EURECOM at the SemStats 2019 Challenge

Thibault Ehrhart[1] and Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France
{thibault.ehrhart,raphael.troncy}@eurecom.fr

Abstract. In this paper, we present two contributions for the SemStats 2019 Challenge. First, we developed the SIRENE ontology for modeling the official database of French enterprises (legal units) and establishments (local units) and we study the coverage of this dataset in Wikidata. Second, we developed a web-based application for visualizing the public database of facilities which has been previously enriched using a tourism and culture knowledge graph.

Keywords: Ontology modeling, data interlinking, knowledge graph, visualization, Wikidata

1 Introduction

This paper is a response to the SemStats 2019 Challenge, a competition based on datasets published by statistical offices. The focus, this year, was datasets provided by the French National Institute of Statistics (INSEE). The goal is to demonstrate an original and helpful usage of the data, using Semantic Web technologies.

Our first contribution has targeted the Sirene Track of the challenge. We proposed an ontology for modeling the data that re-uses a number of well-known vocabularies. Furthermore, we study the coverage of the data in Wikidata (Section 2). Our second contribution has targeted the BPE Track of the challenge. We developed an interactive web-based application for visualizing the database of facilities. We enrich this information using a knowledge graph made of data extracted from numerous location-based social networks.

2 Sirene Track

2.1 Dataset overview

Sirene is the French directory managed by INSEE which assigns a SIREN number to French enterprises, and and a SIRET number to their establishments. The Sirene track challenge consists in proposing a RDF model for this data. The Sirene dataset is divided into 5 files:

– **(1) StockUniteLegale**, one of the two main files of the dataset with StockEtablissement. It contains all active and ceased companies in their current

state in the directory. A legal unit is a legal entity governed by public or private law. This legal entity can be: a legal person whose existence is recognized by law independently of the persons or institutions that own it or who are members of it; or a natural person, who, as an independent, can carry on an economic activity.
- **(2) StockEtablissement**, the second main file, which contains all active and closed establishments in their current state in the directory.
- **(3) StockEtablissementLiensSuccession**, the list of predecessors and successors of establishments.
- **(4) StockUniteLegaleHistorique**, a set of values of certain variables historized in the Sirene directory for all the companies.
- **(5) StockEtablissementHistorique**, a set of values of certain variables historized in the Sirene directory for all the establishments.

The data is saved in CSV format and is updated on a monthly basis. Ceased businesses and closed establishments are included, providing access to Sirene data since 1973.

2.2 Re-using popular vocabularies

We need a model that can represent all the data in the Sirene database. For this, we re-used existing vocabulary, which we expanded when necessary. Our modeling work was initially based on *euBusinessGraph*, an ontology made to represent the basic informations of a company. It uses several other vocabularies, including W3C Org[1], W3C RegOrg[2], FOAF[3], schema.org[4], and ADMS[5].

W3C Org is an ontology designed to publish information about organizations and organizational structures. It is intended to provide a generic and reusable basic ontology that can be expanded or specialized for use in particular situations.

W3C RegOrg is a vocabulary used to represent registered organizations. It is an extension of the W3C Org ontology and is designed to describe organizations that have acquired legal status through a formal registration process, typically in a national or regional register. This ontology focuses only on companies, and excludes natural persons.

2.3 SKOS controlled vocabulary

We also defined a controlled vocabulary to represent the legal categories and the employee groups. The vocabulary of legal categories is organized in hierarchical form. A 3-level hierarchy corresponds to the existing one from the data provided by Sirene. The URI of the entity is based on the code of the legal category.

[1] `https://www.w3.org/TR/vocab-org/`
[2] `https://www.w3.org/TR/vocab-org/`
[3] `http://xmlns.com/foaf/spec/`
[4] `https://schema.org/`
[5] `https://www.w3.org/TR/vocab-adms/`

Listing 1.1. Samples from the legal categories vocabulary

```
<http://sirene.eurecom.fr/categorie-juridique/> a skos:ConceptScheme ;
    rdfs:label "Catégories juridiques" @fr ;
    rdfs:comment "La nomenclature des catégories juridiques retenue
        dans la gestion du répertoire Sirene, répertoire officiel d'
        immatriculation des entreprises et des établissements, a été é
        laborée sous l'égide du comité interministériel Sirene.\n\nC'
        est une nomenclature àvocation inter-administrative, utilisée
        aussi dans la gestion du Registre du Commerce et des Sociétés.
        Elle sert de référence aux Centres de Formalités des
        Entreprises (CFE) pour recueillir les déclarations des
        entreprises."@fr ;
    dct:created "2019-10-01"^^xsd:date ;
    dct:modified "2019-10-01"^^xsd:date .

<http://sirene.eurecom.fr/categorie-juridique/5> a skos:Concept ;
    skos:inScheme <http://sirene.eurecom.fr/categorie-juridique/> ;
    skos:prefLabel "Société commerciale"@fr .

<http://sirene.eurecom.fr/categorie-juridique/54> a skos:Concept ;
    skos:broader <http://sirene.eurecom.fr/categorie-juridique/5> ;
    skos:inScheme <http://sirene.eurecom.fr/categorie-juridique/> ;
    skos:prefLabel "Société àresponsabilité limitée (SARL)"@fr .

<http://sirene.eurecom.fr/categorie-juridique/5422> a skos:Concept ;
    skos:broader <http://sirene.eurecom.fr/categorie-juridique/54> ;
    skos:inScheme <http://sirene.eurecom.fr/categorie-juridique/> ;
    skos:prefLabel "SARL immobilière pour le commerce et l'industrie (
        SICOMI)"@fr .
...
```

The employee group vocabulary uses the `schema:QuantitativeValue` class and contains intervals of the number of employees, with a minimum value and a maximum value. There are 16 employee groups defined by Sirene[6].

Listing 1.2. Example of employee group

```
<http://sirene.eurecom.fr/tranche-effectif/11> a schema:
    QuantitativeValue ;
  schema:minValue "10"^^xsd:int ;
  schema:maxValue "19"^^xsd:int .
```

[6] `https://www.sirene.fr/sirene/public/variable/tefen`

2.4 Sirene ontology and URI pattern

We started by creating a mapping between the properties defined in the dataset with those available in the different ontologies.

Legal units are mapped on `rov:RegisteredOrganization` by reusing the properties defined in this vocabulary. The URI of the legal unit is composed of the base URI followed by the SIREN number of the unit (e.g. `<http://sirene.eurecom.fr/siren/19450855200016>`). The legal category uses the `rov:orgType` property and points to the category URI, as defined in our SKOS controlled vocabulary. The employee group value is mapped to `schema:numberOfEmployees` and points to the URI of the employee group as defined in our ontology.

Establishments are mapped to `rov:RegisteredOrganization` and `org:Site`. The URI of the establishment is composed of the base URI followed by the SIRET number of the establishment (e.g. `<http://sirene.eurecom.fr/siret/32517500032>`). The establishment's address is mapped to the `org:siteAddress` property which points to a URI made from the establishment's URI followed by `/address` (e.g. `<http://sirene.eurecom.fr/siret/32517500032/address>`). The link between the legal unit and the establishment is represented by the `org:hasSite` property. If `etablissementSiege` is set to `true`, then the link is also represented by the `org:hasRegisteredSite` property, which indicates that this is the primary site legally registered by the organization.

Organizational changes are mapped to `org:ChangeEvent`, where the properties `org:originalOrganization` and `org:resultingOrganization` are set to the URIs of the original and the resulting establishments. The URI of the succession link is composed of the URI database followed by a unique identifier generated from the SIRET numbers (e.g. `<http://sirene.eurecom.fr/event/32517500032-12345678901>`).

Since none of the existing ontologies covered the complete scope we needed, we reused them where possible, and we created an extension called `sirene:UniteJuridique`, in the base URI `http://sirene.eurecom.fr/ontology#`.

Listing 1.3. Definition of UniteJuridique

```
sirene:UniteJuridique a owl:Class ;
    rdfs:isDefinedBy <http://sirene.eurecom.fr/ontology#> ;
    rdfs:label "Unité Juridique"@fr ;
    rdfs:isDefinedBy sirene: .
```

This `owl:Class` is also complemented with 37 properties that are based on the name of the variables from the Sirene dataset.

Listing 1.4. List of properties from Sirene Ontology for the UniteJuridique class

```
sirene:identifiantAssociationUniteLegale
```

```
sirene:nicSiegeUniteLegale
sirene:nombrePeriodesUniteLegale
sirene:economieSocialeSolidaireUniteLegale
sirene:categorieEntreprise
sirene:caractereEmployeurUniteLegale
sirene:anneeEffectifsUniteLegale
sirene:anneeCategorieEntreprise
sirene:statutDiffusionUniteLegale
sirene:unitePurgeeUniteLegale
sirene:activitePrincipaleEtablissement
sirene:activitePrincipaleRegistreMetiersEtablissement
sirene:anneeEffectifsEtablissement
sirene:caractereEmployeurEtablissement
sirene:codeCedexEtablissement
sirene:codeCedex2Etablissement
sirene:codeCommuneEtablissement
sirene:codeCommune2Etablissement
sirene:codePaysEtrangerEtablissement
sirene:codePaysEtranger2Etablissement
sirene:denominationUsuelleEtablissement
sirene:distributionSpecialeEtablissement
sirene:distributionSpeciale2Etablissement
sirene:etablissementSiege
sirene:etatAdministratifEtablissement
sirene:indiceRepetitionEtablissement
sirene:indiceRepetition2Etablissement
sirene:nic
sirene:nombrePeriodesEtablissement
sirene:nomenclatureActivitePrincipaleEtablissement
sirene:statutDiffusionEtablissement
sirene:transfertSiege
sirene:continuiteEconomique
```

The data has been enriched with other sources by linking legal units and establishments with data from `http://entreprise.data.gouv.fr`. We have materialized this link usign the `owl:sameAs` property. The link points to `https://entreprise.data.gouv.fr/etablissement/<identifier>`, where `<identifier>` corresponds to the SIREN number for legal units, or the SIRET number for establishments.

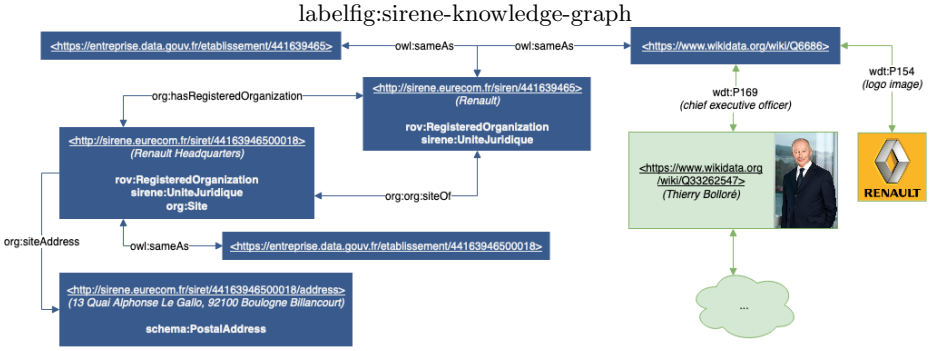
The following diagram shows an example of a materialized legal unit and the relationship with its establishment.

2.5 Studying Sirene coverage in Wikidata

We extracted the data from the Wikidata knowledge base using a SPARQL query to retrieve the entities with properties `P1616` (SIREN number) and `P3215`

Fig. 1. An example of a legal unit and its establishment modeled using the Sirene ontology for the French company Renault

labelfig:sirene-knowledge-graph



(SIRET number). We get about 41k registered organizations and 374 registered establishments in Wikidata. We then link the entities together using their registration number. In the end, we get a list of links to the Wikidata pages of 40984 companies and 374 establishments, which are materialized thanks to the `owl:sameAs` property.

Listing 1.5. Example of entity linking between a legal unit from Sirene and a Wikidata page

```
<http://sirene.eurecom.fr/siren/19450855200016>
    owl:sameAs <https://www.wikidata.org/wiki/Q13334> .
```

3 BPE Track

3.1 Dataset overview

The permanent facilities database (or BPE for "Base de données Permanente des Installations") provides information on the level of facilities and services provided by a territory to its population. It lists over 2.5 million installations of a wide range of different types with their main features, most of which are geolocated.

The datasets provided for the challenge are separated into 3 folders:

1. bpe2018-facilities: contains data for each facility, in RDF format.
2. bpe2018-codelists: the code lists used, expressed in SKOS.
3. bpe2018-geo-quality: metadata on geolocation quality. The quality level is established according to the following rules:

- good: the difference of the coordinates (X, Y) provided with the reality of the ground is less than 100m;
- acceptable: the maximum deviation of the coordinates (X, Y) provided with the reality of the ground is between 100m and 500m;
- bad: the maximum deviation of the coordinates (X, Y) provided with the reality of the field is greater than 500m and random imputations could be made.

The facilities data contains information about the creation date, category, commune number, and geolocation of each facility. The category refers to a SKOS controlled vocabulary that contains 3 levels of categories with 7 first level categories, 27 second level categories, and 187 third level categories. Geolocation uses Lambert-93 projection [7]. In order to facilitate the computation of the distances and the visualization of the results, we decided to convert the geographic coordinates to WGS84 (World Geodetic System 1984), a frequently used coordinate system format.

3.2 City Moove Knowledge Graph

City Moove is a knowledge base specialized in the domain of tourism and city exploration. It contains descriptions of events, places, transportation facilities and social activities, collected from numerous local and global data providers. The entities in the knowledge base are deduplicated, interlinked and enriched using semantic technologies [1]. The query endpoint is available at `https://kb.city-moove.fr/sparql`.

The data model used in the City Moove knowledge base is based on a set of ontologies: DOLCE+DnS Ultralite[8], schema.org[9], Dublin Core[10], LODE[11], Location Core [12], Geo[13], Transit[14], Media Annotations[15], and Topo[16]. In addition to using these ontologies, there is a system of categories that apply to both events and activities, and points of interest, using both the label and category description, as well as all the instances belonging to these categories. The result is represented using the SKOS language and in particular the axioms `skos:closeMatch` and `skos:broadMatch`. This vocabulary has 480 place categories.

During our experiment, we focused particularly on one of the largest areas available in the City Moove knowledge base which is the French Riviera, with

[7] `https://geodesie.ign.fr/?p=72&page=site_lambert93`
[8] `http://ontologydesignpatterns.org/ont/dul/DUL.owl`
[9] `http://schema.org/`
[10] `http://purl.org/dc/elements/1.1/`
[11] `http://linkedevents.org/ontology/`
[12] `http://www.w3.org/ns/locn/`
[13] `http://www.w3.org/2003/01/geo/wgs84_pos#`
[14] `http://vocab.org/transit/terms/`
[15] `http://www.w3.org/ns/ma-ont#`
[16] `http://data.ign.fr/def/topo#`

nearly 339k locations collected to date. The dataset of the BPE contains 70k facilities on the Côte d'Azur.

3.3 Enriching BPE data using social media

We started by defining a mapping between the categories from BPE and those from the City Moove knowledge base. Across all categories of the BPE, we have managed to map 59 of them with at least or more categories of City Moove. We have materialized these relations through RDF triples using the `owl:sameAs` property.

Listing 1.6. Samples from the mapping between BPE categories and City Moove categories

```
<http://data.linkedevents.org/kos/3cixty/touristinformationcenter>
<http://www.w3.org/2002/07/owl#sameAs>
<http://beta.id.insee.fr/codes/territoire/typeEquipement/G104> .

<http://data.linkedevents.org/kos/3cixty/bank>
<http://www.w3.org/2002/07/owl#sameAs>
<http://beta.id.insee.fr/codes/territoire/typeEquipement/A203> .

<http://data.linkedevents.org/kos/3cixty/postoffice>
<http://www.w3.org/2002/07/owl#sameAs>
<http://beta.id.insee.fr/codes/territoire/typeEquipement/A206> .
...
```

In order to enrich the data of the BPE with those of the City Moove knowledge base, we must first link the entities based on properties common to both sets of data. For this, we use the geographical position and the mapping of the categories. The objective is to calculate a similarity score between each entity, by minimizing the score obtained. The distance is calculated using the Haversine formula. The weight of the geographical quality is defined as follows: 1.0 if the quality is bad, 0.8 if the quality is acceptable, 0.6 if the quality is good.

Given the low number of links on the category mapping, we set the weight to 0.1, in order to favor geographic distance rather than categorization. The formula for calculating the similarity score can be summarized as follows:

$$score = (distanceInMeters * geoWeight) + (catMatch * catWeight) \quad (1)$$

where: $score$ is the similarity score, $distanceInMeters$ is the distance in meters between the two geographic positions, $geoWeight$ is the weight of the geographic quality, $catMatch$ is equal to 0.0 when the categories match, or 1.0 otherwise and $catWeight$ is the weight of category mapping.

The scores obtained are then normalized in order to be contained in an interval between 0 and 1, where 1 corresponds to the best score, and 0 to the

worst score. Finally, the results are converted into RDF using the Expressive Declarative Ontology Alignment Language (EDOAL) format[17], which makes it possible to represent the relations between two entities in the form of RDF triples.

Listing 1.7. Example of an alignment between a facility from BPE and a place from the City Moove knowledge base

```
<http://bpe.eurecom.fr/alignment/967> a align:Alignment;
    align:map [
  a align:Cell;
  align:entity1 <http://beta.id.insee.fr/territoire/equipement/14729731
      >;
  align:entity2 <http://data.linkedevents.org/location/86688656-84d6
      -3971-8467-5f78b6cfb7ab>;
  align:measure "1"^^xsd:float;
  align:relation "="
].
```

The properties `align:entity1` and `align:entity2` contain the URI of each entity, while `align:measure` contains the similarity score obtained in previous steps, and `align:relation` describes the kind of relation between the two entities. In the example of Listing 1.7, the entities are considered as perfectly equal.

3.4 Visualizing enriched BPE data

In order to be able to explore the results obtained, we have developed a web application presenting the user with a map of the world with each BPE device represented as a marker. The color of each marker is based on the second-level category given by the BPE. Only reconciled facilities with a minimum score of 0.8 are shown on the map.

When moving the mouse over a marker, a popup appears with the label, category and photo of the reconciled place. The data is queried directly from the City Moove knowledge base in real time using a Federated SPARQL Query[18] which allows for executing queries distributed over different SPARQL endpoints.

Listing 1.8. Query being used to retrieve the data of a given facility from both the BPE graph and the City Moove knowledge base

```
SELECT ?ent1 ?ent2 ?geo ?capacite ?typeNotation ?typeNotationLabel ?
    businessType ?businessTypeLabel ?label ?poster ?streetAddress ?
    measure WHERE {
```

[17] http://alignapi.gforge.inria.fr/edoal.html
[18] https://www.w3.org/TR/sparql11-federated-query/

Fig. 2. Interactive web application enabling to visualize enriched BPE data, available at `http://sirene.eurecom.fr/bpe/`


```
{
    SELECT ?ent1 ?ent2 ?measure WHERE {
        GRAPH <http://semstats.eurecom.fr/bpe/alignments> {
            <${uri}> a align:Alignment .
            <${uri}> align:map ?map .
            ?map align:entity1 ?ent1 .
            ?map align:entity2 ?ent2 .
            ?map align:relation "=" .
            ?map align:measure ?measure .
            FILTER (?measure >= "0.8"^^xsd:float)
        }
    }
    ORDER BY DESC(?measure)
    LIMIT 1
}
GRAPH <http://semstats.eurecom.fr/bpe/facilities> {
    OPTIONAL { ?ent1 ibpe:capacite ?capacite . }
    ?ent1 dcterms:type ?type .
    GRAPH <http://semstats.eurecom.fr/bpe/codelists> {
        ?type skos:notation ?typeNotation .
        ?type skos:prefLabel ?typeNotationLabel .
    }
}
```

```
    SERVICE <https://kb.city-moove.fr/sparql> {
        ?ent2 rdfs:label ?label .
        ?ent2 geo:location/locn:geometry ?geo .
        ?ent2 locationOnt:businessType ?businessType .
        OPTIONAL { ?businessType skos:prefLabel ?businessTypeLabel . }
        OPTIONAL { ?ent2 lode:poster/ma-ont:locator ?poster . }
        OPTIONAL { ?ent2 schema:location/schema:streetAddress ?
            streetAddress . }
    }
}
```

4 Conclusion

In this paper, we tackled two challenges offered by SemStats 2019. We first proposed a way to model the data from the Sirene database by reusing popular ontologies from W3C and the euBusinessGraph H2020 projet. This allows us to connect and enrich the data using the technologies associated with Linked Data, as we have shown by linking Wikidata pages with Sirene entities based on the SIREN and SIRET numbers. Moreover, this could be used to enrich the Wikidata database by filling up existing pages that don't have the SIREN number yet.

We also showed how existing RDF data could be interlinked with other data sources, by using entity matching techniques. We were then able to create a prototype of a web application to showcase the usage of multiple Linked Data sources. The source code to the Sirene track and BPE track challenges are available on GitHub at `https://github.com/D2KLab/insee/`.

References

1. Troncy, R., Rizzo, G., Jameson, A., Corcho, O., Plu, J., Palumbo, E., Hermida, J.C.B., Spirescu, A., Kuhn, K.D., Barbu, C., Rossi, M., Celino, I., Agarwal, R., Scanu, C., Valla, M., Haaker, T.: 3cixty: Building comprehensive knowledge bases for city exploration. Journal of Web Semantics (JWS) **46-47**, 2–13 (2017)