

---

# Web Sémantique

## Projet d'enrichissement de données libres

---

<b>I- Introduction</b>	<b>1</b>
<b>II- Transformation du dataset en données sémantiques</b>	<b>2</b>
II-1 Outil de sémantisation utilisé : OpenRefine	2
II-2 Intégration des données : langage GREL	3
II-3 Résultat des transformations : Turtle	4
<b>III- Requêtes effectuées sur le dataset</b>	<b>5</b>
<b>IV- Liaison à un autre dataset</b>	<b>7</b>
<b>V- Ontologie OWL/RDFS et inférences</b>	<b>7</b>
<b>VI- Liaison des données avec le cloud linked data</b>	<b>9</b>
<b>VII- Métadonnées VOID</b>	<b>9</b>
<b>Liens</b>	<b>9</b>

# I- Introduction

L'objectif de ce projet était de choisir un dataset de l'Enseignement Supérieur et de la Recherche (<https://data.enseignementsup-recherche.gouv.fr>) afin de le "sémantifier".

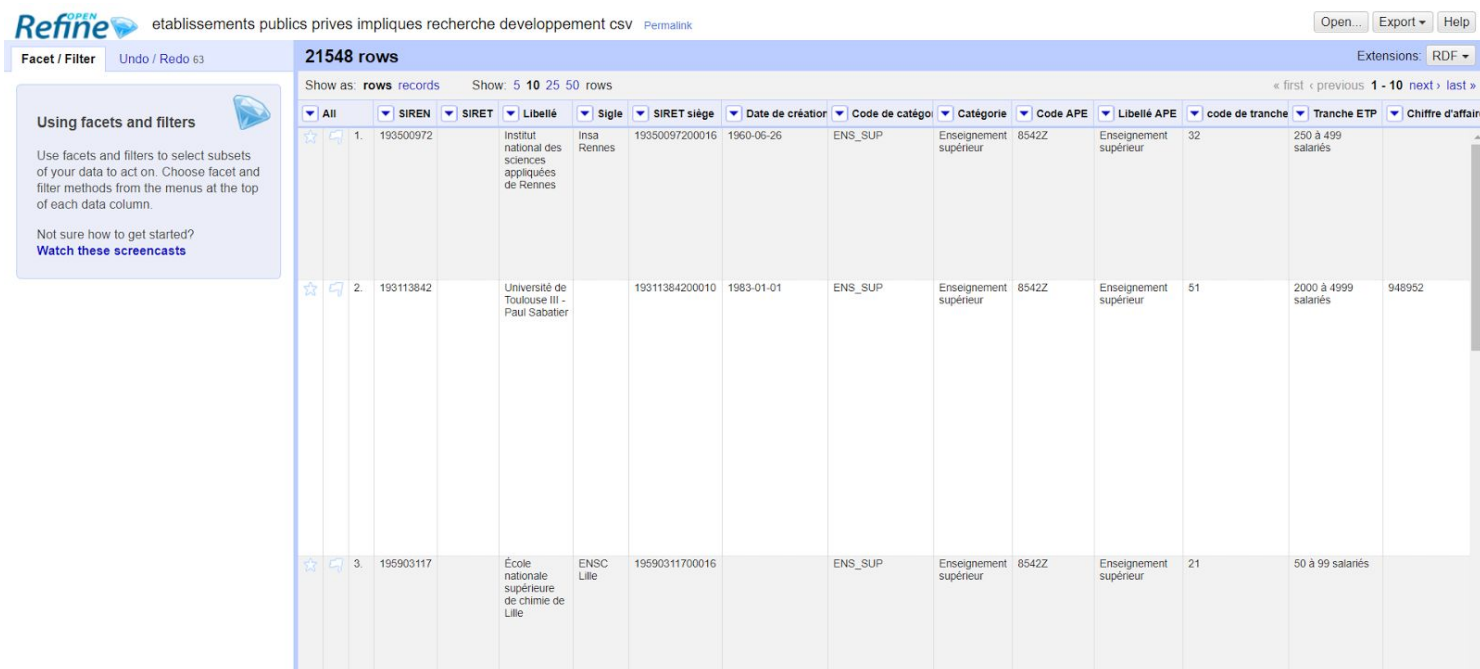
Nous avons choisi de travailler sur les [Établissements publics et privés impliqués dans la recherche et développement](#). Ce dataset regroupe donc des informations sur les entreprises privées/publiques ainsi que les différentes institutions de l'enseignement qui participent à la recherche et au développement en France. Les informations sur le type d'activités liées à la recherche n'est pas précis mais permet néanmoins des requêtes intéressantes et, de plus, les données sémantiques peuvent être liés à certains datasets des autres groupes pour obtenir des informations plus complètes et pertinentes.

	SIREN	SIRET	Libellé	Sigle	SIRET siège	Date de création	Code de catégorie	Catégorie
1	193500972		Institut national des sciences appliquées	Insa Rennes	19350097200016	26 juin 1960	ENS_SUP	Enseignement supérieur
2	193113842		Université de Toulouse III - Paul Sabatier		19311384200010	1 janvier 1983	ENS_SUP	Enseignement supérieur
3	195903117		École nationale supérieure de chimie de	ENSC Lille	19590311700016		ENS_SUP	Enseignement supérieur
4	196244016		Université d'Artois		19624401600016	7 novembre 1991	ENS_SUP	Enseignement supérieur
5	194200937		École nationale d'ingénieurs de Saint-Éti	Enise	19420093700010	1 janvier 1983	ENS_SUP	Enseignement supérieur
6	195935598		Université Lille 1 - Sciences technologies	USTL	19593559800019	10 août 1983	ENS_SUP	Enseignement supérieur
7	193401320		École nationale supérieure d'architecture		19340132000018	23 novembre 1966	ENS_SUP	Enseignement supérieur
8	193801412		École nationale supérieure d'architecture	Ensag	19380141200019	28 juillet 1966	ENS_SUP	Enseignement supérieur
9	193301991		École d'architecture de Bordeaux		19330199100017	6 juillet 1990	ENS_SUP	Enseignement supérieur
10	193500899		École nationale supérieure d'architecture		19350089900029	1 août 1966	ENS_SUP	Enseignement supérieur
11	196917744		Université Claude Bernard - Lyon 1		19691774400019	9 juin 1970	ENS_SUP	Enseignement supérieur
12	197300500		Université de Savoie	UDS	19730050000015	26 juin 1960	ENS_SUP	Enseignement supérieur

*Aperçu des données du dataset*

## II- Transformation du dataset en données sémantiques

### II-1 Outil de sémantisation utilisé : OpenRefine



etablissements publics prives impliquees recherche developpement csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 63

21548 rows Extensions: RDF

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

All	SIREN	SIRET	Libellé	Sigle	SIRET siège	Date de création	Code de catégo	Catégorie	Code APE	Libellé APE	code de tranche	Tranche ETP	Chiffre d'affaire
1.	193500972		Institut national des sciences appliquées de Rennes	Insa Rennes	19350097200016	1960-06-26	ENS_SUP	Enseignement supérieur	8542Z	Enseignement supérieur	32	250 à 499 salariés	
2.	193113842		Université de Toulouse III - Paul Sabatier		19311384200010	1983-01-01	ENS_SUP	Enseignement supérieur	8542Z	Enseignement supérieur	51	2000 à 4999 salariés	948952
3.	195903117		École nationale supérieure de chimie de Lille	ENSC Lille	19590311700016		ENS_SUP	Enseignement supérieur	8542Z	Enseignement supérieur	21	50 à 99 salariés	

OpenRefine - Open Source (<http://openrefine.org/>)

OpenRefine est une application web tournant sur un serveur local. Elle permet de raffiner facilement un grand volume de données, les rendant plus homogènes et utilisables. L'application possède une extension pour transformer des données "plates" - de type CSV dans le cas de notre dataset - en format RDF pour le web sémantique. Il s'agit de matcher les différentes colonnes du tableau CSV avec des propriétés (prédicats) pour que l'application s'occupe de générer un fichier au format RDF/XML ou Turtle.

## II-2 Intégration des données : langage GREL

Certaines données du dataset ont dû être modifiées afin de pouvoir plus aisément les associer à des prédicats existants. Par exemple, l'adresse de chaque établissement était au départ découpée en de nombreux champs tels que le type de voie, le numéro, etc. Il était préférable de regrouper ces données en une seule colonne, l'adresse. Il en a été de même pour les coordonnées géographiques de ces établissements. De plus, tous les champs n'ont pas été utilisés, pour leur manque d'intérêt principalement.

**Add column based on column Numéro de voie**

New column name:

☒ set to blank ☐ store error ☐ copy value from original column

Expression:

Language:

**Preview** History Starred Help

row	value
1.	20 20 Avenue des Buttes de Coesmes
2.	118 118 Route de Narbonne
3.	null Avenue Mendeleiev
4.	9 9 Rue du Temple
5.	58 58 Rue Jean Parot

OK Cancel

*Fusion de colonnes avec le langage GREL*

## II-3 Résultat des transformations : Turtle

```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix schema: <http://schema.org/> .
@prefix geo: <http://rdf.insee.fr/geo/> .
@prefix wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix local: <http://localhost:3333/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dbprop: <http://dbpedia.org/property/> .

_:node1c01q71bgx1 a schema:Organization ;
    schema:foundingDate "1960-06-26" ;
    foaf:isPrimaryTopicOf "https://fr.wikipedia.org/wiki/Institut_national_des_sciences_appliqu%C3%A9es_de_Rennes" ;
    schema:legalName "Institut national des sciences appliquées de Rennes" ;
    dbprop:town "Rennes" ;
    schema:postalCode "35700" ;
    dbprop:location "20 Avenue des Buttes de Coesmes " ;
    dbprop:website "http://www.insa-rennes.fr/" ;
    wgs84:lat "48.1216" ;
    wgs84:long "-1.63297" ;
    dbo:wikiPageExternalLink "https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-principaux-etablissement";
    dc:subject "Participant d'un projet financé par l'ANR;Déposant de brevet (base de l'INPI);PIA - Initiative d'excellence " ;
    local:SIREN "193500972" ;
    local:tranche_etp "250 à 499 salariés" ;
    local:categorie_juridique "Établissement public national à caractère scientifique culturel et professionnel" ;
    local:categorie_org "Enseignement supérieur" ;
    dbprop:region "Bretagne" ;
    geo:Departement "Ille-et-Vilaine" .

_:node1c01q71bgx2 a schema:Organization ;
    schema:foundingDate "1983-01-01" ;
    foaf:isPrimaryTopicOf "https://fr.wikipedia.org/wiki/Universit%C3%A9_Toulouse-III-Paul-Sabatier" ;
    schema:legalName "Université de Toulouse III - Paul Sabatier" ;
    dbprop:town "Toulouse" ;
    schema:postalCode "31400" ;
    dbprop:location "118 Route de Narbonne " ;
    dbprop:website "http://www.univ-tlse3.fr/" ;
```

### Aperçu des données formulées en Turtle

Chaque établissement présent dans le fichier CSV devient un blank node contenant les champs (propriétés) que nous avons décidé de conserver. Nous avons créé quatre propriétés: *SIREN*, *tranche\_etp*, *categorie\_juridique* et *categorie\_org*, n'ayant pas pu trouver de vocabulaire déjà existant qui corresponde parfaitement à ces types de données.

### III- Requêtes effectuées sur le dataset

Une fois les données formatées, il est possible d'effectuer un grand nombre de requêtes SPARQL dessus.

- Nombre d'établissements participants à la recherche, triés par type d'établissement:

```
SELECT ?categorie (count(?e) as ?nbE)
WHERE {
    ?e local:categorie_org ?categorie .
}
GROUP BY ?categorie
```

Résultat:

categorie	nombre
"Entreprise publique"	95
"Entreprise privée"	3011
"Micro-entreprise"	8738
"EPIC"	28
"Enseignement supérieur"	229
"EPST"	8
"Autre établissement de l'Etat"	282
"Institution sans but lucratif"	1786
"Secteur agricole"	111
"Petite ou moyenne entreprise"	4419
"Organisation internationale privée"	149
"Organisation internationale publique"	3
"Grande entreprise"	720
"Collectivité territoriale"	71
"Entreprise de taille intermédiaire"	1898



- Établissements participant à la recherche par ville:

```
SELECT ?ville (count(?e) as ?nbE)
WHERE {
    ?e dbprop:town ?ville .
}
GROUP BY ?ville
```

Résultat (partiel):

ville	nombre
"Famars"	1
"Eybens"	8
"Arc-sous-Cicon"	1
"Wettolsheim"	1
"Vouneuil-sur-Vienne"	1
"Le Chesnay"	10
"Bihorel"	1
"La Léchère"	1
"Darnétal"	1
"Béruuges"	1
"Phalempin"	2
"Martigné-Ferchaud"	1
"Veldhoven"	2
"Condé-sur-Vire"	2
"Bonneuil-sur-Marne"	2
"Peyrestortes"	1
"Gérardmer"	2
"Pomacle"	3

## IV- Liaison à un autre dataset

Dans le cadre de ce projet, nous lierons nos données avec le dataset [Principales institutions exécutant ou finançant la recherche \(hors établissements d'enseignement supérieur\)](#) du groupe Nathan SALAUN, Antoine MAGNIN, Martin LAVILLE. Leur dataset comprend moins d'établissements mais présente plus d'informations pour ces derniers, notamment leurs programmes de recherche.

## V- Ontologie OWL/RDFS et inférences

Étant donné que nous n'avons que très peu utilisé nos propres vocabulaires, il n'était pas nécessaire de recréer une ontologie complète mais seulement pour les vocabulaires que nous avons dû créer. On peut tout de même déterminer des inférences à partir des ontologies déjà existantes, par exemple:

```
<owl:class rdf:about="http://schema.org/Organization">
  <rdfs:label xml:lang="en">Organization</rdfs:label>
  <rdfs:subclassof rdf:resource="http://schema.org/Thing">
    <rdfs:comment xml:lang="en">An organization such as a school, NGO, corporation, club, etc.</rdfs:comment>
  </rdfs:subclassof>
</owl:class>

<owl:class rdf:about="http://schema.org/Thing">
  <rdfs:label xml:lang="en">Thing</rdfs:label>
  <rdfs:comment xml:lang="en">The most generic type of item. </rdfs:comment>
</owl:class>
```

Ici, on voit que les ressources de type *Organization* sont également de type *Thing* (une classe de base dans schema.org)

Remarque: Tous les établissements composant nos données sont des "Organizations", dans l'état actuel, leur type est défini dans une propriété. Il serait possible, lors de la génération des données en RDF d'attribuer dynamiquement la classe qui correspond le mieux (*EducationalOrganization* par exemple pour une Université). L'extension d'OpenRefine ne permet une configuration aussi poussée.



```

<!-- définition des propriétés -->
<owl:datatypeproperty rdf:about="http://localhost:3333/SIREN">
  <rdfs:label xml:lang="fr">SIREN</rdfs:label>
  <rdfs:comment xml:lang="fr">Numéro SIREN d'une organisation</rdfs:comment>
  <rdfs:domain rdf:resource="http://schema.org/Organization"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"></rdfs:range>
</owl:datatypeproperty>

<owl:datatypeproperty rdf:about="http://localhost:3333/tranche_etp">
  <rdfs:label xml:lang="fr">tranche_etp</rdfs:label>
  <rdfs:comment xml:lang="fr">Nombre de salariés d'une organisation (intervalle)</rdfs:comment>
  <rdfs:domain rdf:resource="http://schema.org/Organization"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"></rdfs:range>
</owl:datatypeproperty>

<owl:datatypeproperty rdf:about="http://localhost:3333/categorie_juridique">
  <rdfs:label xml:lang="fr">categorie_juridique</rdfs:label>
  <rdfs:comment xml:lang="fr">Catégorie juridique d'une organisation</rdfs:comment>
  <rdfs:domain rdf:resource="http://schema.org/Organization"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"></rdfs:range>
</owl:datatypeproperty>

<owl:datatypeproperty rdf:about="http://localhost:3333/categorie_org">
  <rdfs:label xml:lang="fr">categorie_org</rdfs:label>
  <rdfs:comment xml:lang="fr">Type d'organisation</rdfs:comment>
  <rdfs:domain rdf:resource="http://schema.org/Organization"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"></rdfs:range>
</owl:datatypeproperty>

```

### *Définition de nos propriétés*

Les propriétés que nous avons définies ne s'appliquent que sur des classes de type *Organization*, et n'ont pour objet que des littéraux.

Voir [schema.owl](#) (GitHub)

## VI- Liaison des données avec le cloud linked data

Un grand nombre des établissements que nous avons dans notre dataset sont probablement déjà présents dans d'autres datasets du *cloud linked data*, de ce fait, il serait parfaitement envisageable de les lier par l'intermédiaire de leur code SIREN (en trouvant une propriété correspondante) ou simplement de leur raison sociale. On pourrait ainsi facilement exécuter des requêtes fédérées pour avoir plus d'informations sur les établissements. À titre d'exemple, l'Institut National des Sciences Appliquées de Rennes existe sur wikidata à l'adresse <https://www.wikidata.org/wiki/Q1934614>.

## VII- Métadonnées VOID

```
:Dataset a void:Dataset;  
  foaf:homepage <http://localhost/>;  
  dcterms:title "Etablissements publics et privés impliqués dans la recherche et le développement";  
  dcterms:description "Liste des établissements publics et privés impliqués dans la recherche et le développement en France";  
  dcterms:contributor "Quentin Mazoua";  
  dcterms:contributor "Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche";  
  dcterms:source <https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-etablissements-publics-prives-impliques-recherche-developpement/>;  
  dcterms:modified "2017-11-29"^^xsd:date;  
  dcterms:license <https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf>  
.
```

Le vocabulaire Void permet de définir des métadonnées pour décrire le dataset et ses contributeurs.

## Liens

### Dataset original:

<https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-etablissements-publics-prives-impliques-recherche-developpement/>

### Slides de présentation des étapes 1 & 2:

[https://docs.google.com/presentation/d/1m\\_3GwvpfPmbcvHaslXAasyFYXKiP3soOkeHqFsl6KrE/edit?usp=drivesdk](https://docs.google.com/presentation/d/1m_3GwvpfPmbcvHaslXAasyFYXKiP3soOkeHqFsl6KrE/edit?usp=drivesdk)

Repo GitHub du projet: [https://github.com/quentinmazoua/web\\_semantique\\_m1](https://github.com/quentinmazoua/web_semantique_m1)