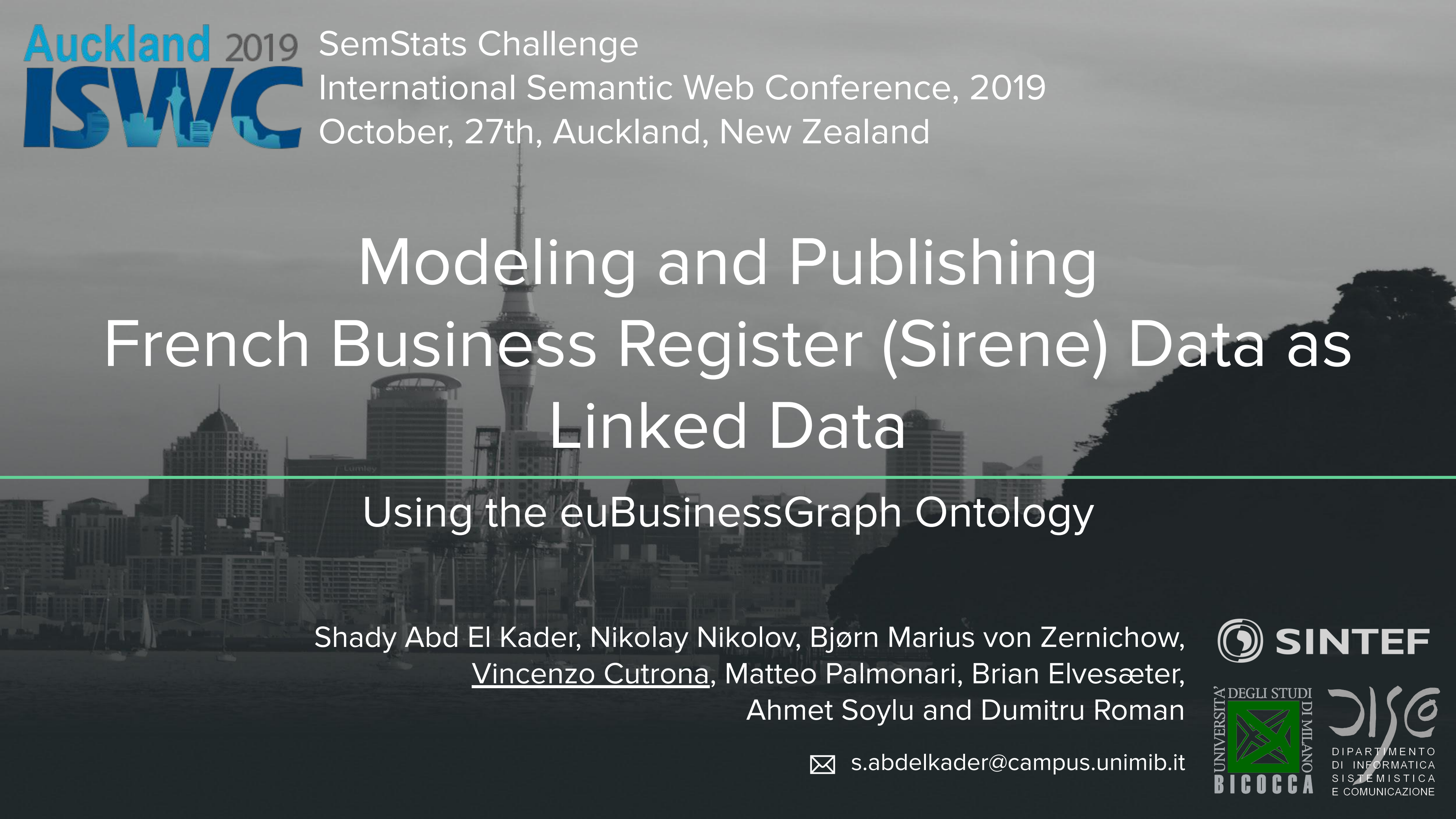


SemStats Challenge
International Semantic Web Conference, 2019
October, 27th, Auckland, New Zealand

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

SINTEF

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Outline

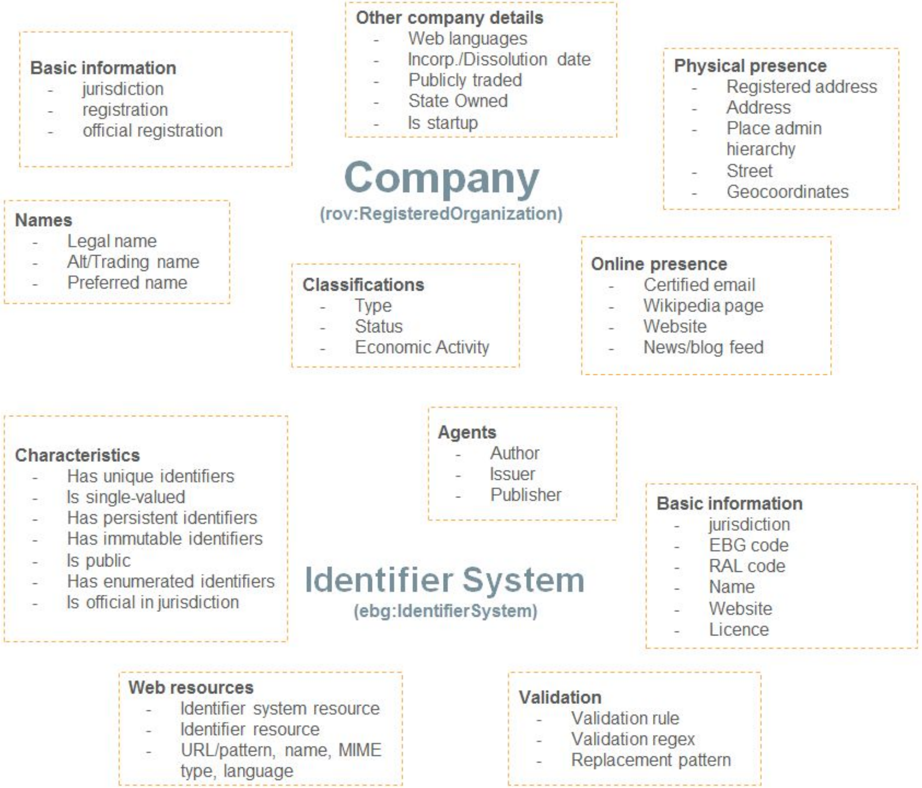
- Introduction
- The euBusinessGraph Ontology
  - Overview
  - Extensions for the Sirene challenge
- Sirene data RDF mapping
  - Design
  - Implementation
- Use cases
  - Data publication
  - Reconciliation and Extension
- Summary and Outlook

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)



Company
(rov:RegisteredOrganization)

Basic information
- jurisdiction
- registration
- official registration

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

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

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

Classifications
- Type
- Status
- Economic Activity

Online presence
- Certified email
- Wikipedia page
- Website
- News/blog feed

Identifier System
(ebg:IdentifierSystem)

Characteristics
- Has unique identifiers
- Is single-valued
- Has persistent identifiers
- Has immutable identifiers
- Is public
- Has enumerated identifiers
- Is official in jurisdiction

Agents
- Author
- Issuer
- Publisher

Basic information
- jurisdiction
- EBG code
- RAL code
- Name
- Website
- Licence

Web resources
- Identifier system resource
- Identifier resource
- URL/pattern, name, MIME type, language

Validation
- Validation rule
- Validation regex
- Replacement pattern

Typical use of the euBusinessGraph Ontology

Sources

National registers

Gazettes

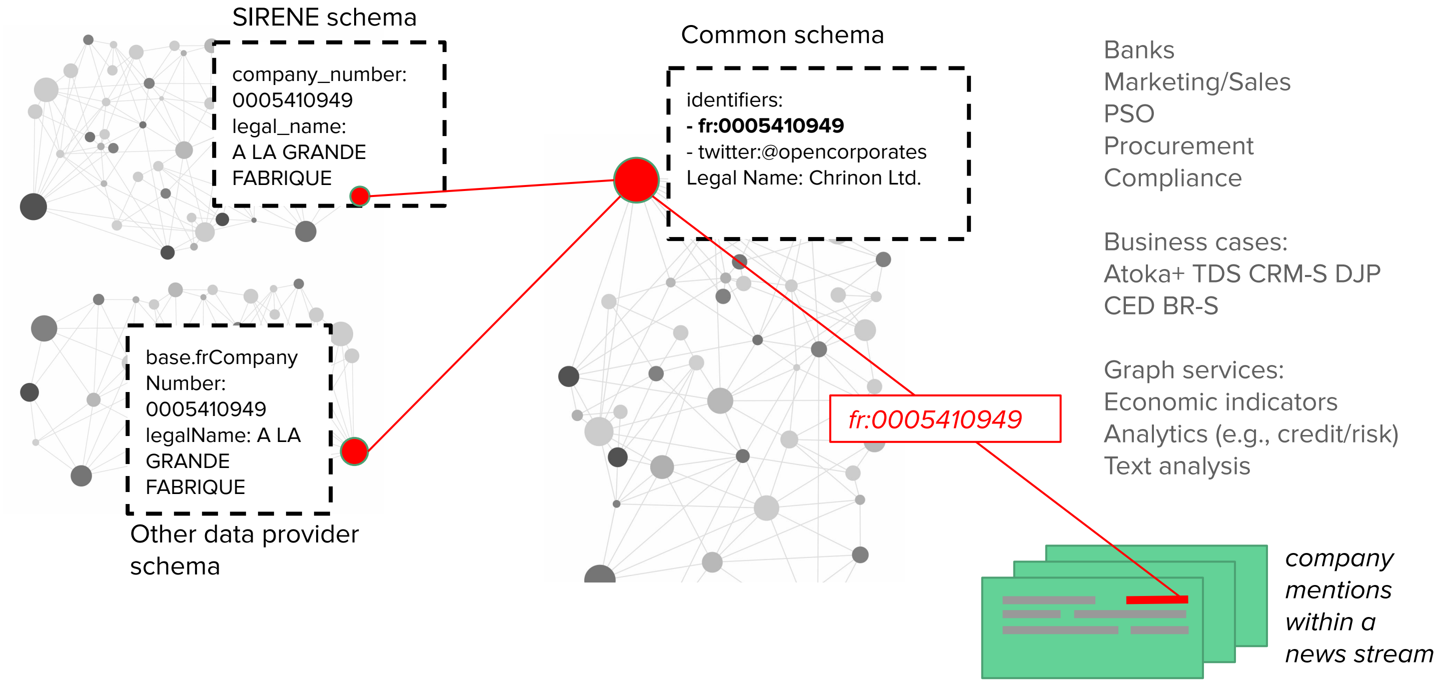
Specialised registers (e.g., start-ups)

Websites

Social media accounts

Data providers

Graph operator



SIRENE schema

company_number: 0005410949
legal_name: A LA GRANDE FABRIQUE

base.frCompany Number: 0005410949
legalName: A LA GRANDE FABRIQUE

Other data provider schema

Common schema

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

company mentions within a news stream

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

Extending the euBusinessGraph Ontology

The Sirene dataset focuses on the description of:

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

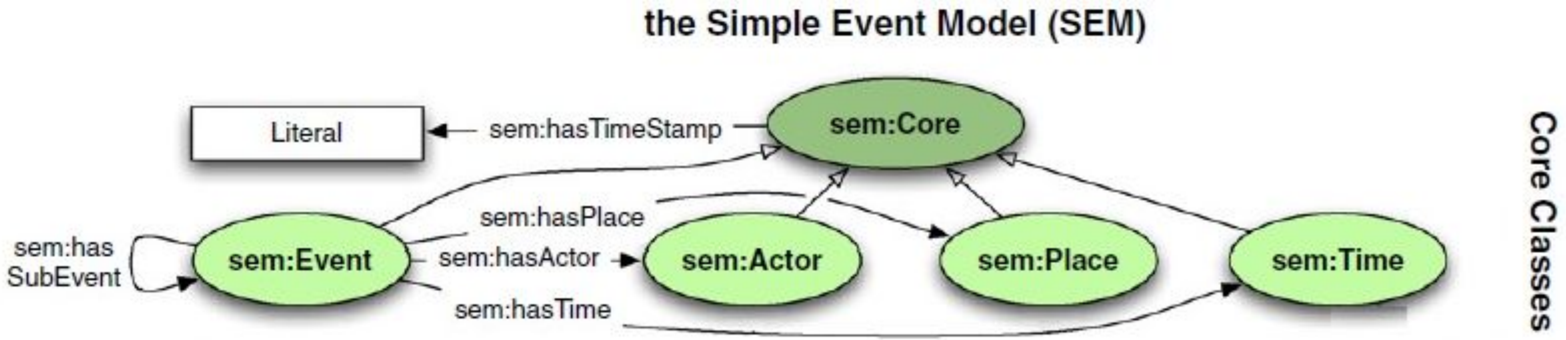
The euBusinessGraph ontology mainly covers **basic company information**

A few extensions were needed to describe key Sirene entities:

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

Events Model

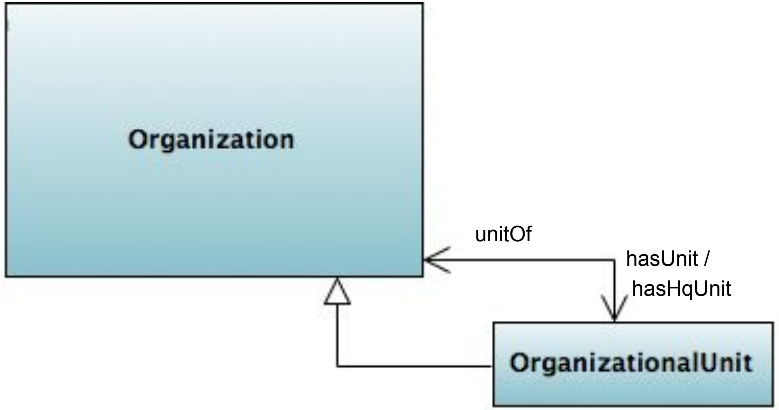
- **Events** are modeled based on the **Simple Event Model (SEM)***
  - Flexible model
  - Easily adaptable to different kinds of events
- **SEM** provides classes and relations that describe generic events
  - Extended with a new property “eubg:eventValue” useful to track different events of the same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



*http://semanticweb.cs.vu.nl/2009/11/sem

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

Other company details ✗

- Web languages
- Incorp./Dissolution date
- Publicly traded
- State Owned
- Is startup

Company

(rov:RegisteredOrganization)

Classifications ✓

- Type
- Status
- Economic Activity

Physical presence ✓

- Registered address
- Address
- Place admin hierarchy
- Street
- Geocoordinates

Online presence ✗

- Certified email
- Wikipedia page
- Website
- News/blog feed

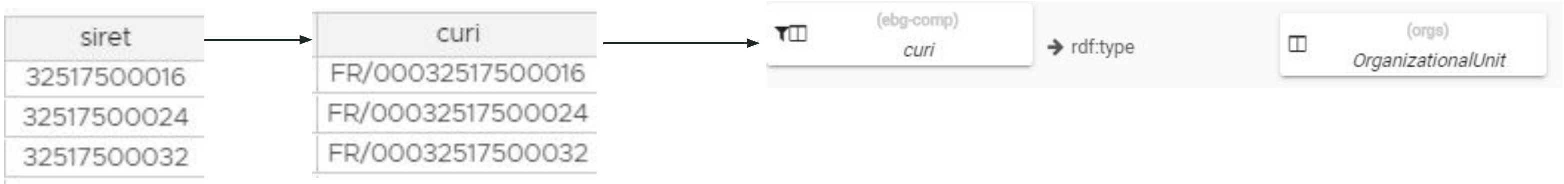
Sirene data mapping to the semantic model (extended euBusinessGraph Ontology)

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”)

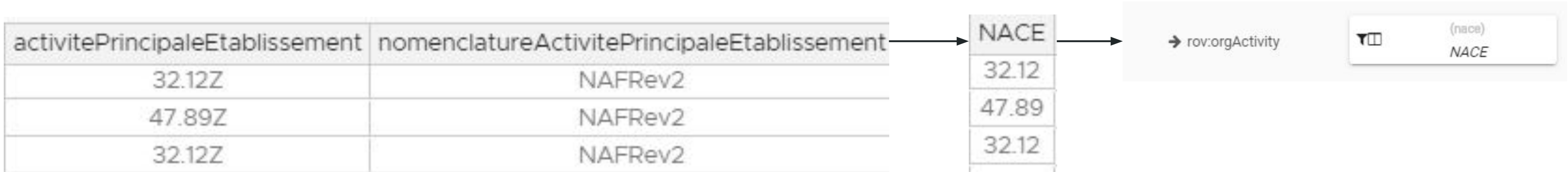
Example #1: Company Information



siret
32517500016
32517500024
32517500032

curi
FR/00032517500016
FR/00032517500024
FR/00032517500032

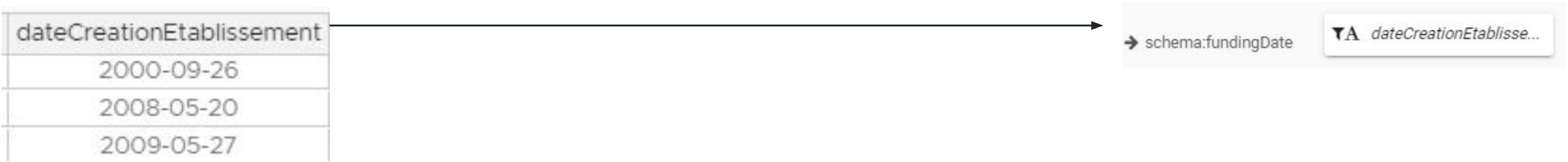
(ebg-comp) *curi* → rdf:type (orgs) *OrganizationalUnit*



activitePrincipaleEtablissement	nomenclatureActivitePrincipaleEtablissement
32.12Z	NAFRev2
47.89Z	NAFRev2
32.12Z	NAFRev2

NACE
32.12
47.89
32.12

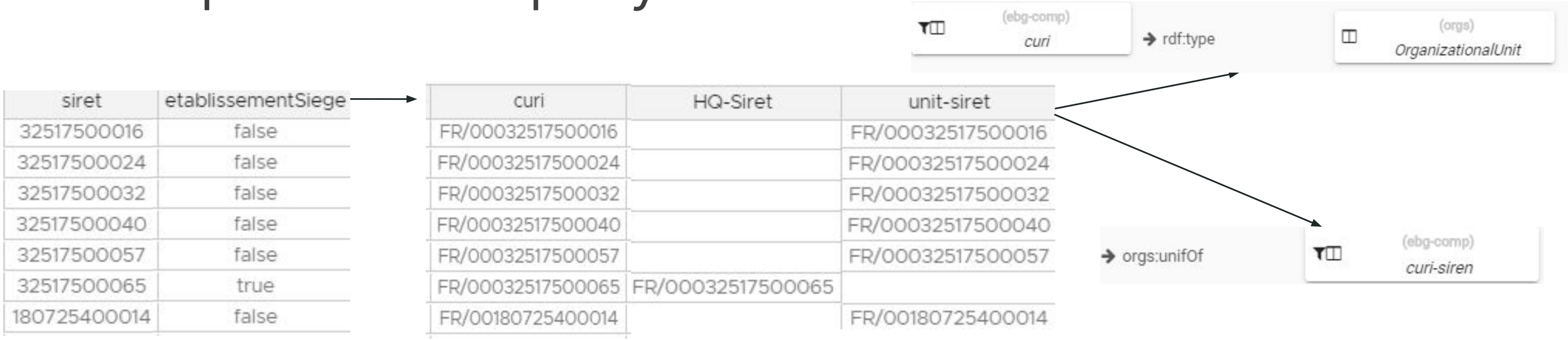
→ rov:orgActivity (nace) *NACE*



dateCreationEtablissement
2000-09-26
2008-05-20
2009-05-27

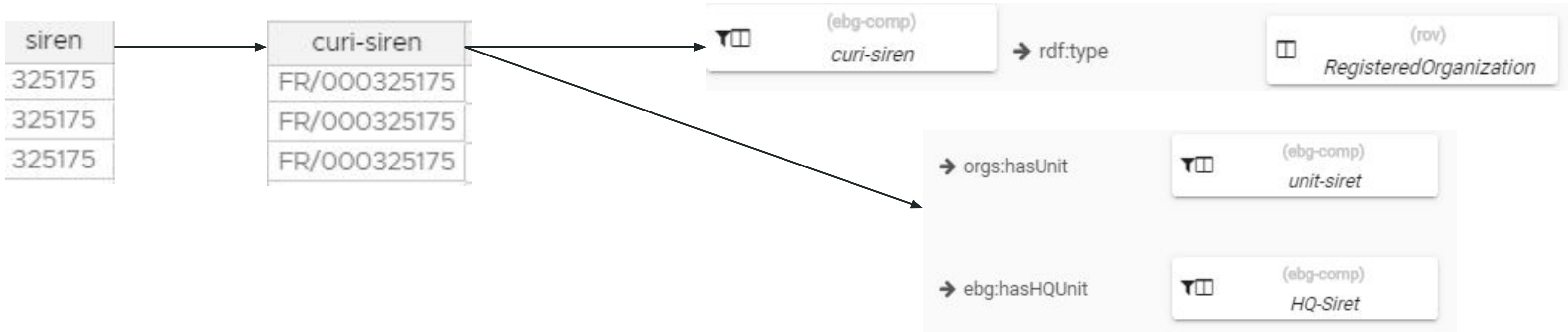
→ schema:fundingDate *dateCreationEtablisse...*

Example #2: Company Relations



siret	etablissementSiege
32517500016	false
32517500024	false
32517500032	false
32517500040	false
32517500057	false
32517500065	true
180725400014	false

curi	HQ-Siret	unit-siret
FR/00032517500016		FR/00032517500016
FR/00032517500024		FR/00032517500024
FR/00032517500032		FR/00032517500032
FR/00032517500040		FR/00032517500040
FR/00032517500057		FR/00032517500057
FR/00032517500065	FR/00032517500065	
FR/00180725400014		FR/00180725400014

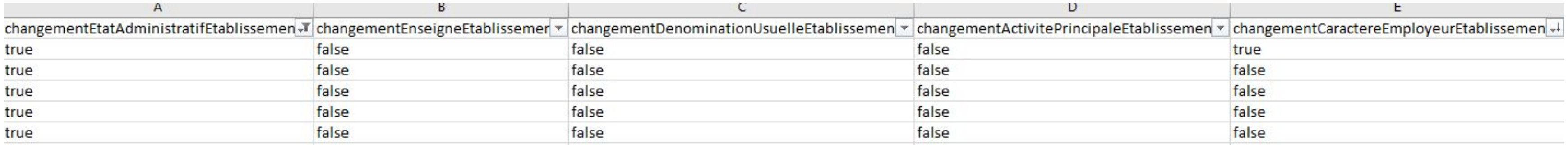


siren
325175
325175
325175

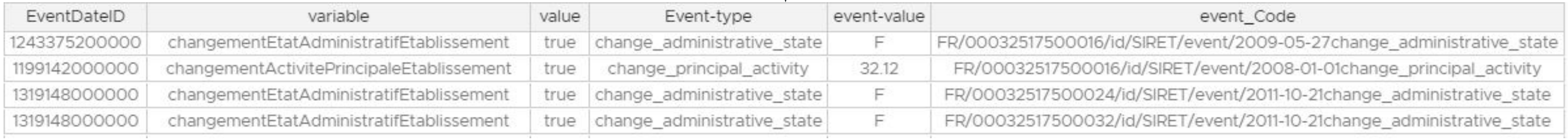
curi-siren
FR/000325175
FR/000325175
FR/000325175

Example #3: Company Events

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



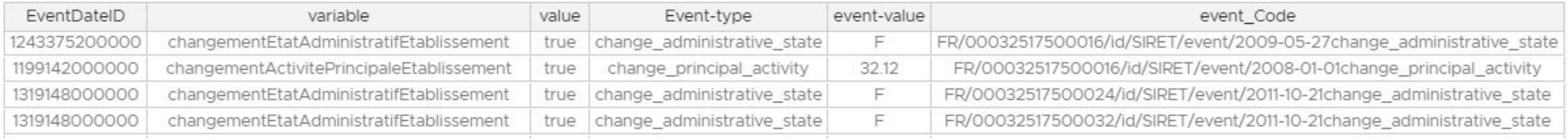
A	B	C	D	E
changementEtatAdministratifEtablissemen	changementEnseigneEtablissemen	changementDenominationUsuelleEtablissemen	changementActivitePrincipaleEtablissemen	changementCaractereEmployeurEtablissemen
true	false	false	false	true
true	false	false	false	false
true	false	false	false	false
true	false	false	false	false
true	false	false	false	false



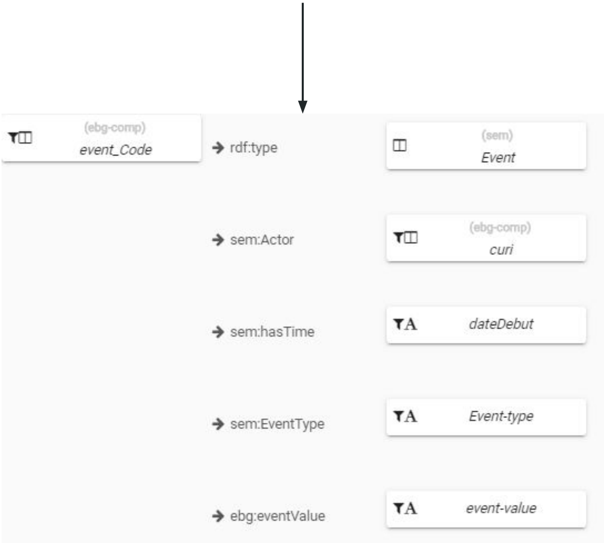
EventDateID	variable	value	Event-type	event-value	event_Code
1243375200000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500016/id/SIRET/event/2009-05-27change_administrative_state
1199142000000	changementActivitePrincipaleEtablissement	true	change_principal_activity	32.12	FR/00032517500016/id/SIRET/event/2008-01-01change_principal_activity
1319148000000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500024/id/SIRET/event/2011-10-21change_administrative_state
1319148000000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500032/id/SIRET/event/2011-10-21change_administrative_state

Example #3: Company Events (cont')

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



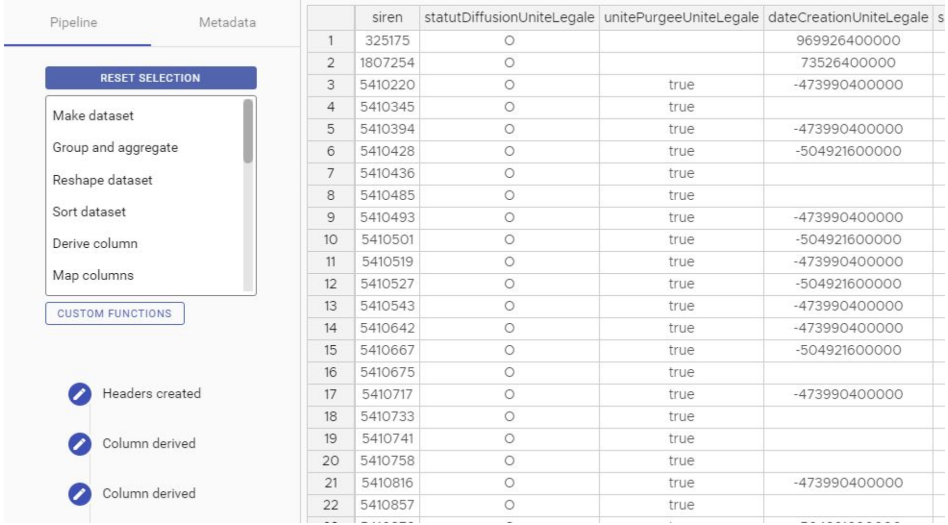
EventDateID	variable	value	Event-type	event-value	event_Code
1243375200000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500016/id/SIRET/event/2009-05-27change_administrative_state
1199142000000	changementActivitePrincipaleEtablissement	true	change_principal_activity	32.12	FR/00032517500016/id/SIRET/event/2008-01-01change_principal_activity
1319148000000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500024/id/SIRET/event/2011-10-21change_administrative_state
1319148000000	changementEtatAdministratifEtablissement	true	change_administrative_state	F	FR/00032517500032/id/SIRET/event/2011-10-21change_administrative_state



Implementation

Transformations and mappings are designed with **Grafterizer 2.0**, the data transformation tool available in DataGraft (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



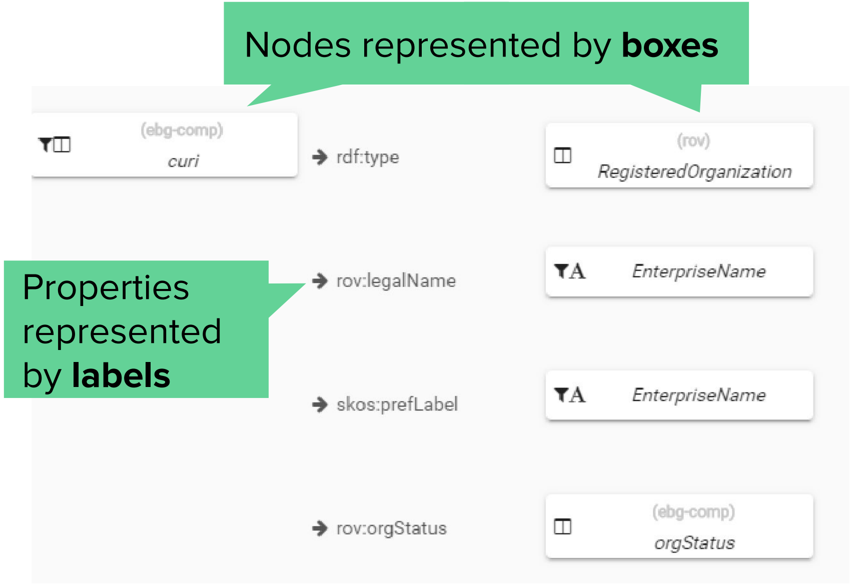
	siren	statutDiffusionUniteLegale	unitePurgeeUniteLegale	dateCreationUniteLegale	s
1	325175	O		969926400000	
2	1807254	O		73526400000	
3	5410220	O	true	-473990400000	
4	5410345	O	true		
5	5410394	O	true	-473990400000	
6	5410428	O	true	-504921600000	
7	5410436	O	true		
8	5410485	O	true		
9	5410493	O	true	-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		

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


Nodes represented by **boxes**

Properties represented by **labels**

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/

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:

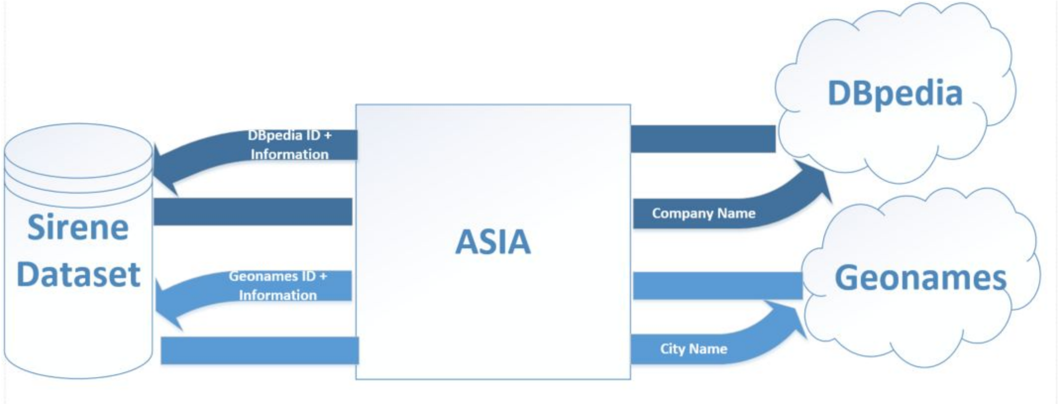
- **Reconciliation** steps, which link values in table to identifiers in external knowledge bases
- **Extension** steps, which add new columns containing values fetched from a third-party source, using identifiers to query the source

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:

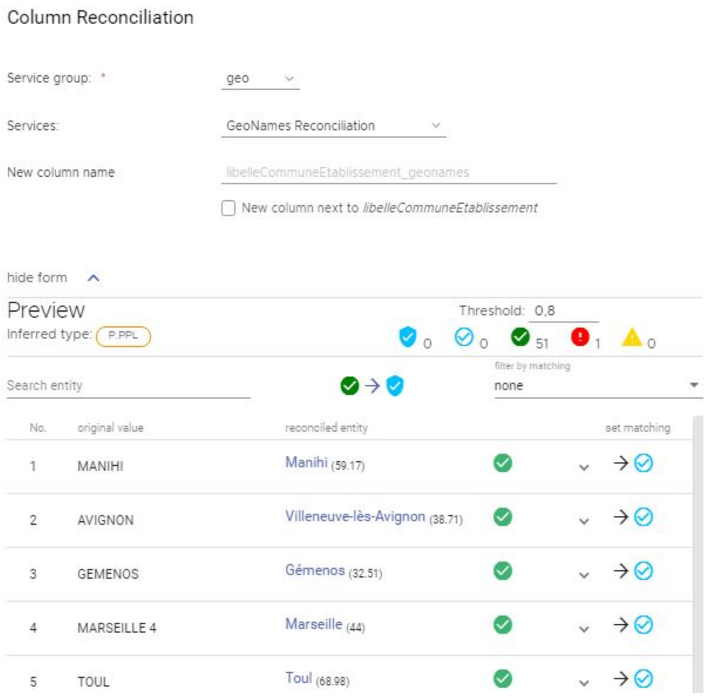
- Company names, to reconcile against DBpedia
- City toponyms, to reconcile against GeoNames



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



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

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)	✓	→ ✓

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!