Modeling and Publishing French Business Register (Sirene) Data as Linked Data Using the euBusinessGraph Ontology

Shady Abd El Kader[1,2], Nikolay Nikolov[1], Bjørn Marius von Zernichow[1],
Vincenzo Cutrona[2], Matteo Palmonari[2], Brian Elvesæter, Ahmet Soylu and
Dumitru Roman[1]

[1] SINTEF AS, Oslo, Norway
{firstname.lastname}@sintef.no
[2] University of Milan-Bicocca, Milano, Italy
{lastname}@disco.unimib.it

Abstract. Sirene is the official database of French enterprises (legal units) and establishments (local units). In response to the SemStats 2019 Call for Challenge to model and provision the Sirene data as RDF, we propose the use of the euBusinessGraph ontology as a basis for modeling and publishing Sirene data as Linked Data. In this paper we discuss the suitability of the euBusinessGraph ontology to cover key Sirene entities, present extensions to the euBusinessGraph ontology developed to accommodate the modelling needs of the Sirene key entities. Furthermore, we describe the Linked Data publication process, covering the needed Sirene data transformations and RDFization, as well as the technology stack supporting the process. The result is key entities data from Sirene published as Linked Data, together with a reproducible process for generation of Linked Data from Sirene data.

Keywords: Sirene, Linked Data, euBusinessGraph

1 Introduction

Basic company data (e.g., company name(s), incorporation date, registered addresses, ownership and related entities, etc.) are the basis of many data value chains that different sectors depend on (e.g., business information, marketing and sales sector). Basic company data is typically recorded, managed, and made available by national business registers. Unfortunately, to date, there is no commonly agreed upon and widely used standard on how to make basic company data available across countries, in a machine-readable format that could enable easier processing and integration of basic company information.

The euBusinessGraph project[1] initiated a process to harmonize basic company data from various company data providers by developing a light-weight ontology for basic company data–the euBusinessGraph ontology[2]–which was used to make available basic company data from a number of jurisdictions from a set of data providers such as OpenCorporates[3], Atoka[4], and the Norwegian Business Register[5] as Linked Data.

In France, company data made available by the French National Institute of Statistics and Economic Studies (Insee) in the form of the Sirene registry[6]. In response to the SemStats 2019 Call for Challenge to model and provision the Sirene data as RDF, in this paper we propose the use of the euBusinessGraph ontology as a basis to model and provision Sirene data as Linked Data. Whereas the euBusinessGraph is meant to cover basic company information, Sirene provides a variety of information that goes beyond basic company information for French companies. In this paper we therefore focus on modeling the key Sirene entities, the relationships between them, and their key attributes using the euBusinessGraph ontology. The main contributions of this paper include the semantic model for capturing key Sirene entities, the tool-supported process for mapping Sirene data to the semantic model, and publication of the resulting data as Linked Data.

The rest of this paper is organized as follows. In Section 2 we provide an overview of modelling key entities of the Sirene data in the euBusinessGraph ontology and the extensions needed to the euBusinessGraph ontology to accommodate the main aspects of the key entities. Section 3 provides details on the mapping of Sirene data to the ontology discussed in Section 2, and the technology stack used to realize the mappings. Section 4 details the data publication process and discusses a data enrichment scenario for the Sirene data enabled by the proposed data publication process. Finally, Section 5 summarizes the paper.

2 Extending the euBusinesssGraph ontology with events to cover key Sirene entities

The Sirene dataset focuses on the description of the legal units, their establishments and the changes they have had since their creation. In order to propose a suitable modelling of the Sirene data, the existing euBusinessGraph ontology for the description of basic company information was used as a basis, to which extensions have been made in order to capture key Sirene entities. The final model Fig. 1 was used to map the dataset derives from three ontologies: *euBusinessGraph*, *the Simple Event Model (SEM)*[7], and the *Organization Ontology*[8]. While the initial model of the *euBusi-*

[1] http://eubusinessgraph.eu

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

[3] https://opencorporates.com

[4] https://atoka.io

[5] https://www.brreg.no

[6] http://www.sirene.fr

[7] https://semanticweb.cs.vu.nl/2009/11/sem

[8] https://www.w3.org/TR/vocab-org

nessGraph ontology aimed to map and describe the present state of a company with its economic/organizational information the final model[9] with these new implementations aims to consider also the hierarchical relationships and the historicity of events that occurred to the organization both Legal Unit and Establishment.

The euBusinessGraph ontology. The central model for dataset mapping is the euBusinessGraph ontology. This ontology was chosen since it was built for harmonizing basic company data in the euBusiessGraph project and covered information found in the Sirene data. More specifically, the euBusinessGraph covered the following aspects:

- Capture the concept of a company and represent different types of companies
- Represent company jurisdictions and registration information
- Capture company contact information, such as the address and other locations
- Capture social data of companies, such as their websites (together with Web languages), RSS/Atom feeds and Wikipedia URLs (not used for the Sirene Challenge)

The euBusinessGraph ontology reuses a large number of ontologies/taxonomies for capturing basic company data, including EU Core Vocabs (W3C Org[10], RegOrg[11], Location[12], Person[13]), schema.org[14], ADMS ontology[15], Dublin Core[16], Financial Industry Business Ontology (FIBO)[17], Global Legal Entity Identifier[18], and Nomenclature des Activités Économiques (NACE)[19]. Since none of the existing ontologies covers the complete scope the euBusinessGraph ontology, we reused where possible and extended the ontology where necessary. The euBusinessGraph model represents a static view of the companies and had a lack of a relational/hierarchical component with possible branches/establishments, for this reason to map the Siren Dataset which have both historical and hierarchical data the model had to be extended even more

Event Description ontology. In order to map the Sirene dataset, it was necessary to extend the euBusinessGraph ontology with the “organizational events" in the Sirene data. Historical events data mapping is not supported by the euBusinessGraph model, which made it necessary to link euBusinessGraph ontology with a model that could describe the events.The SEM ontology, which we chose for the extension with company events, provides classes and relations that can be used to describe generic events. The SEM model can be used to define domain-specific events, such as the company events present in the Sirene dataset and is extensible to cover other types of events. Furthermore, the Simple Event Model was chosen due to its high flexibility

[9] https://github.com/eldysha/ebg-siren
[10] https://www.w3.org/TR/vocab-org
[11] https://www.w3.org/TR/vocab-regorg
[12] https://www.w3.org/ns/locn
[13] https://joinup.ec.europa.eu/release/core-person-vocabulary/100
[14] https://schema.org
[15] https://www.w3.org/TR/vocab-adms
[16] https://www.dublincore.org/specifications/dublin-core
[17] https://spec.edmcouncil.org/fibo
[18] https://www.gleif.org/en/about-lei/introducing-the-legal-entity-identifier-lei
[19] https://ec.europa.eu/eurostat/ramon/ontologies/nace.rdf

and adaptability to different kinds of events, other ontologies of this type are more structured, so less flexible for different dataset as for example to structures such as org:ChangeEvent[20] that only describes events that drastically change a company as for example the merging of two companies. The Simple Event Model ontology was therefore a good starting point but some changes had to be created to the properties of the event, in fact the siren dataset even if it has a fixed number of types of happenings, the single happening can have different value from another with the same type (for example a typology of event change of main activity NACE can have as many outputs as possible NACE) it was therefore decided to extend Simple Event Model ontology with the value of the event. To achieve this, we added a new property that relates the event and the company. This relation partly existed since in the Simple Event Model ontology the event has an "actor" (the organization), but with the new implementation also the "actor" has an event in this way it will be possible to find an event starting from the organization, a feature that did not exist before in the Simple Event Model.

Division Legal Unit and establishment. Another necessary addition for the euBusinessGraph ontology in order to fit with the Sirene Dataset was the conceptual separation between legal unit and establishments. This feature was not present in the initial model, since the euBusinessGraph model was created to describe companies as entities without elaborating on establishments. The euBusinessGraph ontology is based on a number of ontologies and vocabularies that describe company data. A suitable representation of the separation between legal units and establishments is present in the Organization Ontology. In this ontology there is a differentiation between the company and what is called the organizational unit, thus making it possible to describe the two types of entities and their relations. The relationships that were used to enrich the euBusinessGraph ontology relate Establishments and Legal Units bi-directionally - i.e., Establishment with Legal Unit through the org:hasUnit and Legal Unit to Establishment through org:isUnitOf. In order to additionally support the mapping of headquarter units, we extended the euBusinessGraph ontology with the ebg:hasHQUnit that is only used for establishment headquarters.

Connecting the ontologies and Sirene key entities. Fig. 1 shows a synthetic model of the mapping of the Siren Dataset. The double triangular brackets (<< >>) describe entity types, the labels on links between nodes describe the links between entities, the qualified templated names on top of entities describe the URI construction of that entity assuming the euBusinessGraph base URI as default, where a prefix is not defined. The templated qualified names use brackets to denote data that needs to be plugged from the mapped dataset - (jurisdiction) will map to "FR" in the case of the Sirene dataset as the data describes only French companies; the (id) is the identifier of the mapped entity (e.g., a company Siren code).

The Central entities of the model are `<<LegalUnit/Establishment>>` representing organizations and their units. For brevity, we depict both entities as one node in our diagram as their mappings share the major conceptual relationships to other entities and differ only in some attributes and attribute values as the geographical location (municipality, address, postal code etc) which is present in the establishment

[20] https://www.w3.org/TR/vocab-org/#class-changeevent

but not in the legal unit or else the Establishment can be defined as headquarter or a simple establishment this attribute in is not suitable for the Legal Unite since it does not exist such distinction. The `<<LegalUnit/Establishment>>` entity describes the economical/organizational form of the LegalUnit/Establishment as the name (legal name, trade name, preferred name), the classifications (type, start-up type, state owned type, status, economic activity), and other details (incorporation or dissolution date, its online presence as website or Wikipedia page). The `<<LegalUnit/Establishment>>` is associated with two other entities relating to the organization, the `<<Address>>` where the main attributes are the geospatial information and the `<<Identifier>>`, which describes an identifier and its relation to the identifier system used to generate it in the case of Sirene dataset - for the legal unit the SIREN code was used to generate the identifier in Fig. 1 so the identifier for the legal unit is ebg-comp:FR/SIREN where FR is ISO code for France and Siren is the SIREN code of the legal unit. For the Establishment the identifier is similar but with the SIRET code instead of the SIREN, so it is ebg-comp:FR/SIRET. `<<Event>>` describes an event that has occurred to an organization.

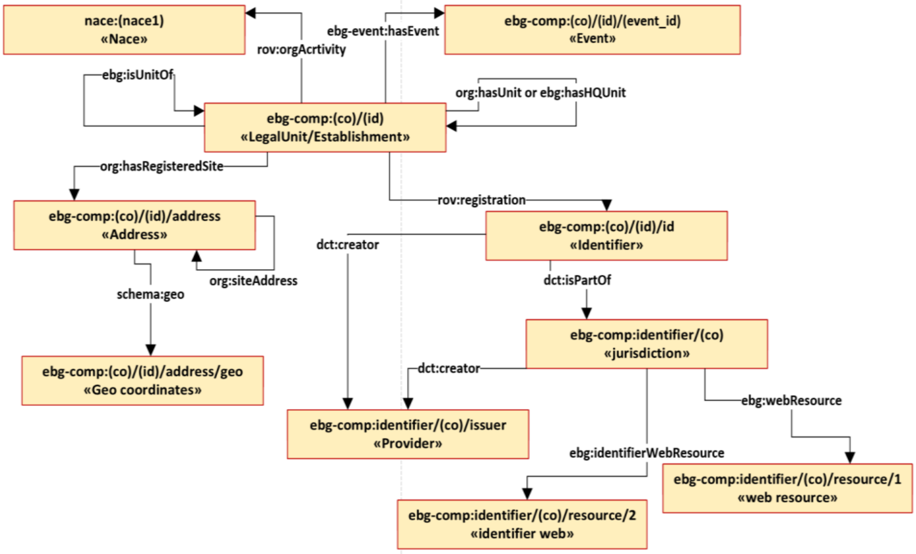


Fig. 1. High-level representation of the extended ontology

3 Data transformation

3.1 Mapping of Sirene data to the semantic model

For the mapping phase it was decided to map the five files separately, generate the RDF files, and then using the same URIs across the different mappings of the files link them properly in an RDF database. The Sirene Dataset contains five files.

(1) **StockUniteLegale**, is one of the two main files (together with **StockEtablissement**) of the Dataset, it describes punctually and anachronistically the

status of the Legal unit, depicting the state of the Legal Unit in the present. the attributes mapped in this file are mainly data that allow to recognize the company, distinguish it from the others and describe its economical/organizational status.

(2) **StockEtablissement**, like file **StockUniteLegal** this file describes the registry and the current state of the establishments, contains information as the address, if it is the head office, the legal unit to which it refers and other information that will be explored in the next section.

(3) **StockUniteLegaleHistorique**, the file describes the changes that a legal unit has undergone, from the possible closure of the activity to the change of the main activity or the change of location. This file has been put in relation to the others thanks to the identifier of the legal unit.

(4) **StockEtablissementHistorique**, like file **StockUniteLegaleHistorique** describes the changes that an establishment has undergone, this file is mainly related to the file **2.** But instead of describing the establishment in an anachronistic way it describes its changes over time.

(5) **StockEtablissementLiensSuccession** describes the passage of an establishment from one Siret code to another, this dataset has been considered as a two-sided event: firstly, the initial establishment can be considered an actor that is sold and becomes a new establishment. Secondly, the new establishment can be described as the actor of an acquisition, in fact it is considered as the purchaser of the old factory. It was therefore decided to consider file **StockEtablissementLiensSuccession** in two different perspectives. **(5a)** Firstly, whereby the subject is the new establishment, the event is the acquisition and the value of the acquisition is the old establishment. **(5b)** And secondly, whereby the subject is the old establishment, the event is the selling and the value is to whom it has been sold, the new establishment. Using this approach, it is possible to capture all the succession movements of the establishments both starting from the old and from the new establishment.

Table 1 describes which main entities of Fig. 1 have been used to map the data in the files. The newly added entities - Event, Head Quarter, organizational unit - have been useful to describe events over time and the relations between Legal Unit and establishment.

Table 1. Column headers refer to files, rows the entities, the intersection shows which entity was mapped for each file

	1.	**2.**	**3.**	**4.**	**5a.**	**5b.**
UniteLegal	Yes	Yes	Yes	Yes	No	No
Etablissement/Organizational unit	No	Yes	Yes	Yes	Yes	Yes
Event	No	No	Yes	Yes	Yes	Yes
HeadQuarter/siege	No	Yes	Yes	Yes	No	No
Identifier/jurisdiction	Yes	Yes	Yes	Yes	Yes	Yes
Address	No	Yes	No	No	No	No
NACE	Yes	Yes	No	No	No	No

3.2 Example description of the full mapping of the StockEtablissement file

To better understand how the files were mapped here is an example with the StockEtablissement file. As previously described, the StockEtablissement file describes the establishment of a company. In order to map this organization, it has been necessary to make some transformations and create different attributes not initially present in the dataset. In general, the mapping consists of three main entities and their attributes. The three major units are **Establishments**, **Establishment Identifiers** and **Establishment Addresses**.

Establishments correspond to the `<<LegalUnit/Establishment>>` entities in Fig. 1. The URI of the entity is created by concatenating the base URI of the euBusinessGraph company entities (http://data.businessgraph.io/company/) the abbreviation of the country of jurisdiction (FR) and the identification code of the establishment (i.e., the siret code). The type of the entity (RDFS[21] type) is set to the OrganizationalUnit class from the Organization Ontology. The company-related data in the table are mapped as property-object pairs as follows: data in the *enseigne1Etablissment* column represents the company name and if it exists it can be mapped to preferred name of the model through the skos:prefLabel property. The *etat- AdministratifEtablissement* column describes the state of the establishment and through transformation (mapping the names in the cells to the enumeration of possible values) it has been mapped to the value of the properties rov:orgStatus and ebg:orgStatusText. Values in the column *activiteprincipaletablissement* describe the principal activity of the establishment, which, through a mapping, are made compatible with the NACE typology are linked through the properties rov:orgActivity and ebg:orgStatusText. *dateCreationEtablissement* contains the date of founding of the establishment and has been mapped as a literal through the property schema:fundingDate. The address and identifier of the company are entities (described below) that are linked using the properties orgs:hasRegisteredSite and rov:registration. The parent legal unit of each establishment is derived from the SIREN code associated with it and linked through the orgs:unifOf property (one of the extensions to the euBusinessGraph ontology).

Establishment Identifier entities are derived from the base company URI of euBusinessGraph, jurisdiction (FR) and the Siret code: ebg-comp:(co)/(id)/id (e.g., the qualified name "ebg-comp:FR/123456789100/id"). Through this identifier it is possible to recognize the establishment even among other identifier providers. The identifier entity has a type of adms:identifier. Apart from the data in the files, we augment the mapping with additional attributes that relate to the identifier system - e.g., the indication of the jurisdiction (France) or the indication of the dataset provider (siren.fr), the identifier system unique URI in the euBusinessGraph scheme. Establishment Address provides the address of the establishment and is formed by Establishment URI ebg-comp:(co)/(id)/address (e.g., "ebg-comp:FR/ 123456789100/address").

Establishment Addresses are compound entities that describe the address of a company. In our mapping, they are typed as both locn:Address and orgs:Site and have a self-reference via orgs:siteAddress. The *LibelleCommuneEtablissment* column is mapped through locn:postName to Locality/City/Settlement in the mapping model

[21] https://www.w3.org/TR/rdf-schema

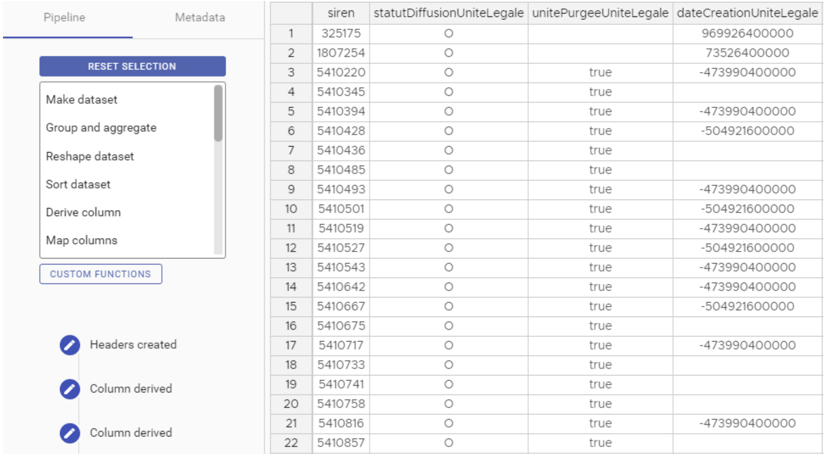
and indicates the city of the establishment. The column *codePostalEtablissement* describes the postal code and is connected via locn:postCode to the Establishment Address entity. We formulate a street address attribute created by combining *libelleVoieEtablissement, typeVoieEtablissement, typeVoieEtablissement* - e.g., instead of the three separate values “av.” “Cesar” “32” we concatenate into “Cesar avenue, 32”. The street address is associated to the Establishment Address entity through locn:thoroughfare. The full address of the company includes the city and postal code and is composed of the street address along with the values of the columns *LibelleCommuneEtablissment* (commune) and *codePostalEtablissement* (postal code). The full address is associated to the Establishment Address entity through the property locn:fullAddress. In order to successfully map the relations of the company with the establishment, we relate the URIs of the companies (based on the SIREN code) to the URIs of the Establishments (based on the SIRET code). Apart from that, depending on whether the *etablissementSiege* is true or false (i.e., if the establishment is the company headquarters or not), the establishment URI-siren refers to Establishment URI with ebg:hasHQUint or orgs:hasUnit respectively.

3.3 Realization

For the transformation of the files we used the DataGraft platform [1] and the Grafterizer 2.0 data transformation tool [2]. DataGraft allows for managing different types of assets such as files, transformations and SPARQL endpoints so that they can be shared and reused. Grafterizer 2.0 uses a batch approach for transforming tabular (CSV) data into RDF triples. Thereby, a sample dataset is uploaded to the DataGraft platform to be used to define a tabular transformation (e.g., for generating URIs, filtering, cleaning up data) and RDF mapping. The definition of the tabular transformation and RDF mapping can then be compiled into an executable JAR file, which can process data at scale for the full dataset.

We created 7 transformations for the 6 files. Each transformation processes one of the files and the file that contains data about company acquisitions (StockEtablissementLiensSuccession) was processed by two different transformations in order to map triples for buyers and sellers of enterprises from the two perspectives (one transformation creates the entities for the buyer and one for the seller). The transformations consist of tabular pre-processing of the files and a graph mapping. The tabular transformations prepare the input data for generating URIs for the entities, deriving the ontological enumerated entities (e.g., the state of a company - dissolved, inactive, liquidated, etc. have specific URIs in our solution) and filtering out redundant rows or columns. The tabular transformation UI of Grafterizer is shown in Fig. 2.

Tabular transformations in Grafterizer are represented by a pipeline of consecutive steps that are applied to the input dataset. The pipeline can include custom functions written in Clojure (e.g., for creating custom textual URIs for the entities).



	siren	statutDiffusionUniteLegale	unitePurgeeUniteLegale	dateCreationUniteLegale
1	325175	O		969926400000
2	1807254	O		73526400000
3	5410220	O	true	-473990400000
4	5410345	O	true	
5	5410394	O	true	-473990400000
6	5410428	O	true	-504921600000
7	5410436	O	true	
8	5410485	O	true	
9	5410493	O	true	-473990400000
10	5410501	O	true	-504921600000
11	5410519	O	true	-473990400000
12	5410527	O	true	-504921600000
13	5410543	O	true	-473990400000
14	5410642	O	true	-473990400000
15	5410667	O	true	-504921600000
16	5410675	O	true	
17	5410717	O	true	-473990400000
18	5410733	O	true	
19	5410741	O	true	
20	5410758	O	true	
21	5410816	O	true	-473990400000
22	5410857	O	true	

Fig. 2**.** Tabular transformation UI in Grafterizer 2.0

The graph mapping template is used to generate RDF data based on the transformed tabular data. Part of the Graph template from the transformation of legal units (StockUniteLegale) is shown in Fig**.** 3.

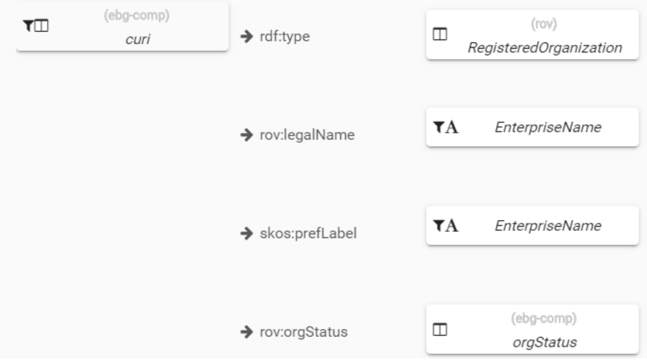


Fig. 3. Graph mapping template in Grafterizer 2.0

In Grafterizer, a node of a triple is represented by a box and properties, as the labels between two boxes. A node can be either a URI node, a literal or a blank node. Nodes can be populated either with free-defined text (or URI), or by using a cell value from a specified column. In the example in Fig. 3, the root node representing the company (ebg-comp:Establishment URI) is associated with four properties - rdf:type, rov:legalName, skos:prefLabel and rov:orgStatus. The rov:RegisteredOrganization is a URI node with a free-defined text value (the URI of the RegisteredOrganization class in the Registered Organization Vocabulary[22]). On the other hand, the EnterpriseName node is a literal node that is populated by the value of the EnterpriseName column and represents the legal name of an organisation in our mapping.

[22] https://www.w3.org/TR/vocab-regorg

4 Use cases

4.1 Data publication

The full dataset provided in the Sirene challenge amounts to approximately 16GB of CSV data. As mentioned in section 3.3, we used an approach, whereby the full dataset is transformed in batch through a transformation that is generated on a sample dataset in the Grafterizer 2.0 tool. To achieve this, we followed the data wrangling concept developed by the EW-Shopp project [3]. Thereby, the input dataset has been split into smaller chunks of approximately 200 thousand rows and transformed in parallel. The transformation itself was realized using a Docker[23] image that executes a transformation and stores the result in an output folder. The transformation job was managed by a Rancher[24] container orchestration system that allows for the distribution of Docker containers across a set of nodes. The nodes were connected through Ethernet fabric and share a distributed file system (GlusterFS[25]), which is used to share files and store intermediate results. The data workflow consisted of two steps - one that splits the data and one that transforms the data. The data transformations themselves were scaled up to 30 instances of the Docker image and distributed across a cluster of 8 nodes with a total of 56 CPU cores and 272GB of RAM. The resulting RDF dataset consists of N-triples[26] formatted files that contain fully qualified names (the reason for the large size of the output) of all nodes and amounts to approx. 450GB of data. The total number of triples in the resulting dataset is approx. 3 billion. The result of the distributed transformation has been made available at https://sirene-data.sintef.cloud. Results are organized in folders, which contain the triples that correspond to each input file.

4.2 Reconciliation and extension

Once the mapping was completed it was decided to integrate the dataset with other information to analyze whether it was possible to improve the quality of the data, for example preparing them to be used with other different data sets to do this it was decided to try to make a semantic enrichment with the help a function of ASIA. ASIA is tool for data enrichment natively integrated with Grafterizer front-end and back-end [4]. Users can enrich a given dataset using semantic table annotation functionalities from a UI, with suggestions provided by a set of schema-level alignment and instance-level reconciliation services then manually validated by the user. An enrichment task is performed by an arbitrary composition of functions of two different kinds: reconciliation and extension. Each function is implemented as a service. Reconciliation services link values that occur in the table to identifiers in external knowledge bases. ASIA currently supports different reconciliation services including

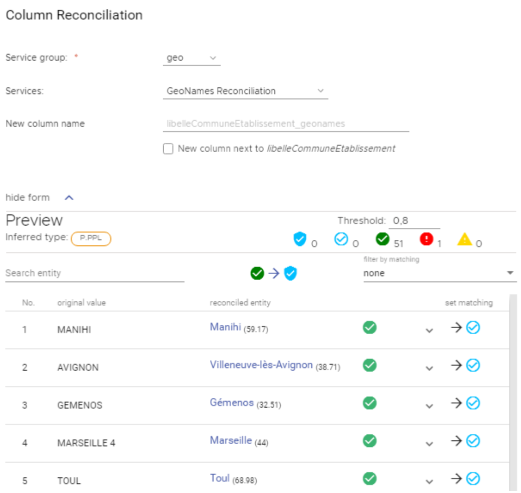
[23] https://www.docker.com
[24] https://rancher.com
[25] https://www.gluster.org
[26] https://www.w3.org/TR/n-triples

one for Geonames[27] and one for DBpedia based on the cross-lingual named entity linking service Wikifier[28]. Extension services add new columns and fill them in with values fetched from a third-party source, using identifiers (possibly obtained after a reconciliation step) to query the source. ASIA currently supports different extension services including Geonames. The input data has been enriched with ASIA services using two kinds of information available in the dataset:

- Geographical information about the companies
- Company names

For Geographical information, it was possible to enrich the data with identifiers from the Geonames knowledge base. In particular, it has been possible to reconcile the geographical information from the attribute reporting municipality names to entities of type PPL[29], e.g., by reconciling "Marseilles" with its Geonames identifier, and the geographical information from the attribute describing French populated place (PPL), then using geonames extension service it was possible to enrich the data with other entities of type ADM1, ADM2, ADM3 and ADM4, i.e., entities representing fourth-order administrative divisions. In this way it is possible to link the input data to Geonames and fetch more data from Geonames using extension services. The reconciliation results for a data sample have been validated using the ASIA reconciliation UI depicted in Fig 4. While validating the results, the user specifies a threshold for the reconciliation service, which is then applied in batch mode like the other transformations specified with Grafterizer.

An attempt to reconcile company names to DBpedia companies using Wikifier has been made, but results were not satisfactory. The dataset contains several unipersonal companies, and many of those companies are recognized by the name and surname of the owner. This yields to too many false positives (homonymous companies) and false negatives (most of these companies are not represented in DBpedia) as a user can check from the same UI. After checking the quality of the suggestions from the UI this reconciliation step has been discarded. To overcome the above limitations there are two possible approaches: a pragmatic one is to use Wikidata data extension service, querying the service using the Siren number, to fetch data for those companies represented in



Column Reconciliation

Service group: * geo

Services: GeoNames Reconciliation

New column name libelleCommuneEtablissement_geonames

□ New column next to *libelleCommuneEtablissement*

Preview

Threshold: 0,8

Inferred type: P.PPL

Search entity

filter by matching: none

No.	original value	reconciled entity		set matching
1	MANIHI	Manihi (59.17)	✓	→ ✓
2	AVIGNON	Villeneuve-lès-Avignon (38.71)	✓	→ ✓
3	GEMENOS	Gémenos (32.51)	✓	→ ✓
4	MARSEILLE 4	Marseille (44)	✓	→ ✓
5	TOUL	Toul (68.98)	✓	→ ✓

Fig 4. Screenshot of the Grafterizer+ASIA tool for data reconciliation

[27] https://www.geonames.org

[28] http://wikifier.org

[29] https://www.geonames.org/export/codes.html

Wikidata; a second approach would be to use an entity linking service for a data source specialized on company data (e.g., OpenCorporates knowledge base). However, previous work on this problem suggest us that additional information like addresses would be needed to obtain reliable results for company name disambiguation [5].

5 Summary and outlook

This paper contributes to the SemStats 2019 Call for Challenge to model and provision the Sirene data as RDF. We proposed the use of the euBusinessGraph ontology as a basis for modeling and publishing Sirene data as Linked Data. We described the modelling process as well as the Linked Data publication process. The result is key entities data from Sirene published as Linked Data, together with a reproducible process for generation of Linked Data from Sirene data.

The proposed semantic model based on the euBusinessGraph ontology was able to capture the key elements of an organization, its relationship to establishments and branches, as well as events. It however left out various attributes in that were not considered strictly necessary for an organizational/economic description, for example the *StatutDiffusionEtablissement* (an agreement to share data) or the *UnitLegalSex* (which describes the genre of the owner)[30]. As part of future work we plan to have a wider coverage of Sirene data attributes in the semantic model.

References

1. Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., & Moynihan, R. (2018). DataGraft: One-stop-shop for open data management. Semantic Web, 9(4), 393-411.
2. Sukhobok, D., Nikolov, N., Pultier, A., Ye, X., Berre, A., Moynihan, R., & Roman, D. (2016, May). Tabular data cleaning and linked data generation with Grafterizer. In European Semantic Web Conference (pp. 134-139). Springer, Cham.
3. Nikolov, N., Ciavotta, M., & De Paoli, F. (2018, September). Data wrangling at scale: the experience of EW-shopp. In Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings (p. 32). ACM.
4. Cutrona, V., Ciavotta, M., De Paoli, F., & Palmonari, M. (2019). ASIA: a Tool for Assisted Semantic Interpretation and Annotation of Tabular Data. In Proceeding of ISWC (Posters & Demos). LNCS. To appear.
5. Maurino, A., Rula, A., von Zenichow B. M., Soto Gomez, M., Elvesæter, B., & Roman, D. (2019). 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'19). ACM.

[30] A list with the Sirene attributes captured in the semantic model (and the associated mappings) can be found at https://github.com/eldysha/ebg-siren/blob/master/siren%20mapping%20attributes.xlsx.