Semantic Web 13 (2022) 41–68
DOI 10.3233/SW-210424
IOS Press

The euBusinessGraph ontology: A lightweight ontology for harmonizing basic company information

Dumitru Roman [a,*], Vladimir Alexiev [b], Javier Paniagua [c], Brian Elvesæter [a], Bjørn Marius von Zernichow [a], Ahmet Soylu [d], Boyan Simeonov [b] and Chris Taggart [e]

[a] *SINTEF AS, Norway*
E-mails: dumitru.roman@sintef.no, brian.elvesater@sintef.no, BjornMarius.vonZernichow@sintef.no
[b] *Ontotext, Bulgaria*
E-mails: vladimir.alexiev@ontotext.com, boyan.simeonov@ontotext.com
[c] *SpazioDati, Italy*
E-mail: paniagua@spaziodati.eu
[d] *OsloMet – Oslo Metropolitan University, Norway*
E-mail: ahmet.soylu@oslomet.no
[e] *OpenCorporates, UK*
E-mail: chris.taggart@opencorporates.com

Editor: Oscar Corcho, Universidad Politécnica de Madrid, Spain
Solicited reviews: Vojtěch Svátek, Prague University of Economics and Business, Czech Republic; Two anonymous reviewers

Abstract. Company data, ranging from basic company information such as company name(s) and incorporation date to complex balance sheets and personal data about directors and shareholders, are the foundation that many data value chains depend upon in various sectors (e.g., business information, marketing and sales, etc.). Company data becomes a valuable asset when data is collected and integrated from a variety of sources, both authoritative (e.g., national business registers) and non-authoritative (e.g., company websites). Company data integration is however a difficult task primarily due to the heterogeneity and complexity of company data, and the lack of generally agreed upon semantic descriptions of the concepts in this domain. In this article, we introduce the euBusinessGraph ontology as a lightweight mechanism for harmonising company data for the purpose of aggregating, linking, provisioning and analysing basic company data. The article provides an overview of the related work, ontology scope, ontology development process, explanations of core concepts and relationships, and the implementation of the ontology. Furthermore, we present scenarios where the ontology was used, among others, for publishing company data (business knowledge graph) and for comparing data from various company data providers. The euBusinessGraph ontology serves as an asset not only for enabling various tasks related to company data but also on which various extensions can be built upon.

Keywords: Company data, open data, linked data, ontology, business knowledge graph

1. Introduction

Corporate information, including basic company information (e.g., name(s), incorporation data, registered addresses, ownership and related entities, etc.), financials (e.g., balance sheets, ratings, etc.) as well as contextual data (e.g., cadastral data on corporate properties, geo data, personal data about directors and shareholders, public tenders data, etc.) are the foundation that many data value chains depend upon in different

[*] Corresponding author. E-mail: dumitru.roman@sintef.no.

sectors. The most evident examples of sectors are the business information sector, the marketing and sales sector and the business publishing industry. At the same time, the use of company data is extremely significant in many other business sectors and societal activities including transparency and accountability [20].

Recently, a number of initiatives have been established to harmonise and increase the interoperability of corporate and financial data across national borders, including public initiatives such as the Global Legal Entity Identification System – GLEIS,[1] Bloomberg's open FIGI system for securities,[2] as well as long-established proprietary initiatives such as the Dun & Bradstreet DUNS number.[3] Other notable initiatives include the European Business Register (EBR),[4] which aims to federate several national business registers in order to offer a unique point of access, and BREX,[5] which "wraps" the EBR, extends its country coverage and offers a pricing model to access the underlying data. Additionally there are established and widespread adopted standardisation systems in the area of company financials (e.g., official deposited and public balance sheets data, which is in most cases exchanged in the XBRL format[6]). However, due to various reasons including technical, operational and organizational limitations, the systems and data sources mentioned above are mostly fragmented across borders, limited in scope and size,[7] and siloed within specific business communities with limited accessibility from outside their originating sectors. For example, register exchanges only offer access to official national registry data, not linked to any other contextual datasets (i.e., there is no obvious way of following a link from a company's registered data to a tender it has won in another country), nor among themselves across countries (which means that there is no "machine-readable" and easy way to follow, for example, a shareholding relationship from an individual to companies in two different countries).

As a result, collecting and aggregating information about a business entity from several public sources (official and non-official ones, such as public tender registries, press mentions of companies and related entities, cadastral records, etc.), and especially across borders and languages is a tedious and very expensive task which renders many potential business models non-feasible. As a step in addressing this challenge, governments and other public bodies are increasingly publishing open data about firmographics and contextual databases, which reference companies. For example, the UK, Norway, France, and Denmark make the public records about companies available as open data, and other countries have different degrees of openness for their company registries.[8] Examples of contextual databases include the EU TED (Tenders Electronic Daily) public procurement notices,[9] gazette notices, Horizon 2020 project data,[10] and Structural Funds.[11] Unfortunately, firmographics datasets are not yet fully harmonised and interoperable because data differs widely in semantics from one source to the other, and due to data formats ranging from UK's five star Linked Data [18] to poorly accessible and poorly documented ones. Furthermore, contextual databases are not linked to the company registries and they still use different identifier systems or, in some cases, no identifiers at all. Private businesses are also producers of valuable company-related data, which is seldom linked to the public sources mentioned above. For example, media publishers often reference businesses and legal entities by name (hence ambiguously) even within their digital publications (with the exception of traded company tickers, which are sometimes used by specialised financial publishers), because there isn't any widespread markup schema to annotate a digital reference to a company, nor a standardised way of accessing its information once it is unambiguously identified. As a result, it is extremely expensive, time consuming and error prone to find, interpret and reconcile these data from private sector sources. One of the immediate consequences is that the business information sector is very cost-inefficient in itself, which is reflected in a lack of transparency and efficiency of the markets. Nevertheless, the most relevant consequence

[1] https://www.gleif.org

[2] https://en.wikipedia.org/wiki/Financial_Instrument_Global_Identifier

[3] http://www.dnb.com/duns-number.html

[4] http://www.ebr.org

[5] https://brex.io

[6] https://www.xbrl.org

[7] Less than 1.6M companies worldwide were assigned a Legal Entity Identifier (LEI) number as of December 2019 (https://search.gleif.org) and these are only used in financial transactions of certain kind.

[8] https://index.okfn.org/dataset/companies and http://registries.opencorporates.com.

[9] https://ted.europa.eu

[10] https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects

[11] https://cohesiondata.ec.europa.eu

in this context is that these inefficiencies severely harm digital innovation across sectors, which is often introduced by small and agile actors (e.g., startups, civil society organizations) who lack the capacity to invest time and resources in overcoming these problems.

In this article, we follow the established approach for harmonizing and integrating data based on ontologies (e.g., [5,16]). In particular, we develop an ontology – the euBusinessGraph ontology – for harmonising and integrating *basic* company information.[12] The ontology is meant to be used as a key mechanism for aggregating, linking, provisioning and analysing company-related data. In this paper we provide an overview of the related work, ontology scope, ontology development process, explanations of core concepts and relationships, implementation of the ontology, and examples of scenarios where the ontology was used, among others, for publishing company data (business knowledge graph) and for comparing data from various company data providers. The main challenge this paper addresses is related to striking the right balance between a semantic model for basic company information that is too complex and hard to understand and a simplistic least common denominator model, while at the same time exploiting proper mechanisms to reuse numerous related ontologies.

The remainder of the article is organised as follows. Section 2 provides an overview of related work and ontologies relevant to company-related data. Section 3 describes the euBusinessGraph ontology development process, covering the scope, requirements, and the development approach. Section 4 gives an overview of the core concepts and relations in the euBusinessGraph ontology, together with details about the realization of the ontology. Section 5 provides examples of the usage of the ontology. Finally, Section 6 concludes this article and outlines possible future work.

2. Related work

Several ontologies and data models were described in the literature and have relevance to capturing the structure and complexity of company-related data. In what follows, we look specifically at works dealing with *basic* information about companies, covering organizational structures of companies, economic classifications of companies, company identification schemes, and locations of companies.[13] This includes actual ontologies and vocabularies, and also several initiatives and data models relevant in the development of the euBusinessGraph ontology for basic company information.

The ontologies and vocabularies discussed in this section either insufficiently cover basic company information or are too complex due to many ontological commitments. Nevertheless, as we shall see below, relevant ontologies and data models were partly re-used and/or provided inspiration in the development of the euBusinessGraph ontology.

2.1. *Organizational structure*

The W3C Organization Ontology (ORG) [29] is a W3C recommendation since 2014. It aims to capture information about organizations (companies and institutions), including governmental organizations. It primarily captures organizational structure (e.g., sub-organizations and classification), reporting structure (e.g., roles and posts), location information (e.g., sites and addresses), and organizational history (e.g., merger and renaming). ORG is highly generic and designed as a core ontology, capturing general concepts and encouraging extensions for specific domains. It has been reused by other ontologies such as PPROC [25] in the procurement domain.

The e-Government Core Vocabularies [32] were developed in order to provide a minimum level of semantic interoperability for e-Government systems as part of the SEMIC (Joinup) community[14] and the ISA program[15] of the European Commission. They include basic concepts about legal entities, locations, persons, public services, public organizations, and public services. The Core Public Organization Vocabulary (CPOV) and the Core Business Vocabulary (CBV) are the most relevant vocabularies in our context. After ini-

[12] By *basic* we mean company information that is usually covered in Trade Registers, but excluding change tracking (e.g., when a headquarters address changes) and documents (e.g., financial returns) due to the typical unavailability of such data. Examples of *non-basic* information can include transactions (investments, M&A, IPO), company relations (parent/subsidiary, branch, competitor, supplier/client, joint venture), etc.

[13] An overview of all ontologies and vocabularies that were reused in the euBusinessGraph ontology (including those not specifically dealing with basic company information) are discussed in Section 4 (with a summary provided in Table 1).

[14] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic

[15] https://ec.europa.eu/isa2

tial development by EC SEMIC and ISA2, they were transferred to W3C for standardization. The CBV was revised and formally adopted by W3C in 2013 as the Registered Organization Vocabulary (RegOrg).[16] RegOrg is an extension of ORG for describing organizations that have gained legal entity status through a formal registration process, typically in a national or regional register.

The Popolo Project defines data interchange formats and data models in the context of the Open Government initiative.[17] A set of concepts and relations are provided for capturing persons and organizations and the relationships between them (e.g., membership properties). A vocabulary for describing organizations is also provided. This vocabulary reuses terms from the ORG ontology and adds some new ones (e.g., other name, area, and contact detail).

The Application Profile of the Organization Ontology (ORG-AP-OP) was developed by the Publications Office of the European Union and supports its Whoiswho service.[18] It provides actual contact information for staff working at the European Institutions. It is concerned with people and the roles they play in the actual institutions. Similarly, in 2015, the ISA Programme of the EC initiated the development the Core Public Service Vocabulary and its Application Profile (CPSV-AP) [42]. However, it defines a number of terms closely related to CPOV, such as the administrative level, the type of organization, and its home page.

The Schema.org initiative [17] is spearheaded by the big four search engines, Google, Yahoo, Bing and Yandex, and is a collaborative effort to create, maintain, and promote schemas for structured data on the Internet. It is highly reusable since it makes few ontological commitments in order to cater to a truly global audience of millions of Web sites. Schema.org considers schemas as a set of types arranged in a hierarchy and associated with a set of properties. The core vocabulary is currently composed of 614 types and 902 properties. The "Organization" concept is among one of the commonly used types (among with, e.g., person, product, event) and models businesses (e.g., type, contact, etc.) and marketing aspects (e.g., logo, social profile, etc.).

2.2. *Financial and economic*

The Financial Industry Business Ontology (FIBO) [7] is a joint effort of the Enterprise Data Management Council (EDMC) and the Object Management Group (OMG), aiming to go beyond a mere dictionary and capture the semantics of the business domain from a financial perspective. FIBO formalizes entities such as companies, directors, ownership and control relations, business registers, monetary amounts, debts, obligations, contracts, and financial instruments. It is composed of a large number of smaller ontologies, with a modular perspective, each of which models a specific financial area [24]. The result is a large and very complex set of ontologies for the financial industry consisting of 11 core domains and 49 modules made available in more than 400 ontology files.

There are a number of classification vocabularies to specify the kind of economic activity such as International Standard Industrial Classification of All Economic Activities (ISIC) [11], which is a United Nations industry classification system, and European Commission's NACE [14], which is preferred in the context of European interoperability.

The Entity Legal Forms Code List[19] (ISO 20275) by GLEIF provides a world-wide list for types of business entities including a translation to English. Wikipedia's Legal Entity Types[20] also provides approximate equivalents in the company law of English-speaking countries.

2.3. *Company identification and location*

The Global Legal Entity Identifier Foundation (GLEI) established a registration structure to issue Legal Entity Identifiers (LEI) to legal entities participating in financial transactions. The LEI structure is standardized as ISO 17442 [19]. LEI includes two code lists that are relevant in the context of basic company information, that is registration authorities list including 651 national official registers with their descriptions such as authority code, jurisdiction, and website; and, entity legal form code resolving variant names for each valid legal form within a jurisdiction to a single code per legal form.

[16] https://www.w3.org/TR/vocab-regorg
[17] http://www.popoloproject.com/specs
[18] http://whoiswho.europa.eu
[19] https://www.gleif.org/en/about-lei/code-lists/iso-20275-entity-legal-forms-code-list
[20] https://en.wikipedia.org/wiki/List_of_legal_entity_types_by_country

The Business Registers Interconnection System (BRIS) interconnects business registers across Europe and provides a single (though limited) company search form.[21] The list of legal forms, list of national registers, and the pan-European company identifier (which is formed by register and company identifiers) are relevant for capturing basic company information.

With respect to capturing various forms of locations for companies, several initiatives are relevant. Eurostat has established a unified hierarchy of regions across the EU, EFTA and Candidate Countries. It consists of a nomenclature of Territorial Units for Statistics (NUTS) [15] and Local Administrative Units (LAU).[22] NUTS and LAU are important geographic resources since a significant amount of 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); and, NUTS and LAU provide a uniform hierarchy, whereas the administrative hierarchy varies greatly in different countries.

The EU ISA2 Location Core Vocabulary [13] aims at describing any place in terms of its name, address or geometry through a minimum set of classes and properties. It integrates with the Business (i.e., RegOrg) and Person Core Vocabularies of ISA2.

GeoVocab.org[23] provides vocabularies for geospatial modelling. This includes vocabularies NeoGeo Geometry Ontology for describing geographical regions and NeoGeo Spatial Ontology for describing topological relations between features.

Finally, GeoNames[24] provides a free geographical database covering all countries and containing over eleven million place names. It includes data elements such as administrative regions and settlements, and physical places.

2.4. *Other relevant initiatives*

In addition to well known initiatives such as FOAF,[25] Dublin Core[26] and DBPedia,[27] there are other ontologies, vocabularies and initiatives that are relevant in the context of modelling basic company information, including:

- ADMS ontology [10] describes various interoperability assets, including XML schemas, generic data models, code lists, taxonomies, dictionaries, vocabularies. ADMS is relevant in our context since we aggregate free company datasets from various company data providers.
- Vocabulary of Interlinked Datasets (VoID) [1] provides terms and patterns for describing RDF datasets and could be used in a variety of situations such as data discovery, cataloging and archiving of datasets.
- Simple Knowledge Organization System (SKOS) [4] offers a vocabulary for expressing the basic structure and content of concept schemes. This is essential for example for company classification (e.g., type and status).
- The IANA language code registry[28] uses ISO 639-1, 639-2 and 639-3 language codes (2 and 3-letter codes) and extends it with additional info (script, region of use, dialect). It can be consumed more easily from a Google sheet generated in Feb 2018.[29] Language tags are relevant in our context as some information (e.g., company names, street addresses) may be available in different languages.
- Person Core Vocabulary[30] aims at describing natural persons with a minimum set of classes and properties and is developed under the ISA Programme of the European Union. It is essential for representing people for example playing different roles in an organization.
- The Simple Event Model ontology (SEM) [41] is created for modelling events in a variety of domains and it is relevant for capturing different events in the lifetime of a company.

3. euBusinessGraph ontology development

In order to design the euBusinessGraph ontology, we applied common techniques recommended by well established ontology development methods [8,26]. We used a bottom-up approach by identifying the scope and user group of the ontology, requirements, and

[21] https://e-justice.europa.eu

[22] https://ec.europa.eu/eurostat/web/nuts/local-administrative-units

[23] http://geovocab.org

[24] http://www.geonames.org

[25] http://xmlns.com/foaf/spec

[26] https://dublincore.org

[27] https://wiki.dbpedia.org

[28] https://www.iana.org/assignments/language-tags/language-tags.xml

[29] https://docs.google.com/open?id=1M1yv9aBUmc-NyCJX69vOLUmH2uIglSwmDwgRgByI1AI

[30] https://www.w3.org/ns/person

ontological and non-ontological resources (some of which are referred to in Section 2).

One of the main resources used during the ontology development was company data that was provided by four company data providers and that needed to be harmonized before further processing. The data providers were OpenCorporates,[31] SpazioDati,[32] Brønnøysund Register Centre,[33] and Ontotext.[34] The data made available by the data providers originally came from both official sources (e.g., national and regional company registers) and unofficial sources (e.g., the corporate web, business-centric news aggregators and social networks). In the following we provide a brief description of the data provisioned by the four data providers:

- OpenCorporates provides core company data on over 180 million entities, obtained from more than 130 company registers around the world. The data is sourced only from official public sources and full provenance is provided. The depth of data varies from jurisdiction to jurisdiction, sometimes including officers, industry codes, even occasionally shareholders and ultimate beneficial owners.
- SpazioDati integrates detailed up-to-date company and contact information on legal entities in Italy and the United Kingdom. Their dataset contains basic firmographics about more than 11 million business entities in both jurisdictions and information about 13 million directors and managers. Data comes from both authoritative sources (e.g., Registro imprese, the Italian Register of Companies and all the regional chambers of commerce) and non-authoritative sources (e.g., company websites, social media accounts, and business-centric news websites).
- Brønnøysund Register Centre (Brønnøysundregistrene) maintains the Norwegian Central Coordinating Register for Legal Entities (Enhetsregisteret)[35] – a database that contains information on all legal entities in Norway such as commercial enterprises and governmental agencies. It also includes business sole proprietorships, associations and other economic entities without registration duty that have chosen to join the register on a voluntary basis.
- Ontotext extracted data from the Bulgarian Trade Register. This register provides a centralized database whose purpose is to facilitate the start-up of businesses in Bulgaria, as well as to curb corruption practices.

These data sources were analyzed to determine the scope and requirements of the ontology. They cover official company information in Bulgaria, Norway, Italy and the United Kingdom, with additional unofficial information for the later two jurisdictions.

3.1. Scope and requirements

After an analysis of the data provided by the different providers and the information available therein, we identified the major concerns that the ontology should address. Figure 1 provides an overview of the different types of information found during the data analysis, organized according to the type of entity being described (*Registered Organization* and *Officer*). In addition, the ontology needed to cover the description of dataset offerings by individual data providers (*Dataset*) and the description of identifier systems used to uniquely identify companies (*Identifier System*).

We identified target domains for our ontology, which primarily map to the business information sector, the marketing and sales sector, and the business publishing industry interested in new innovative data-driven products and services. Users working with data in these domains will benefit from a common representation that covers the types of information contributed by the different data providers. This common representation will also ease the task of data providers and aggregators who need to validate, transform and clean the data by providing a single ontology to target. The fact that there is a single ontology that provides a common representation will also benefit service developers who need to reference company information to implement their services. To this end, the ontology has to capture the properties of the different identifiers that can be used to link the different entities being represented, providing machine readable descriptions for the identifier systems in use, including support for describing rules for validation and normalization of company and company-related identifiers.

Taking into account the needs of the intended users of the ontology and after the analysis of the data provided, we elicited the ontology requirements following the Neon methodology [39] for requirements specification, and considering two kinds of requirements.

[31] https://opencorporates.com
[32] http://spaziodati.eu
[33] http://www.brreg.no
[34] https://www.ontotext.com
[35] https://data.brreg.no/enhetsregisteret/oppslag/enheter

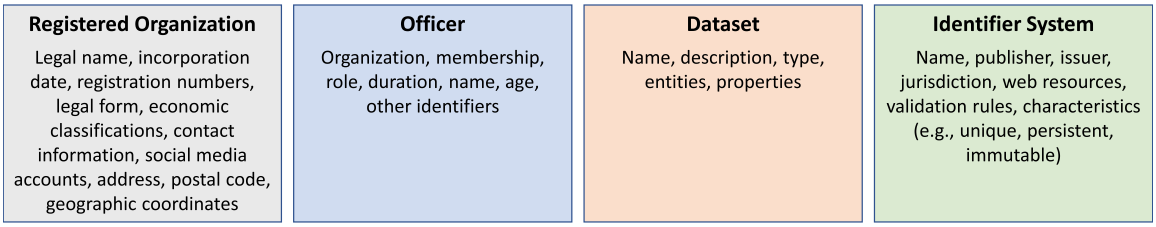
Registered Organization
Legal name, incorporation date, registration numbers, legal form, economic classifications, contact information, social media accounts, address, postal code, geographic coordinates
Officer
Organization, membership, role, duration, name, age, other identifiers
Dataset
Name, description, type, entities, properties
Identifier System
Name, publisher, issuer, jurisdiction, web resources, validation rules, characteristics (e.g., unique, persistent, immutable)

Fig. 1. Overview of the scope of the euBusinessGraph ontology.

Firstly, we consider functional requirements dealing with the scope of data and groups of competency questions that the data should be able to answer. The functional requirements include:

(1) Capture the concept of a company, representing the different types or legal forms that companies can take, their jurisdictions and registration information, legal and alternative names, official and secondary locations, prevalent economic activity, web keywords and social media accounts, among others;
(2) Capture the concept of company officers, their roles and officerships, including temporal information to be able to represent these officerships through time;
(3) Provide machine-readable descriptions of the properties of the different systems of identifiers available to external applications and services, so that algorithms can be developed to select and prioritise the most suitable identifiers for a task (this includes provisioning of validation and cleaning rules for identifiers to help their usage); and
(4) Provide data advertising and extensibility features, including description of additional properties of company and company-related entities that are not covered by the model but are available from company data providers as unique or differentiating features.

Examples of groups of competency questions that the data should be able to answer include:

(1) What companies are relevant to the search keyword "Opel"?
(2) What are the jurisdictions and legal types of companies matching that search keyword?
(3) What companies match an industry classification such as "Automotive companies"?
(4) What are alternative names for the company "Opel Group GmbH"?
(5) What jurisdiction does the company "Opel Group GmbH" belong to?
(6) What is the official address of "Opel Group GmbH"?
(7) Does the company "Opel Group GmbH" have other locations (additional addresses)?
(8) What companies are located in "Rüsselsheim am Mein"?
(9) What are the economic activities registered for the company "Opel Group GmbH"?
(10) Is the company "Opel Group GmbH" publicly traded? Is it a startup? Is it government-owned?
(11) How many companies in a given industry classification are available in each country, NUTS statistical region, province, county, and city?
(12) What are the points of contact (phone, email) of "Opel Group GmbH"?
(13) What is the online presence (web, email, blog) of "Opel Group GmbH"?
(14) What is the Wikipedia page of the company "Opel Group GmbH"?
(15) Who are the officers of "Opel Group GmbH" (including historical timeline)?
(16) What is the legal company type of "Opel Group GmbH"?
(17) What is the current status of "Opel Group GmbH"?
(18) Which jurisdictions are covered by which data provider datasets?
(19) What is the number of companies and persons in each data provider dataset?
(20) What additional properties are available for "Opel Group GmbH" from different data providers?
(21) What additional properties are available from different data providers, and what is their coverage across companies?

Secondly, we consider non-functional requirements dealing with the general requirements or aspects that

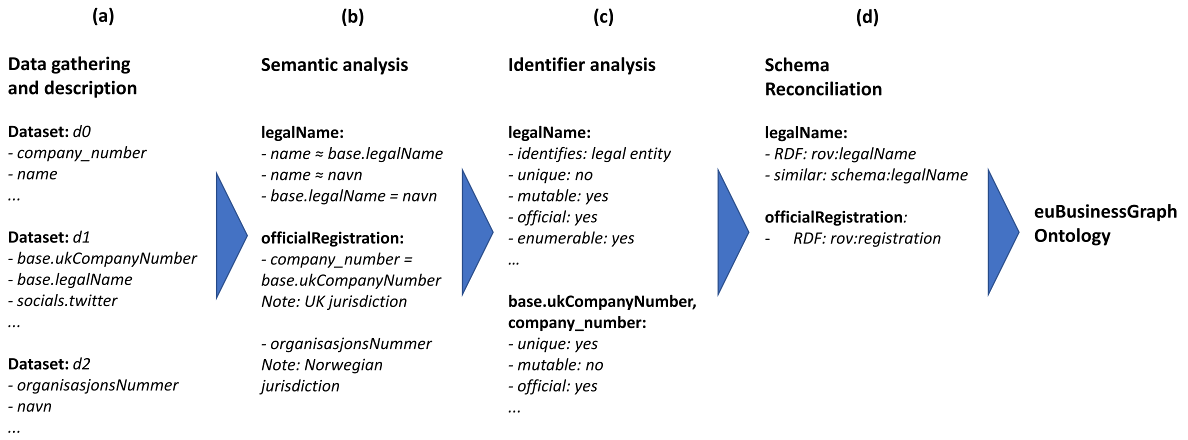

Fig. 2. Phases of the euBusinessGraph ontology development process.

the ontology should fulfill. Foremost amongst these is **reuse**:[36]

(1) Reuse existing ontologies as often as possible, in order to reduce effort and promote the use of the integrated data;
(2) Make the developed ontology as easy to reuse as possible;
(3) Use simple and pragmatic mechanisms with low "ontological commitment" rather than complex ontological mechanisms (see Section 3.3); and
(4) Support the integration of all company data provided by at least one data provider, spanning authoritative and non-authoritative sources and modelled in different ways under a single representation schema.

3.2. *Ontology development*

The ontology development process was guided by the need to harmonize and integrate datasets with different sets of attributes, different representations for the same entity and in some cases close but not entirely similar semantics. Figure 2 depicts the four phases of the ontology development process in which we (a) gathered data from all company data providers that include natural language descriptions and example instances of each data attribute they provided, (b) analyzed attribute descriptions, refining them with additional notes describing their scope and using this information to group similar attributes, (c) analyzed identifiers and their identifier systems to produce machine readable descriptions of their properties, and (d) carried out manual reconciliation with the aim to reuse existing vocabularies.

There are differences in the types of information available from source to source (e.g., one dataset contains only official information from the national registers, while another integrates contact information parsed from company websites), differences in the way the same bit of information is represented by each provider (e.g., addresses as strings or as complex objects with separate attributes for street number, name and municipality) and differences in semantics for closely related concepts that may appear to be the same (e.g., information about officerships and their durations that contain references to possibly ambiguous officer names versus log entries that link person identification numbers to roles in different companies through time).

In the first phase of the ontology development process, as shown in Fig. 2(a), each data provider provided a description of the dataset they shared. This data analysis focused on identifying the different attributes present and the way in which they were represented. Each attribute was described, adding notes and example uses that clarified the semantics as deemed appropriate. In this phase we already identified similar or even *same-as* candidates (e.g., *company_number*, *base.ukCompanyNumber*, *organisasjonsNummer* in Fig. 2(a)). Moreover, each provider specified to which extent a particular attribute was shared, in one of three modalities: (i) fully available, (ii) fully available to perform entity matching, but not available in any other case, and (iii) fully available for matching but available in reduced form for other purposes (e.g., address in-

[36] Six of the nine scenarios in the "Neon Book" Chapter 2.3 (Nine Scenarios for Building Ontology Networks) revolve around reuse.

formation without street numbers). Analyzing the descriptions provided in the previous phase, we identified a common subset shared by all contributed datasets. This common subset contained attributes that represented the same or very similar concepts in all datasets, which allowed us to group attributes from different providers accordingly (see similar attributes grouped under the *legalName* label across different providers in Fig. 2(b)).

In the next phase, exemplified in Fig. 2(c), we performed a different analysis to assess the suitability of each attribute to work as an identifier of the instance it described. The analysis contained a heterogeneous group of attributes with identifying characteristics: identifiers for geographical entities, legal entities, company headquarters and secondary sites, company websites, among others. Within the provided data, we found several ways to identify an instance in a group of similar instances (e.g., registration numbers and legal names are two different and useful ways to identify a company). Some identifiers are ambiguous in nature, such as company names, while others can be used to uniquely refer to a company, as is often the case with company registration numbers. The expectation is that the former will often be found in unstructured texts while the latter will be useful to annotate those unstructured texts to link to the corresponding instance being referred to. Some identifiers belong to official registers while others are self-issued and not centralized (e.g., websites). Some identifiers are subject to particular geographic jurisdictions (e.g., company registrations in local trade registers), or belong to special registers that attest that companies belong to a certain class (e.g., register of startup companies). In other cases, identifiers simply indicate the database in which the company information can be found (e.g., identification codes issued by data providers such as OpenCorporates, codes issued by other companies that aggregate company data such as Dun & Bradstreet), the website of a company or the various associated social network identifiers (e.g., a company's Facebook page or Twitter handle).

In light of the varied nature of the identifiers available, it was determined that the semantic model should also represent key aspects of the different identifier systems in use. These key aspects should encode expectations of the identifiers issued under each system and provide readily available rules to aid in validation and transformation of these identifiers. The expectations should help to determine the suitability of a particular indicator for common use cases that included publishing, reconciliation and matching within unstructured text. Additionally, the semantic model should provide links to information about issuing authorities and maintainers, revisions, databases and other resources.

In the last phase of the development process, as exemplified in Fig. 2, we searched within existing vocabularies for all the concepts identified in the common subset aiming to reuse whenever possible. Examples of reuse from appropriate ontologies include W3C Org, RegOrg, Location, Person (not W3C), Schema.org and ADMS datasets and identifiers.

Differences in the ways each provider decided to share the various attributes present in their datasets made it necessary to understand the scope of the ontology as early in the process as possible. In this way, it was possible to determine what to cover while having a clear path for extensibility.

3.3. Reuse approach

One reuse approach is to create own terms (classes and properties), and tie them to existing ontologies using semantic mapping properties (e.g., `equivalentClass`, `subClassOf`, `equivalentProperty`, `subPropertyOf`). This has the benefit of a single name space, which may make it easier to produce data conforming to the ontology. Such approach is often used in very large ontologies, e.g., Wikidata or Schema.org. Using such semantic mappings has its cost, since it complicates inferencing and querying requirements. For lighter-weight ontologies such as euBusinessGraph, we prefer to reuse terms from existing ontologies directly. We also prefer to define our own terms only when we cannot find appropriate terms in existing ontologies. For example, we added the following classes (see Sections 4.2.1 and 4.2.3):

- `IdentifierSystem` with its characteristics: Who created an identifier system and when, characteristics (Boolean flags), validation regex, etc. We use `adms:Identifier` but it only has a link to the issuing organization, not to the system the identifier is part of.
- `WebResource`: A URL described with its purpose, MIME type and language. Schema.org has appropriate properties, but one needs to use an overly abstract class.
- `IdentifierWebResource`: A templatized URL that can return information about a company when its identifier is substituted, and is again de-

scribed with its purpose, MIME type and language. Wikidata has a similar feature: its external-id properties and the meta-property `formatter Url`, but it is not easy to reuse.

Furthermore, we want our own terms to be easily reusable. This is facilitated when terms do not carry a lot of "ontological baggage", such as deep class hierarchies or strong bindings of properties to classes. RDFS defines the properties `rdfs:domain` and `rdfs:range`, which are strict and "prescriptive": according to RDFS semantics, every time the described property is used, its subject (respectively object) gets the domain (respectively range) as type. This also makes these RDFS properties "monomorphic": they should take single values, otherwise the subject/object will get multiple types, which are usually not intended. E.g., if "name" is applicable to Person and Organization with the axiom `:name rdfs:domain :Person, :Organization`, entities with "name" will become both Person and Organization, which is an often made mistake. Ontology engineers overcome this problem by introducing abstract super-classes (e.g., Actor or even Nameable), using `owl:union` to make a disjunction of several classes, or using `owl:Restriction` to bind the property to the class locally. But all of these approaches complicate the ontology, increase "ontological commitment", and ultimately make ontology reuse harder.

Schema.org describes many real-world entities, is applicable in a wide variety of domains, and integrates data from a huge number of providers and domains. The Web Data Commons crawl from 2019-12[37] found 44B triples about 14B entities at 934M pages from 12M web domains, and the majority of that data is in Schema.org. To cope with this web-scale integration of data, Schema uses properties `schema:domainIncludes` and `schema:rangeIncludes`, which are advisory and "descriptive" (describe properties applicable to a class, without being exclusive), therefore polymorphic: an axiom like `:name schema:domainIncludes :Person, :Organization` doesn't cause any unintended types to be inferred. In addition to Schema.org, the same approach is used in the Web of Things initiative.[38]

This approach has much lower "ontological commitment" and we find that it enables more flexible reuse and combination of different ontologies, so it is appropriate in the company data domain, where data comes from a large variety of providers. Rather than using complicated OWL mechanisms, we prefer to use RDF Shapes to validate incoming data from data providers.

4. Ontology overview

The euBusinessGraph ontology is composed of 20 classes, 33 object properties, and 57 data properties (see Table 1) that make it possible to represent basic company-related data. Figure 3 gives an overview of the ontology, depicting the main classes and their relationships (i.e., object properties). The ontology covers the following areas:

(1) **Registered Organization**: The focal point of the ontology is companies that are registered as legal entities. Companies gain legal entity status by the act of registration. The class **`Registered Organization`** is used to represent such a company. A company can have several **`Sites`**, for which the official registered site where legal papers can be served is captured by the object property `hasRegisteredSite`. A site can have an **`Address`**. Moreover, a company can have several different **`Resources`** associated in order to capture, e.g., `url` and `email` information.

(2) **Identifier System**: A company can have several **`Identifiers`**, for which the official registration is captured by the object property `registration`. An identifier is part of an **`Identifier System`**. Both the **`Identifier`** and the **`IdentifierSystem`** can have a `creator` of either a type **`Person`** or a type **`Organization`**.[39] The **`IdentifierSystem`** also has additional **`IdentifierWebResources`** and **`Web Resources`** information associated.

(3) **Officer**: A company has associated officers, e.g., directors. The class **`Membership`** is used to as-

[37] http://webdatacommons.org/structureddata/2019-12/stats/stats.html

[38] See https://www.w3.org/2019/wot/td and try curl-Haccept:text/turtle https://www.w3.org/2019/wot/td.

[39] Note: We use **`RegisteredOrganization{rov}`** for companies, and **`Organization{schema}`** for identifier system creators, identifier issuers, and data providers (let's call them "auxiliary orgs"). Whether an auxiliary org is also a registered company is not a relevant concern for the ontology, because the number of auxiliary orgs is very small.

Table 1

Prefixes, namespaces, and count of classes and properties used in the euBusinessGraph ontology

Prefix	Ontology (namespace)	Classes	Object properties	Data properties
adms	Asset Description Metadata Schema (http://www.w3.org/ns/adms#)	1	1	
dbo	DBpedia (http://dbpedia.org/ontology/)			1
dc	DCMI Metadata Terms – Elements (http://purl.org/dc/elements/1.1/)			1
dct	DCMI Metadata Terms (http://purl.org/dc/terms/)		2	1
ebg	**The euBusinessGraph Ontology (http://data.businessgraph.io/ontology#)**	**4**	**8**	**24**
locn	ISA Programme Location Core Vocabulary (http://www.w3.org/ns/locn#)	1	2	7
org	The Organization Ontology (http://www.w3.org/ns/org#)	3	7	
person	Core Person Vocabulary (http://www.w3.org/ns/person#)	1		1
ramon	Reference and Management of Nomenclatures (http://rdfdata.eionet.europa.eu/ramon/ontology/)	1		3
rdf	Resource Description Framework (http://www.w3.org/1999/02/22-rdf-syntax-ns#)	1		
rov	Registered Organization Vocabulary (http://www.w3.org/ns/regorg#)	1	4	1
schema	Schema.org (http://schema.org/)	2	6	14
sioc	SIOC Core Ontology (http://rdfs.org/sioc/ns#)		1	
skos	Simple Knowledge Organization System RDF Schema (http://www.w3.org/2004/02/skos/core#)	2		3
time	Time Ontology in OWL (http://www.w3.org/2006/time#)	2	2	1
void	Vocabulary of Interlinked Datasets (http://rdfs.org/ns/void#)	1		
	Total number of entities	**20**	**33**	**57**

sociate officer data. It connects a **`Registered Organization`** with a **`Person`** through a **`Role`**.

(4) **Dataset**: Finally, in order to capture information about datasets that are offered by company data providers, we include the class **`Dataset`** that can have relevant **`WebResources`** information associated.

Further details about the **Registered Organization**, **Identifier System**, **Officer** and **Dataset** ontology areas, covering the full set of classes, object properties and data properties, are given in Sections 4.1, 4.2, 4.3 and 4.4 respectively. Moreover, Section 4.5 presents validation rules for the ontology.

The class diagrams (depicting the ontology classes, object properties and data properties) and the object diagrams (depicting instances of the ontology classes and properties) in this section were created using the Graphical Ontology Editor (OWLGrEd).[40] An overview of the graphical elements in OWLGrEd for visualizing ontologies can be found in [6]. OWLGrEd expresses classes, namespaces, object properties, data properties and their data types, as well as cardinality in a visual manner. The notation **`Registered Organization{rov}`** on a class refers to the term **`RegisteredOrganization`** defined in the namespace `rov`. The notation `legalName{rov}: string{xsd}[1..*]` on a data property refers to the term `legalName` defined in the namespace `rov`, that

[40] http://owlgred.lumii.lv

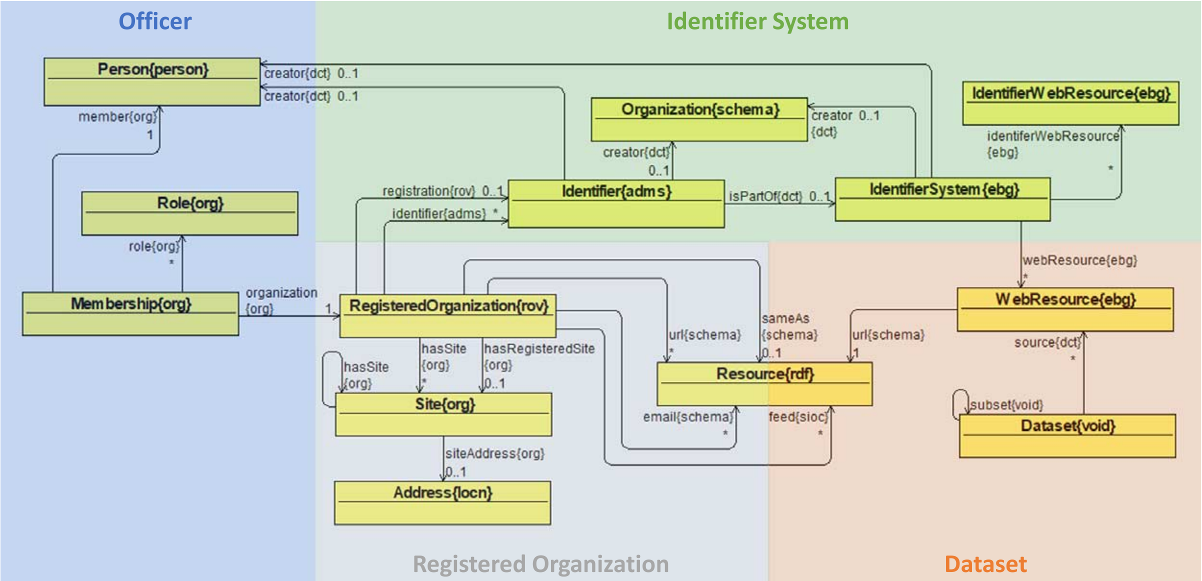

Fig. 3. euBusinessGraph ontology overview: main classes and their relationships.

has the datatype `string` defined in the namespace `xsd`, and a cardinality of `1..*` (i.e., one or more). In the following descriptions of the ontology, we omit the namespace for classes since the context is given. However, we list the namespace for object and data properties as a class may reuse properties from different namespaces.

We reused classes and properties from existing ontologies and nomenclatures where appropriate in order to build the ontology. Table 1 lists the prefixes, namespaces, and count of classes, object and data properties used in the euBusinessGraph ontology, including those reused from the other ontologies. Looking at the count of classes and properties per ontology, it can be seen that the euBusinessGraph ontology adds relatively few, which are mostly terms around describing identifier systems. The novelty of the proposed lightweight semantic model lies in the careful combination and reuse of terms from existing ontologies, and its expression as a detailed model or application profile and RDF Shapes.

Availability of the ontology and related materials. The ontology, datasets and examples described in this paper are released as open source[41] on the euBusinessGraph GitHub repository.[42] The repository contains the ontology source file,[43] the ontology reference documentation[44] generated with pyLODE,[45] and the sources for the full example[46] used throughout this article. Additional materials related to the ontology include a semantic model with informative descriptions [3], a poster [2], and the ontology home page.[47]

4.1. Registered organization

Registered organizations are the main entities for which information is captured in the euBusinessGraph ontology. The ontology is not concerned with unregistered informal groups. Registered organizations gain legal entity status by the act of registration and are distinct from the broader concept of organizations, groups or, in some jurisdictions, sole traders. Figure 4 shows the classes and properties for representing core data about a registered organization. The class **`RegisteredOrganization`** contains names and other basic information about an organization such as `legalName{rov}` and `juris-`

[41] The ontology is released under the Open Data Commons Attribution License (ODC-By).

[42] https://github.com/euBusinessGraph/eubg-data

[43] https://raw.githubusercontent.com/euBusinessGraph/eubg-data/master/model/ebg-ontology.ttl

[44] https://rawcdn.githack.com/euBusinessGraph/eubg-data/master/ontology/doc.html

[45] https://github.com/RDFLib/pyLODE

[46] https://github.com/euBusinessGraph/eubg-data/tree/master/example

[47] https://www.eubusinessgraph.eu/eubusinessgraph-ontology-for-company-data

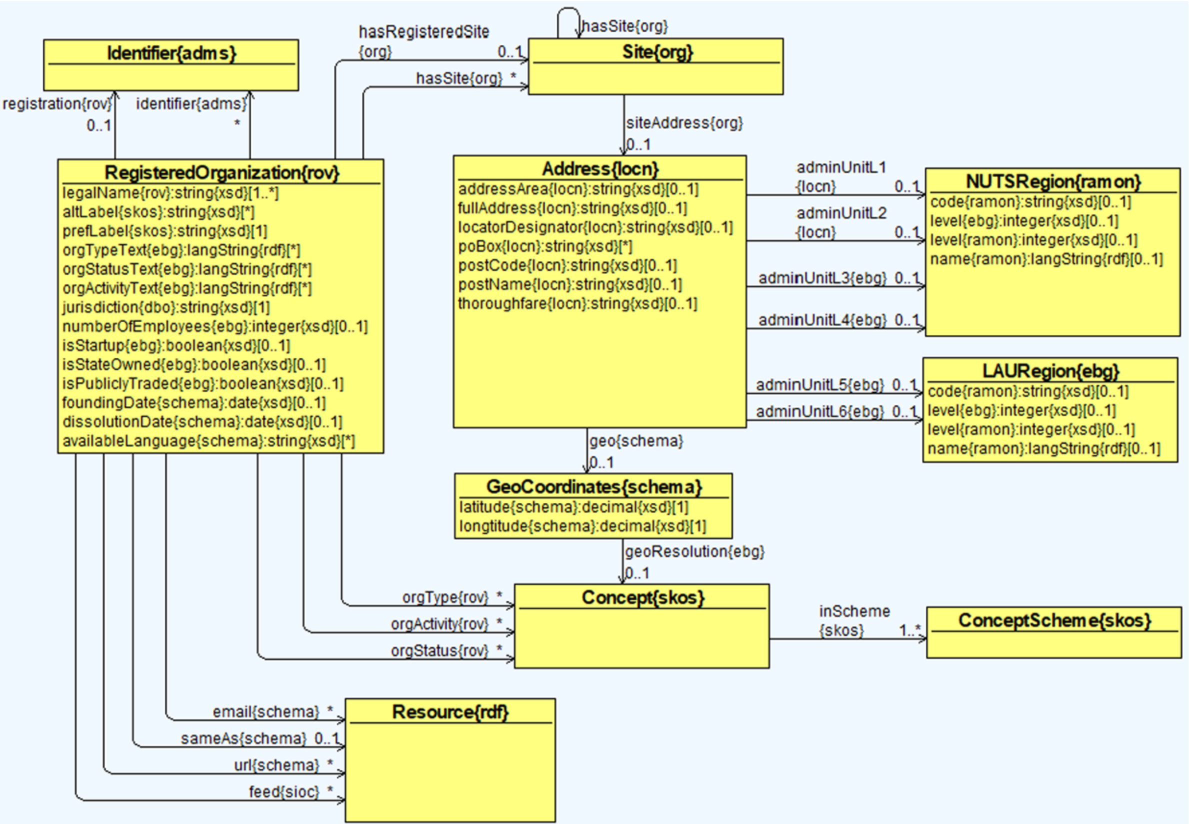
Identifier{adms}
hasRegisteredSite {org} 0..1
hasSite{org}
Site{org}
hasSite{org} *
registration{rov} 0..1
identifier{adms} *
siteAddress{org} 0..1
RegisteredOrganization{rov}
legalName{rov}:string{xsd}[1..*]
altLabel{skos}:string{xsd}[*]
prefLabel{skos}:string{xsd}[1]
orgTypeText{ebg}:langString{rdf}[*]
orgStatusText{ebg}:langString{rdf}[*]
orgActivityText{ebg}:langString{rdf}[*]
jurisdiction{dbo}:string{xsd}[1]
numberOfEmployees{ebg}:integer{xsd}[0..1]
isStartup{ebg}:boolean{xsd}[0..1]
isStateOwned{ebg}:boolean{xsd}[0..1]
isPubliclyTraded{ebg}:boolean{xsd}[0..1]
foundingDate{schema}:date{xsd}[0..1]
dissolutionDate{schema}:date{xsd}[0..1]
availableLanguage{schema}:string{xsd}[*]
Address{locn}
addressArea{locn}:string{xsd}[0..1]
fullAddress{locn}:string{xsd}[0..1]
locatorDesignator{locn}:string{xsd}[0..1]
poBox{locn}:string{xsd}[*]
postCode{locn}:string{xsd}[0..1]
postName{locn}:string{xsd}[0..1]
thoroughfare{locn}:string{xsd}[0..1]
adminUnitL1 {locn} 0..1
adminUnitL2 {locn} 0..1
adminUnitL3{ebg} 0..1
adminUnitL4{ebg} 0..1
NUTSRegion{ramon}
code{ramon}:string{xsd}[0..1]
level{ebg}:integer{xsd}[0..1]
level{ramon}:integer{xsd}[0..1]
name{ramon}:langString{rdf}[0..1]
adminUnitL5{ebg} 0..1
adminUnitL6{ebg} 0..1
LAURegion{ebg}
code{ramon}:string{xsd}[0..1]
level{ebg}:integer{xsd}[0..1]
level{ramon}:integer{xsd}[0..1]
name{ramon}:langString{rdf}[0..1]
geo{schema} 0..1
GeoCoordinates{schema}
latitude{schema}:decimal{xsd}[1]
longtitude{schema}:decimal{xsd}[1]
geoResolution{ebg} 0..1
orgType{rov} *
orgActivity{rov} *
orgStatus{rov} *
Concept{skos}
inScheme {skos} 1..*
ConceptScheme{skos}
email{schema} *
sameAs{schema} 0..1
url{schema} *
feed{sioc} *
Resource{rdf}

Fig. 4. Registered organization: main classes and properties.

diction{dbo} (see Section 4.1.1), supports different types of classifications such as orgActivity{skos}, orgType{skos} and orgStatus{skos}) (see Section 4.1.2). An organization can have several online resources associated such as email{schema} (see Section 4.1.3). A registered organization has a public site/address where legal papers can be served, and possible other sites/addresses. The sites/addresses are represented using the classes **Site** and **Address** (see Section 4.1.4). The object property registration{rov} denotes the identifier of a company. The identifier system is described in further details in Section 4.2.

4.1.1. Names and other basic information

The ontology adopts two different name types for a registered organization, namely formal legal names and informal alternative names, e.g., a trading name. In addition we code a single name as the preferred name of the organization. The **Registered Organization** class has the following data properties to record names:

- legalName{rov}: The legal name of the company, i.e., the official name of a company. A company 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, i.e., conversion of a legal name in one alphabet to another, e.g., from Russian to Latin.
- altLabel{skos}: Alternative names, e.g., an informal or popular name of the company. We also use this for former names.
- prefLabel{skos}: A single preferred name of a company.

The ontology defines the following data properties for capturing additional basic information about an organization:

- jurisdiction{dbo}: Jurisdiction in which the company is registered.
- numberOfEmployees{ebg}: The number of employees in the company.

- `isStartup{ebg}`: Whether the company is a startup.
- `isStateOwned{ebg}`: Whether this company is owned by the government, a government agency, municipality, city or other public entity. In many cases it is not possible to compute this attribute without access to a shareholder register, so it may be missing.
- `isPubliclyTraded{ebg}`: Whether the company is publicly traded (listed at a stock exchange).
- `foundingDate{schema}`: Date when the company was created.
- `dissolutionDate{schema}`: Date the company was dissolved or removed from register.
- `availableLanguage{schema}`: Languages used by the company.

4.1.2. *Classifications*

Three types of classifications are defined in the ontology for representing the company type, status and economic activity of a **`RegisteredOrganization`**:

- `orgType{rov}`: Company type (legal form of the entity). There is no set of company types that is standardized across jurisdictions. Each jurisdiction will thus have a limited set of recognized company types. These should be expressed in a consistent manner in a SKOS concept scheme. Values are taken from the euBusinessGraph company type concept scheme[48] that covers jurisdictions NO, UK, IT and BG defined in collaboration with the data providers.
- `orgTypeText{ebg}`: Company type (legal form of the entity) given in the form of free text.
- `orgStatus{rov}`: The operational and/or legal registration status of the entity, e.g., whether a company is active or not. There is no globally accepted list of company statuses. For inactive, some providers look at hard evidence (i.e., that the company was deregistered), others at dissolution date in the past, or an extended period of inactivity (dormant). Because of this, a user cannot assume that active and inactive are opposites. A best practice for recording status levels is to use the relevant jurisdiction's terms and to encode these in a SKOS concept scheme. Values are taken from the euBusinessGraph company status concept scheme[49] that covers jurisdictions NO, GB and BG, and statuses from data providers OpenCorporate, SpazioDati and LEI. This concept scheme was defined in collaboration with the data providers.
- `orgStatusText{ebg}`: Company status as it comes from a data provider (free text).
- `orgActivity{rov}`: Economic activity is recorded using a controlled vocabulary based on EC NACE 2. Values are taken from the euBusinessGraph NACE concept scheme[50] which implements the NACE 2 vocabulary.
- `orgActivityText{ebg}`: Economic activity of the organization (free text).

The nomenclature value (SKOS concept) is used in faceting and semantic search. The free-text value is used to provide additional detail and facilitate full-text search. Many IT systems include such redundant info: both nomenclature (codified) fields and free-text fields with additional detail or nuance. E.g., in the museum data domain, CDWA[51] and LIDO[52] have "indexing" vs "display" properties.

4.1.3. *Online resources*

We represent commonly used electronic resources and channels (website, Wikipedia, email, news feed) as specific object properties of a company pointing to a **`Resource`** class:

- `email{schema}`: Email that is officially registered and with the same validity as certified mail.
- `sameAs{schema}`: Wikipedia page pertaining to the company.
- `url{schema}`: Website pertaining to the company or URL of a web resource.
- `feed{sioc}`: URL of RSS/Atom feed pertaining to the company.

4.1.4. *Sites and addresses*

Physical presence of companies is defined via addresses. We model address in a structured way using a set of attributes such as country, macroregion, province, etc. Addresses may have geographic loca-

[48] https://raw.githubusercontent.com/euBusinessGraph/eubg-data/master/data/lookups/EBG-company-type.ttl

[49] https://github.com/euBusinessGraph/eubg-data/blob/master/data/lookups/EBG-company-status.ttl

[50] https://raw.githubusercontent.com/euBusinessGraph/eubg-data/master/data/NACE/nace.ttl

[51] https://www.getty.edu/research/publications/electronic_publications/cdwa

[52] http://cidoc.mini.icom.museum/working-groups/lido/lido-technical/specification

tions specified with a different resolution level. Least precise geographic location are resolved at the level of a country, while most precise are geographical points that specify location up to a street and house number. We also enable data providers to provide full addresses in the form of a free text, which is essentially a string that combines all attributes together into a human-readable format. To provide RDF binding for the attributes, we considered two ontologies. From the ISA Programme Location Core Vocabulary we reused structured attributes such as `fullAddress{locn}` that specifies the full address in a free-text form. To represent geographic coordinates, Schema.org was used as it provides a simpler way to model geographic coordinates via two properties (`latitude{schema}` and `longitude{schema}`).

We distinguish between registered, and other kinds of addresses. Many jurisdictions have the concept of registered address, i.e., the legal address where summons, subpoenas and other legal documents can be sent. An address is modelled using the **`Site`** and **`Address`** classes. A **`Site`** of a company is connected using the object property `hasSite{org}`. A registered site is additionally connected using the object property `hasRegisteredSite{org}`. A **`Site`** connects to an **`Address`** through the object property `siteAddress{org}`.

The class **`Address`** represents a mailing or physical address of the company and has the following properties:

- `fullAddress{locn}`: Full address, free text.
- `adminUnitL1{locn}`: Country of the address.
- `adminUnitL2{locn}`: NUTS1 region of the address.
- `adminUnitL3{ebg}`: NUTS2 region of the address.
- `adminUnitL4{ebg}`: NUTS3 region of the address.
- `adminUnitL5{ebg}`: LAU1 region of the address. Some countries (e.g., Bulgaria) use both LAU1 and LAU2 levels. Others (e.g., Italy) use only LAU2.
- `adminUnitL6{ebg}`: LAU2 region of the address.
- `postName{locn}`: Locality/city/settlement of the address, free text.
- `addressArea{locn}`: Part of a city, village or neighbourhood.
- `thoroughfare{locn}`: Street name (and optionally number).
- `locatorDesignator{locn}`: Street number and/or building name.
- `postcode{locn}`: Postal code of the address.
- `poBox{locn}`: Some addresses are associated with a PO box instead of a street address.

NUTS values are assigned using the EU NUTS classification as Linked Data (NUTS-RDF) datasets.[53] The NUTS-RDF datasets cover 34 European countries and use the `NUTSRegion` class to represent the NUTS regions. In order to represent the lower-level LAU regions we introduced the `LAURegion` class and created our own set of LAU-RDF datasets[54] covering 32 jurisdictions (including all of the EU and EEA), 26 languages, and both LAU territorial levels (lau4, lau5). LAU-RDF datasets were created from the official Eurostat Excel spreadsheet for 2016[55] for EU, and our own research on some other jurisdictions.

4.1.5. Example

Figure 5 is an object diagram depicting how the ontology is used to represent company data about the legal entity SpazioDati. Each object (depicted as a green rectangle) is an instance of a class defined in the ontology. The objects have data properties according to the class definitions. The data properties are assigned values depicted using the notation `data property = "value"`. Some properties are mandatory (multiplicity of `1..`) whereas others are optional (cardinality of `0..` or `*`). Not all information about a company is available from a data provider. Thus an object will only contain the data properties that we are able to retrieve from the data provider. This may vary greatly from data provider to data provider, and from jurisdiction to jurisdiction.

4.2. Identifier system

Mechanisms to identify companies in various data sources are essential in integration of data about companies across data sources. A proper understanding of what kind of systems of identifiers can be used for companies is thus necessary in this context. We analyzed various types of identifiers commonly used for companies and collected various properties of the systems they are part of. We modelled identifiers and iden-

[53] http://nuts.geovocab.org

[54] https://github.com/euBusinessGraph/eubg-data/tree/master/data/LAU/rdf

[55] https://ec.europa.eu/eurostat/documents/345175/501971/EU-28_LAU_2016

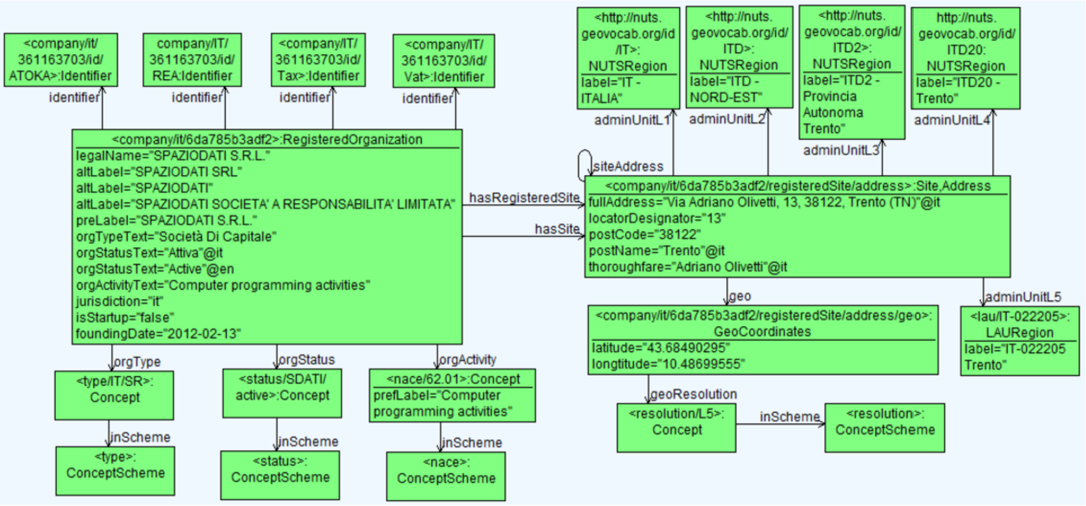
<company/it/361163703/id/ATOKA>:Identifier
company/IT/361163703/id/REA:Identifier
<company/IT/361163703/id/Tax>:Identifier
<company/IT/361163703/id/Vat>:Identifier
identifier
<http://nuts.geovocab.org/id/IT>:NUTSRegion
label="IT - ITALIA"
<http://nuts.geovocab.org/id/ITD>:NUTSRegion
label="ITD - NORD-EST"
<http://nuts.geovocab.org/id/ITD2>:NUTSRegion
label="ITD2 - Provincia Autonoma Trento"
http://nuts.geovocab.org/id/ITD20:NUTSRegion
label="ITD20 - Trento"
adminUnitL1
adminUnitL2
adminUnitL3
adminUnitL4
<company/it/6da785b3adf2>:RegisteredOrganization
legalName="SPAZIODATI S.R.L."
altLabel="SPAZIODATI SRL"
altLabel="SPAZIODATI"
altLabel="SPAZIODATI SOCIETA' A RESPONSABILITA' LIMITATA"
preLabel="SPAZIODATI S.R.L."
orgTypeText="Società Di Capitale"
orgStatusText="Attiva"@it
orgStatusText="Active"@en
orgActivityText="Computer programming activities"
jurisdiction="it"
isStartup="false"
foundingDate="2012-02-13"
hasRegisteredSite
hasSite
siteAddress
<company/it/6da785b3adf2/registeredSite/address>:Site,Address
fullAddress="Via Adriano Olivetti, 13, 38122, Trento (TN)"@it
locatorDesignator="13"
postCode="38122"
postName="Trento"@it
thoroughfare="Adriano Olivetti"@it
geo
adminUnitL5
<company/it/6da785b3adf2/registeredSite/address/geo>:GeoCoordinates
latitude="43.68490295"
longtitude="10.48699555"
<lau/IT-022205>:LAURegion
label="IT-022205 Trento"
geoResolution
<resolution/L5>:Concept
inScheme
<resolution>:ConceptScheme
orgType
orgStatus
orgActivity
<type/IT/SR>:Concept
<status/SDATI/active>:Concept
<nace/62.01>:Concept
prefLabel="Computer programming activities"
inScheme
<type>:ConceptScheme
<status>:ConceptScheme
<nace>:ConceptScheme

Fig. 5. Example of company representation for SpazioDati.

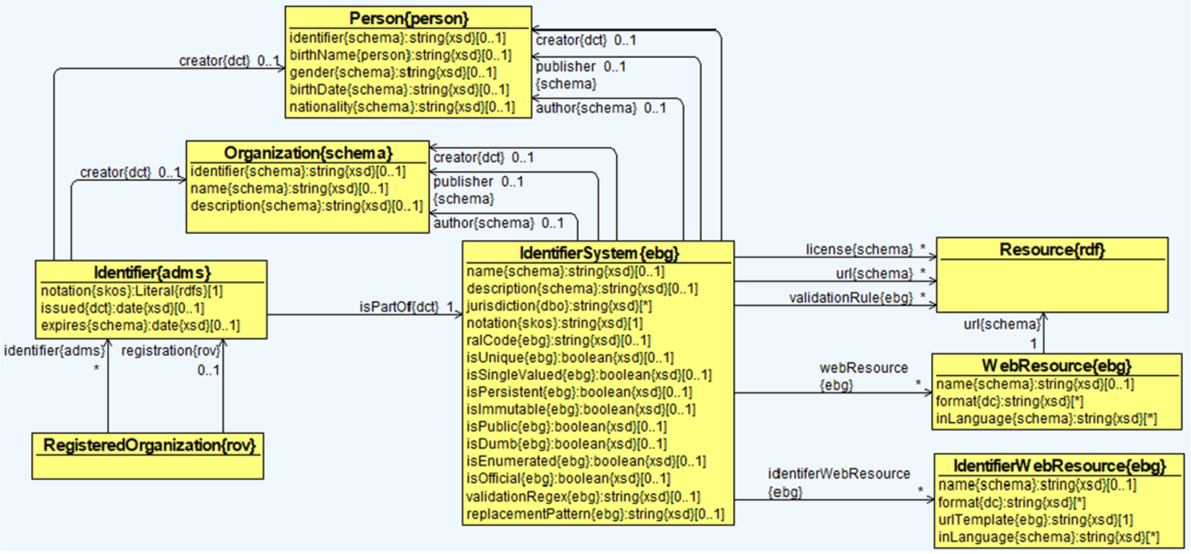
Person{person}
identifier{schema}:string{xsd}[0..1]
birthName{person}:string{xsd}[0..1]
gender{schema}:string{xsd}[0..1]
birthDate{schema}:string{xsd}[0..1]
nationality{schema}:string{xsd}[0..1]
creator{dct} 0..1
publisher 0..1 {schema}
author{schema} 0..1
creator{dct} 0..1
Organization{schema}
identifier{schema}:string{xsd}[0..1]
name{schema}:string{xsd}[0..1]
description{schema}:string{xsd}[0..1]
creator{dct} 0..1
publisher 0..1 {schema}
author{schema} 0..1
creator{dct} 0..1
Identifier{adms}
notation{skos}:Literal{rdfs}[1]
issued{dct}:date{xsd}[0..1]
expires{schema}:date{xsd}[0..1]
isPartOf{dct} 1
identifier{adms} *
registration{rov} 0..1
RegisteredOrganization{rov}
IdentifierSystem{ebg}
name{schema}:string{xsd}[0..1]
description{schema}:string{xsd}[0..1]
jurisdiction{dbo}:string{xsd}[*]
notation{skos}:string{xsd}[1]
ralCode{ebg}:string{xsd}[0..1]
isUnique{ebg}:boolean{xsd}[0..1]
isSingleValued{ebg}:boolean{xsd}[0..1]
isPersistent{ebg}:boolean{xsd}[0..1]
isImmutable{ebg}:boolean{xsd}[0..1]
isPublic{ebg}:boolean{xsd}[0..1]
isDumb{ebg}:boolean{xsd}[0..1]
isEnumerated{ebg}:boolean{xsd}[0..1]
isOfficial{ebg}:boolean{xsd}[0..1]
validationRegex{ebg}:string{xsd}[0..1]
replacementPattern{ebg}:string{xsd}[0..1]
license{schema} *
url{schema} *
validationRule{ebg} *
Resource{rdf}
url{schema} 1
webResource {ebg} *
WebResource{ebg}
name{schema}:string{xsd}[0..1]
format{dc}:string{xsd}[*]
inLanguage{schema}:string{xsd}[*]
identiferWebResource {ebg} *
IdentifierWebResource{ebg}
name{schema}:string{xsd}[0..1]
format{dc}:string{xsd}[*]
urlTemplate{ebg}:string{xsd}[1]
inLanguage{schema}:string{xsd}[*]

Fig. 6. Classes, object properties and data properties for representing identifier systems and identifiers.

tifier systems explicitly in the ontology as shown in Fig. 6.

A **RegisteredOrganization** can have several **Identifiers** issued by different issuers for different purposes. This is modelled by having each company identifier belong to an **IdentifierSystem** (see Section 4.2.1). In this way, we can differentiate between an “official registration” in official business registers and “alternative registrations” in other kinds of registers. While they have the same nature, only the former can be used to uniquely identify a company in each jurisdiction, and to confirm existence of the company as a legal entity in this jurisdiction. Other registrations may not be unique or persistent. The ontology models the different cases through properties that describe the lifecycle of each identifier issued and by encoding a series of characteristics of the identifier system to which the identifier belongs (see Section 4.2.2). Additionally, we model Web resources (see Section 4.2.3) that are frequently found for identifier systems such as search endpoints, templates for building identifier URLs (through which company information can be reached) and other resources that describe the system’s rules. Finally, the model captures the representation of different agents (see Section 4.2.4) that are in charge of setting and maintaining rules, issuing identifiers and publishing identifier databases.

4.2.1. Identifier and identifier system

The **`Identifier`** class represents a company identifier. It has the following object and data properties:

- `isPartOf{dct}`: System the identifier is a part of.
- `creator{dct}`: The issuer of the identifier. In many countries there is a single registry although in others, such as Spain and Germany, multiple registries exist. If the system has an issuer, in most cases the identifier issuer will coincide with that issuer.
- `notation{skos}`: Literal value of the identifier.
- `issued{dct}`: Date when the identifier was issued.
- `expires{schema}`: Date when the identifier expires.

The **`IdentifierSystem`** class represents a system managed by a publisher (e.g., a register or agency) that is used to issue identifiers to companies. Many registers keep several identifier systems. There can be three different types of agents related to a system. This is modelled using three different object properties:

- `author{schema}`: The author who is in charge of specifying the rules and organization of the system.
- `creator{dct}`: The issuer who issues identifiers and then keeps them in a database (register).
- `publisher{schema}`: The publisher who publishes the identifier database (register) in some form.

4.2.2. Identifier system properties and characteristics

Identifier systems have some basic properties:

- `name{schema}`: Name of the identifier system.
- `description{schema}`: Description of the identifier system.
- `jurisdiction{dbo}`: Jurisdiction to which the identifier system applies.
- `notation{skos}`: Short mnemonic code for the identifier system, used in its URL. Also used in identifier URLs that are part of the system. Issued locally by euBusinessGraph. For identifier systems published by the sole or preferred official register in a jurisdiction, we use the jurisdiction code (e.g., “BG”, “GB”). For others, if the identifier system has no explicit name, we use a short mnemonic code of the publisher: upper-case for company registers (e.g., “OCORP” for OpenCorporates, “SDATI” for SpazioDati, “BRC” for Brønnøysund Register Centre, “RAL”, “EU”, “BRIS”), mixed-case for social network registers (e.g., “Twitter”, “Facebook”).
- `ralCode{ebg}`: GLEI RAL code for the identifier system.
- `url{schema}`: Various websites of the identifier system and/or its associated issuer and register, e.g., home page, search, download.
- `license{schema}`: License that applies to the system.
- `webResource{ebg}`: Web resource(s) associated with an identifier system.
- `identiferWebResource{ebg}`: Identifier Web resource(s) associated with an identifier system.

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

- `isUnique{ebg}`: Whether each identifier in the system relates to only one entity.
- `isSingleValued{ebg}`: Whether each entity has only one identifier in the system.
- `isPersistent{ebg}`: Whether identifiers cannot be removed from the register (e.g., when a company is dissolved).
- `isImmutable{ebg}`: Whether identifiers cannot change.
- `isPublic{ebg}`: Whether identifiers from the system are available for public use: consulting, search or download.
- `isDumb{ebg}`: “Intelligent” or “smart” identifiers contain built-in “intelligence” (semantic information) embedded in the identifier. This is increasingly considered bad practice, 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.
- `isEnumerated{ebg}`: Whether the system has an issuer, and issued identifiers are kept in a database (register). For example, every official register is enumerated, while websites are not enumerated.

- `isOfficial{ebg}`: Whether the system is considered the official one in all jurisdictions in which it applies.

Identifier systems are associated with some properties that can be useful for identifier validation:

- `validationRule{ebg}`: URL providing human or machine-readable rule(s) for validating identifiers in the system.
- `validationRegex{ebg}`: Regular expression for validating identifier values of that system.
- `replacementPattern{ebg}`: Pattern to use together with the `validationRegex{ebg}` to normalize identifier values by removing optional decorations. For example, "$1$2$3" can be used together with the regular expression "(\d2)-?(\d3)-?(\d4)" to extract the pure digits from a DUNS number spelled with optional dashes, e.g., to transform "36-032-1459" into "360321459".

4.2.3. *Web resources*

A Web resource is a URL complemented with a MIME type to specify what the URL is about. These web resources are used for identifier systems (e.g., to provide the search or download URL) and per-company, as a URL template in which to substitute the identifier value. There can be several MIME types because some URLs return various resource types using content negotiation. The class **`WebResource`** has the following object and data properties:

- `url{schema}`: URL of the Web resource.
- `name{schema}`: Name or short (generic) description of the resource.
- `format{dc}`: MIME type(s) of the resource. If several are provided, the server must provide all these resource types using content negotiation.
- `inLanguage{schema}`: Language of the Web resource.

The class **`IdentifierWebResource`** has the mandatory data property `urlTemplate{ebg}` in addition to the three data properties defined for `WebResource`. The property `urlTemplate{ebg}` specifies a template that can be used uniformly to build URLs for all identifiers in the system. The template value can have placeholders that should be interpreted as follows:

- If it has a placeholder `{}`, substitute the identifier value there.
- If it has placeholders like `$1`, `$2`,..., substitute the groups extracted by the `validationRegex` of the `IdentifierSystem`.

4.2.4. *Agents*

We represent an agent using either a **`Person`** or **`Organization`** class, depending on the type of agent. For both types, we define the `identifier{schema}` data property which can be assigned a textual identifier or a URL value. For **`Organization`**, we additionally assign values to the data properties `name{schema}` and `description{schema}`. For **`Person`**, we introduce a set of data properties (see Section 4.3 for further details).

4.2.5. *Example*

An example of an identifier system is shown in Fig. 7, illustrating the OpenCorporates identifier system for which OpenCorporates is the publisher and the official UK identifier system for which Companies House is the publisher.

4.3. *Officer*

An officer is a natural person (as opposed to a legal person) that has a high-level management role in a company, e.g., executives and directors, and other important roles (e.g., secretary, legal council and treasurer). Officers have the authority to act on behalf of the corporation, including contract authority. Officers can also be shareholders. However, since few jurisdictions have rules for beneficial ownership reporting, and even those who do still have very little data about it,[56] shareholders were considered out of scope for the euBusinessGraph ontology.

We use the membership model[57] of the W3C Organization Ontology in a straightforward way to represent officer data as shown in Fig. 8. An officer is represented using a **`Person`** class for which the properties `identifier{schema}` and `birthName{person}` are mandatory. The identifier may come from official registries or be derived from these. Additionally, other properties may be present such as `gender{schema}`, `birthDate{schema}` and `nationality{schema}`.

A **`Membership`** describes the relation between an officer and the company in which they occupy a position. The **`Role`** defines the position the officer fulfills according to the membership. Ideally, the roles should be defined according to a SKOS concept scheme. We have not defined a global set of officer roles as this may vary per jurisdiction and/or provider. Thus,

[56] https://www.openownership.org/

[57] https://www.w3.org/TR/vocab-org/#membership-roles-posts-and-reporting

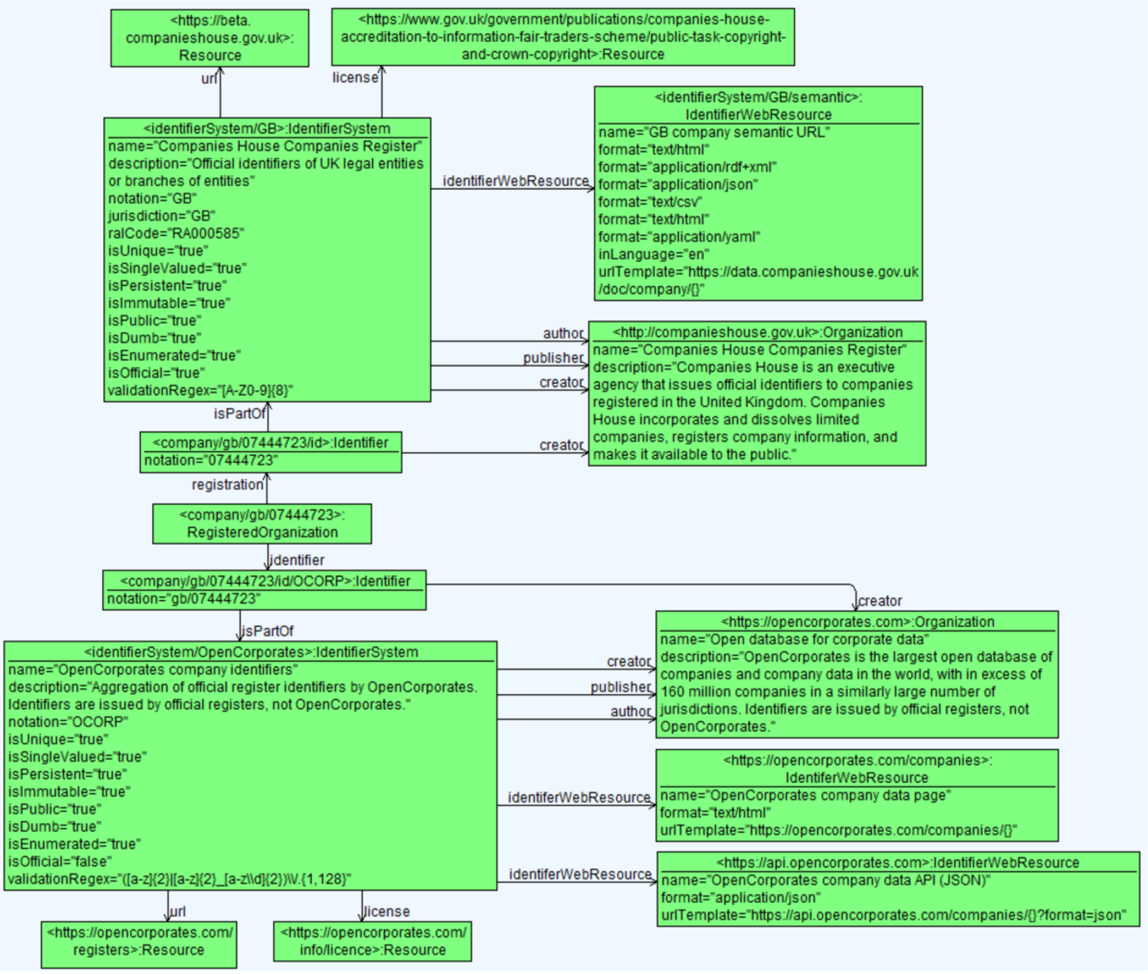
<https://beta.companieshouse.gov.uk>:Resource
<https://www.gov.uk/government/publications/companies-house-accreditation-to-information-fair-traders-scheme/public-task-copyright-and-crown-copyright>:Resource
url
license
<identifierSystem/GB/semantic>:IdentifierWebResource
name="GB company semantic URL"
format="text/html"
format="application/rdf+xml"
format="application/json"
format="text/csv"
format="text/html"
format="application/yaml"
inLanguage="en"
urlTemplate="https://data.companieshouse.gov.uk/doc/company/{}"
<identifierSystem/GB>:IdentifierSystem
name="Companies House Companies Register"
description="Official identifiers of UK legal entities or branches of entities"
notation="GB"
jurisdiction="GB"
ralCode="RA000585"
isUnique="true"
isSingleValued="true"
isPersistent="true"
isImmutable="true"
isPublic="true"
isDumb="true"
isEnumerated="true"
isOfficial="true"
validationRegex="[A-Z0-9]{8}"
identifierWebResource
author
publisher
creator
<http://companieshouse.gov.uk>:Organization
name="Companies House Companies Register"
description="Companies House is an executive agency that issues official identifiers to companies registered in the United Kingdom. Companies House incorporates and dissolves limited companies, registers company information, and makes it available to the public."
isPartOf
<company/gb/07444723/id>:Identifier
notation="07444723"
creator
registration
<company/gb/07444723>:RegisteredOrganization
identifier
<company/gb/07444723/id/OCORP>:Identifier
notation="gb/07444723"
creator
isPartOf
<https://opencorporates.com>:Organization
name="Open database for corporate data"
description="OpenCorporates is the largest open database of companies and company data in the world, with in excess of 160 million companies in a similarly large number of jurisdictions. Identifiers are issued by official registers, not OpenCorporates."
<identifierSystem/OpenCorporates>:IdentifierSystem
name="OpenCorporates company identifiers"
description="Aggregation of official register identifiers by OpenCorporates. Identifiers are issued by official registers, not OpenCorporates."
notation="OCORP"
isUnique="true"
isSingleValued="true"
isPersistent="true"
isImmutable="true"
isPublic="true"
isDumb="true"
isEnumerated="true"
isOfficial="false"
validationRegex="([a-z]{2}|[a-z]{2}_[a-z\\d]{2})\/.{1,128}"
creator
publisher
author
identiferWebResource
<https://opencorporates.com/companies>:IdentiferWebResource
name="OpenCorporates company data page"
format="text/html"
urlTemplate="https://opencorporates.com/companies/{}"
identiferWebResource
<https://api.opencorporates.com>:IdentifierWebResource
name="OpenCorporates company data API (JSON)"
format="application/json"
urlTemplate="https://api.opencorporates.com/companies/{}?format=json"
url
license
<https://opencorporates.com/registers>:Resource
<https://opencorporates.com/info/licence>:Resource

Fig. 7. Example of representing the OpenCorporates identifier system and the Companies House official UK identifier system.

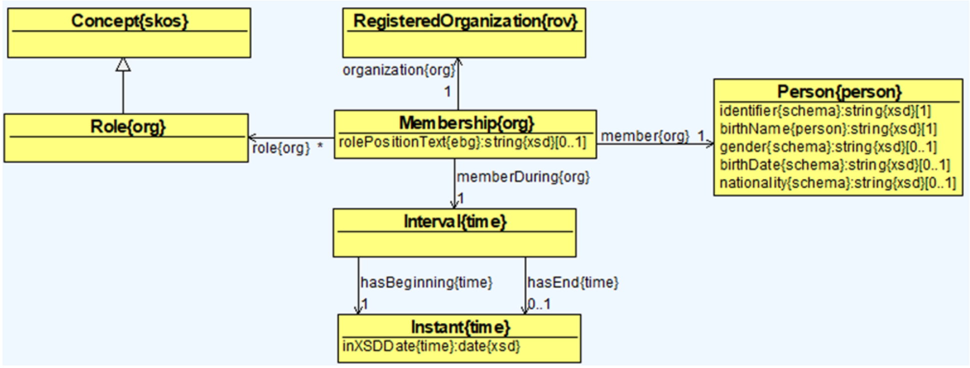
Concept{skos}
RegisteredOrganization{rov}
organization{org}
1
Role{org}
role{org} *
Membership{org}
rolePositionText{ebg}:string{xsd}[0..1]
member{org} 1
Person{person}
identifier{schema}:string{xsd}[1]
birthName{person}:string{xsd}[1]
gender{schema}:string{xsd}[0..1]
birthDate{schema}:string{xsd}[0..1]
nationality{schema}:string{xsd}[0..1]
memberDuring{org}
1
Interval{time}
hasBeginning{time}
1
hasEnd{time}
0..1
Instant{time}
inXSDDate{time}:date{xsd}

Fig. 8. Classes, object properties and data properties for representing officers.

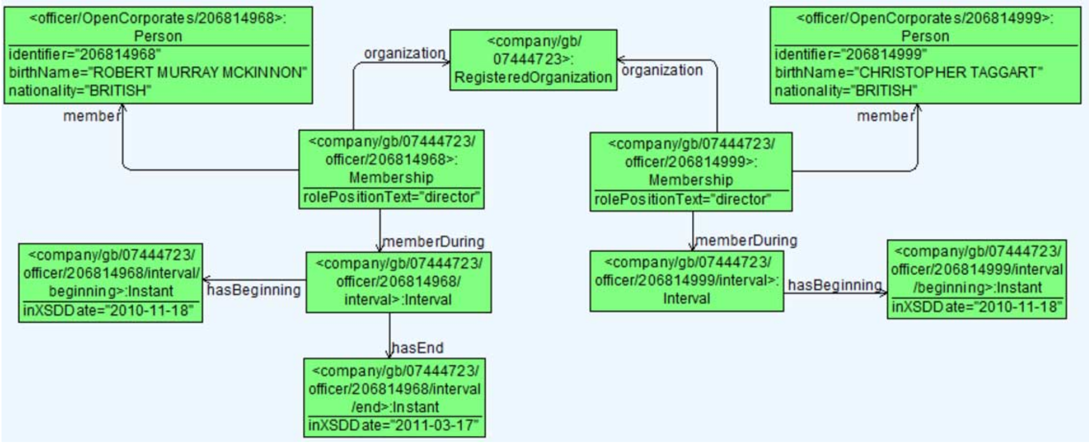
<officer/OpenCorporates/206814968>: Person
identifier="206814968"
birthName="ROBERT MURRAY MCKINNON"
nationality="BRITISH"
member
organization
<company/gb/07444723>: RegisteredOrganization
organization
<officer/OpenCorporates/206814999>: Person
identifier="206814999"
birthName="CHRISTOPHER TAGGART"
nationality="BRITISH"
member
<company/gb/07444723/officer/206814968>: Membership
rolePositionText="director"
<company/gb/07444723/officer/206814999>: Membership
rolePositionText="director"
memberDuring
memberDuring
<company/gb/07444723/officer/206814968/interval/beginning>:Instant
inXSDDate="2010-11-18"
hasBeginning
<company/gb/07444723/officer/206814968/interval>:Interval
<company/gb/07444723/officer/206814999/interval>: Interval
hasBeginning
<company/gb/07444723/officer/206814999/interval/beginning>:Instant
inXSDDate="2010-11-18"
hasEnd
<company/gb/07444723/officer/206814968/interval/end>:Instant
inXSDDate="2011-03-17"

Fig. 9. Example of officer representation for the company OpenCorporates.

we also introduced the data property rolePositionText{ebg} in the **Membership** class in order to capture the role as free text.

The membership interval is defined by the memberDuring{org} object property that points to an Interval. The interval has a beginning and an end date. For open intervals only the beginning is mandatory. These dates are defined by the class Instant which has the data property inXSDDate{time}.

4.3.1. *Example*

An example of officer roles using the free text data property rolePositionText{ebg} for the company OpenCorporates is shown in Fig. 9.

4.4. *Dataset*

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

- Publisher, source, last modified, license, home page, download distribution, etc.
- Subsets of data by kind of entity (e.g., companies vs. addresses), field coverage (which fields are included in which subsets), and entity characteristics (e.g., Italian companies, startups, startups in Italy).
- Count of entities in a dataset or subset.

After an analysis of various dataset description ontologies, we decided on using VOID with some extensions (see Fig. 10). VOID describes RDF datasets in terms of entities, classes, properties, triples, partitions (e.g., triples having a particular class), etc. A **Dataset** has a subset{void} relation that is used to describe a dataset polyhierarchy. For each data provider we can capture their full dataset and the respective subsets. For each **Dataset** the publisher{dct}, type{dct} and license{dct} have to be captured.

4.4.1. *Example*

Figure 11 shows an example of the datasets provided by SpazioDati. The dataset <dataset/SDATI> consists of two subsets, namely <dataset/SDATI/IT> and <dataset/SDATI/GB>. For each subset we specify the number of entities and the properties that are available.

4.5. *Validation rules*

In order to ensure that data can be correctly published according to the ontology, we devised a set of data validation rules that are associated with the ontology. The types of validations rules considered are as follows:

- **Data completeness:** Specifies that a given set of business attributes must be present (e.g., attribute legalName must be available).
- **Accuracy** Describes that data values must be correct (e.g., values of attribute jurisdiction must be included in the list of recognized nations available on Wikipedia[58]).

[58] https://en.wikipedia.org/wiki/List_of_sovereign_states

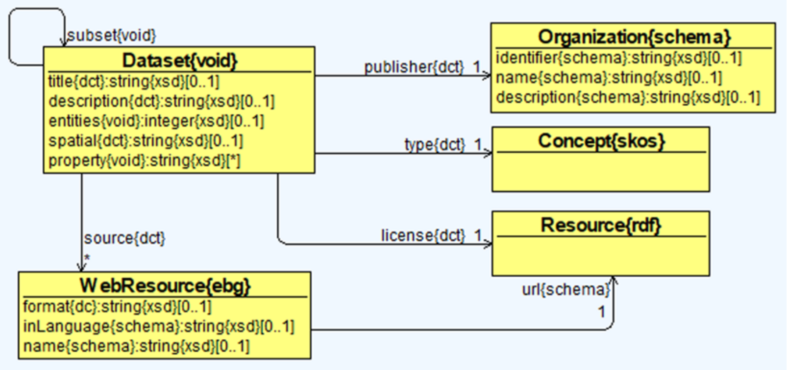
subset{void}
Dataset{void}
title{dct}:string{xsd}[0..1]
description{dct}:string{xsd}[0..1]
entities{void}:integer{xsd}[0..1]
spatial{dct}:string{xsd}[0..1]
property{void}:string{xsd}[*]
publisher{dct} 1
Organization{schema}
identifier{schema}:string{xsd}[0..1]
name{schema}:string{xsd}[0..1]
description{schema}:string{xsd}[0..1]
type{dct} 1
Concept{skos}
license{dct} 1
Resource{rdf}
source{dct}
*
WebResource{ebg}
format{dc}:string{xsd}[0..1]
inLanguage{schema}:string{xsd}[0..1]
name{schema}:string{xsd}[0..1]
url{schema}
1

Fig. 10. Classes, object properties and data properties for representing datasets.

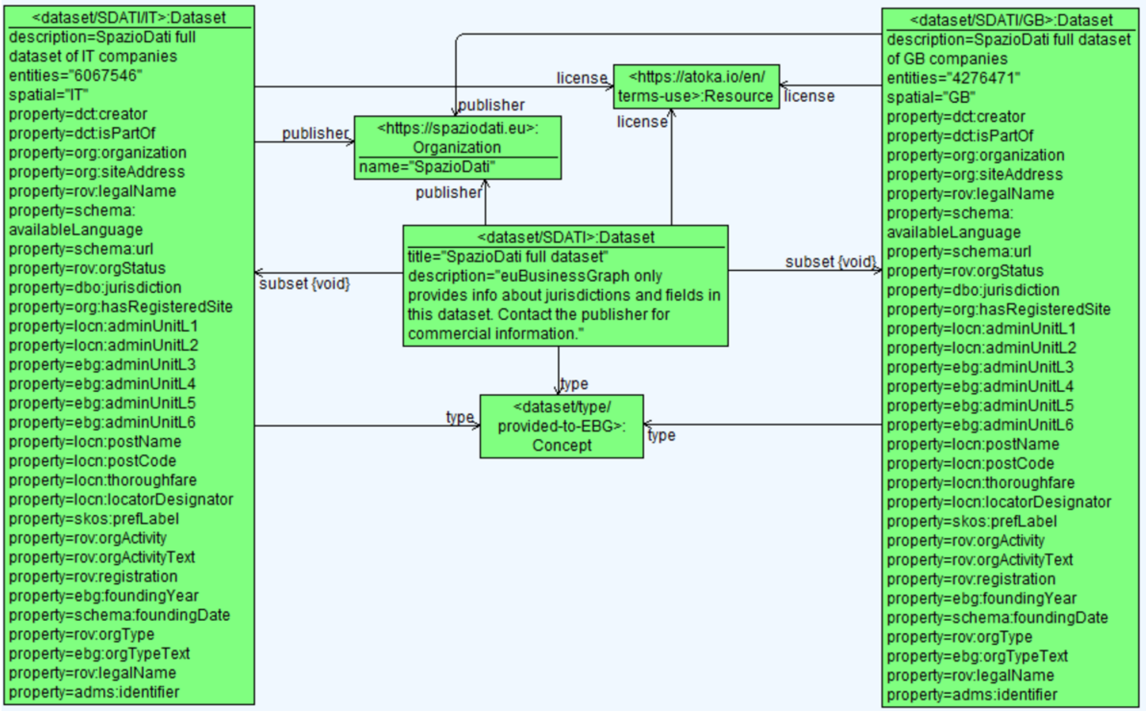
<dataset/SDATI/IT>:Dataset
description=SpazioDati full dataset of IT companies
entities="6067546"
spatial="IT"
property=dct:creator
property=dct:isPartOf
property=org:organization
property=org:siteAddress
property=rov:legalName
property=schema:availableLanguage
property=schema:url
property=rov:orgStatus
property=dbo:jurisdiction
property=org:hasRegisteredSite
property=locn:adminUnitL1
property=locn:adminUnitL2
property=ebg:adminUnitL3
property=ebg:adminUnitL4
property=ebg:adminUnitL5
property=ebg:adminUnitL6
property=locn:postName
property=locn:postCode
property=locn:thoroughfare
property=locn:locatorDesignator
property=skos:prefLabel
property=rov:orgActivity
property=rov:orgActivityText
property=rov:registration
property=ebg:foundingYear
property=schema:foundingDate
property=rov:orgType
property=ebg:orgTypeText
property=rov:legalName
property=adms:identifier
license
<https://atoka.io/en/terms-use>:Resource
license
license
publisher
publisher
<https://spaziodati.eu>:Organization
name="SpazioDati"
publisher
<dataset/SDATI>:Dataset
title="SpazioDati full dataset"
description="euBusinessGraph only provides info about jurisdictions and fields in this dataset. Contact the publisher for commercial information."
subset {void}
subset {void}
type
type
type
<dataset/type/provided-to-EBG>:Concept
<dataset/SDATI/GB>:Dataset
description=SpazioDati full dataset of GB companies
entities="4276471"
spatial="GB"
property=dct:creator
property=dct:isPartOf
property=org:organization
property=org:siteAddress
property=rov:legalName
property=schema:availableLanguage
property=schema:url
property=rov:orgStatus
property=dbo:jurisdiction
property=org:hasRegisteredSite
property=locn:adminUnitL1
property=locn:adminUnitL2
property=ebg:adminUnitL3
property=ebg:adminUnitL4
property=ebg:adminUnitL5
property=ebg:adminUnitL6
property=locn:postName
property=locn:postCode
property=locn:thoroughfare
property=locn:locatorDesignator
property=skos:prefLabel
property=rov:orgActivity
property=rov:orgActivityText
property=rov:registration
property=ebg:foundingYear
property=schema:foundingDate
property=rov:orgType
property=ebg:orgTypeText
property=rov:legalName
property=adms:identifier

Fig. 11. Example of datasets provided by SpazioDati.

- **Precision:** Specifies that all data values for a business attribute must be as precise as required by the attribute's business requirements, intended meaning, intended usage, and precision in the real world.
- **Consistency:** Specifies that certain business attributes must follow a given pattern (e.g., `age` and `dateOfBirth` attributes are connected by the following rule `age` = INT(YEARFRAC(`dateOfBirth`,TODAY())).
- **Temporal dimension:** Refers to the temporal dimension of data, such as volatility (the average time between update of data), timeliness (the average age of values), or currency (when data is entered in the system). An example of such a rule would be "the last modification date of attribute `companyRevenue` must be more recent than a year ago".

There are several possible ways to describe data validation rules, ranging from an algorithmic style such

```
ebgsh:Company a sh:NodeShape;
  sh:targetClass rov:RegisteredOrganization;
  sh:closed true;
  sh:nodeKind sh:IRI;
  sh:pattern "^http://data.businessgraph.io/company/[A-Z]{2}/.+/";
  sh:property [sh:path rov:legalName;
    sh:or ([sh:datatype xsd:string] [sh:datatype rdf:langString]);
    sh:not ([sh:pattern "^_|_\$|_{2}"]); sh:minCount 1];
  sh:property [sh:path rov:orgActivity;
    sh:nodeKind sh:IRI;
    sh:pattern "^http://data.businessgraph.io/nace/.+"];
```

Fig. 12. Example of SHACL shape used to validate EBG RDF company data.

as: `legalName EXISTS AND len(trim (legalName)) <> 0` to a semantic based definition by using the SHACL [22] (Shapes Constraint Language) notation. SHACL is a language for validating RDF data graphs against a set of conditions that are provided as shapes and other constructs expressed in the form of an RDF graph (i.e., a shapes graph). ShEx [28] (Shape Expression) is a similar high-level language that can be used to validate RDF graph data. Both SHACL and ShEx use RDF syntax, and share the mechanisms of shape constraints, node constraints, property constraints, cardinalities, and logical operators. Examples of SHACL and ShEx shapes for the euBusinessGraph ontology are available in the Github repository.[59] Figure 12 shows an example of how SHACL validation shapes can be defined for a company URI node and two corresponding attributes (i.e., `legalName` and `orgActivity`). The `legalName` pattern requires the legal name to be canonicalized, i.e., not have leading, trailing or consecutive spaces (denoted as underscores below).

The ontology itself is not limited to European companies, but to ensure maximum compatibility across European data, our data provider rules and RDF Shapes choose specific EU-relevant thesauri. We use NACE for industrial classification (`rov:orgActivity`), NUTS+LAU for geographic regions (properties of `locn:Address`), specific SKOS schemes for Legal Type and Status. These can be easily adapted to other nomenclatures or knowledge bases: there is a multitude of industrial classifications (NAICS and the older SIC for North America; TRBC and GICS for stock indexes, etc.), and GeoNames is more appropriate for geographic regions in a global setting. For example, the ONTO CG ontology that builds upon euBusinessGraph uses more general nomenclatures.

5. Examples of use of the euBusinessGraph ontology

We present examples of how the euBusinessGraph ontology was used. We will first describe the approach on how the ontology was used to harmonize and make available company data from various data providers, resulting in the development of a business knowledge graph (Section 5.1 and Section 5.2). We will then show how this knowledge graph was used in the euBusinessGraph marketplace for basic company data – a place where data consumers can search, analyse, and compare data from various providers (Section 5.3). Finally, we provide an example on how the ontology was used in the area of public procurement (Section 5.4), and how it was extended in the domain of financial transactions (Section 5.5).

5.1. Overview of data mapping approach

In order to develop the euBusinessGraph knowledge graph harmonizing data from various data providers, we devised a data mapping approach that was used to convert company data from CSV and JSON sources into RDF conforming to the ontology. In the following, we describe the mapping notation and provide specific examples showing how the mapping rules were used. Actual mappings for data are publicly available via the DataGraft platform[60] [30,31].

The first step of the mapping process is to select attributes (e.g., `base.legalName`) from the original data source (e.g., JSON file from data provider), and

[59] https://github.com/euBusinessGraph/eubg-data/tree/master/model

[60] https://datagraft.io

Table 2

Mapping parameters defined for each JSON data attribute

Mapping parameter	Data provider's JSON data attribute
id	id
legalName	base.legalName
jurisdiction	base.country
ORGTYPE	base.legalForms[*].name
ORGACTIVITY	base.ateco[*].code
COUNTRY	base.registeredAddress.state
MACROREGION	base.registeredAddress.macroregion
REGION	base.registeredAddress.region
PROVINCE	base.registeredAddress.province
MUNICIPALITY	base.registeredAddress.municipality
lat	base.registeredAddress.lat
lon	base.registeredAddress.lon
LATLONPREC	base.registeredAddress.latlonPrecision

Table 3

Helper functions used to create base URIs

Helper function	Definition	Comments
ebg-comp	http://data.businessgraph.io/company	Base company URI
curi	**ebg-comp**/*jurisdiction*/*id*	Company URI
ciduri	curi/id	Company identifier URI
cadruri	curi/address	Company address URI
guri	cadruri/geo	Geographic coordinate URI

construct parameter names (e.g., legalName) so that we can reference the attribute values in the definition of the mapping functions, as exemplified in Table 2. When defining the mappings, we assume that the input data is a set of attribute-value pairs. Mapping parameters in Table 2 that are specified as lower-case italic letters refer to a string or number value (e.g., *legalName* refers to "SpazioDati S.R.L" in the data provider's raw data source files), while parameters denoted in upper-case letters refer to SKOS concept schemes that were defined as part of the RDF generation process. As an example of the use of concept schemes, the mapping parameter ORGACTIVITY will refer to a URI that uses a classification vocabulary to represent the data attribute (e.g., the URI <nace:62.01> uses a controlled vocabulary[61] to describe NACE economic activities for a company).

Next, Table 3 defines a set of helper functions for a subset of base URIs that will be used to map JSON data to RDF. The helper functions improve readability of mapping rules by reducing the text needed to refer to a specific URI. As an example, the helper function **curi** refers to the actual URI http://data.businessgraph.io/company/IT/361163703. To produce this URI, mapping parameters listed in italic (e.g., *jurisdiction* and *id*) will be replaced by the actual values (e.g., "IT" and "361163703") from the source JSON data. Furthermore, the mapping definitions may contain input parameters denoted in bold that refer to another function that was defined as part of the mapping process (e.g., **ebg-comp** points to the URI http://data.businessgraph.io/company). After the set of helper functions were defined, mapping rules were constructed for each of the data provider JSON attributes listed in Table 2. The resulting mapping rules are described in Table 4.

5.2. *Infrastructure for the knowledge graph generation*

A data provisioning infrastructure was developed to onboard data from various data providers. Using this approach, data source files from data providers were processed and mapped to the euBusinessGraph ontology using the mapping process discussed in the pre-

[61] https://github.com/euBusinessGraph/eubg-data/blob/master/data/NACE/nace.ttl

Table 4
Mapping functions for a subset of company data attributes

Scope of mapping function	Definition	Comments
Company URI node	<**curi**> rdf:type rov:RegisteredOrganization .	Company class
	<**curi**> rov:registration <**ciduri**> .	Company identifier triple
	<**curi**> org:hasRegisteredSite <**cadruri**> .	Company address triple
	<**curi**> schema:geo <**guri**> .	Company geo-coordinate triple
	<**curi**> rov:legalName "*legalName*" .	Legal name
	<**curi**> dbo:jurisdiction "*jurisdiction*" .	Jurisdiction
	<**curi**> rov:orgType ORGTYPE .	Organization type
	<**curi**> rov:orgActivity ORGACTIVITY .	Economic activity
Identifier URI node	<**ciduri**> rdf:type adms:Identifier .	Identifier class
	<**ciduri**> skos:notation "*id*" .	Identifier value
Address URI node	<**cadruri**> rdf:type locn:Address .	Address class
	<**cadruri**> rdf:type org:Site .	Adress type
	<**cadruri**> org:siteAddress <**cadruri**> .	Self reference
	<**cadruri**> locn:adminUnitL1 COUNTRY .	Country
	<**cadruri**> locn:adminUnitL2 MACROREGION .	Macro region
	<**cadruri**> ebg:adminUnitL3 REGION .	Region
	<**cadruri**> ebg:adminUnitL4 PROVINCE .	Province
	<**cadruri**> ebg:adminUnitL5 MUNICIPALITY .	Municipality
Geo-coordinate URI node	<**guri**> rdf:type schema:GeoCoordinates .	Geolocation class
	<**guri**> schema:latitude *lat* .	Latitude
	<**guri**> schema:longitude *lon* .	Longitude
	<**guri**> ebg:geoResolution LATLONPREC .	Geo-oordinate resolution

vious section. After transforming each dataset from a tabular format (i.e., CSV or JSON) to RDF, the resulting data was published to one named graph for each data provider jurisdiction in an enterprise semantic graph database, GraphDB,[62] hosted by Ontotext.

GraphDB is a service component on the Ontotext Platform.[63] The platform also includes important features such as data mutations, user management (Fusion Auth), access control, deployment and monitoring. In addition to GraphDB, the data provisioning infrastructure includes a set of data ingestion services and data preparation tools that can be used to simplify data cleaning and transformation from the various sources. The services include data interlinking tools for data transformation, enrichment, interlinking, and metadata generation processes in order to publish the business graph data as Linked Data.

The core process of knowledge graph creation is executed by using the cloud-based data management platform DataGraft. Grafterizer[64] [40] is a framework (part of DataGraft) for interactive data cleaning and transformation, and RDF knowledge graph generation that is used together with the tabular annotation tool ASIA[65] [9] and ABSTAT[66] [27] to map company data to the euBusinessGraph ontology. Finally, the RDF triples are published as a knowledge graph in GraphDB. Grafterizer, ASIA and ABSTAT were used to clean, transform, enrich and convert tabular data to RDF as part of the business knowledge graph construction.

The next section describes how the published knowledge graph was used to populate a marketplace for company data.

[62] http://graphdb.ontotext.com
[63] http://platform.ontotext.com
[64] https://www.eubusinessgraph.eu/grafterizer-2-0
[65] https://www.eubusinessgraph.eu/asia-2
[66] https://www.eubusinessgraph.eu/abstat

5.3. The euBusinessGraph marketplace

A main motivation behind the development of a data marketplace for basic company data is the democratisation of the company information market, currently dominated by a few large international players (e.g., Bisnode[67]) that create a market barrier for smaller company data providers like OpenCorporates and SpazioDati. The intention of the marketplace is to enable such smaller players to join a common ecosystem to promote their data offerings, and for data consumers to have a central point where they could easily compare company data offerings. A public prototype of the data marketplace application[68] developed to showcase the use of the euBusinessGraph ontology is available online.[69]

The ontology was used in the marketplace to realize functionality for a) full-text advanced search and detailed faceted search for exploration of the company knowledge graph, b) analytics services such as data aggregation and visualization (e.g., company activities per city), and c) search for company news articles, and search for company events.

5.4. Use of the euBusinessGraph ontology in the public procurement domain

Public procurement accounts for a substantial part of the public investment and global economy and therefore there is a need for better insight into, and management of government spending. In this respect, national, regional, local, and EU-wide public procurement portals were established to publish procurement notices regarding the purchase of work, goods or services from companies by public authorities in order to increase transparency, economic activity, and competitiveness [33]. However, the technical landscape is quite scattered and there are no common data formats and models used for exposing such data uniformly allowing advanced analytics and analysis, such as for fraud and trend detection. To this end, the euBusinessGraph ontology was used in the procurement domain, in the context of the project TheyBuyForYou (TBFY),[70] for integrating public procurement and company data into the TBFY knowledge graph [34,36]. The resulting knowledge graph allows browsing, visualising, and analysing public EU-wide procurement data and enables a variety of business cases built on top of it by various stakeholders, such as buyers, suppliers, and policy makers.

The data integrated includes procurement data provided by OpenOpps,[71] and company data provided by OpenCorporates. OpenOpps has gathered over 2M tender documents from more than 300 publishers through Web scraping and by using open APIs and provides the resulting data in Open Contracting Data Standard (OCDS),[72] while OpenCorporates uses its own adhoc schema. These two datasets are integrated through an ontology network. An ontology for procurement data was developed based on the OCDS standard [37] and the euBusinessGraph ontology was used for representing the company data. The two datasets are integrated through a reconciliation process [38]. Suppliers appearing in tender data are matched against company data provided by OpenCorporates. The matched company data is extracted and ingested to the TBFY knowledge graph [35]. The current release of the TBFY knowledge graph includes 129M triples as of October 2020 and more data will be ingested.

5.5. Use of the euBusinessGraph ontology for financial transactions

Company-related economic information is crucial to many business operations. It empowers customer relationship management, acquisition of new clients, marketing campaigns, supply chain management, market analysis, competitive intelligence, mergers and acquisitions, etc. In this respect, the euBusinessGraph ontology was used for matching and linking company-related economic information within the context of Ontotext's Intelligent Matching and Linking of Company Data (CIMA) project.[73] CIMA aims to use AI/ML technologies for linking and harmonizing company-related business data from various sources. The project applies machine learning, semantic modeling and integration, entity matching, automatic classification, logical inference to make data richer, better harmonized, integrated, interlinked and easier to use. As part of the project, Ontotext is creating a Company Knowledge Graph (ONTO-CG) for demo purposes by integrating data from open and a few proprietary datasets. The emphasis of the project is on finan-

[67] http://www.bisnode.com

[68] https://www.eubusinessgraph.eu/the-marketplace

[69] http://marketplace.businessgraph.io

[70] http://theybuyforyou.eu

[71] https://openopps.com

[72] https://standard.open-contracting.org/latest/en

[73] https://www.ontotext.com/cima

cial data, industrial classification, company size and important observations such as annual sales and number of employees.

6. Conclusion and outlook

The analysis of existing initiatives in the area of interoperability of company-related data revealed the fact that harmonization of company data was far from a solved problem. Company data by different providers is very heterogeneous. We argued for the importance of harmonised basic company data as a key enabler for different value chains in various sectors that depend on company information.

In this article, we described the euBusinessGraph ontology for harmonizing basic company data as a lightweight mechanism for aggregating, linking, provisioning and analysing basic company data. It reuses numerous ontologies and adds extra properties and classes (e.g., IdentifierSystem) to describe the full scope of basic company information. The main challenge this paper addressed is related to finding the right balance between a semantic model for basic company information that is too complex and hard to understand (for an example of such model see FIBO) and a simplistic least common denominator model, while at the same time exploiting proper mechanisms to reuse numerous related ontologies.

The euBusinessGraph ontology was developed following standard practices in ontology development, identifying the scope and competency questions with different stakeholders, identifying and reusing existing ontologies, and publishing the ontology according to existing best practices for Linked Data vocabulary publishing. We provided an overview of the ontology scope, the ontology development process, explanations of core concepts and relationships, and the implementation of the ontology. Furthermore, we provided examples where the ontology was used, among others, for publishing company data and for comparing company data from various data providers.

An important aspect of our approach is the use of `schema:domainIncludes` and `schema:rangeIncludes` (following Schema.org and Web of Things ontologies), which represent low ontological commitment and improve ontology reuse. Furthermore, we provide RDF Shapes (SHACL and SHEX) for validation of RDF data. The novelty of the proposed lightweight semantic model lies in the careful combination and reuse of terms from existing ontologies, and its expression as a detailed model, including application profile, RDF Shapes and data provider mapping documentation.

The euBusinessGraph ontology serves now as an asset not only for enabling various tasks related to basic company data but also on top of which more specific extensions can be built upon. As an example of such an extension, initial efforts have been made to capture events that happen during the lifetime of a company [21] and for representing the French register data in RDF [12,21]. In addition to possible extensions of the ontology, other interesting directions for future work can be envisioned. For example, interlinking harmonized data from various data providers is an interesting topic for future work (preliminary work on interlinking company data harmonised using the euBusinessGraph ontology is reported in [23]). Extending the ontology with classification datasets for additional jurisdictions (e.g., Germany) will further increase the relevance of the business graph, and enable more precise queries to be executed on the harmonized data. This harmonization process includes describing supplementary identifier systems for company entities and officers for new data providers, as well as creating additional classification schemes for NACE, NUTS, LAU, organization types and organization status.

In the context of the TheyBuyForYou project, the ontology will be used as a core component of the proposed procurement knowledge graph and the ontology network. Currently, on the one hand, more data is being reconciled and ingested into the TBFY knowledge graph and on the other hand more research and development work is being undertaken in order to improve the reconciliation process matching supplier data against company data. Essentially, it will demonstrate how one can integrate disparate but relevant data sources, pose interesting queries that were otherwise not possible to answer, and create new business scenarios. In the CIMA (ONTO-CG) project, the euBusinessGraph semantic model is extended to cover financial transactions, and prototypes and exploitable systems are built using the Ontotext Platform to query RDF data integrated from numerous sources.

Acknowledgements

The work in this article was partly funded by the EC H2020 projects euBusinessGraph (grant 732003), EW-Shopp (grant 732590), TheyBuyForYou (780247), CIMA (grant BG16RFOP002-1.005-0168-C01), and further reused in InnoRate (821518). Special thanks to the members of the euBusinessGraph project consortium for stimulating discussions around various aspects of basic company information, especially to Tatiana Tarasova, Fredrik Seehusen, and David Norheim for their initial involvement in the development of the ontology.

References

[1] K. Alexander, R. Cyganiak, M. Hausenblas and J. Zhao, Describing linked datasets with the VoID vocabulary, World Wide Web Consortium (W3C), 2011. https://www.w3.org/TR/void/.

[2] V. Alexiev, A. Kiryakov and P. Tarkalanov, euBusinessGraph: Company and economic data for innovative products and services, in: *Proceedings of the 13th International Conference on Semantic Systems (Semantics 2017)*, 2017. http://rawgit2.com/webdata/SEMANTiCS2017-posters/master/papers_final/163_Alexiev/index.html.

[3] V. Alexiev, T. Tarasova, J. Paniagua, C. Taggart, B. Elvesaeter, F. Seehusen, D. Roman and D. Norheim, euBusinessGraph Semantic Data Model, euBusinessGraph Consortium, 2018. https://docs.google.com/document/d/1dhMOTlIOC6dOK_jksJRX0CB-GIRoiYY6fWtCnZArUhU/edit.

[4] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber and E. Summers, Key choices in the design of simple knowledge organization system (SKOS), *Journal of Web Semantics* **20** (2013), 35–49. doi:10.1016/j.websem.2013.05.001.

[5] S.K. Bansal and S. Kagemann, Integrating big data: A semantic extract-transform-load framework, *IEEE Computer* **48**(3) (2015), 42–50. doi:10.1109/MC.2015.76.

[6] J. Barzdins, K. Cerans, R. Liepins and A. Sprogis, Advanced ontology visualization with OWLGrEd, in: *Proceedings of the 8th International Workshop on OWL: Experiences and Directions (OWLED 2011)*, CEUR Workshop Proceedings, Vol. 796, CEUR-WS.org, 2011. http://ceur-ws.org/Vol-796/owled2011_submission_7.pdf.

[7] M. Bennett, The financial industry business ontology: Best practice for big data, *Journal of Banking Regulation* **14**(3) (2013), 255–268. doi:10.1057/jbr.2013.13.

[8] O. Corcho, M. Fernández-López and A. Gómez-Pérez, Ontological engineering: Principles, methods, tools and languages, in: *Ontologies for Software Engineering and Software Technology*, C. Calero, F. Ruiz and M. Piattini, eds, Springer, Berlin, 2006, pp. 1–48. doi:10.1007/3-540-34518-3_1.

[9] V. Cutrona, M. Ciavotta, F.D. Paoli and M. Palmonari, ASIA: A tool for assisted semantic interpretation and annotation of tabular data, in: *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) Co-Located with 18th International Semantic Web Conference (ISWC 2019)*, CEUR Workshop Proceedings, Vol. 2456, CEUR-WS.org, 2019, pp. 209–212. http://ceur-ws.org/Vol-2456/paper54.pdf.

[10] M. Dekkers, Asset Description Metadata Schema (ADMS), World Wide Web Consortium (W3C), 2013. https://www.w3.org/TR/vocab-adms/.

[11] Department of Economic and Social Affairs, International Standard Industrial Classification of All Economic Activities (ISIC), United Nations, 2008. https://unstats.un.org/unsd/classifications/Econ/isic.

[12] T. Ehrhart and R. Troncy, EURECOM at SemStats 2019, in: *Proceedings of Semantic Statistics (SemStats 2019)*, 2019.

[13] EU ISA Programme Core Vocabularies Working Group, ISA Programme Location Core Vocabulary, World Wide Web Consortium (W3C), 2015. https://www.w3.org/ns/locn.

[14] Eurostat, Statistical classification of economic activities in the European Community (NACE), European Commission, 2008. https://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-07-015.

[15] Eurostat, *Methodological manual on territorial typologies*, European Commission, 2019. https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-18-008. doi:10.2785/930137.

[16] M. Giese, A. Soylu, G. Vega-Gorgojo, A. Waaler, P. Haase, E. Jiménez-Ruiz, D. Lanti, M. Rezk, G. Xiao, Ö.L. Özçep and R. Rosati, Optique: Zooming in on big data, *IEEE Computer* **48**(3) (2015), 60–67. doi:10.1109/MC.2015.82.

[17] R.V. Guha, D. Brickley and S. Macbeth, Schema.org: Evolution of structured data on the web, *Communications of the ACM* **59**(2) (2016), 44–51. doi:10.1145/2844544.

[18] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool, 2011.

[19] ISO/TC 68/SC 8 Technical Committee, Financial services – Legal entity identifier (LEI), International Organization for Standardization (ISO), 2019. https://www.iso.org/standard/75998.html.

[20] M. Janssen, D. Konopnicki, J.L. Snowdon and A. Ojo, Driving public sector innovation using big and open linked data (BOLD), *Information Systems Frontiers* **19**(2) (2017), 189–195. doi:10.1007/s10796-017-9746-2.

[21] S.A.E. Kader, N. Nikolov, B.M. von Zernichow, V. Cutrona, B.E.M. Palmonari, A. Soylu and D. Roman, Modeling and publishing French business register (Sirene) data as linked data using the euBusinessGraph ontology, in: *Joint Proceedings of the International Workshops on Sensors and Actuators on the Web, and Semantic Statistics Co-Located with 18th International Semantic Web Conference (ISWC 2019)*, CEUR Workshop Proceedings, Vol. 2549, CEUR-WS.org, 2019. http://ceur-ws.org/Vol-2549/article-06.pdf.

[22] H. Knublauch and D. Kontokostas (eds), Shapes constraint language (SHACL), World Wide Web Consortium (W3C), 2017. https://www.w3.org/TR/shacl/.

[23] A. Maurino, A. Rula, B.M. von Zernichow, M.S. Gomez, B. Elvesæter and D. Roman, Modelling and linking company data in the euBusinessGraph platform, in: *Proceedings of the 5th Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets (DSMM 2019)*, ACM, 2019, pp. 1–6. doi:10.1145/3336499.3338012.

[24] M. McDaniel and V.C. Storey, Evaluating domain ontologies: Clarification, classification, and challenges, *ACM Computing Survey* **52**(4) (2019), 70:1–70:44. doi:10.1145/3329124.

[25] J.F. Muñoz-Soro, G. Esteban, O. Corcho and F. Seron, PPROC, an ontology for transparency in public procurement, *Semantic Web* **7**(3) (2016), 295–309. doi:10.3233/SW-150195.

[26] N.F. Noy and D.L. McGuinness, Ontology development 101: A guide to creating your first ontology, Technical report, Stanford Medical Informatics, 2001.

[27] R.A.A. Principe, B. Spahiu, M. Palmonari, A. Rula, F.D. Paoli and A. Maurino, ABSTAT 1.0: Compute, manage and share semantic profiles of RDF knowledge graphs, in: *Proceedings of the ESWC 2018 Satellite Events*, A. Gangemi, A.L. Gentile, A.G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J.Z. Pan and M. Alam, eds, LNCS, Vol. 11155, Springer, 2018, pp. 170–175. doi:10.1007/978-3-319-98192-5_32.

[28] E. Prud'hommeaux, J.E. Labra Gayo and H. Solbrig, Shape expressions: An RDF validation and transformation language, in: *Proceedings of the 10th International Conference on Semantic Systems (SEM 2014)*, ACM, 2014, pp. 32–40. doi:10.1145/2660517.2660523.

[29] D. Reynolds (ed.), The organization ontology, World Wide Web Consortium (W3C), 2014. https://www.w3.org/TR/vocab-org/.

[30] D. Roman, M. Dimitrov, N. Nikolov, A. Putlier, D. Sukhobok, B. Elvesæter, A. Berre, X. Ye, A. Simov and Y. Petkov, Datagraft: Simplifying open data publishing, in: *Proceedings of the ESWC 2016 Satellite Events*, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer and C. Lange, eds, LNCS, Vol. 9989, Springer, 2016, pp. 101–106. doi:10.1007/978-3-319-47602-5_21.

[31] D. Roman, N. Nikolov, A. Putlier, D. Sukhobok, B. Elvesæter, A. Berre, X. Ye, M. Dimitrov, A. Simov, M. Zarev, R. Moynihan, B. Roberts, I. Berlocher, S. Kim, T. Lee, A. Smith and T. Heath, DataGraft: One-stop-shop for open data management, *Semantic Web* **9**(4) (2018), 393–411. doi:10.3233/SW-170263.

[32] Semantic Interoperability Community, e-Government Core Vocabularies, European Commission – ISA Programme, 2019. https://joinup.ec.europa.eu/solution/e-government-core-vocabularies.

[33] E. Simperl, Ó. Corcho, M. Grobelnik, D. Roman, A. Soylu, M.J.F. Ruíz, S. Gatti, C. Taggart, U.S. Klima, A.F. Uliana, I. Makgill and T.C. Lech, Towards a knowledge graph based platform for public procurement, in: *Proceedings of the 12th International Conference on Metadata and Semantic Research (MTSR 2018)*, E. Garoufallou, F. Sartori, R. Siatri and M. Zervas, eds, CCIS, Vol. 846, Springer, 2018, pp. 317–323. doi:10.1007/978-3-030-14401-2_29.

[34] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F.Y. Martinez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, T.C. Lech and D. Roman, Enhancing public procurement in the European Union through constructing and exploiting an integrated knowledge graph, in: *Proceedings of 19th International Semantic Web Conference (ISWC 2020)*, J.Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, LNCS, Vol. 12507, Springer, 2020, pp. 430–446. doi:10.1007/978-3-642-33876-2_35.

[35] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F.Y. Martinez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, T.C. Lech and D. Roman, Integrating and analysing public procurement data through a knowledge graph: A demonstration in a nutshell, in: *Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice Co-Located with 19th International Semantic Web Conference (ISWC 2020)*, CEUR Workshop Proceedings, Vol. 2721, CEUR-WS.org, 2020. http://ceur-ws.org/Vol-2721/paper492.pdf.

[36] A. Soylu, Ó. Corcho, E. Simperl, D. Roman, F.Y. Martínez, C. Taggart, I. Makgill, B. Elvesæter, B. Symonds, H. McNally, G. Konstantinidis, Y. Zhao and T.C. Lech, Towards integrating public procurement data into a semantic knowledge graph, in: *Proceedings of the Posters and Demonstrations Session of 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018)*, CEUR Workshop Proceedings, Vol. 2262, CEUR-WS.org, 2018. http://ceur-ws.org/Vol-2262/ekaw-poster-01.pdf.

[37] A. Soylu, B. Elvesæter, P. Turk, D. Roman, Ó. Corcho, E. Simperl, G. Konstantinidis and T.C. Lech, Towards an ontology for public procurement based on the open contracting data standard, in: *Proceedings of the 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society (I3E 2019)*, I.O. Pappas, P. Mikalef, Y.K. Dwivedi, L. Jaccheri, J. Krogstie and M. Mäntymäki, eds, LNCS, Vol. 11701, Springer, 2019, pp. 230–237. doi:10.1007/978-3-030-29374-1_19.

[38] A. Soylu, B. Elvesæter, P. Turk, D. Roman, Ó. Corcho, E. Simperl, I. Makgill, C. Taggart, M. Grobelnik and T.C. Lech, An overview of the TBFY knowledge graph for public procurement, in: *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas)*, CEUR Workshop Proceedings, Vol. 2456, CEUR-WS.org, 2019, pp. 53–56. http://ceur-ws.org/Vol-2456/paper14.pdf.

[39] M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A. Gangemi (eds), *Ontology Engineering in a Networked World*, Springer, Berlin, 2012.

[40] D. Sukhobok, N. Nikolov, A. Pultier, X. Ye, A.J. Berre, R. Moynihan, B. Roberts, B. Elvesæter, M. Nivethika and D. Roman, Tabular data cleaning and linked data generation with grafterizer, in: *Proceedings of the ESWC 2016 Satellite Events*, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer and C. Lange, eds, LNCS, Vol. 9989, Springer, 2016, pp. 134–139. doi:10.1007/978-3-319-47602-5_27.

[41] W.R. van Hage, V. Malaisé, R. Segers, L. Hollink and G. Schreiber, Design and use of the simple event model (SEM), *Journal of Web Semantics* **9**(2) (2011), 128–136. doi:10.1016/j.websem.2011.03.003.

[42] Working Group for Describing Public Services, Core Public Service Vocabulary Application Profile (CPSV-AP), European Commission – ISA^2 Programme, 2016, https://ec.europa.eu/isa2/solutions/core-public-service-vocabulary-application-profile-cpsv-ap_en.