Modeling and Publishing French Business Register (Sirene) Data as Linked Data

Using the euBusinessGraph Ontology

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

s.abdelkader@campus.unimib.it
Outline

- Introduction
- The euBusinessGraph Ontology
  - Overview
  - Extensions for the Sirene challenge
- Sirene data RDF mapping
  - Design
  - Implementation
- Use cases
  - Data publication
  - Reconciliation and Extension
- Summary and Outlook
Introduction

- Company data are the **basis** of many **data value chains**
- Basic company data are typically managed by **national business registers**
- **No standard** exists for harmonizing basic company data
  - Across countries
  - Machine-readable
  - For enabling integration of basic company information
The euBusinessGraph Ontology

- An approach to **harmonize basic company data**
  - Based on several existing vocabularies, such as EU Core Vocabs, schema.org, ADMS Vocab, Dublin Core, and more
- Concepts and relations to describe:
  - Basic company information
  - Systems of identifiers
- Suitable for representing a **snapshot** of companies status (no history)
Typical use of the euBusinessGraph Ontology

**Sources**
- National registers
- Gazettes
- Specialised registers (e.g., start-ups)
- Websites
- Social media accounts

**Data providers**

- SIRENE schema
  - company_number: 0005410949
  - legal_name: A LA GRANDE FABRIQUE
- base.fr Company Number: 0005410949
  - legalName: A LA GRANDE FABRIQUE
- Other data provider schema

**Graph operator**

- Common schema
  - identifiers:
    - fr:0005410949
    - twitter:@opencorporates
  - Legal Name: Chrinon Ltd.

**Data consumers**

**Service providers**

- Banks
- Marketing/Sales
- PSO
- Procurement
- Compliance

**Business cases:**
- Atoka+ TDS CRM-S DJP
- CED BR-S

**Graph services:**
- Economic indicators
- Analytics (e.g., credit/risk)
- Text analysis

**Sources**

- Data providers
- Graph operator
- Data consumers
- Service providers

**Other data provider schema**

- company mentions within a news stream
Extending the euBusinessGraph Ontology

The Sirene dataset focuses on the description of:

- Legal units
- Establishments of legal units
- Legal events occurred since their creation

The euBusinessGraph ontology mainly covers basic company information

A few extensions were needed to describe key Sirene entities:

1. Events (legal changes in companies)
2. Legal unit - establishment relationships
Events Model

- **Events** are modeled based on the **Simple Event Model (SEM)**
  - Flexible model
  - Easily adaptable to different kinds of events
- **SEM** provides classes and relations that describe generic events
  - Extended with a new property “eubg:eventValue” useful to track different events of the same type, but with different value, e.g., change of the address or change of the activity type

Legal Unit - Establishment Relationship

- **Legal unit - establishment relationships** modeled using the **Organization Ontology**
  - Already used in euBusinessGraph
  - Provides concepts to describe relationships between Legal Unit and Establishment:
    - An Establishment is a unit of a Legal Unit
    - A Legal Unit might have an establishment or a HQ establishment

*https://www.w3.org/TR/vocab-org/
Core euBusinessGraph Concepts

**Basic information**
- jurisdiction
- registration
- official registration

**Names**
- Legal name
- Alt/Trading name
- Preferred name

**Event**
- Event Type
- Date
- Event Value

**Classifications**
- Type
- Status
- Economic Activity

**Other company details**
- Web languages
- Incorp./Dissolution date
- Publicly traded
- State Owned
- Is startup

**Company** *(rov:RegisteredOrganization)*

**Physical presence**
- Registered address
- Address
- Place admin hierarchy
- Street
- Geocoordinates

**Online presence**
- Certified email
- Wikipedia page
- Website
- News/blog feed
Sirene data mapping to the semantic model (extended euBusinessGraph Ontology)

For the mapping phase it was decided to:

1. Map the five files **separately** (1+ mappings for each file)
2. Generate the RDF files
3. Use the **same URIs** across different mappings to link their resources in an RDF database

Some of the attributes had a preliminary transformation to better fit the RDF mapping (E.g., “av.”, “Cesar”, “32” cells were concatenated into “Cesar avenue, 32”)

Example #1: Company Information
Example #2: Company Relations
# Example #3: Company Events

https://datagraft.io/shad/transformations/rdf-new_stocketablissementhistorique_utf8/edit

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>changementEstatAdministratifEtablissement</td>
<td>changementEnseigneEtablissement</td>
<td>changementDenominationUsuelleEtablissement</td>
<td>changementActivitePrincipaleEtablissement</td>
<td>changementCaractereEmployeEtablissement</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EventDateID</th>
<th>variable</th>
<th>value</th>
<th>Event-type</th>
<th>event-value</th>
<th>event_Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>124337520000</td>
<td>changementEstatAdministratifEtablissement</td>
<td>true</td>
<td>change_administrative_state</td>
<td>F</td>
<td>FR/000325175000016/ic/SIRET/event/2009-05-27/change_administrative_state</td>
</tr>
<tr>
<td>119914200000</td>
<td>changementActivitePrincipaleEtablissement</td>
<td>true</td>
<td>change_principal_activity</td>
<td>32.12</td>
<td>FR/000325175000016/ic/SIRET/event/2008-01-01/change_principal_activity</td>
</tr>
<tr>
<td>131914300000</td>
<td>changementEstatAdministratifEtablissement</td>
<td>true</td>
<td>change_administrative_state</td>
<td>F</td>
<td>FR/000325175000024/id/SIRET/event/2011-10-21/change_administrative_state</td>
</tr>
<tr>
<td>131914800000</td>
<td>changementEstatAdministratifEtablissement</td>
<td>true</td>
<td>change_administrative_state</td>
<td>F</td>
<td>FR/000325175000032/id/SIRET/event/2011-10-21/change_administrative_state</td>
</tr>
</tbody>
</table>
Example #3: Company Events (cont’)

https://datagraft.io/shad/transformations/rdf-new_stocketablissementhistorique_utf8/edit
Implementation

Transformations and mappings are designed with **Grafterizer 2.0**, the data transformation tool available in DataGraft ([https://datagraft.io](https://datagraft.io))

- **Grafterizer 2.0 uses a batch approach** for transforming tabular data (CSV) into RDF triples
- **DataGraft allows you to manage different types of assets**, such as files, data transformations and SPARQL endpoints
  - Assets can be shared and reused
Implementation (cont’)

The graph mapping is used to generate RDF data from the transformed tabular data.

Mapping elements in Grafterizer:

- Nodes are boxes
  - URI, Literal or Blank
  - Populated with free-defined text or by reading values from a specific column
- Properties are labels between nodes
Use Case #1: Data Publication

- The full dataset provided in the challenge amounts to approx. **16GB**
- We applied the mapping by following the data wrangling concept developed within the **EW-Shopp project**:
  - RDF mapping designed on a sample (Grafterizer 2.0 UI)
  - Script execution on the full dataset at scale (EW-Shopp processing solution)
- The resulting RDF dataset:
  - Contains approx. **3 billion triples** (n-triple format)
  - Amounts to approx. **450GB** (mainly due to fully qualified names)
- Data available at [https://sirene-data.sintef.cloud/](https://sirene-data.sintef.cloud/)
Use Case #2: Reconciliation and Extension

It should be useful to enrich the Siren dataset with additional information.

A table enrichment task is performed by applying an arbitrary sequence of:

- **Reconciliation** steps, which link values in table to identifiers in external knowledge bases.
- **Extension** steps, which add new columns containing values fetched from a third-party source, using identifiers to query the source.
Reconciliation and extension

**ASIA** is a tool that supports the data enrichment, fully integrated with Grafterizer.

We enriched the input data with **ASIA services** by exploiting two kinds of information available in the dataset:

- Company names, to reconcile against DBpedia
- City toponyms, to reconcile against GeoNames
Reconciliation and Extension (cont’)

The enrichment tasks lead to different results:

1. **Company-based enrichment**: it was not satisfactory, because many companies are identified by the name and surname of the owner, leading to many false positives while reconciling names against DBpedia.

2. **Toponyms-based enrichment**: it successfully added information about spatial administrative levels (e.g., ADM1, ADM2, ADM3, ADM4) from GeoNames.
Summary and outlook

- euBusinessGraph as the baseline ontology for company information
  - Extended to capture modelling needs from the Sirene dataset
- The extended euBusinessGraph ontology captures the key company elements represented in the Sirene dataset
  - Some attributes were discarded because not strictly relevant to the organizational/economic description, e.g., StatutDiffusionEtablissement (an agreement to share data), UnitLegalSex (the genre of the company owner)
- Exemplified the use of the resulting ontology in two use cases
- Potential future work: Further extension the euBusinessGraph Ontology to cover all the data attributes described in the Sirene datasets
This work has been funded from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 732003 (euBusinessGraph) and No 732590 (EW-Shopp).

Thank you!