Intelligence Artificielle

un outil au service de l'investigation

Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, Madhulika Mohanty

Inria Saclay & Institut Polytechnique de Paris

Qui sommes nous ?

Nous faisons partie de l'**équipe de recherche CEDAR** (Rich Data Analytics at Cloud scale) à **Inria Saclay**.

Nelly Barret
Doctorante
Inria & IPP

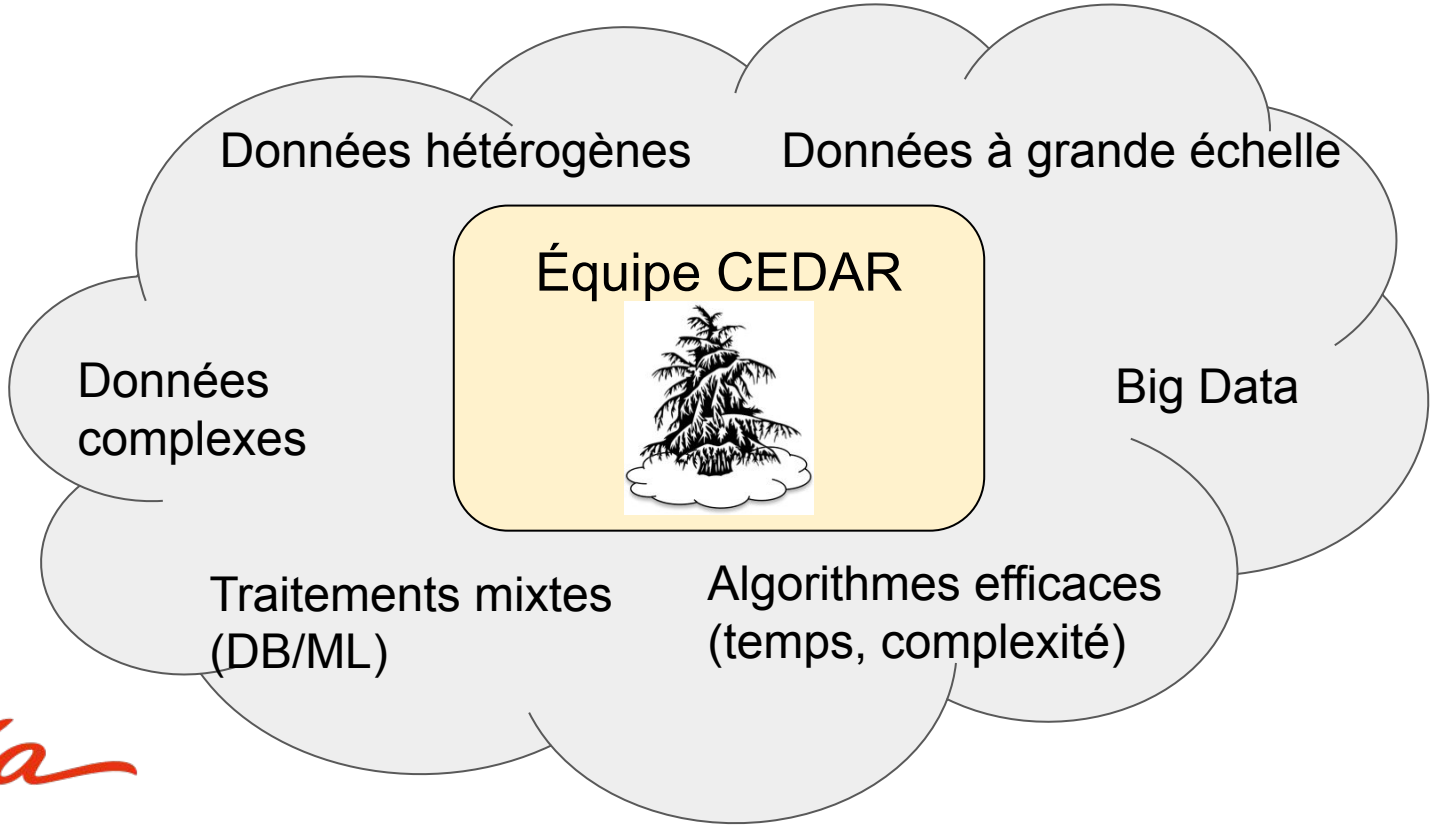
Simon Ebel
Ingénieur recherche
Inria

Théo Galizzi
Ingénieur recherche
Inria

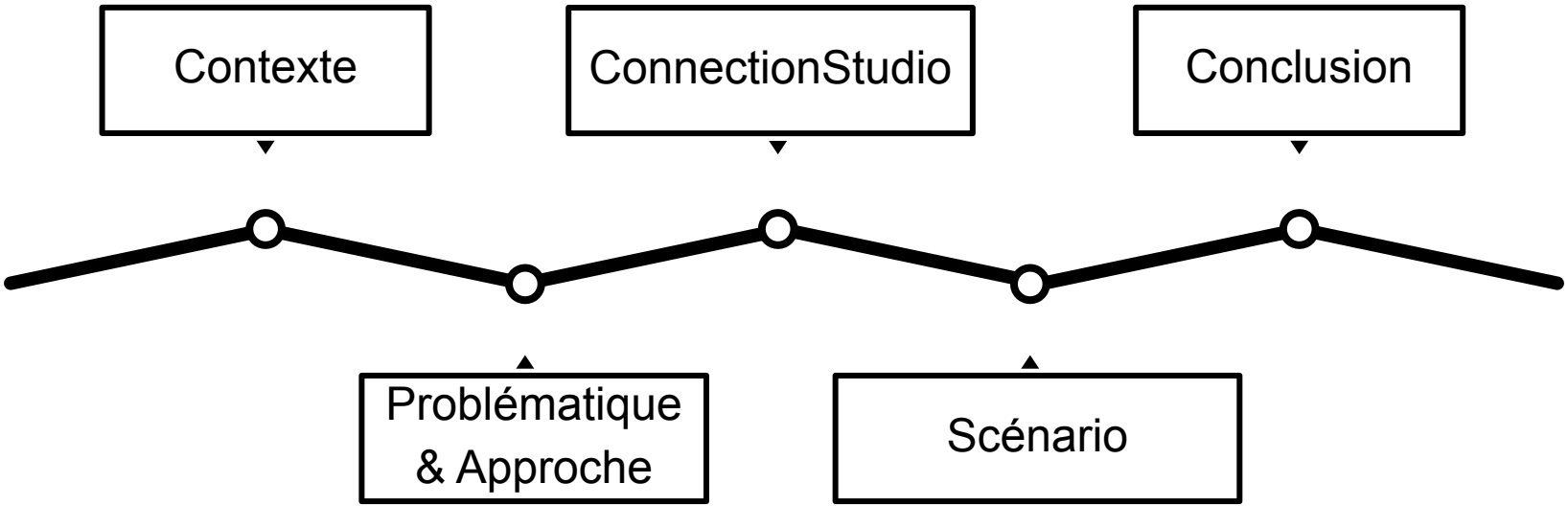
Ioana Manolescu
Chercheure et chef d'équipe
Inria & IPP

Madhulika Mohanty
Post-doctorante
Inria

L’équipe CEDAR



Organisation de l’atelier



Contexte

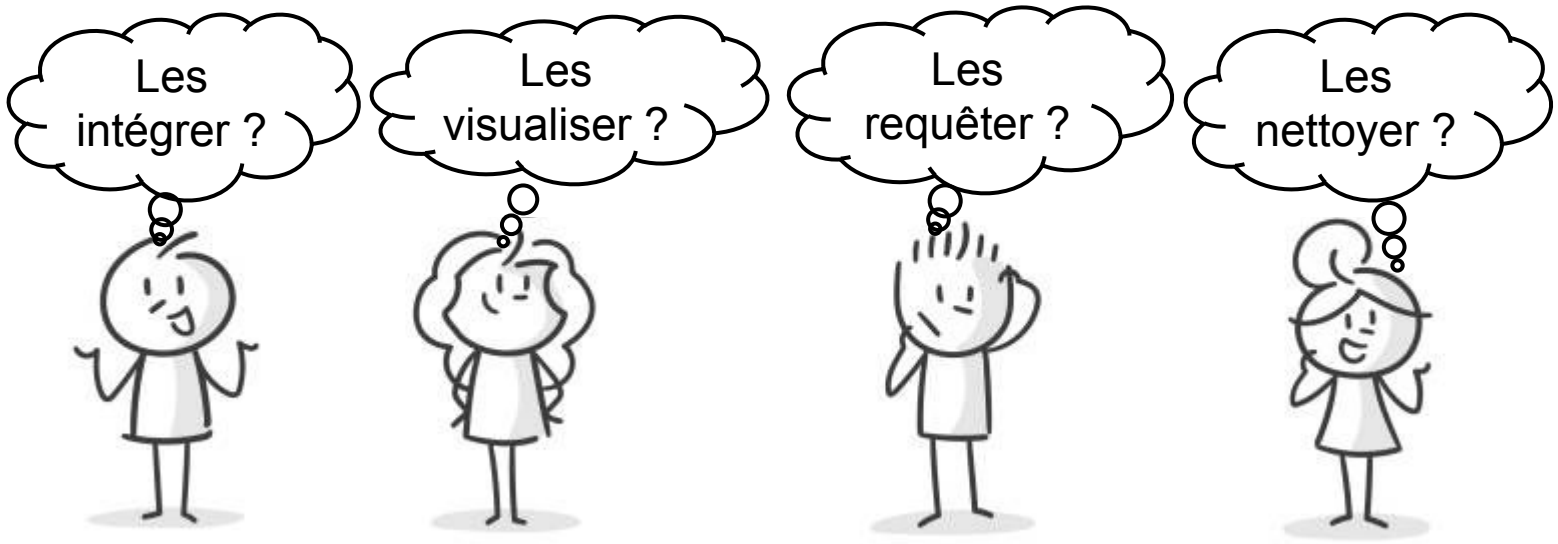
- Les “faits” proviennent de **différentes sources** de données, souvent **larges et complexes**
- Les sources sont de **formats très variés** :

Structurées	**Semi-structurées**	**Non-structurées**
- Bases de données - Tables (Excel, CSV)	- JSON (Web) - XML (Web) - HTML (Web)	- Texte - PDF - RDF (*Open Data*) - Neo4j (*Graphe*)

- Chaque source vient d’un *producteur* → **pas d’uniformisation ni de schéma**
- Les données ne sont pas produites en pensant à l'**utilisateur final**

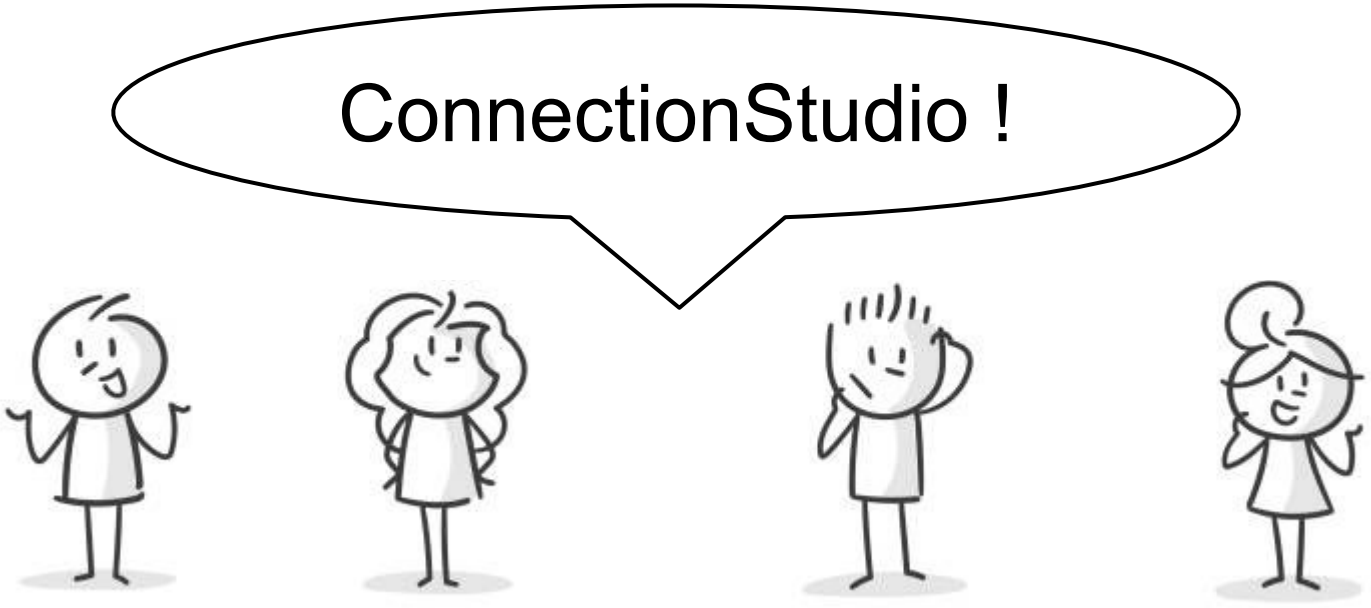
Problématique & approche

Comment faciliter l'investigation de sources très hétérogènes ?



Problématique & approche

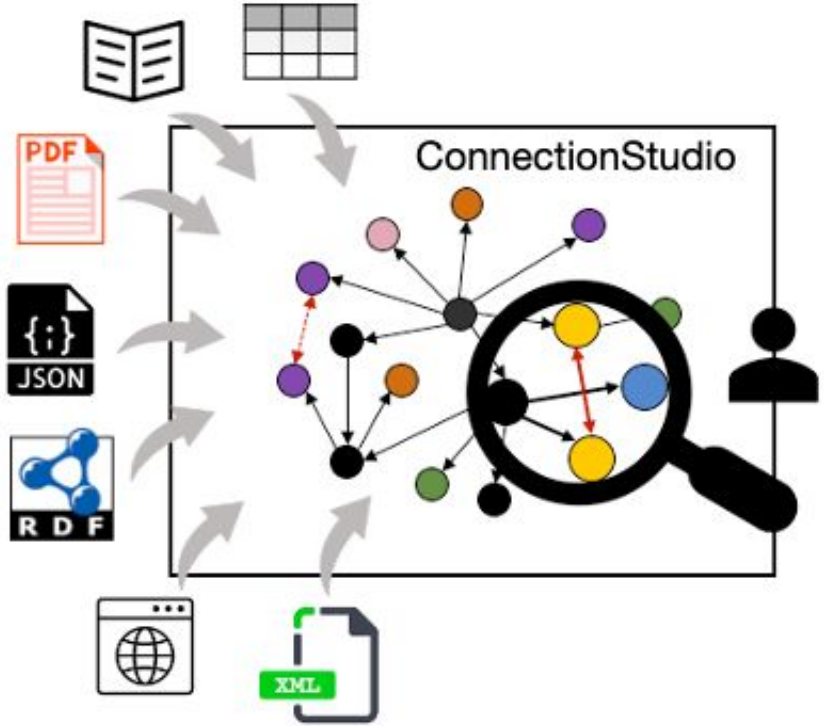
Comment faciliter l'investigation de sources très hétérogènes ?



ConnectionStudio

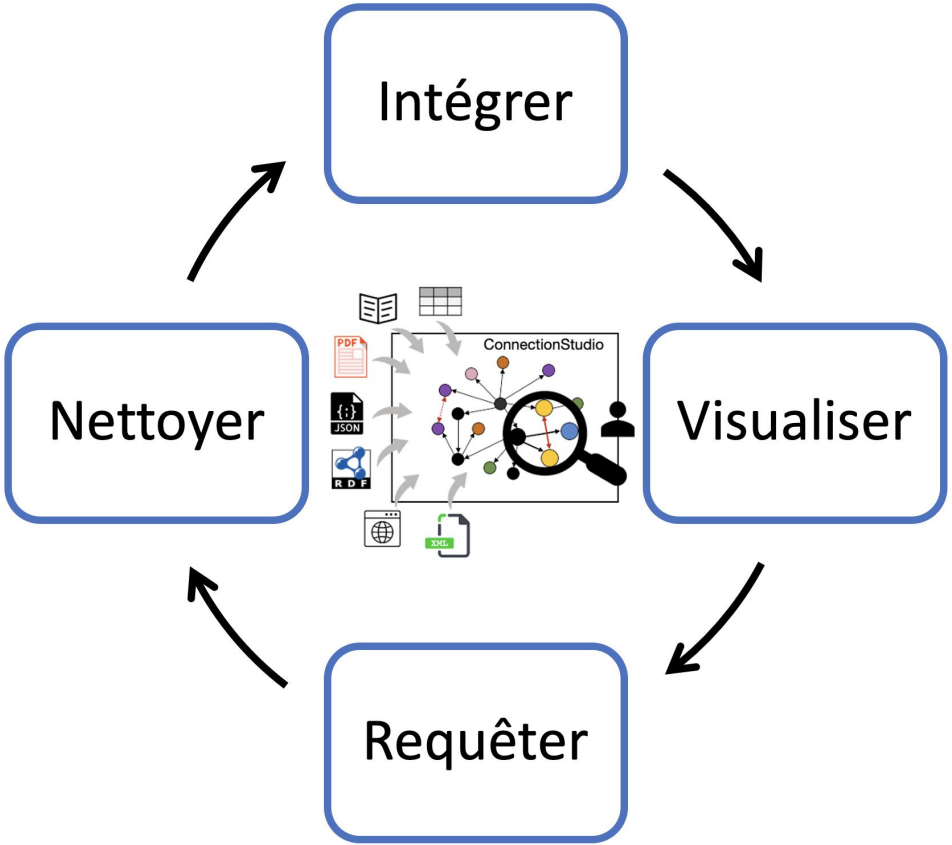
ConnectionStudio est un outil qui :

- Intègre des données hétérogènes de manière unifiée
- Identifie les entités : lieux, personnes, entreprises, …
- Permet de visualiser et interagir avec les données

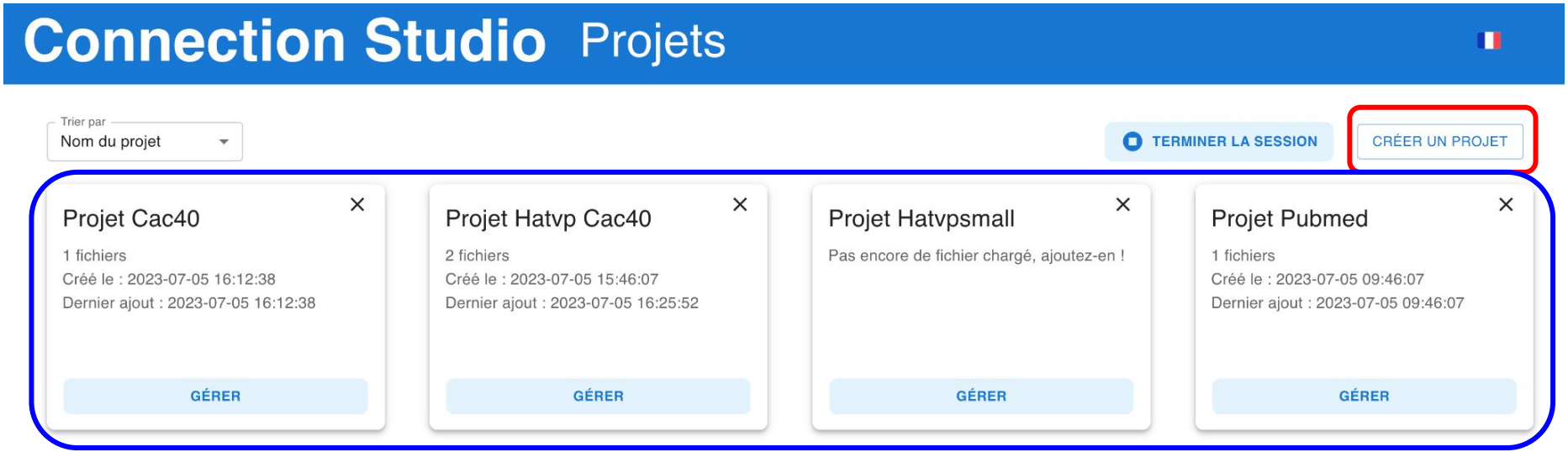


Problématique & approche

Comment faciliter l'investigation de sources très hétérogènes ?



ConnectionStudio



Les projets déjà créés

Un projet = tous les fichiers concernant la même thématique.

Introduction aux données : CAC40

CAC40 : 40 sociétés, de différentes branches d'activités, qui reflètent la tendance globale de l'économie des grandes entreprises françaises.

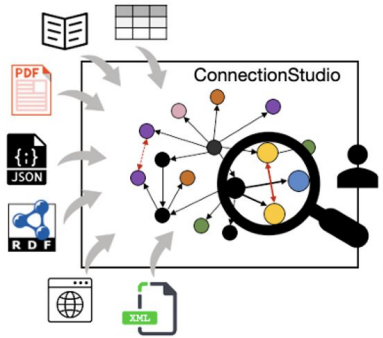
	A	B	C
1	**numero**	**nom_entreprise**	**description_wikipedia**
2	1	AIR LIQUIDE	Air liquide (AL), anciennement L'Air liquide, est un groupe industriel français d'envergure internationale, spécialiste des gaz industriels, c'est-à-dire des gaz pour l'industrie, la santé, l'environnement et la recherche.
3	2	AIRBUS	Airbus est un constructeur d'avion franco-allemand et une coopération industrielle internationale présente dans le secteur aéronautique et spatial civil et militaire. Ses activités les plus importantes sont la construction d'avions de ligne, d'hélicoptères et d'avions militaires. À travers la participation à diverses entreprises, le groupe est engagé dans les lanceurs spatiaux, les satellites artificiels, les missiles et les avions de combat.
4	3	ALSTOM	Alstom est une multinationale française, aujourd'hui spécialisée dans le secteur des transports, principalement ferroviaires (trains, tramways et métros). Alstom faisait partie du groupe Alcatel-Alsthom, nouveau nom de

Intégrer

Visualiser

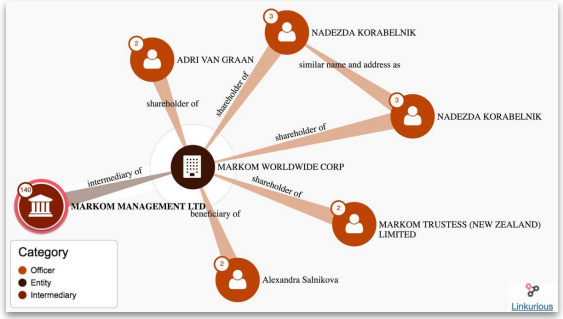
Requêter

Nettoyer

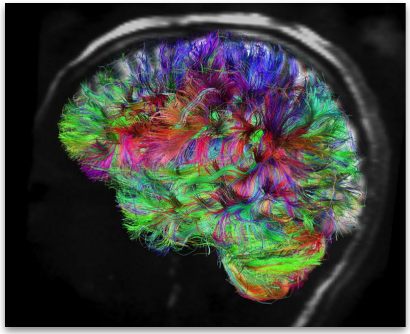


Graphe

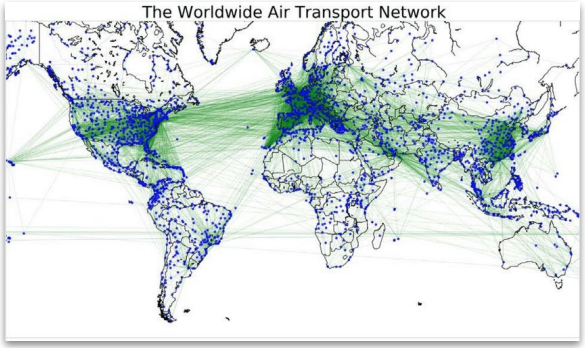
- Le **paradigme “graphe”** décrit :
  - Des objets (*noeuds*)
  - Connectés par des liens (*arêtes*)
- Beaucoup utilisé car très **flexible**



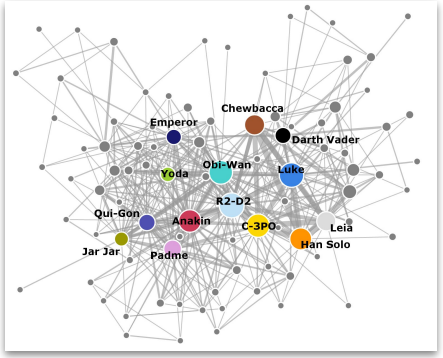
Panama papers (données ICIJ)



Neurones du cerveau humain



Vols internationaux



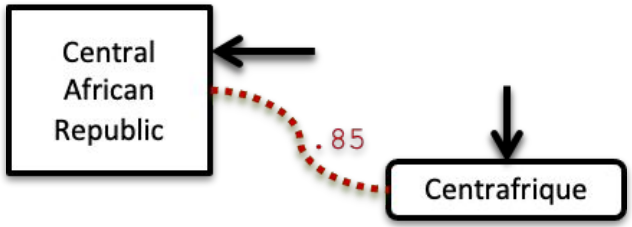
Personnages dans “Star Wars”

Extraction d’entités

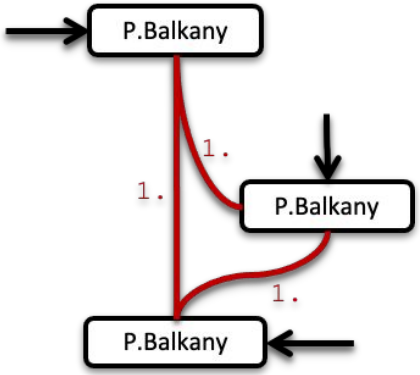
- Extraction d’entités simples :
  - Email, date, site Web, …
- Extraction d’entité nommées
  - Expression linguistique référentielle: personnes, lieux, organisations, …
  - Complexe à identifier → fait par des modèles de langage

Annotation sémantique

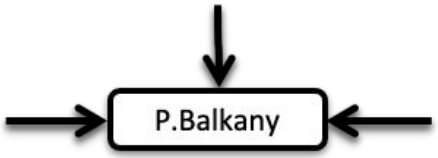
- Connexion entre les entités nommées quand elles sont *similaires*
- Plusieurs similarités : exactes, approximatives
- Unification possible quand similarité totale → - de noeuds et + de connexions



Similarité approximative



Similarité exacte



Fusion des entités identiques

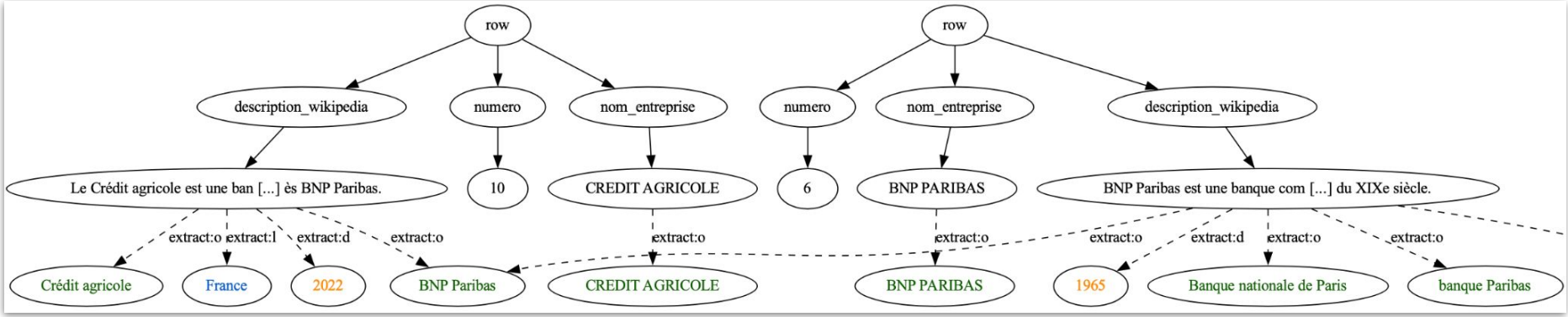
Chargement des sources de données

- Les donnés sont modélisées selon le paradigme “graphe”:
  - Un **noeud** est une **petite portion d’information**
  - Les **liens** connectent les noeuds pour reconstituer **la sémantique logique** du fichier

- L’extraction d’entités nommées est appliqué sur chaque noeud valeur
  - 2 modèles linguistiques : celui de Stanford et celui de FLAIR
  - 2 langues : français ou anglais

- Les noeuds qui contiennent des informations identiques sont fusionnés → connexions entre les sources

Introduction aux données : CAC40



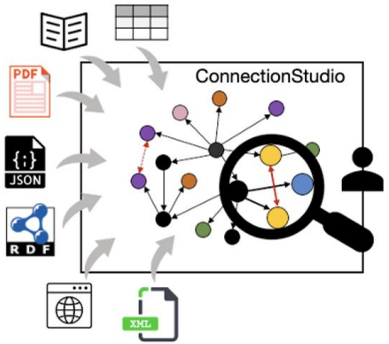
Graphe de données correspondant au CSV du CAC40

Intégrer

Visualiser

Requêter

Nettoyer



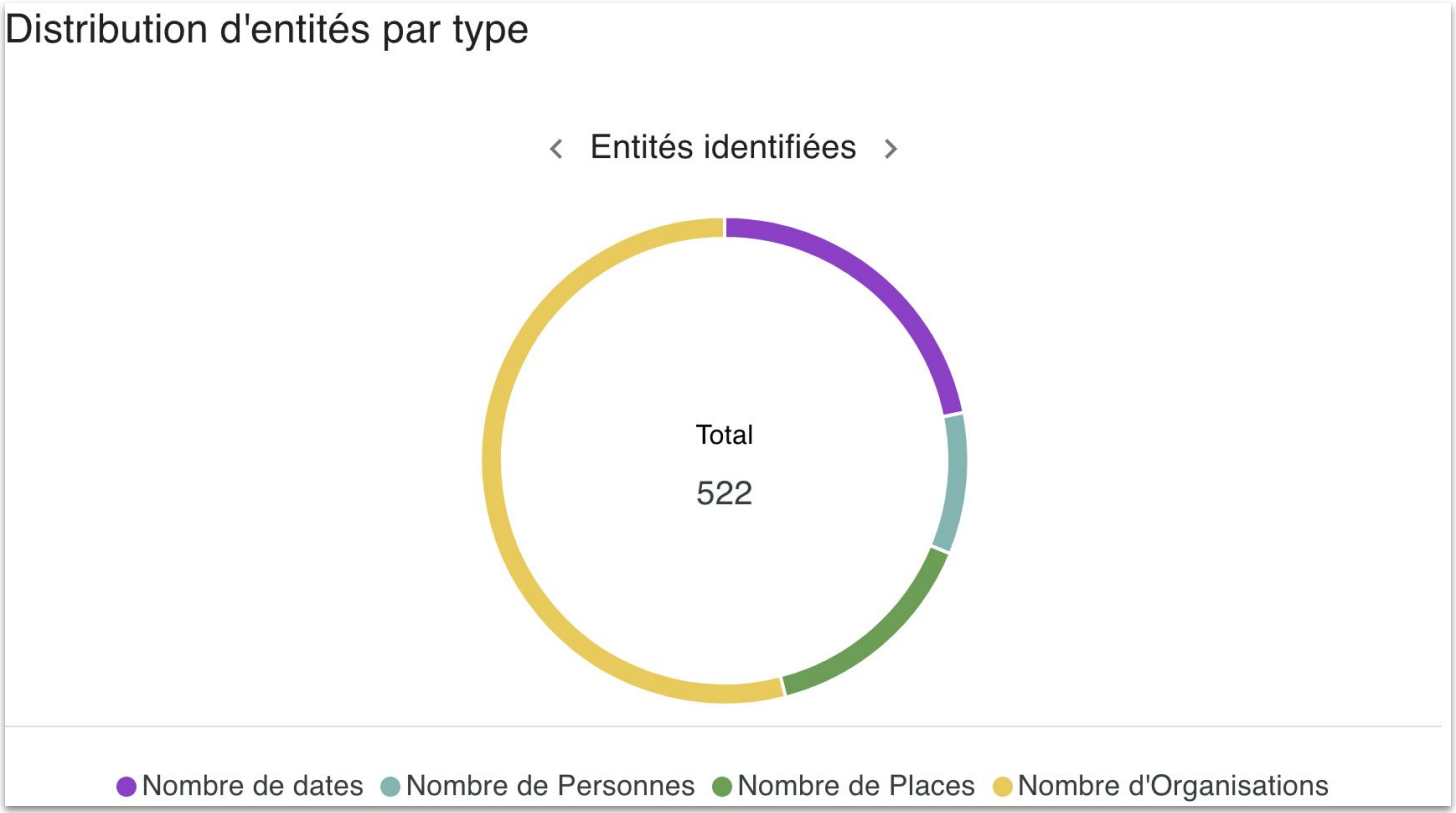
Visualisation 1 : statistiques

Comment avoir une bonne idée des entités détectées ?

- Distribution d'entités par types
- Nuage des 100 entités les plus fréquentes (*tag cloud*)
- Distribution des types d'entités par fichier
- Liste des 100 entités les plus partagées parmi le plus de sources

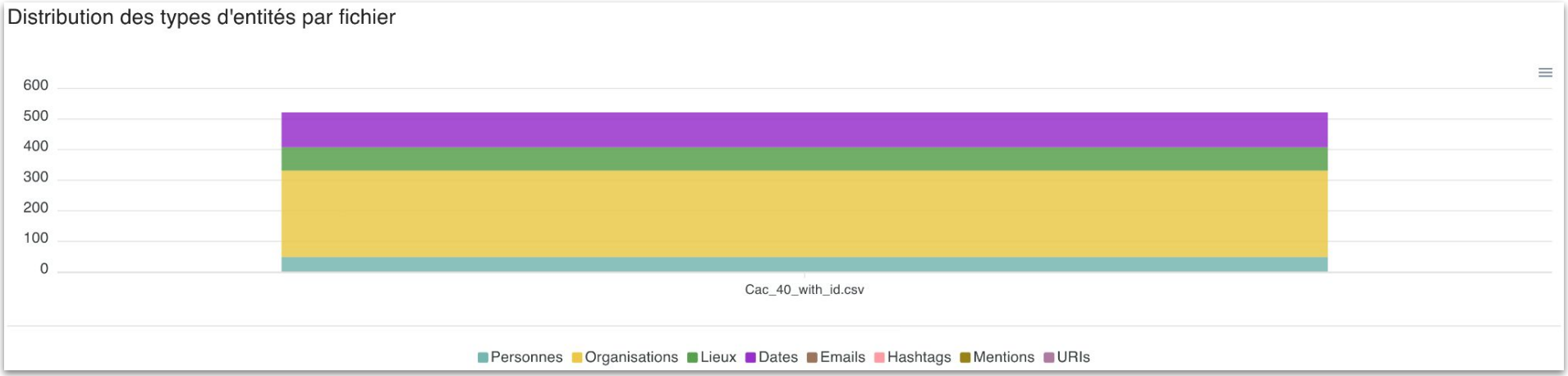
- Tous les graphiques/tableaux sont exportables

Visualisation 1 : statistiques

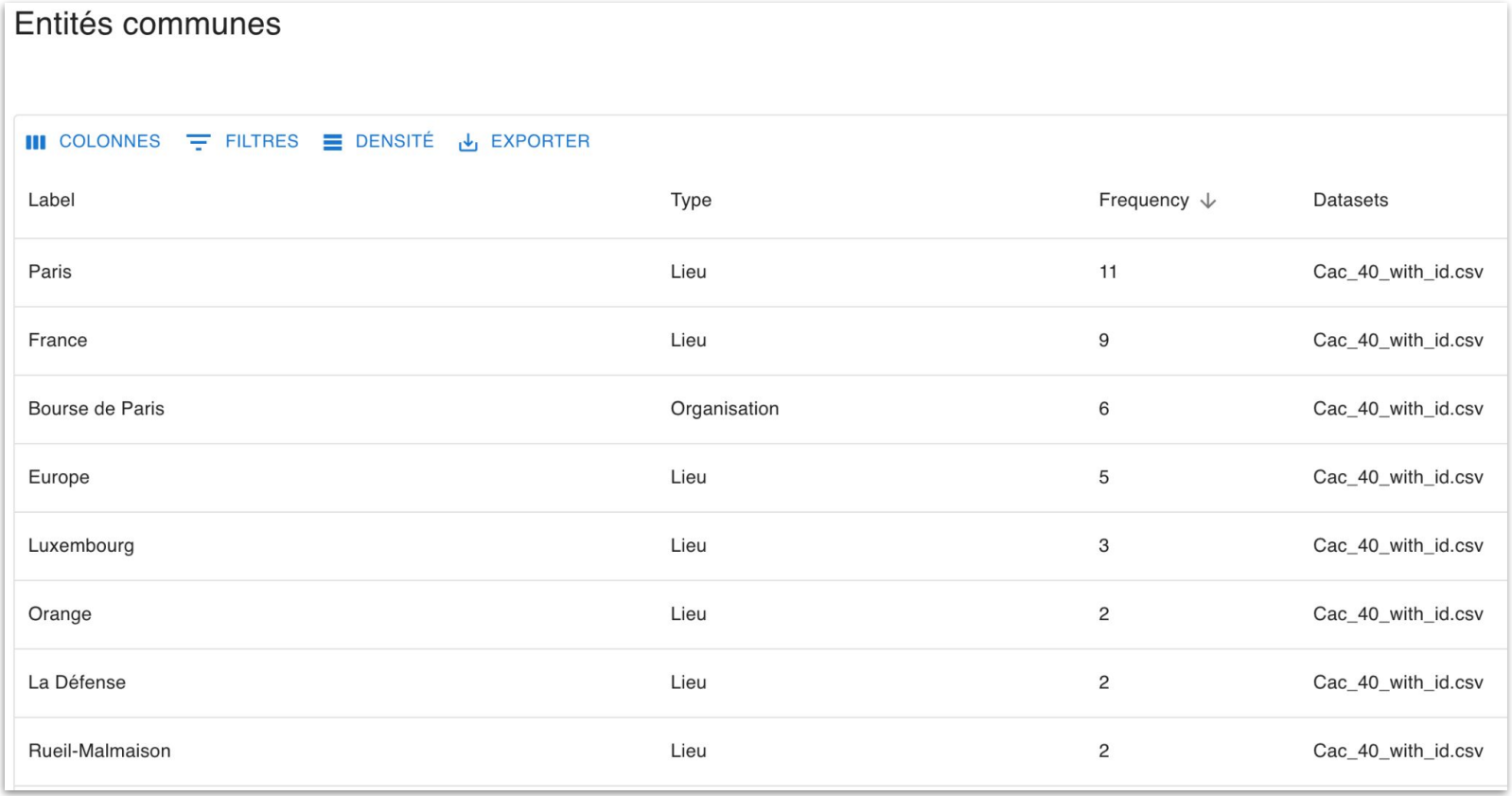


Visualisation 1 : statistiques

Visualisation 1 : statistiques



Visualisation 1 : statistiques



Entités communes

COLONNES FILTRES DENSITÉ EXPORTER

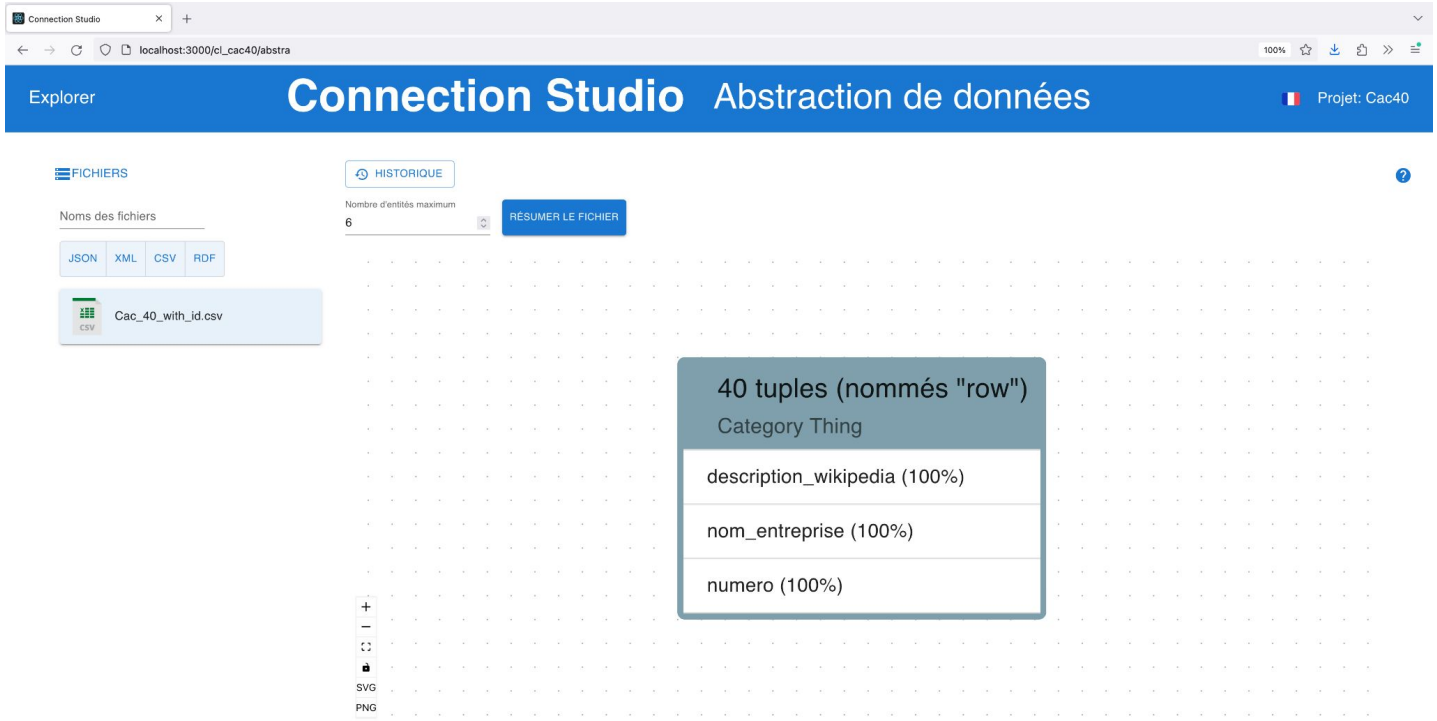
Label	Type	Frequency ↓	Datasets
Paris	Lieu	11	Cac_40_with_id.csv
France	Lieu	9	Cac_40_with_id.csv
Bourse de Paris	Organisation	6	Cac_40_with_id.csv
Europe	Lieu	5	Cac_40_with_id.csv
Luxembourg	Lieu	3	Cac_40_with_id.csv
Orange	Lieu	2	Cac_40_with_id.csv
La Défense	Lieu	2	Cac_40_with_id.csv
Rueil-Malmaison	Lieu	2	Cac_40_with_id.csv

Visualisation 2 : structure d’un fichier

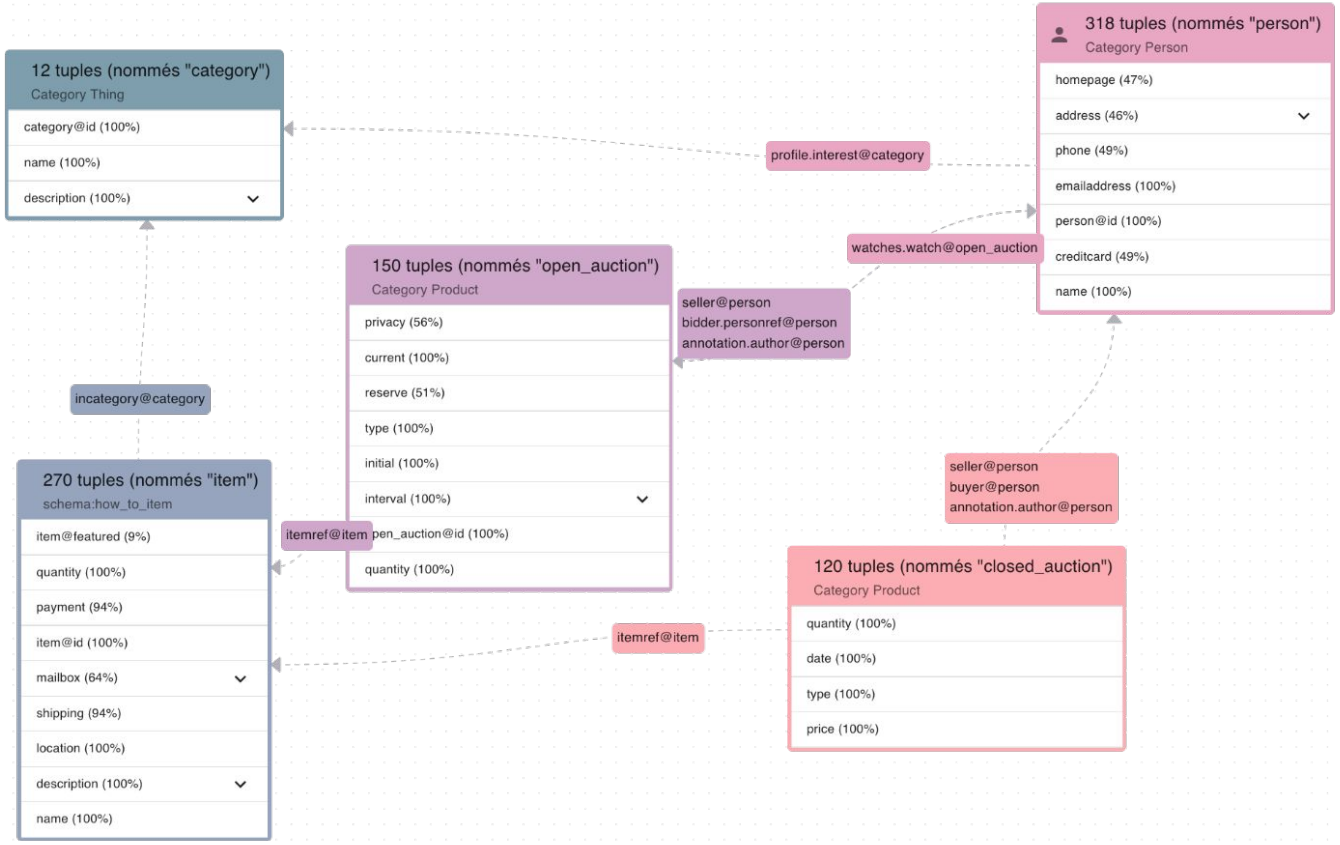
Comment avoir une bonne idée de la structure d’un fichier ?

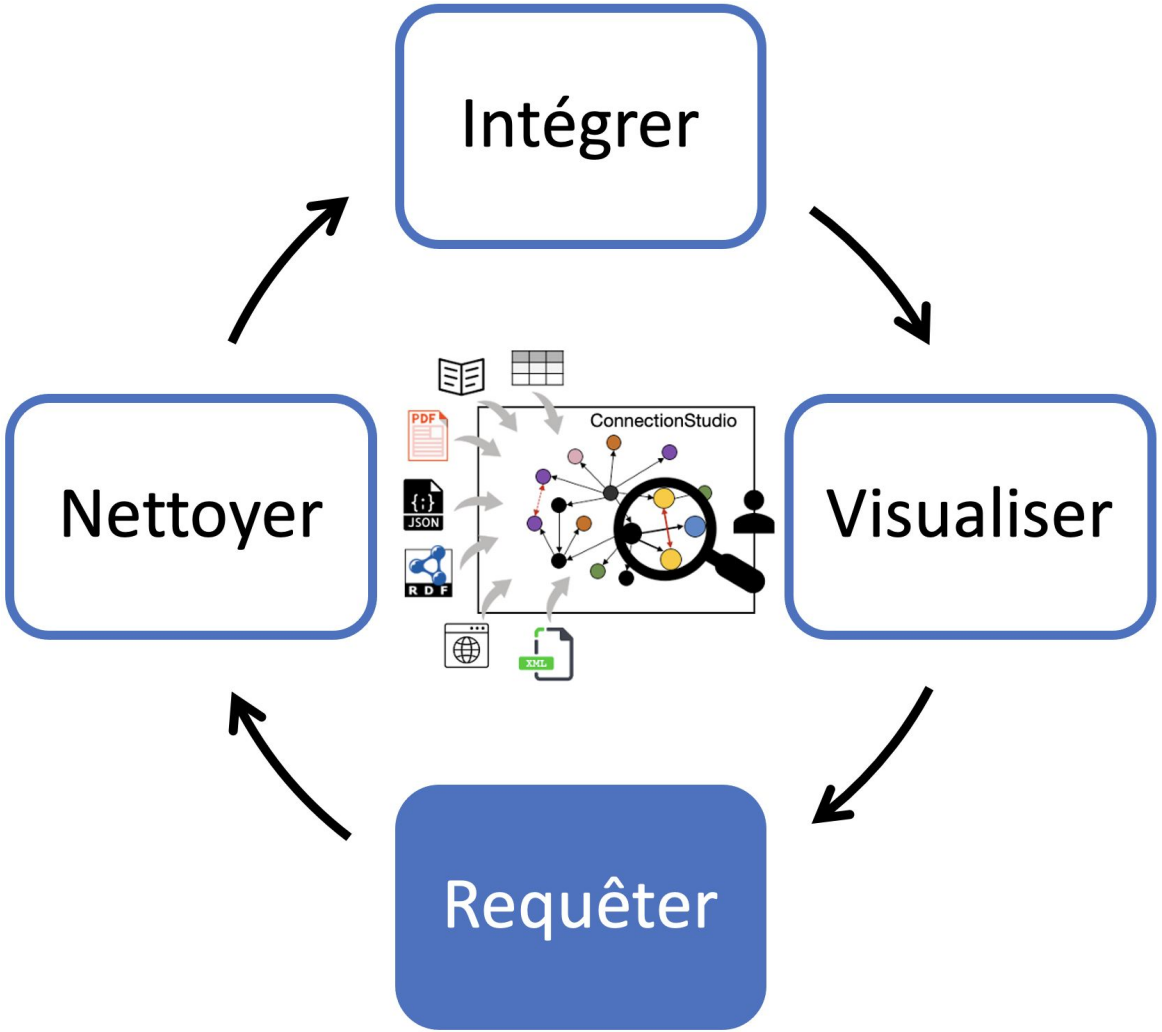
- Une source de données peut être vue comme :
  - Un ou plusieurs ensembles d’entités similaires
  - Un ensemble de relations qui les connectent
- Résultat : une **description orientée utilisateur**

Visualisation 2 : structure d’un fichier



Visualisation 2 : structure d’un fichier




Intégrer
Visualiser
Requêter
Nettoyer
ConnectionStudio
PDF
JSON
RDF
XML

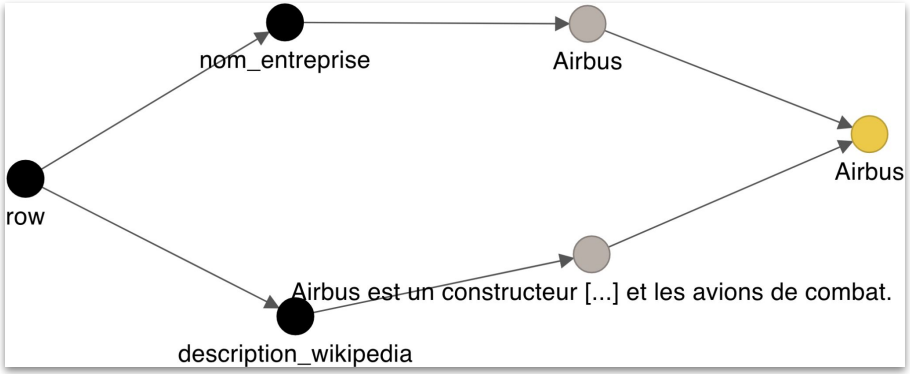
Requêtes

- **Requête** : interrogation des données pour en récupérer un sous-ensemble
- 2 moyens :

	Par **mot(s)-clé(s)**	Avec le **langage SQL**
Requiert	un ou plusieurs mots clés	une requête SQL
Donne	parties du graphe qui contiennent les mots clés	tableau de résultats
Adapté quand/pour	on connaît les mots-clés qui nous intéressent	explorer de grandes données

Requêtes

	Par **mot(s)-clé(s)**	Avec le **langage SQL**
Requiert	un ou plusieurs mots clés	une requête SQL
Donne	parties du graphe qui contiennent les mots clés	tableau de résultats
Adapté quand/pour	on connaît les mots-clés qui nous intéressent	explorer de grandes données



Requête “Airbus” sur le graphe de données

```
cl_cac40=# SELECT id, label
FROM nodes
WHERE label LIKE 'Airbus%';
```

```
 id  |             label
-----+-------------------------------
  20 | Airbus
  24 | Airbus est un constructeur...
 628 | Airbus
(3 rows)
```

Requête “Airbus” en SQL

Requête 1 : Connexions entre entités

Comment trouver comment sont connectées les entités ?

Identifie toutes les connexions entre deux types d’entités :

- Personnes ↔ Personnes
- Personnes ↔ Lieux
- Entreprises ↔ Personnes
- …

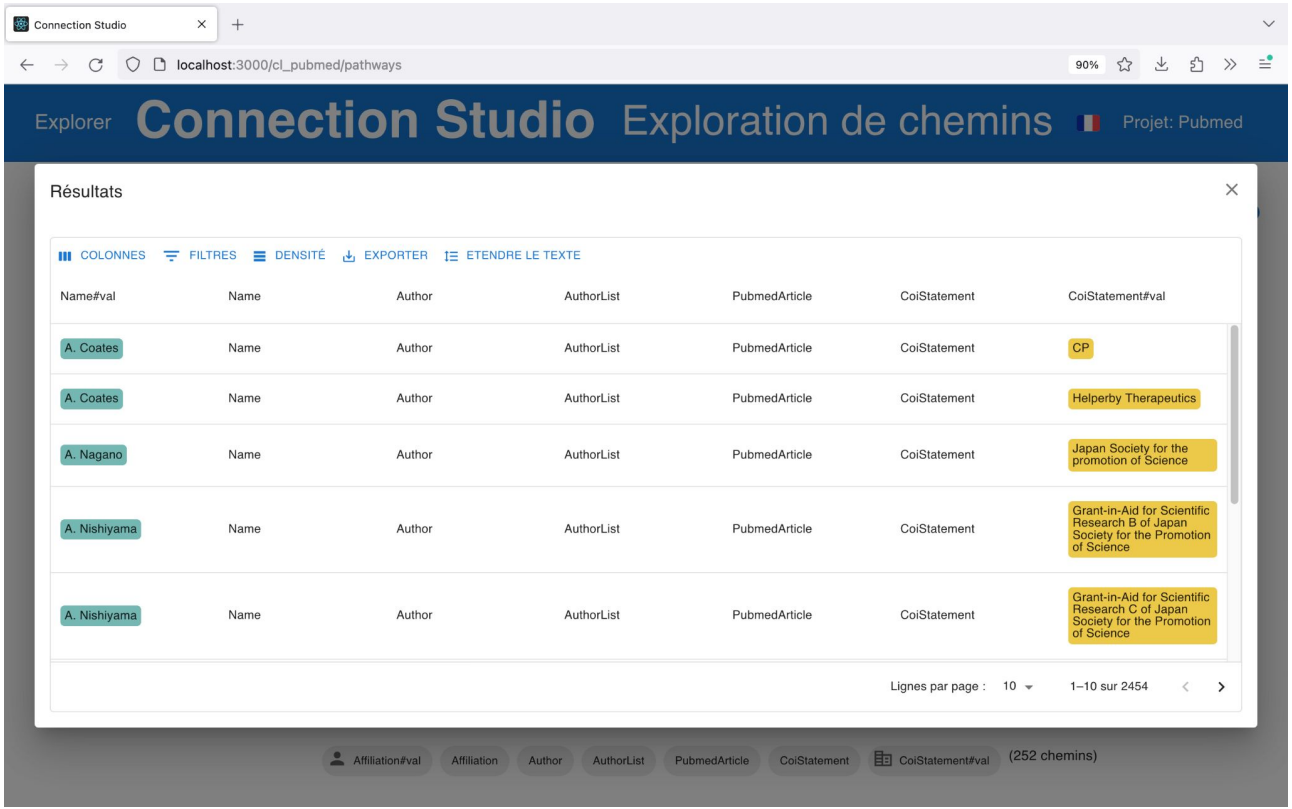
Requête 1 : Connexions entre entités

Résultats

COLONNES | 1 FILTRES | DENSITÉ | EXPORTER | ETENDRE LE TEXTE

nom_entreprise#val	nom_entreprise	row	description_wikipedia	description_wikipedia#val
ArcelorMittal	nom_entreprise	row	description_wikipedia	Europe
ArcelorMittal	nom_entreprise	row	description_wikipedia	Luxembourg
Engie	nom_entreprise	row	description_wikipedia	Bruxelles
Engie	nom_entreprise	row	description_wikipedia	Luxembourg
Engie	nom_entreprise	row	description_wikipedia	Paris
Eurofins	nom_entreprise	row	description_wikipedia	Luxembourg
Eurofins	nom_entreprise	row	description_wikipedia	Nantes

Conflits d’intérêts dans le bio-médical



Requête 2 : recherche par mots-clés

Comment explorer les données via des mots-clés ?

- Chaque mot-clé → un noeud dans le graphe de données
- Résultats → connexions entre ces noeuds
- Paramètres

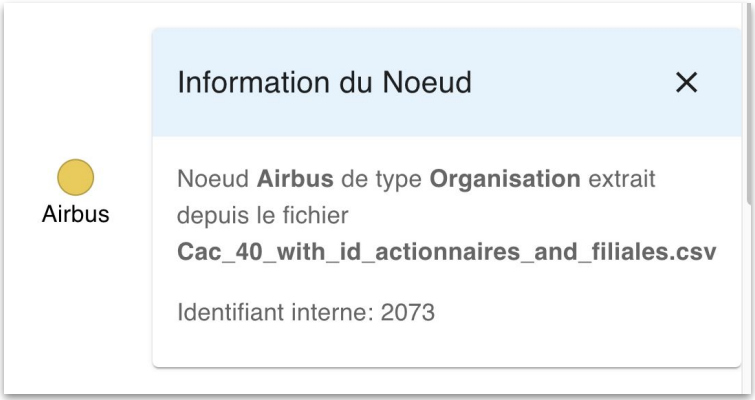
Adapté pour :

- Chercher certains mots-clés
- Explorer le graphe sans en connaître la structure

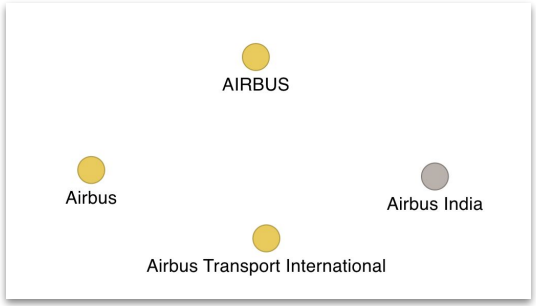
Requête 2 : recherche par mots-clés

Quoi ?	Comment ?	Résultat
Exactement un mot	`exact:Airbus`	Un noeud exactement labellisé **`Airbus`**
Un mot	`Airbus / "Airbus"`	Un noeud qui contient : - **`Airbus`** (non-sensible à la casse) - **`"Airbus"`** (sensible)
Exploration par mot-clé	`explore Capgemini`	Les noeuds labellisés **`Capgemini`** ainsi que leurs voisins
Exactement plusieurs mots	`exact:CREDIT+AGRICOLE`	Les noeuds labellisés exactement et uniquement **`CREDIT AGRICOLE`**
Plusieurs mots	`"Airbus Engie"`	Les noeuds qui contiennent **`Airbus`** et **`Engie`** (dans n'importe quel ordre)
Exploration par type	`type:Person`	Tous les noeuds de type Personne

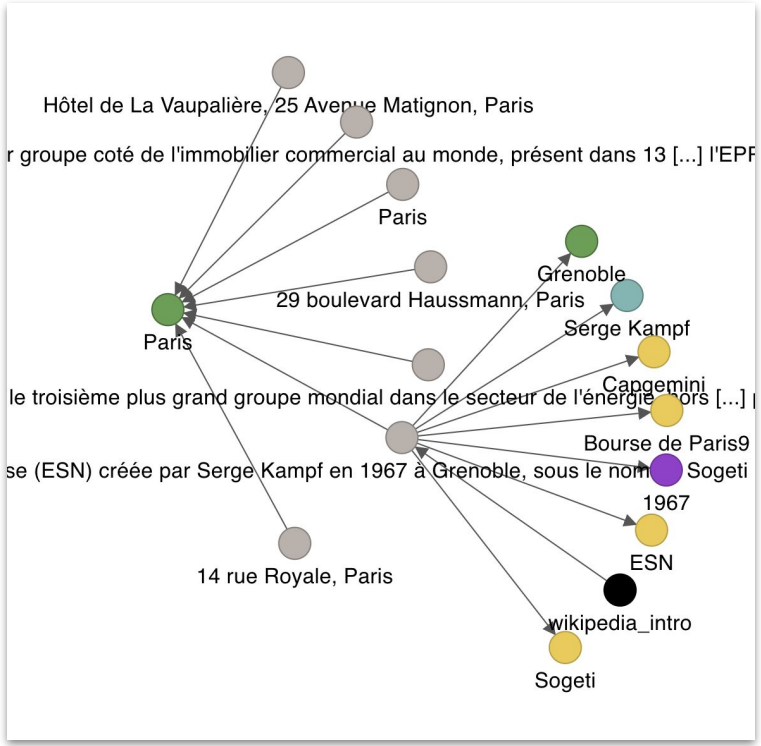
Requête 2 : recherche par mots-clés



Un noeud et ses informations

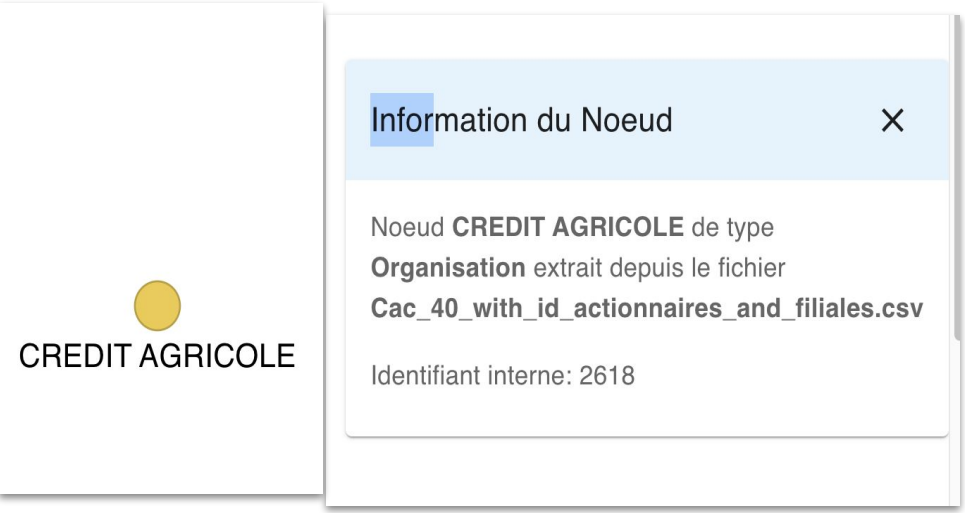


Les noeuds “Airbus”



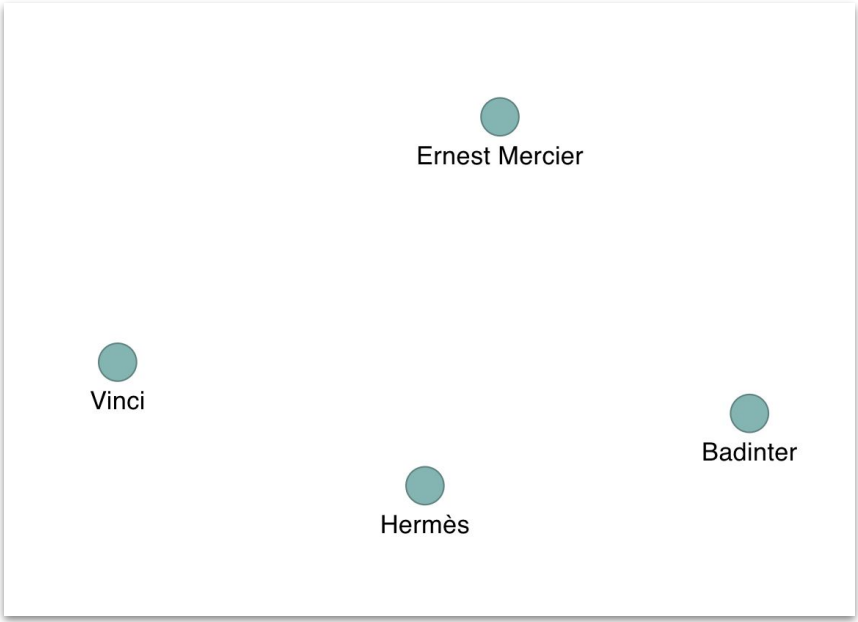
Exploration manuelle

Requête 2 : recherche par mots-clés



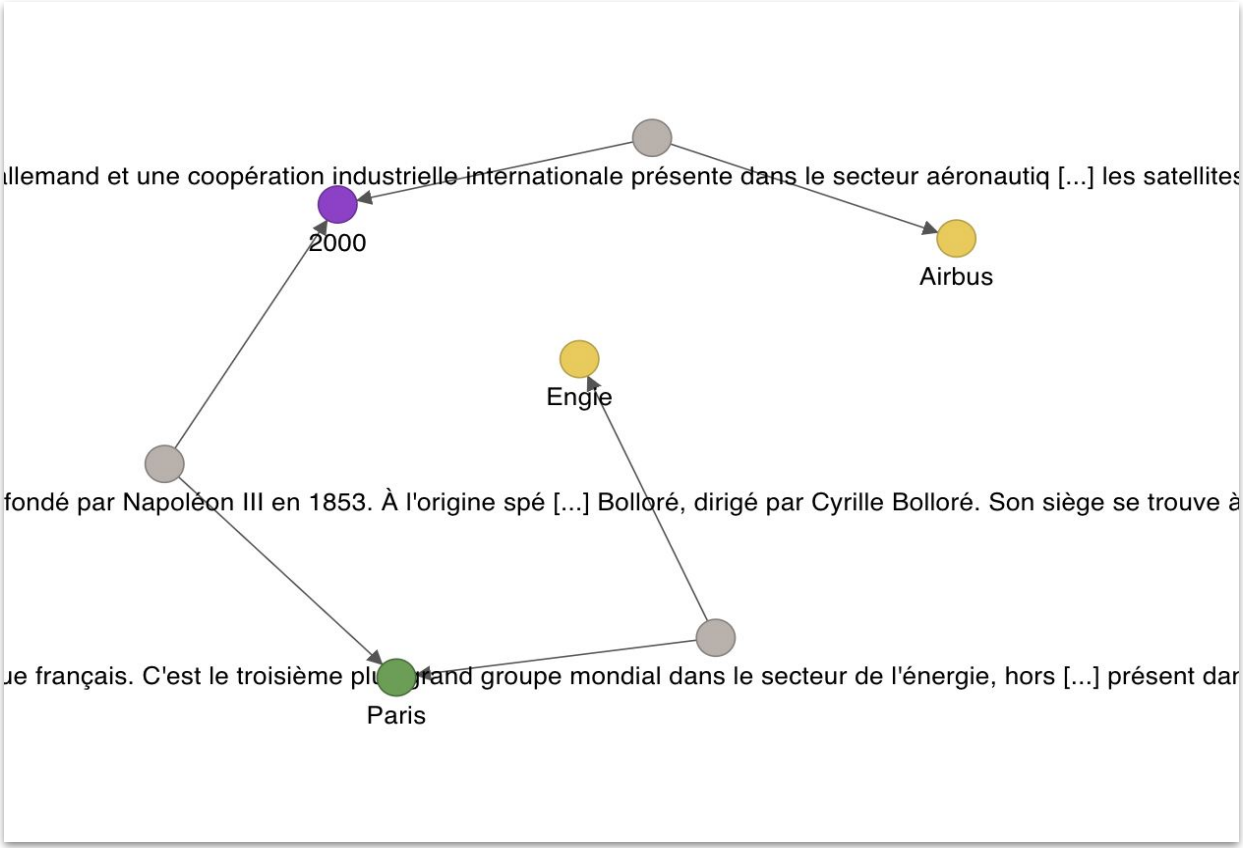
Un noeud et ses informations

Des noeuds de type "Personne"



Requête 2 : recherche par mots-clés

Connexions entre les noeuds labellisés "Airbus" et "Engie" respectivement



Requête 3 : Requêtes utilisateur

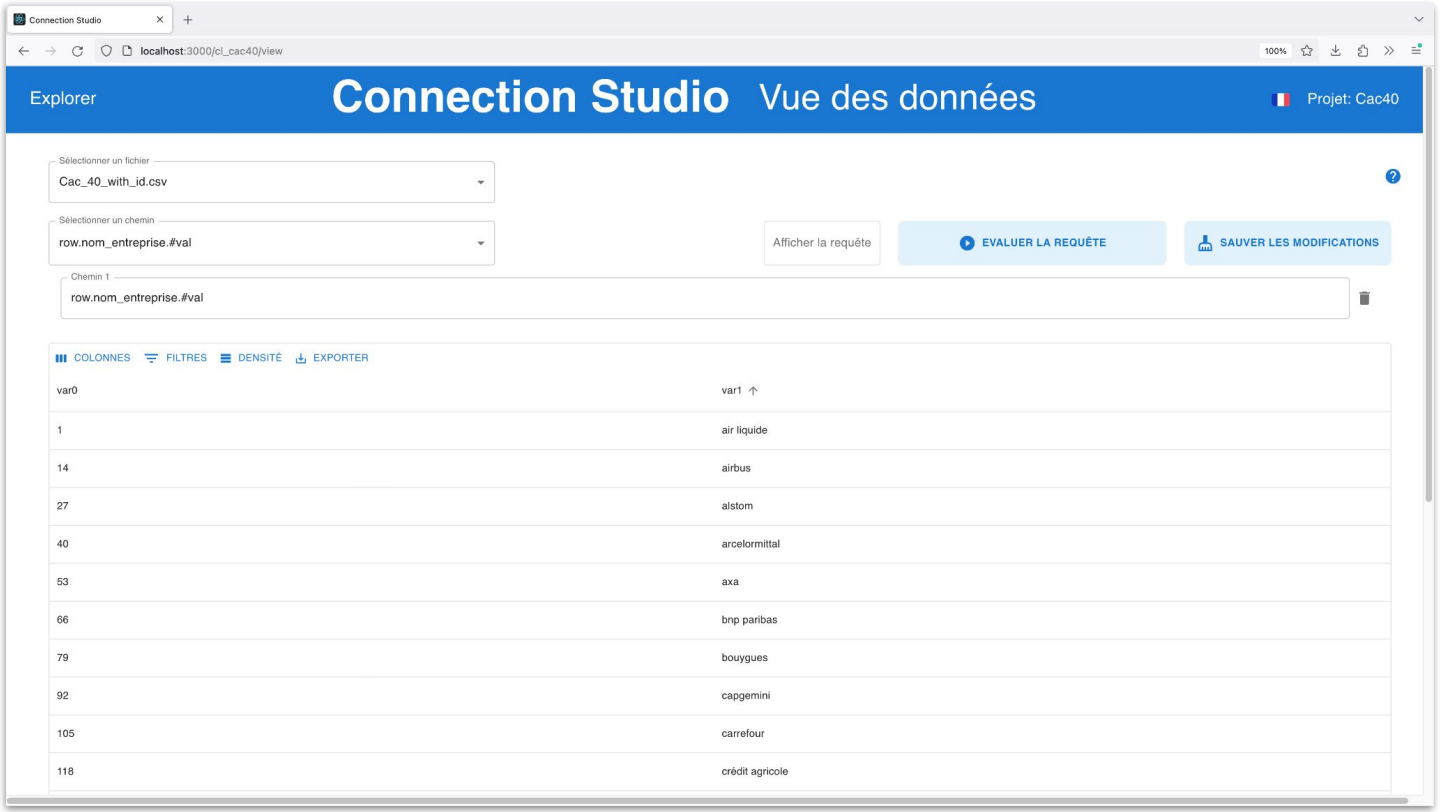
Comment requêter les données de manière personnalisée ?

- **Explorer la base de données** sans compétences en **SQL**
  - SQL : langage utilisé pour interagir avec une base de données
  - Syntaxe spécifique à apprendre → bloquant pour les utilisateurs
- Liste de fragments de données (↔ `SELECT` de SQL)
- Interconnectables entre eux de manière personnalisée (↔ `JOIN` de SQL)

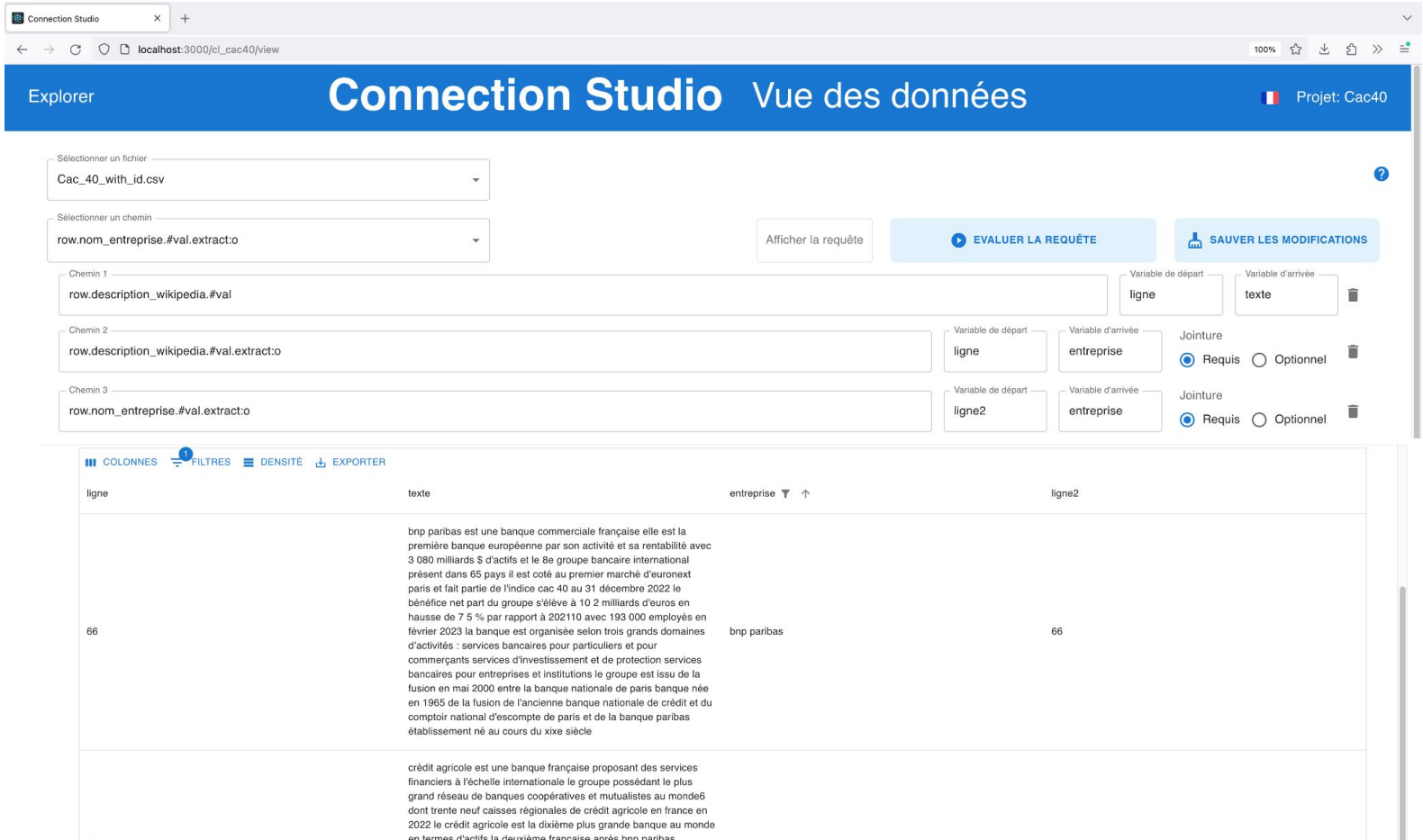
```
SELECT magasin, pays,
COUNT(vente)
FROM magasins, ventes
WHERE date='12/07/2023'
GROUP BY magasin, pays
HAVING COUNT(vente) > 1000€
```

Exemple de requête SQL

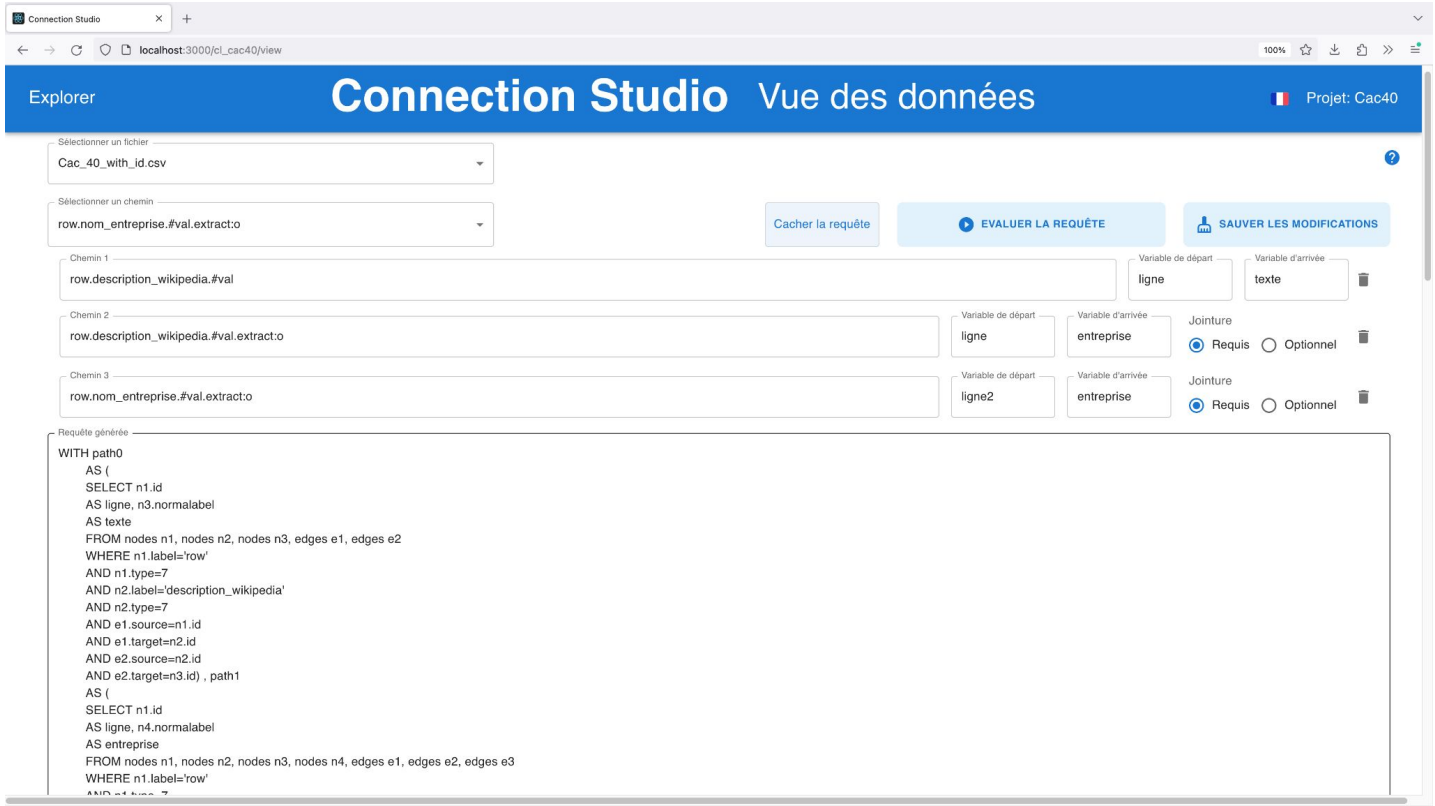
Requête 3 : Requêtes utilisateur



Requête 3 : Requêtes utilisateur



Requête 3 : Requêtes utilisateur

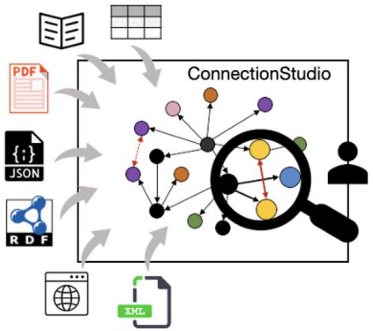


Intégrer

Visualiser

Requêter

Nettoyer

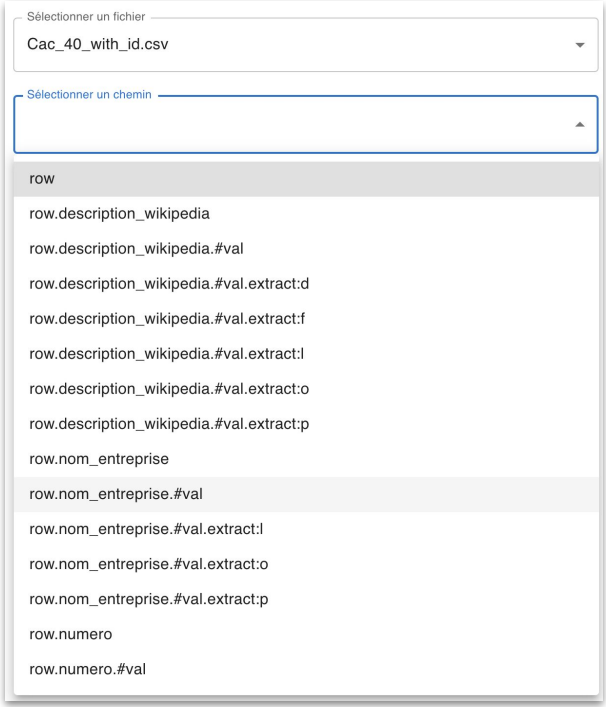


Nettoyage : politiques d’extraction

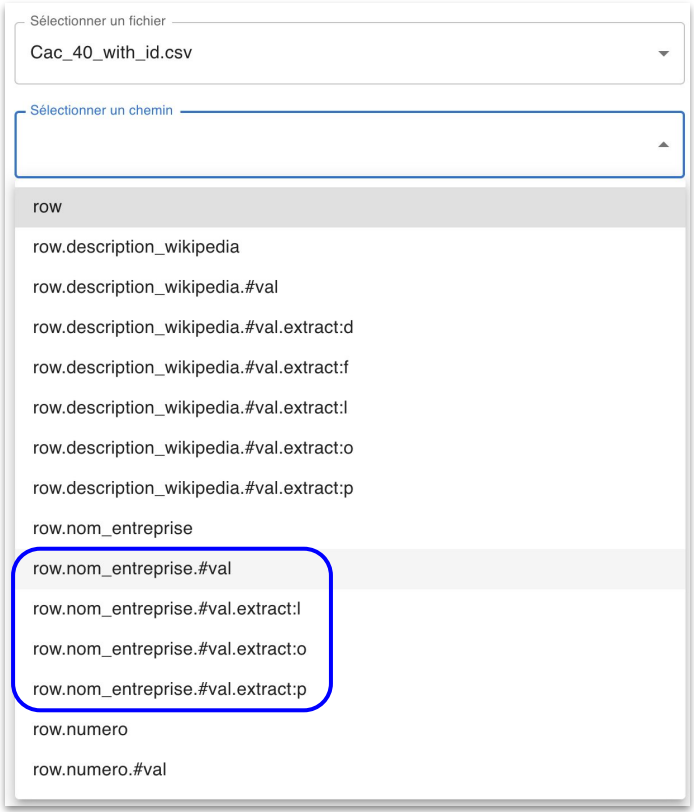
Comment nettoyer et uniformiser les données ?

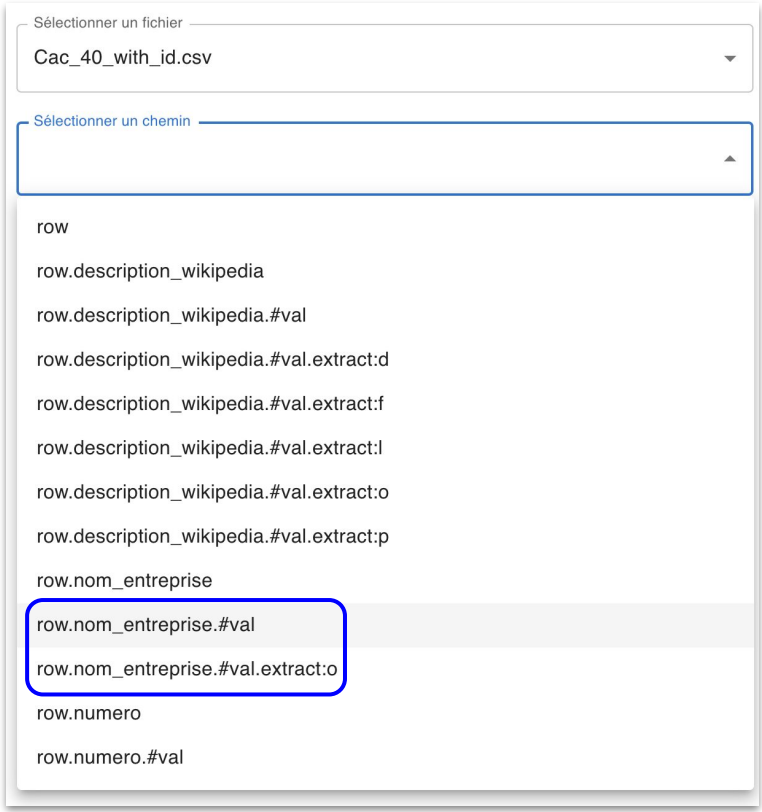
- Nettoyage des données semi-automatique :
  - Nettoyage lors de l’ingestion (i.e. *normalisation* des valeurs)
  - Nettoyage par l’utilisateur après chargement

Politique d’extraction : association d’un chemin
et d’un type d’entité pour aider l’extraction d’entités



Nettoyage : politiques d’extraction





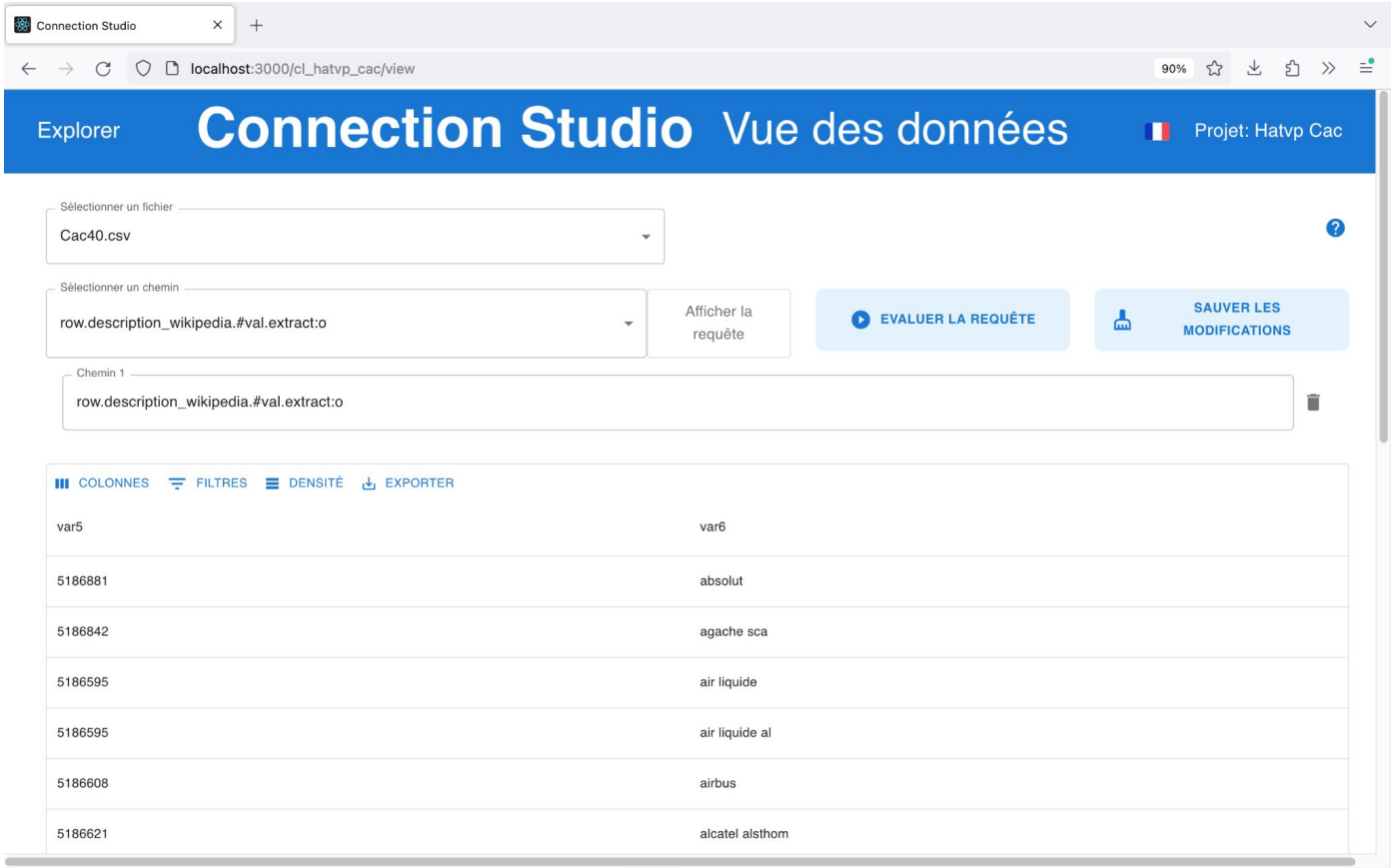
Nettoyage : modification des valeurs

Comment nettoyer et uniformiser les données ?

- Nettoyage des données semi-automatique :
  - Nettoyage lors de l'ingestion (i.e. *normalisation* des valeurs)
  - Nettoyage par l'utilisateur après chargement

- Possibilité de nettoyer les valeurs à la main
- Modifie les valeurs dans la base de données

Nettoyage : modification des valeurs



Scénario : HATVP & CAC40

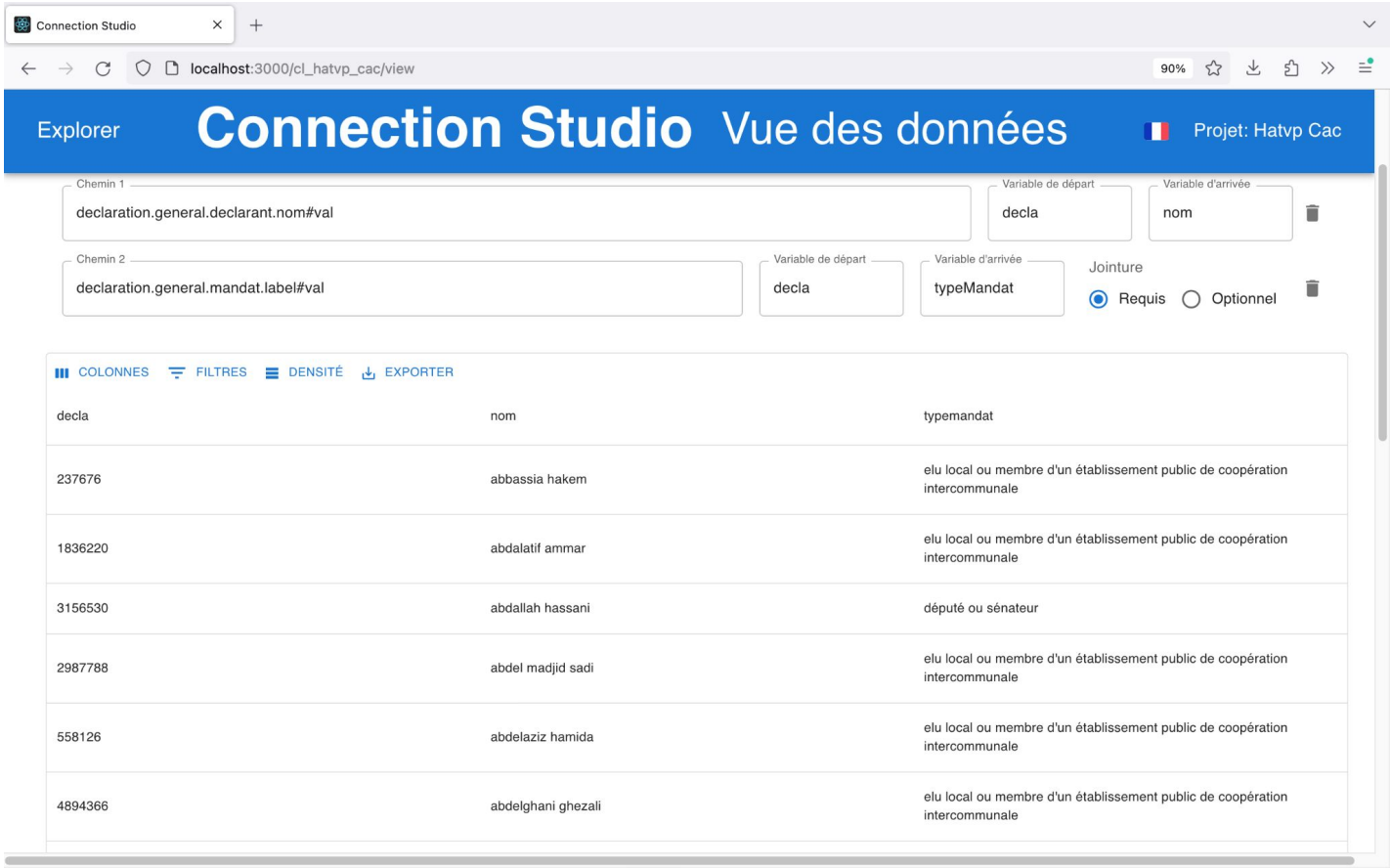
- HATVP : Haute Autorité pour la Transparence de la Vie Publique
- CAC40 : liste des 40 entreprises françaises “les plus influentes”

HATVP → fichier XML de 1,5 millions de lignes

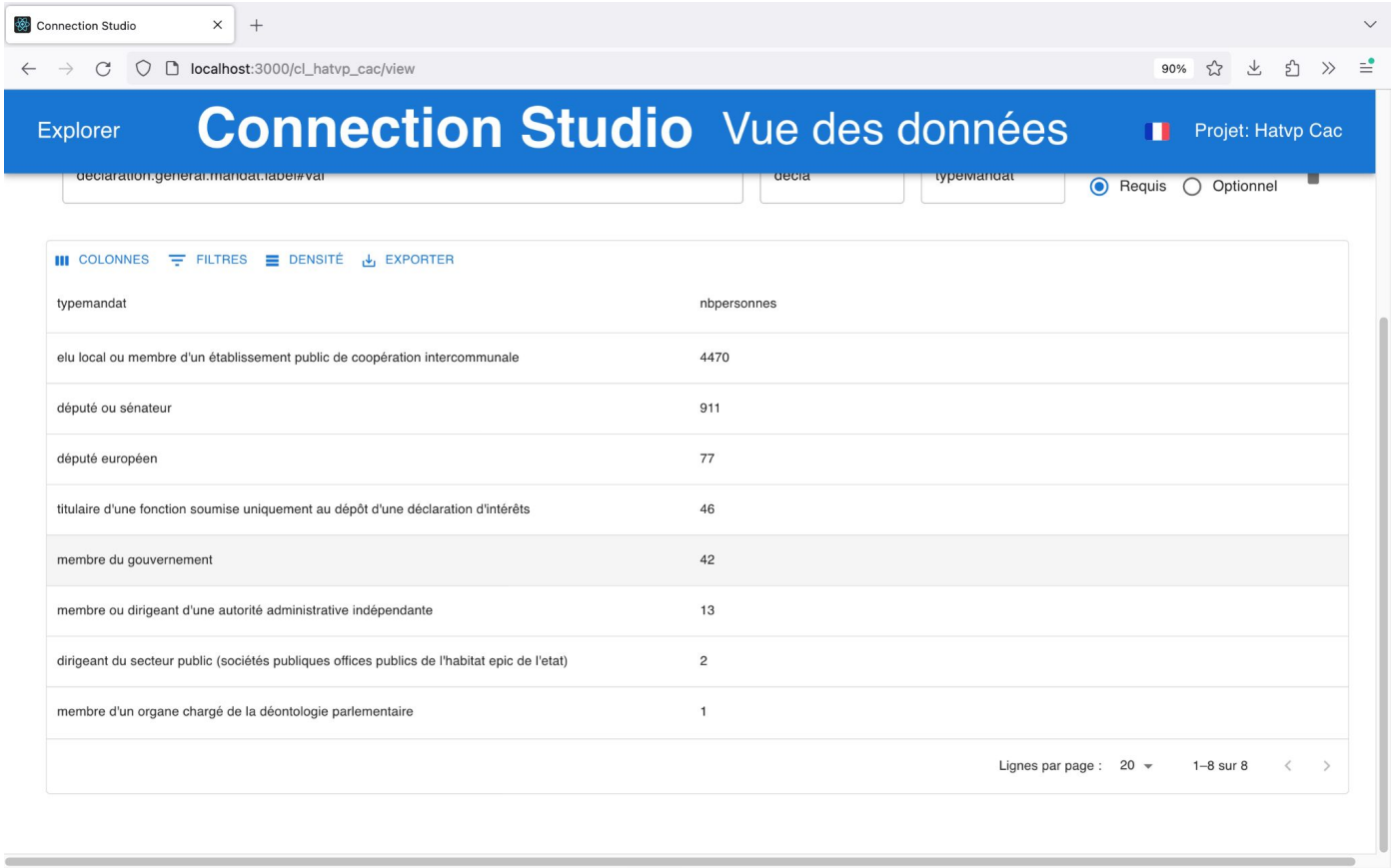
CAC40 → fichier CSV de 40 lignes

Haute Autorité
pour la transparence
de la vie publique

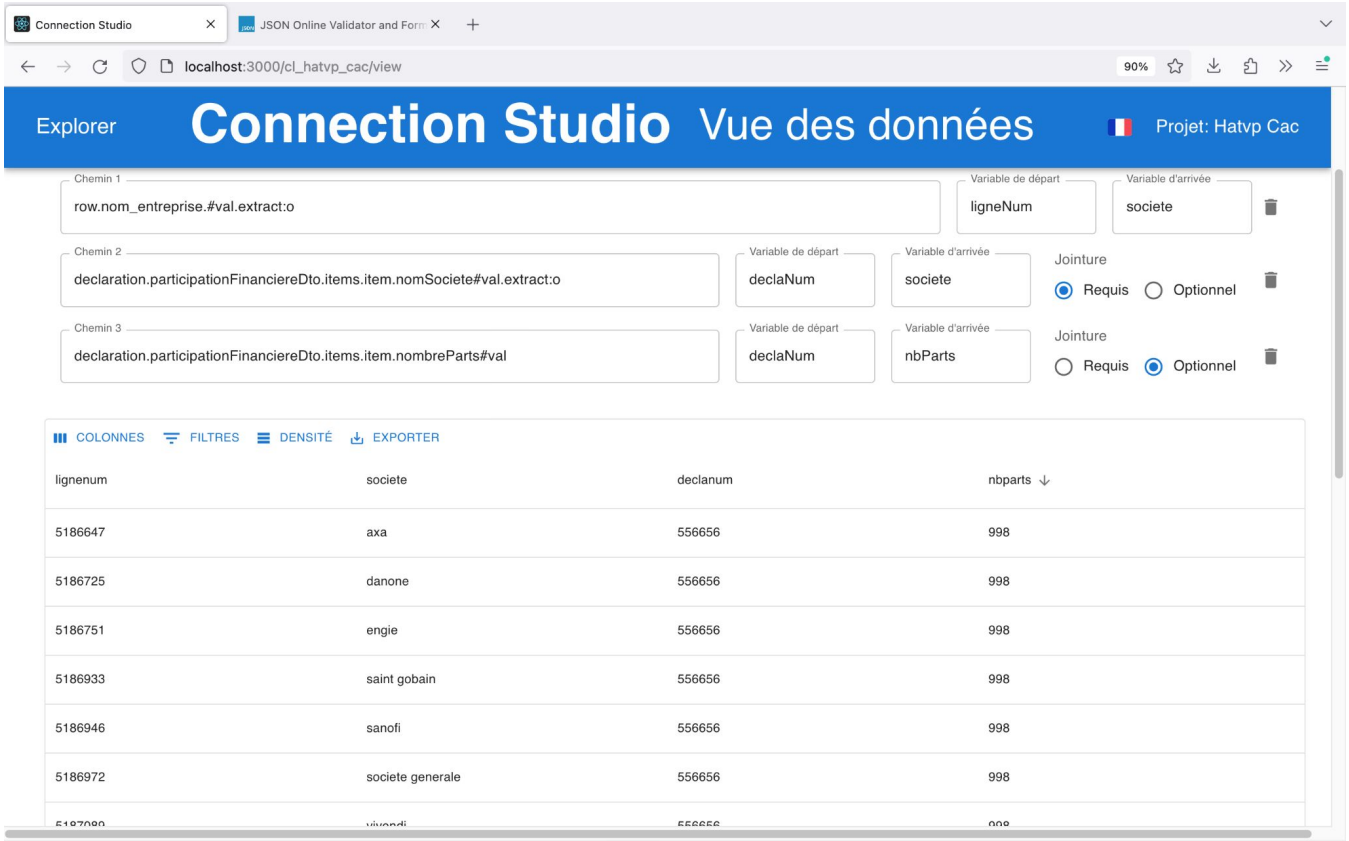
Scénario : HATVP & CAC40



Scénario : HATVP & CAC40



Scénario : HATVP & CAC40



Conclusion

ConnectionStudio :

- intègre des données très hétérogènes de manière unifiée,
- en extrait des entités nommées (personnes, lieux, …),
- permet d'inspecter, requêter et connecter ses données de différentes façons

Site web : https://project.inria.fr/connectionstudio/fr/

Contact : ioana.manolescu@inria.fr