EURECOM @
SemStats 2019 Challenge

Thibault Ehrhart and Raphaël Troncy
Sirene Track

- French directory managed by INSEE which assigns a SIREN number to French enterprises, and a SIRET number to their establishments
- **Goal:** proposing a RDF model for Sirene data
- CSV dataset which includes:
  - All active and ceased companies
  - All open and closed establishments
  - Organizational changes between establishments
Modeling decisions

- Re-using existing ontologies
- Extending ontologies when necessary
- euBusinessGraph Ontology¹
  - Organization (https://www.w3.org/TR/vocab-org/)
  - Registered Organization (https://www.w3.org/TR/vocab-regorg/)
  - FOAF (http://xmlns.com/foaf/spec/)
  - Schema.org (https://schema.org/)
  - ADMS (https://www.w3.org/TR/vocab-adms/)

¹ https://www.eubusinessgraph.eu/eubusinessgraph-ontology-for-company-data/
Controlled Vocabularies

- **Legal Categories**
  - SKOS-based scheme
  - 306 concepts
  - 3 levels of categories

- **Employee Group** (‘*tranches d’effectifs*’)
  - Uses schema:QuantitativeValue
  - 16 levels defined by Sirene¹

---

Sirene Ontology (1)

- **Legal Units**
  - Mapped on rov:RegisteredOrganization
  - URI based on SIREN number
  - Legal category mapped to rov:orgType
  - Staffing level mapped to schema:numberOfEmployees

- **Establishments**
  - Mapped on rov:RegisteredOrganization and org:Site
  - URI based on SIRET number
  - Postal address mapped to org:siteAddress
  - Linked to legal unit via org:hasSite and org:hasRegisteredSite

- **Organizational Changes**
  - Mapped to org:ChangeEvent
  - Properties org:originalOrganization and org:resultingOrganization are set to the URIs of the establishments
None of the existing ontologies covered the complete scope we needed.

We created an extension called **UniteJuridique**
- Prefix: sirene
- Github: [https://github.com/D2KLab/insee/tree/master/sirene/ontology](https://github.com/D2KLab/insee/tree/master/sirene/ontology)

It is declared as an **owl:Class** and contains **37 properties** that are based on the name of the variables from the Sirene dataset.
- Examples:
  - sirene:identifiantAssociationUniteLegale
  - sirene:activitePrincipaleRegistreMetiersEtablissement
  - ...

---

**Sirene Ontology (2)**
Data Enrichment

- Linking legal units and establishments with other sources using $\texttt{owl:sameAs}$.

- Data from entreprise.data.gouv.fr.

- Data from Wikidata
SIREN coverage in Wikidata

- We extracted the data from the Wikidata knowledge graph using a SPARQL query to retrieve the entities with properties P1616 (SIREN number) and P3215 (SIRET number).
- We then link the entities together using their registration number.
- In the end, we obtain a list of links to the Wikidata pages of 40,984 companies and 374 establishments, which are materialized thanks to the owl:sameAs property.
- Dump: 76GB
Example
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.

3 datasets:
- bpe2018-facilities: contains data for each facility, in RDF format.
- bpe2018-codelists: the code lists used, expressed in SKOS.

**Goal:** enriching BPE data with other sources
Knowledge Base: City Moove

- Knowledge base specialized in the domain of tourism and city exploration

- Contains descriptions of events, activities, POIs, transportation facilities and social activities, collected from numerous local and global data providers (tourism offices, social medias, etc.)

- Entities are deduplicated, interlinked and enriched using semantic technologies

- Contains a vocabulary for categories of places, with over 480 categories

- Largest area covered: French Riviera (Côte d'Azur), with nearly 339k locations collected in 2019
Enriching BPE data using social media

- We created a mapping between the 501 categories from BPE and those from the City Moove knowledge base
  - 59 BPE categories were mapped with at least one category from City Moove
  - Relation materialized using the `owl:sameAs` property
- Entity linking based on properties common to both sets of data.
  - Using: the geographical position, and the categories mapping
  - Goal: calculate a similarity score between each entity, by minimizing the score obtained
BPE Entities Linking

Similarity score formula:

\[
\text{score} = (\text{distanceInMeters} \times \text{geoWeight}) + (\text{catMatch} \times \text{catWeight})
\]

- **score** is the similarity score desired
- **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
- **catWeight** is the weight of categories matching (0.1)

Note: scores are normalized to be contained between 0 (worst) and 1 (best)
Finally, the results are converted into RDF using the Expressive Declarative Ontology Alignment Language (EDOAL), which makes it possible to represent the relations between two entities in the form of RDF triples:

```
<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 "="
  ].
```
Showcase: we developed a web app allowing the user to explore the data on a map with each BPE installation being represented as a marker. Quality of the alignment is BAD!

When moving the mouse over a marker, a popup appears with the label, category and photo of the reconciled place, as well as the similarity score.

The data is queried directly from the City Moove knowledge base in real time using a Federated SPARQL Query, which allows for executing queries distributed over different SPARQL endpoints.

http://sirene.eurecom.fr/bpe/
Conclusion

● We created a model of the Sirene database by reusing existing ontologies from W3C and euBusinessGraph

● We linked Wikidata pages with Sirene entities, using the technologies associated with Linked Data. This could also help enriching Wikidata by filling up existing pages that don't have the SIREN number yet

● We showed how existing RDF data could be interlinked with other data sources, by using entity matching techniques and alignment ontologies.

● The source code to the Sirene track and BPE track challenges are available on GitHub: https://github.com/D2KLab/insee/.