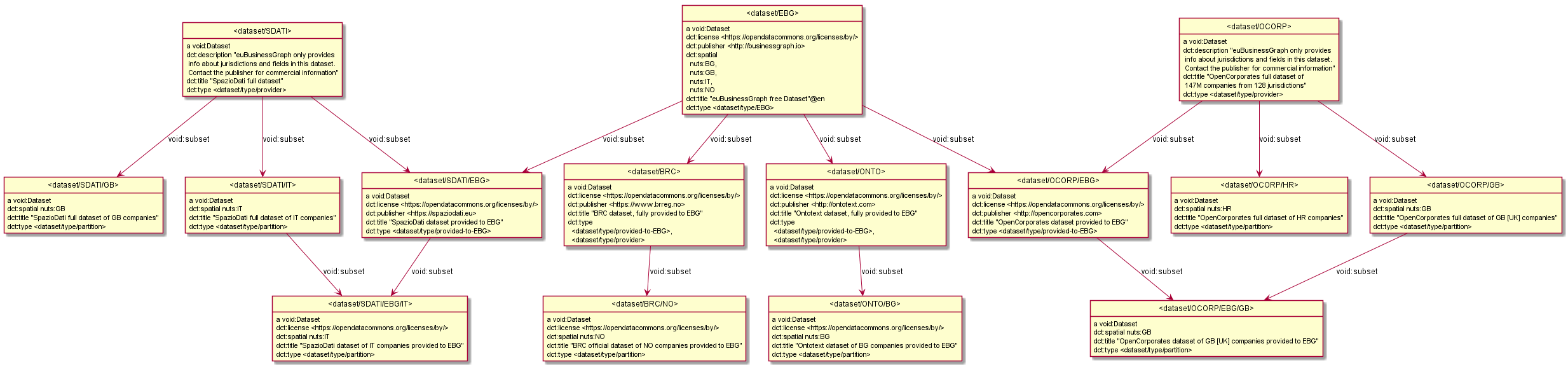
We use the void:subset relation to set out a dataset polyhierarchy:

* Each provider describes their full dataset (eg <dataset/OCORP>), and the respective subset provided to EBG (eg <dataset/OCORP/EBG>). The full dataset may include parts that are not provided to EBG, and are only "advertised" in the EBG marketplace application.
* The EBG dataset (<dataset/EBG>) points to all subsets it aggregates from different providers (<dataset/OCORP/EBG>, <dataset/SDATI/EBG>, <dataset/BRC>, <dataset/ONTO>). Since the last two in this list are fully provided to EBG, there is no need to describe subsets like <dataset/ONTO/EBG>
* A dataset covers one or more jurisdictions. We make a void:subset for each jurisdiction, and for uniformity we do this even if there is only one. We'll capture the precise jurisdiction in a later section. Eg
  + <dataset/OCORP> includes <dataset/OCORP/GB>, <dataset/OCORP/HR>, etc
  + <dataset/ONTO> includes only <dataset/ONTO/BG>
* For each dataset we specify dct:publisher.
* We use dct:type to specify the kind of dataset:
  + <dataset/type/provider>: a provider dataset
  + <dataset/type/provided-to-EBG>: subset provided to EBG
  + <dataset/type/EBG>: the root EBG dataset (aggregation)
  + <dataset/type/partition>: describes a jurisdiction and property partition (see a further section)
* Each dataset must specify a dct:license.
  + **dct:license is mandatory for each <dataset/type/partition> subset** (for both provider and provided-to-EBG datasets), to provision the dataset matrix functionality in the marketplace application.
  + We have used ODC-BY (<https://opendatacommons.org/licenses/by/>), but that is only an example. TODO: EBG should select a specific license, and providers should confirm they're ok with it. "BY" is a bit problematic since giving credit to the specific contributor is not so easy in dataset aggregation scenarios
* Each dct:license URL must have a logo (attached as schema:image) to be displayed in the marketplace application.
  + For ODC-BY we use the OKFN logo
  + CreativeCommons provide widely recognized images for their licenses, eg from [How to Apply a Creative Commons License to Your Work](http://guides.library.queensu.ca/c.php?g=704790&p=5014948) (a Research Guide by Queen's University):



Assume that OCORP has 2 datasets (<dataset/OCORP/GB> and <dataset/OCORP/HR>) of which only the first is provided to EBG (<dataset/OCORP/GB>), but not in the full breadth of data that is available (so we need to describe it separately). Assume that SDATI also has 2 datasets, of which only <dataset/SDATI/EBG/IT> is provided to EBG. Assume that ONTO and BRC have one dataset each, which are fully provided to EBG. Then we end up with the following hierarchy:

(dataset-hierarchy.png from dataset-hierarchy.ttl)



## 4.2 Dataset Metadata

The more descriptive metadata is included about a dataset, the easier it is to discover and use by consumers. Descriptive fields defined in VOID include the following:

* dct:type, dct:license, dct:publisher as described in [Dataset Hierarchy](#_gz749xduhe5r)
* dct:creator: could be the same as dct:publisher, or another party who helped with creating the data
* Source info as described in [Dataset Sources](#_iqgah519zkzu)
* Statistical/partition info as described in [Dataset Partitions](#_57ijy4m8iszq)
* dct:title: name (mandatory)
* dct:description: description (optional)
* dct:subject: topic of the dataset ("companies"). We express it using several LOD resources: <http://dbpedia.org/resource/Company>, <https://www.wikidata.org/entity/Q783794>, <http://vocab.getty.edu/aat/300160084>, <http://eurovoc.europa.eu/4189>
* dct:date: dataset date. Or if more specific dates are desired:
  + dct:created (creation),
  + dct:published (publication),
  + dct:modified (last update of the underlying data). **Providers must provide this field.**
* dct:accrualPeriodicity: how often is the dataset updated and new entries added to it. For the EBG dataset (<dataset>), use freq:irregular (occurs at uneven intervals)
* foaf:homepage, foaf:page: homepage or another web page describing the dataset (also see the use of ebg:WebResource in [Dataset Sources](#_iqgah519zkzu))
* void:sparqlEndpoint: SPARQL endpoint that can be used to query the dataset. Subsets are assumed to be accessible from the same one.
* TODO: void:dataDump: produce and describe dumps in various formats, if desired
* void:exampleResource: several examples of core entities in the dataset
* void:uriSpace: common prefix of the core resources in the dataset. For the EBG dataset, these are "<http://data.businessgraph.io/company/>" and "<http://data.businessgraph.io/person/>". Inherited by subsets.
* void:features: RDF formats the dataset is available in (format:N-Triples, format:RDF\_XML, format:Turtle)
* void:vocabulary: ontologies used in the dataset.
  + For the eBG dataset, those are ebg:, foaf:, locn:, ngeo:, org:, ramon:, rov:, schema:, sioc:, time: .
  + We don't list ontologies used only for dataset metadata (eg void:, format:, freq: )

See data/dataset/dataset-EBG.ttl.

* We don't show a diagram because there's only one node (with a lot of detail).
* In addition to the root company dataset <dataset/EBG>, it describes the subsidiary datasets <dataset/NACE> and <lau>.

## 4.3 Dataset Sources

Each provider dataset should describe the source(s) of data it has used.

* If a dataset omits sources, it should have a single parent, and the sources are assumed to be the same as for the parent.
* Ideally, sources for different jurisdictions should not be mixed together. So there should be separate datasets per jurisdiction, and separate description of sources for each one.

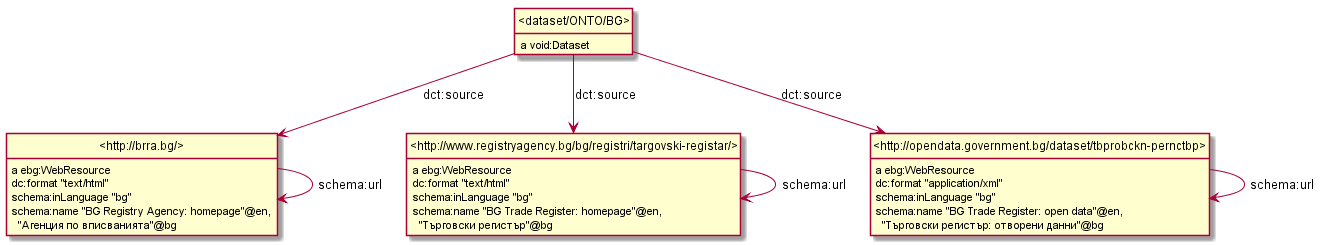
Eg the sources of <dataset/ONTO/BG> include:

* <http://brra.bg>: Registry Agency
* [http://www.registryagency.bg/bg/registri/targovski-registar](http://www.registryagency.bg/bg/registri/targovski-registar/): Trade Register home page
* <http://opendata.government.bg/dataset/tbprobckn-pernctbp>: Trade Register open data

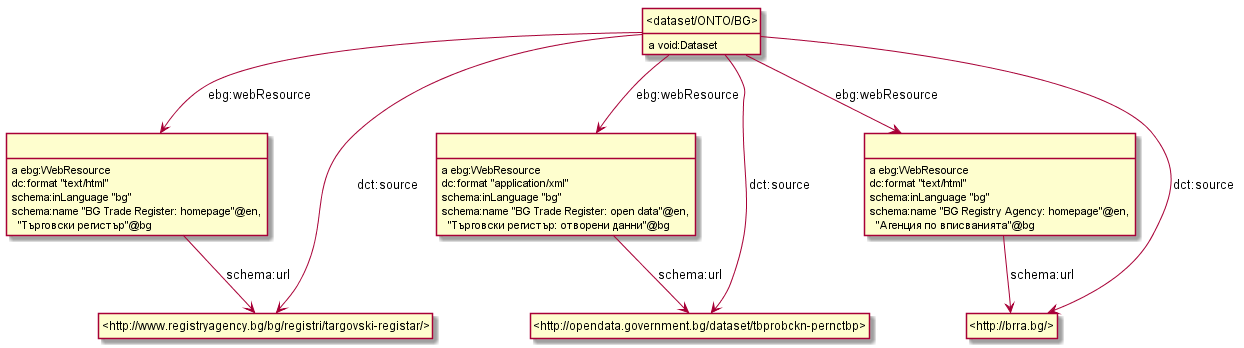
We can describe them:

* Briefly using dct:source, without distinguishing the nature of the source.
* Or with additional detail using [Web Resource](#_f00qpci2owig) (ebg:WebResource)

The example below uses both approaches (dataset-sources.png from dataset-sources.ttl). Note that we have used the actual web URLs as node URLs, and also put them in schema:url, expressed as a self-link.



Another option would be to use blank nodes (or EBG-minted URLs) and connect them using ebg:webResource (dataset-sources1.png from dataset-sources1.ttl). This is a more complex representation, so we don't use it.



## 4.4 Company to Dataset Links

For each company, we'd like to know which provider(s) have provided data about it. Data about the same company may be provided by several providers, e.g. both OCORP and SDATI provide the GB jurisdiction. We use two mechanisms:

* Triple level: each triple is in a graph that reflects the provider dataset it came from.
* Company level: each company has void:inDataset links to the respective provider datasets. These links are made with the SPARQL update query [dataset/inDataset.ru](https://github.com/euBusinessGraph/eubg-data/blob/master/data/dataset/inDataset.ru). These links are used to enable a Provider facet on the EBG Marketplace app.

Eg <company/GB/123456> is provided by both OCORP and SDATI:

graph <provider/ocorp/uk> {

<company/GB/123456> a rov:RegisteredOrganization;

void:inDataset <dataset/OCORP/EBG>;

rov:registration <company/GB/123456/id>;

adms:identifier <company/GB/123456/id/OCORP>.

}

graph <provider/sdati/uk> {

<company/GB/123456> a rov:RegisteredOrganization;

void:inDataset <dataset/SDATI/EBG>;

rov:registration <company/GB/123456/id>;

adms:identifier <company/GB/123456/id/SDATI>.

}

* The company is related by void:inDataset to the OCORP and SDATI datasets provided to EBG (these have dct:publisher pointing to the respective provider)
* All statements including the official ID (rov:registration), company name etc are recorded in both provider graphs.
* The provider-specific IDs (adms:identifier) are recorded only in the respective provider's graph.

TODO: The source of a dataset is often an official register, which is also associated with an identifier system. Therefore we need to express this link.

## 4.5 Dataset Statistics and Partitions

Describing the internal structure of datasets presents endless opportunities. So we present a number of typical examples, and will increase the set as further description needs arise.

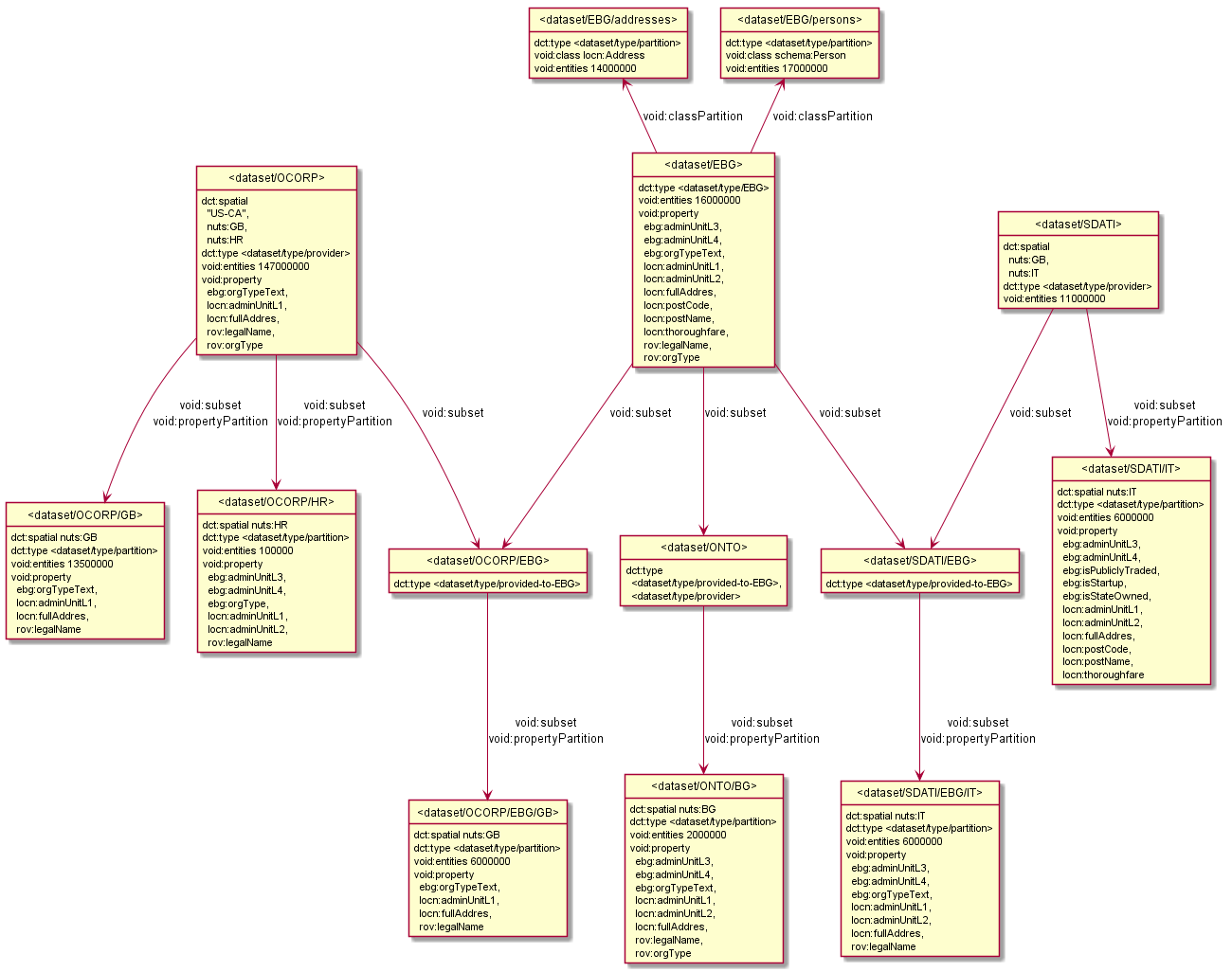
This is done with **partitions**: subsets of a dataset constructed in specific ways. We may use blank nodes for these partitions, and dct:type <dataset/partition> to distinguish them from other kinds of datasets (eg downloadable).

We use the following mechanisms:

* To specify the **jurisdictions** that a dataset covers, we use dct:spatial with a NUTS node for European countries (eg nuts:IT Italy) or literal for others (eg "US-CA" California)
  + This also allows to describe subsets covering only a province, eg nuts:ITH31 Verona.
* To specify the **number** of core entities (i.e. companies) included in a dataset, we use void:entities.
* To describe several **kinds of entities** we use void:classPartition with void:class giving the entity class, and void:entities giving the number.
* To specify **properties** (features) included in a dataset, we use void:propertyPartition with void:property listing all relevant properties (this specifies the dataset has at least some triples using each of those properties; VOID suggests to use only one void:property per partition, but we allow multiple to simplify the representation). No claim is implied that every entity in the dataset has all these properties. A superset should list the union of properties included in its subsets (and may list more, if the subsets don't partition it in full).
* To specify the **number of statements** **with given property**, we use void:propertyPartition with void:property specifying the prop, and void:triples giving the number of statements. If we need to count several specific properties, we need to use several propertyPartitons, with one void:property per partition.

dataset-stats.png (from dataset-stats.ttl) shows:

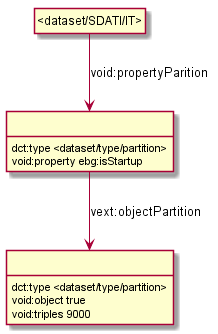
* Total number of companies (main entities) in the overall EBG dataset (<dataset>), its provided subsets (eg <dataset/ONTO>, <dataset/OCORP/EBG), as well as provider datasets (eg <dataset/OCORP>, <dataset/OCORP/GB>)
* Number of entities per class: <dataset/companies> (rov:RegisteredOrganization) as main entity, and optionally <dataset/persons> (schema:Person), <dataset/addresses> (locn:Address) as subsidiary entities.
* Spatial coverage (jurisdictions) of the overall EBG dataset (eg "BG", "GB", "IT", "NO"), its subsets and provider datasets. We show only two for <dataset/OCORP> ("GB" and "HR") but in fact there are more.
  + Note: NUTS and LAU admin units use "UK" eg nuts:UK nuts:UKI nuts:UKI2 nuts:UKI22. However, dbo:jurisdiction and dct:spatial use "GB".
* Features (properties) included in each dataset and/or subset, eg
  + <dataset/OCORP/EBG> includes ebg:orgTypeText (legal form free text), locn:adminUnitL1 (address country), locn:fullAddres (address free text)
  + <dataset/EBG> includes also rov:orgType (legal form nomenclature), locn:adminUnitL2, ebg:adminUnitL3, ebg:adminUnitL4 (administrative place hierarchy), locn:postName, locn:thoroughfare, locn:postCode (address fields). This wider set of features comes from various subsets: there is no implication that all entities have these features.
  + <dataset/SDATI/IT> includes specific Boolean properties ebg:isStartup, ebg:isStateOwned, ebg:isPubliclyTraded, which are not provided to the EBG dataset.
  + **Providers must describe fully the properties included in per-jurisdiction subsets (eg <dataset/OCORP/GB> and per-jurisdiction provided subsets (eg <dataset/OCORP/EBG/GB>).** They may include similar information about their main datasets, but that is not mandatory.
* We use the fact that void:propertyPartition is a sub-property of void:subset thus compatible with it, and use the same node to express both jurisdiction and property partition, connecting that node by both void:propertyPartition and void:subset to the main dataset.



Finally, we introduce one extension of VOID (from the VOID-ext ontology)

* If we also need to specify the **number of statements with given property and object**, we use void:propertyPartition, then vext:objectPartition

Eg dataset-stats1.png (from dataset-stats1.ttl) shows that <dataset//SDATI/IT> includes 9k startup companies. We could use similar mechanisms to express even more complex combinations, eg "number of startup companies in the province of Verona" or what exact properties are provided for these companies.

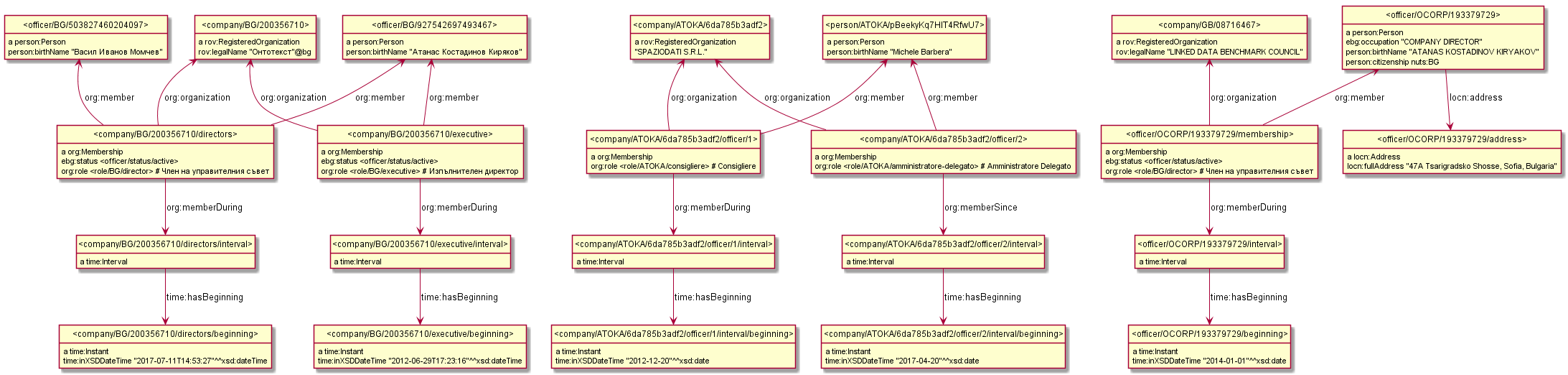


# 5 Officers and Membership

We use the W3C Org Membership model in a straightforward way.

Consider officers.png (made from officers.ttl) that shows the following Officer records:

* one director from GB (Kiryakov at LDBC)
* one executive from BG (Kiryakov at ONTO)
* two directors from BG (Kiryakov and Momtchev at ONTO) using a shared org:Membership node. This works only if the whole board of directors is appointed and discharged at once (as it is at AGMs). To reflect individual directors coming and going, individual Membership records are needed.
* One person (Michele Barbera) with two roles in the organization (ATOKA) that have different roles and intervals, so we use two different Membership nodes.



## 5.1 Officer

To represent officer data we follow [person:]. Only the officer’s name and identifier are mandatory. The identifier may come from official registries or be derived from these. Additionally, other properties may be present such as address, birth date and citizenship.

* Class: Officer
* RDF: person:Person
* URL:
  + If we have the official registration ID of the officer, we use that  
    **officer/<jurisdiction>/<id>**, e.g.   
    the officer officer/BG/927542697493467 (Atanas Kostadinov)
  + Otherwise we use an ID from a data provider (alternative business register)  
    **officer/<provider>/<id>,** e.g.  
    company/ATOKA/pBeekyKq7HIT4RfwU7 (Michele Barbera)
* Description: An officer is a natural person (as opposed to a legal person) [person:] that has a high-level management role in a Company (e.g., the CEO, treasurer and chief financial officer). Despite their high status, they typically serve at the will of the corporate directors, who can fire or replace them. Officers can also be shareholders and directors but don't necessarily have to be. They have the authority to act on behalf of the corporation, including contract authority.
* Scope note:
* Examples: Whether a person is an officer or not is determined by the role they fulfill which must entail management attributes. Among the roles that determine that a person is an officer, we may find:
  + Chief Executive Officer: the head of management on an organisation
  + Chief Operating Officer: responsible for the day-to-day operations of an organisation
  + Chief Financial Officer: in charge of a corporation’s financial risk

### 5.1.1 Officer Name

* Data Property: Birth Name
* Data Type: xsd:string
* Cardinality: 1
* RDF: person:birthName
* Description: The legal name of the person which is usually determined at or around birth, usually persistent and for this reason usually recorded by some official registers. Although birth names tend to be persistent we do not assume this has to be the case.
* Scope Note: transliterated names?
* Rules:
  + Don't allow leading, trailing and two consecutive spaces
* Examples:
  + "Michele Barbera"
  + "Atanas Kostadinov Kiryakov”

### 5.1.2 Officer Address

* Data property: full address
* Data type: rdf:langString or xsd:string
* RDF: [locn:fullAddress](https://www.w3.org/ns/locn#locn:fullAddress)
* Cardinality: 0 or 1
* Description: full address, free text
* Examples: "47A Tsarigradsko Shosse, Sofia, Bulgaria"

## 5.2 Membership

A membership describes the relation between a company officer and the company in which they occupy a position.

* Class: Membership
* RDF: org:Membership
* URL: membership URLs are expressed from the company side of the relation like this:

**<company url>/officer/<id>**, where <id> is a sequential number or type of officer e.g.   
company/BG/927542697493467/executive, company/ATOKA/6da785b3adf2/officer/1.

This works if only the most recent officer records are RDFized (i.e. we have a single Executive and Board of Directors per company). In the general case each Membership depends on person (org:member), organization (org:organization), and interval (org:memberDuring), so a GUID should be used.

* Description: represents the relation between an organisation and a member of that organisation.
* Example: “Michele Barbera became a director at SpazioDati on April 20th, 2012”
* Scope Note: it is acceptable to use the same Membership node for several officers, if they had exactly the same role and membership interval (e.g. in many countries the Board is appointed and discharged at the same interval, namely between AGMs of the company)

### 5.2.1 Membership Company

* Object Property: Membership Company
* RDF: org:organization
* Cardinality: 1
* Range: Company
* Description: Points to the company the membership is about

### 5.2.2 Member Person

* Object Property: Member
* Range / Data Type: Officer
* RDF: org:member
* Cardinality: 1
* Description: the officer that the membership is about
* Examples:
  + <officer/ATOKA/pBeekyKq7HIT4RfwU7> for “Michele Barbera”
  + <officer/BG/927542697493467> for "Атанас Костадинов Киряков"

### 5.2.3 Membership Interval

* Object Property: membership interval
* Range: time:Interval
* Cardinality: 1
* RDF: org:memberDuring
* Description: the interval (beginning/end) of the membership relation. May be an open interval.
* URL: **<membership-URL>/interval**, eg company/BG/200356710/executive/interval

This node includes relations time:hasBeginning and/or time:hasEnd that point to further nodes **<membership-URL>/interval/beginning** and **<membership-URL>/end** having attributes time:inXSDDateTime.

* Scope note: for the common case of missing time, we prefer to use xsd:date rather than completing it with a fake null time, eg "<date>T00:00:00Z"^xsd:dateTime.

### 5.2.4 Membership Role

* Object property: role
* Range: org:Role
* RDF: org:role
* Cardinality: 0 or more
* Description: the role that the officer fulfills according to the membership.
* Scope Note: EBG does not introduce a global set of officer roles. These may vary per jurisdiction and/or provider
* Examples:
  + <role/BG/director>

# 6 SCRATCH

## 6.1 Descriptive Template

* Class / Data Property / Object Property / Lookup Property:
* Range / Data Type
* Cardinality:
* RDF:
* Description:
* Scope Note:
* Rules;
* Examples:

For classes we also include

* URL:

For properties we also include:

* OCORP, BRC, SDATI, etc:
* OCORP example, BRC example, SDATI example

For lookup values, we may include:

* Parent (skos:broader): broader concept (hypernym)

1. [NORMA - The Software!](https://www.ormfoundation.org/files/folders/norma_the_software/default.aspx) The ORM Foundation, accessed 6 August 2017 [↑](#footnote-ref-1)
2. [RDF by Example: rdfpuml for True RDF Diagrams, rdf2rml for R2RML Generation](http://vladimiralexiev.github.io/pres/20161128-rdfpuml-rdf2rml/). Alexiev, V. In Semantic Web in Libraries 2016 (SWIB 16), Bonn, Germany, November 2016. [HTML](http://vladimiralexiev.github.io/pres/20161128-rdfpuml-rdf2rml/index-full.html), [PDF](http://vladimiralexiev.github.io/pres/20161128-rdfpuml-rdf2rml/RDF_by_Example.pdf), [Video](https://youtu.be/4WoYlaGF6DE) [↑](#footnote-ref-2)
3. Cool URIs don't change: <https://www.w3.org/Provider/Style/URI> [↑](#footnote-ref-3)
4. Cool URIs for the Semantic Web, W3C Interest Group Note, 03 December 2008. <https://www.w3.org/TR/cooluris/> [↑](#footnote-ref-4)
5. Designing URI Sets for the UK Public Sector. A report from the Public Sector Information Domain of the CTO Council’s cross-Government Enterprise Architecture. Interim paper, Version 1.0, October 2009, <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf> [↑](#footnote-ref-5)
6. Study on Persistent URIs, with identification of best practices and recommendations on the topic for the Member States and the European Commission, Phil Archer, 2013. <http://philarcher.org/diary/2013/uripersistence/> [↑](#footnote-ref-6)
7. [Schema.org: Evolution of Structured Data on the Web](http://queue.acm.org/detail.cfm?id=2857276). R.V. Guha, Dan Brickley, Steve Macbeth. ACM Queue, November-December 2015 [↑](#footnote-ref-7)
8. Stasinos Konstantopoulos, Angelos Charalambidis, Antonis Troumpoukis, Giannis Mouchakis, and Vangelis Karkaletsis. [The Sevod Vocabulary for Dataset Descriptions for Federated Querying](http://ceur-ws.org/Vol-1927/paper4.pdf). PROFILES 2017 workshop at ISWC 2017. [↑](#footnote-ref-8)
9. Martin Brümmer et al. [DataID: Towards Semantically Rich Metadata for Complex Datasets](http://svn.aksw.org/papers/2014/Semantics_Dataid/public.pdf). In: Proceedings of the 10th International Conference on Semantic Systems. SEM ’14. Leipzig, Germany: ACM, 2014, pp. 84–91 [↑](#footnote-ref-9)
10. Markus Freudenberg, Martin Brümmer, Jessika Rücknagel, Robert Ulrich, Thomas Eckart, Dimitris Kontokostas, Sebastian Hellmann. [The Metadata Ecosystem of DataID](https://svn.aksw.org/papers/2016/MSOR_DataID2/public.pdf), MTSR 2016. [↑](#footnote-ref-10)
11. Mohamed Ben Ellefi, Zohra Bellahsene, John Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, Konstantin Todorov. [RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications](http://www.semantic-web-journal.net/content/rdf-dataset-profiling-survey-features-methods-vocabularies-and-applications). Semantic Web Journal, 2017-08. [↑](#footnote-ref-11)
12. <https://www.capgemini.com/2012/07/why-smart-keys-are-dumb-and-dumb-keys-are-smart/> [↑](#footnote-ref-12)