

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

Inria

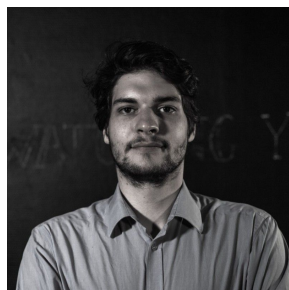


Qui sommes nous ?

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

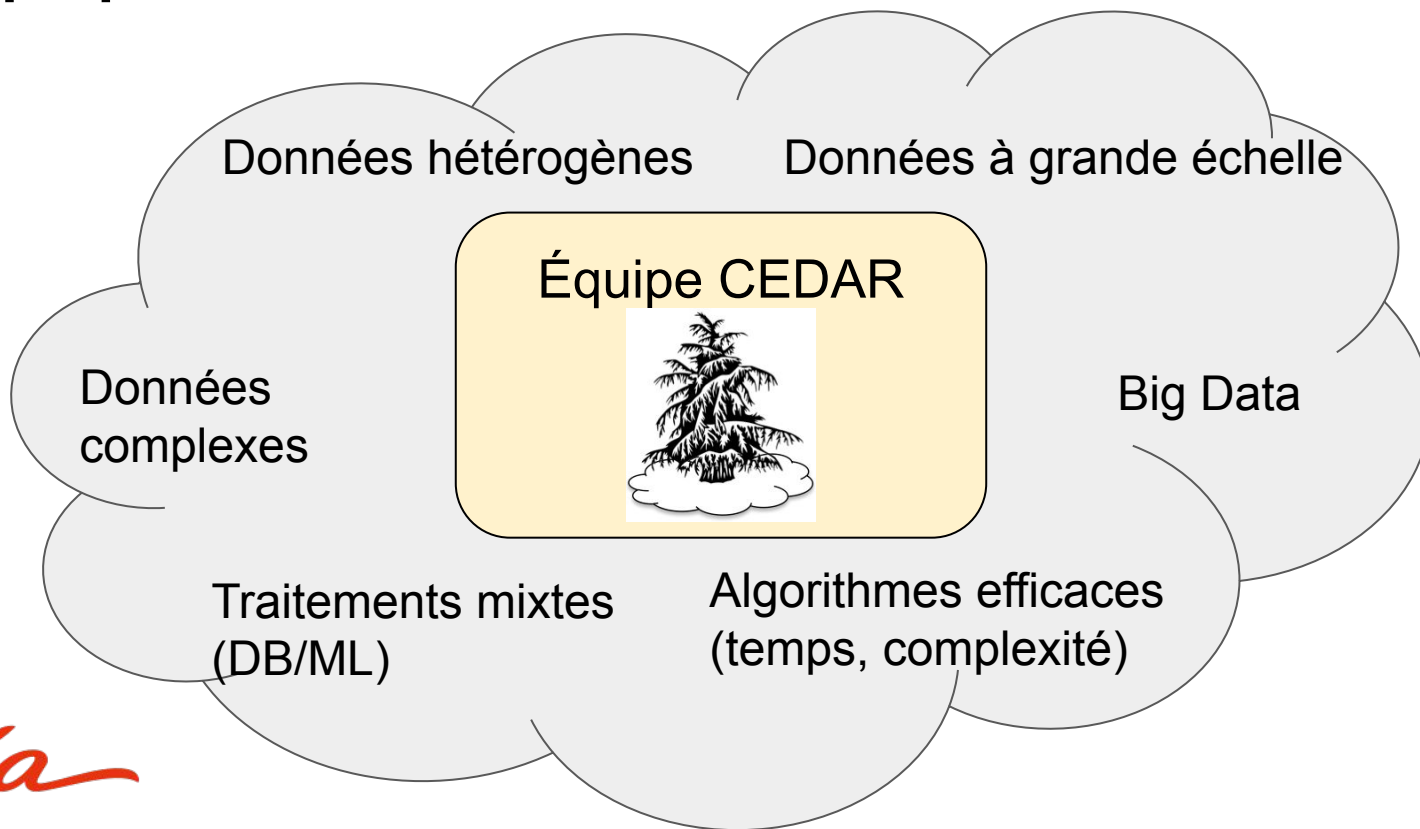


Ioana Manolescu
Chercheure et chef
d'équipe
Inria & IPP

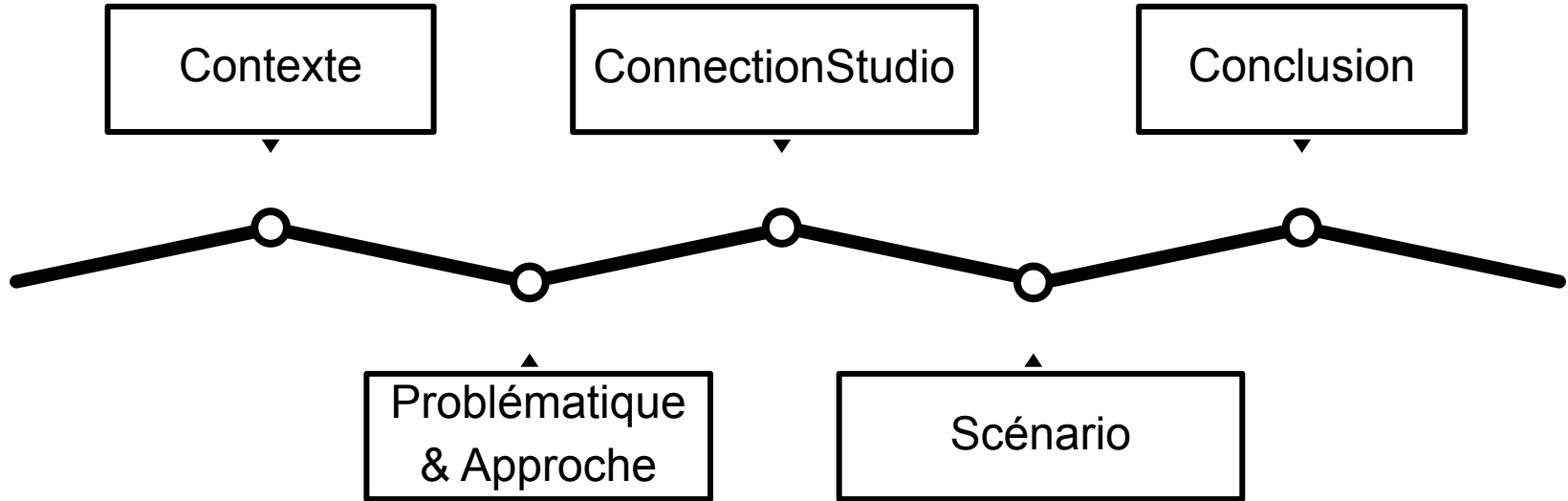


**Madhulika
Mohanty**
Post-doctorante
Inria

L'équipe CEDAR



Organisation de l'atelier



Contexte

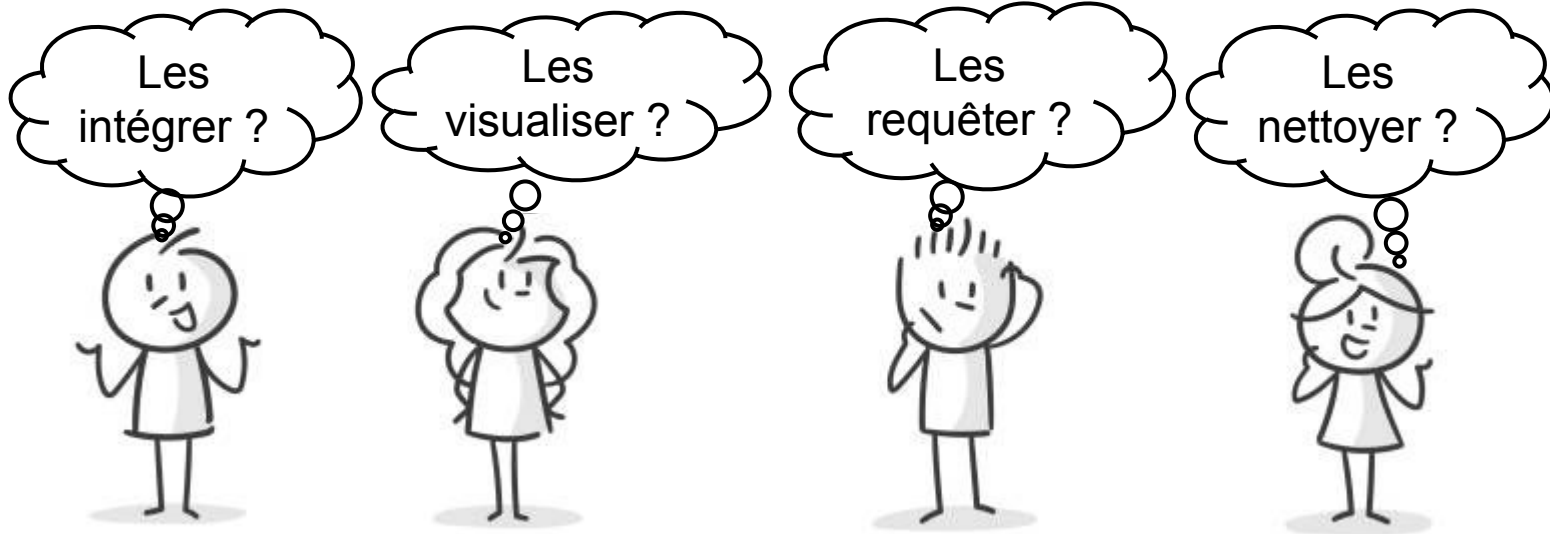
- Les “faits” proviennent de **différentes sources** de données, souvent **larges et complexes**
- Les sources sont de **formats très variés** :

Structurées	Semi-structurées	Non-structurées
<ul style="list-style-type: none">- Bases de données- Tables (Excel, CSV)	<ul style="list-style-type: none">- JSON (Web)- XML (Web)- HTML (Web)	<ul style="list-style-type: none">- Texte- PDF- RDF (<i>Open Data</i>)- Neo4j (<i>Graphe</i>)

- 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 ?

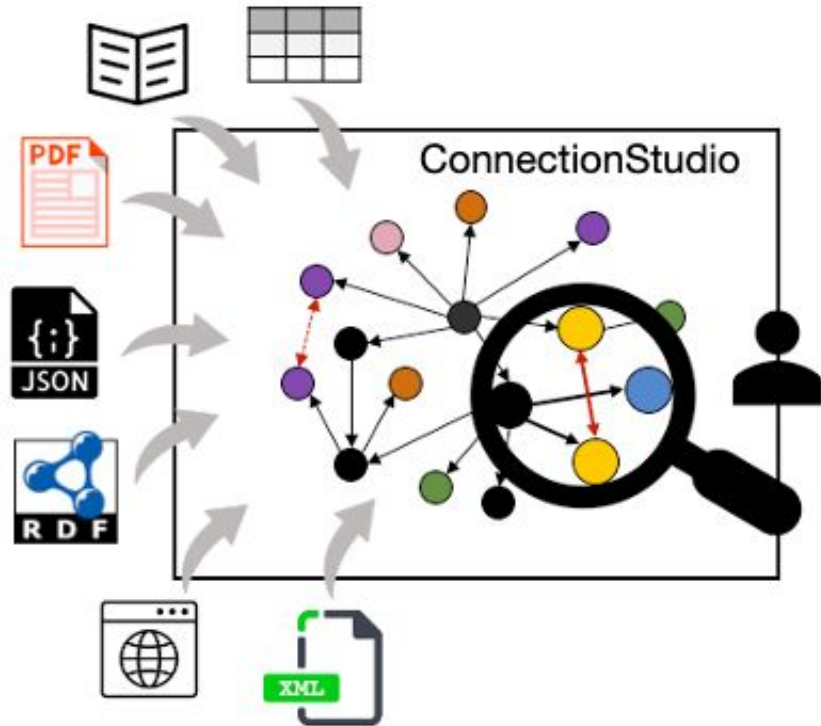
ConnectionStudio !



ConnectionStudio

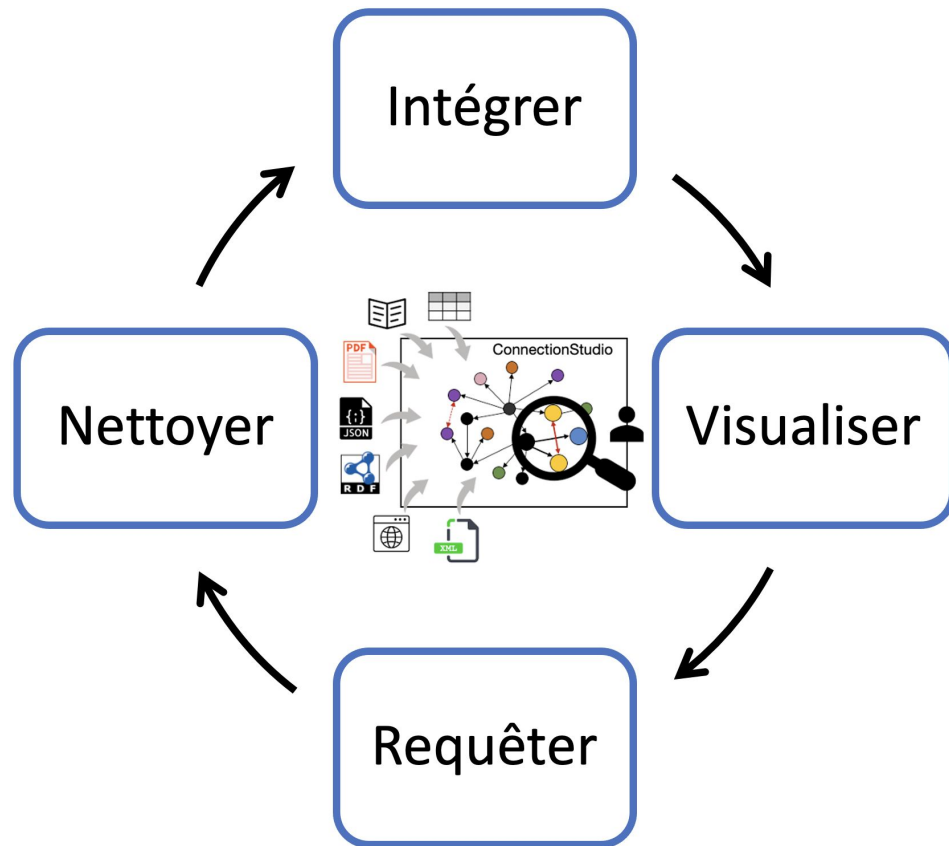
ConnectionStudio est un outil qui :

- 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

Connection Studio Projets



Trier par
Nom du projet

TERMINER LA SESSION

CRÉER UN PROJET

Projet Cac40

1 fichiers

Créé le : 2023-07-05 16:12:38

Dernier ajout : 2023-07-05 16:12:38

GÉRER

Projet Hatvp Cac40

2 fichiers

Créé le : 2023-07-05 15:46:07

Dernier ajout : 2023-07-05 16:25:52

GÉRER

Projet Hatvpsmall

Pas encore de fichier chargé, ajoutez-en !

GÉRER

Projet Pubmed

1 fichiers

Créé le : 2023-07-05 09:46:07

Dernier ajout : 2023-07-05 09:46:07

GÉRER

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.

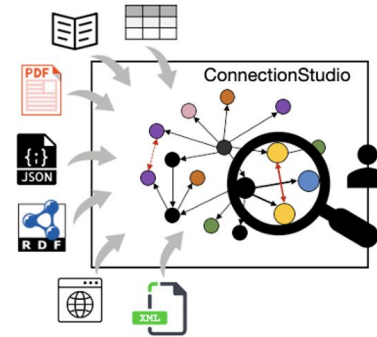
	A	B	C
1	numero	nom_entreprise	description_wikipedia
2	1	AIR LIQUIDE	Air liquide (AL), anciennement L'Air liquide, est un groupe industriel français d'envergure internationale, spécialiste des gaz industriels, c'est-à-dire des gaz pour l'industrie, la santé, l'environnement et la recherche.
3	2	AIRBUS	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

Nettoyer

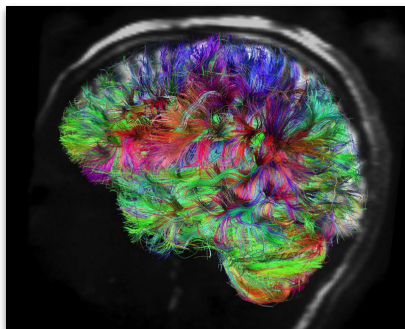
Visualiser

Requêter

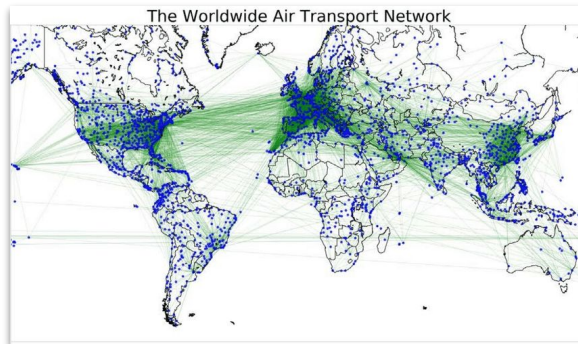


Graphe

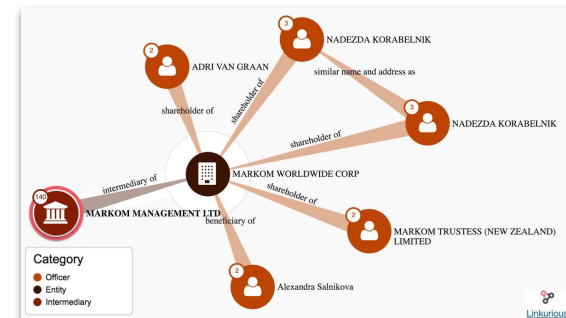
- Le **paradigme “graphe”** décrit :
 - Des objets (*noeuds*)
 - Connectés par des liens (*arêtes*)
- Beaucoup utilisé car très **flexible**



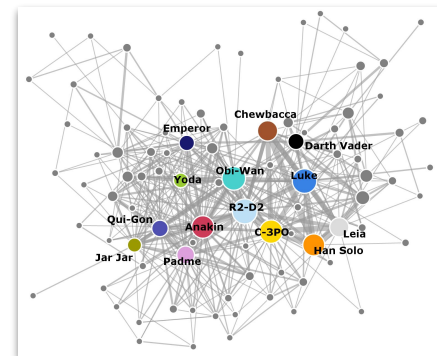
Neurones du cerveau humain



Vols internationaux



Panama papers (données ICIJ)



Personnages dans “Star Wars”

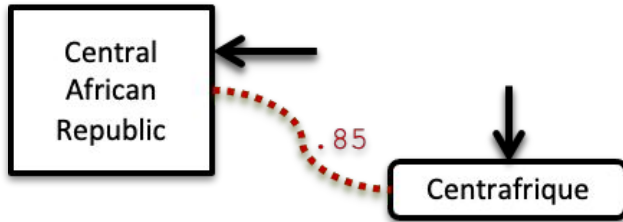
Extraction d'entités

- 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

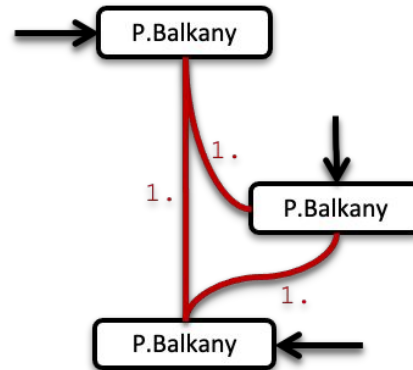
In **December 1903** DATE **the Royal Swedish Academy of Sciences** ORG awarded **Marie** PERSON and **Pierre Curie** PERSON, along with **Henri Becquerel** PERSON, **the Nobel Prize in Physics** WORK_OF_ART.

Annotation sémantique

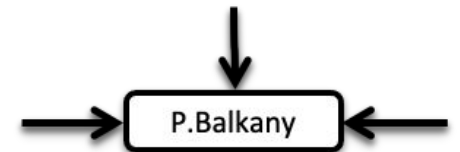
- Connexion entre les entités nommées quand elles sont *similaires*
- Plusieurs similarités : exactes, approximatives
- Unification possible quand similarité totale → - de noeuds et + de connexions



Similarité approximative



Similarité exacte

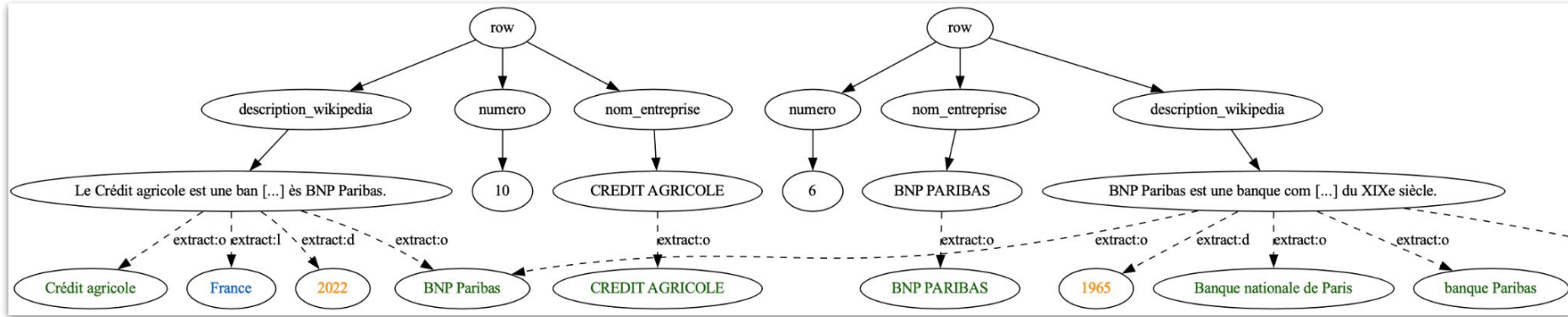


Fusion des entités identiques

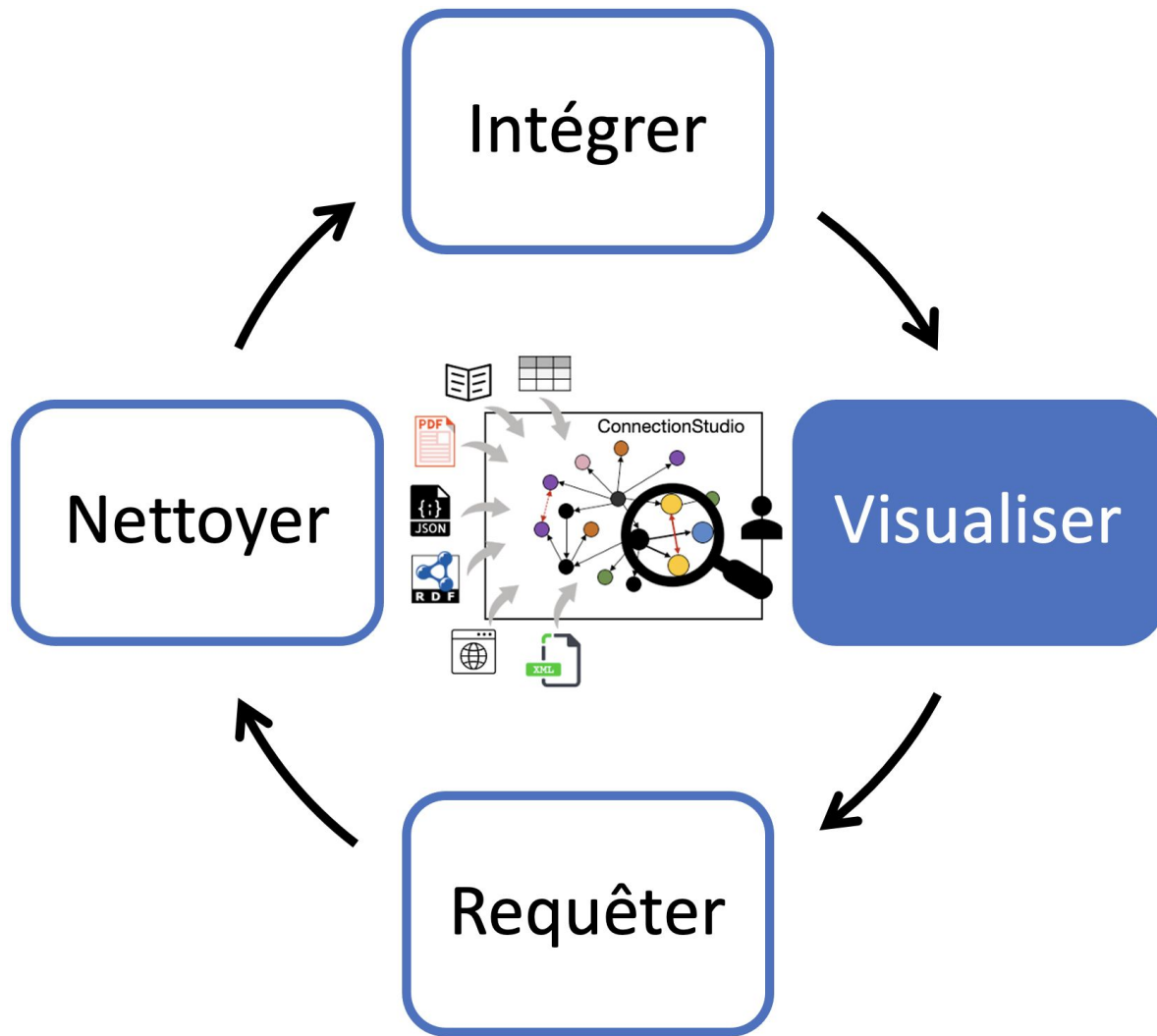
Chargement des sources de données

- Les données 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



Visualisation 1 : statistiques

Comment avoir une bonne idée des entités détectées ?

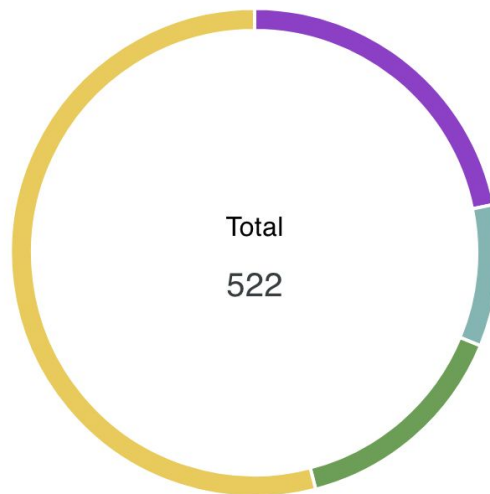
- Distribution d'entités par types
- Nuage des 100 entités les plus fréquentes (*tag cloud*)
- Distribution des types d'entités par fichier
- Liste des 100 entités les plus partagées parmi le plus de sources

- Tous les graphiques/tableaux sont exportables

Visualisation 1 : statistiques

Distribution d'entités par type

< Entités identifiées >



● Nombre de dates ● Nombre de Personnes ● Nombre de Places ● Nombre d'Organisations

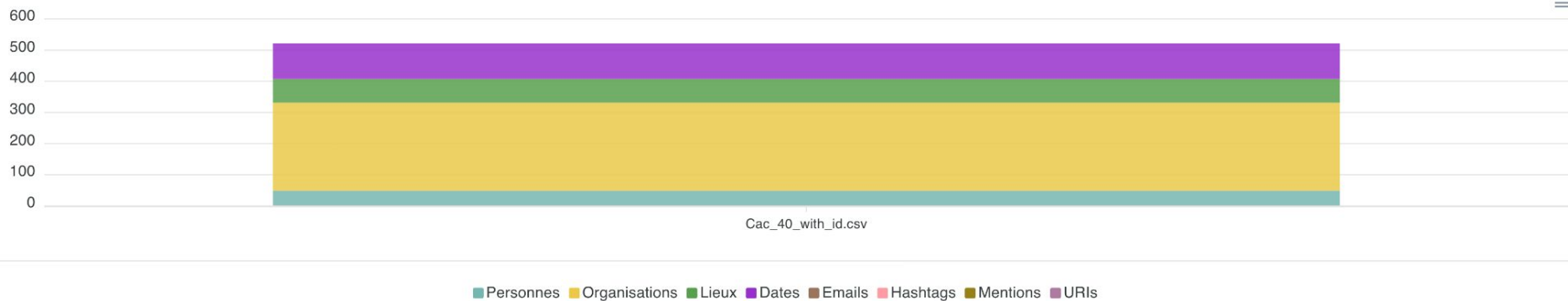
Visualisation 1 : statistiques

Nuage d'entités



Visualisation 1 : statistiques

Distribution des types d'entités par fichier



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**

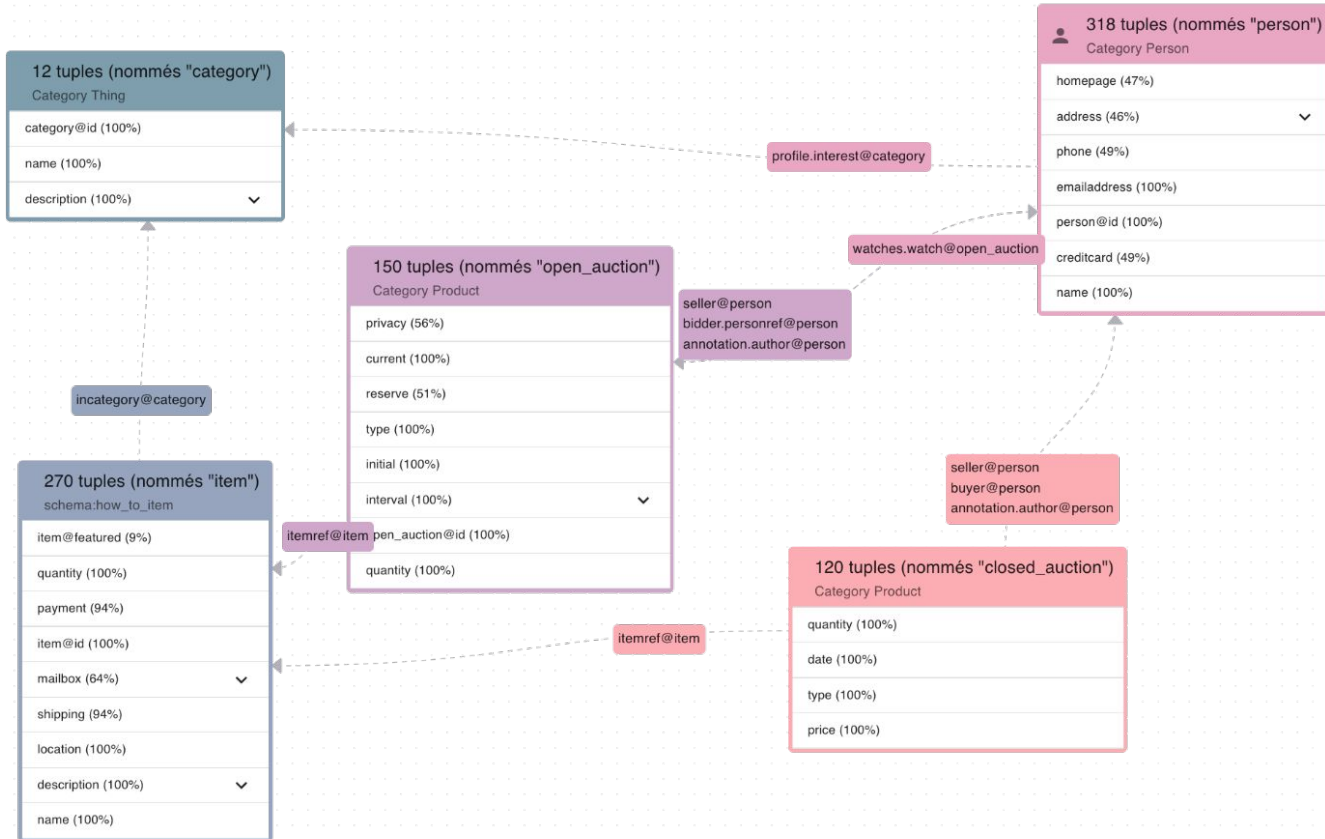
Visualisation 2 : structure d'un fichier

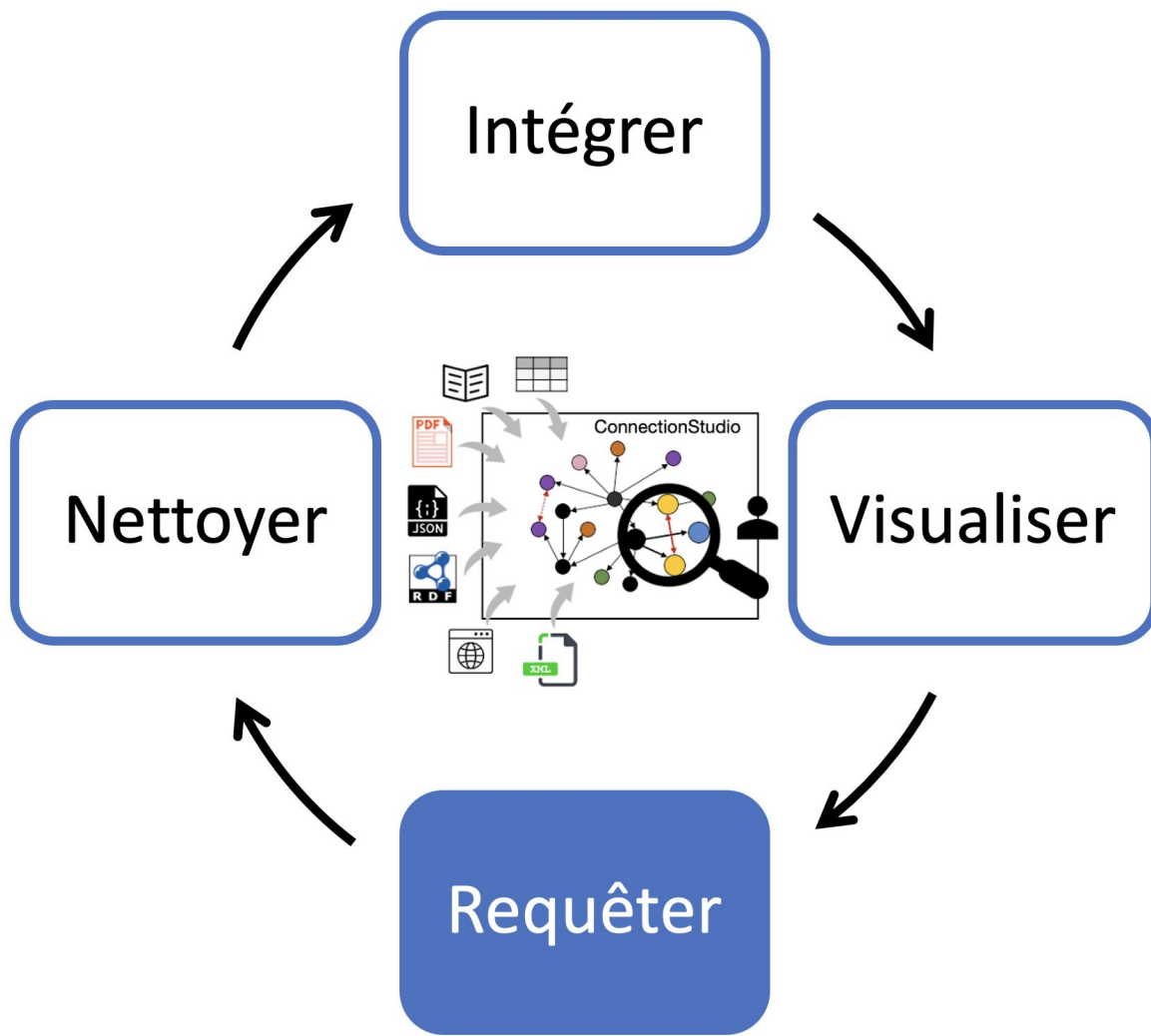
The screenshot shows the Connection Studio web application. The browser address bar indicates the URL is localhost:3000/ci_cac40/abstra. The main header is blue with the text "Connection Studio" and "Abstraction de données". The left sidebar contains a "FICHIERS" section with a list of file formats: JSON, XML, CSV, and RDF. Below this, a file named "Cac_40_with_id.csv" is selected. The main area displays a grid of 40 tuples, each labeled "row". A tooltip is shown over one of the rows, detailing its structure:

- 40 tuples (nommés "row")
- Category Thing
- description_wikipedia (100%)
- nom_entreprise (100%)
- numero (100%)

At the bottom left, there are icons for zooming (+, -, double arrows) and saving the visualization as SVG or PNG.

Visualisation 2 : structure d'un fichier





Intégrer

Nettoyer

Visualiser

Requête

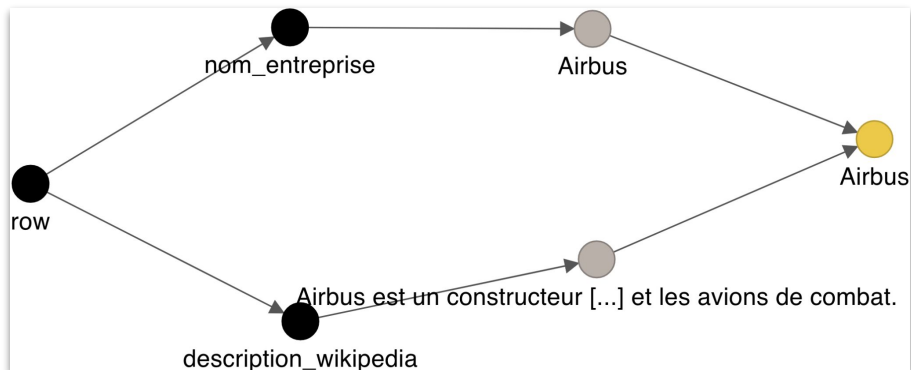
Requêtes

- **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

The screenshot shows the Connection Studio web application interface. The browser address bar indicates the URL is localhost:3000/cl_pubmed/pathways. The application header includes the title 'Exploration de chemins' and 'Projet: Pubmed'. The main content area is titled 'Résultats' and displays a table of search results. The table has columns for Name, Author, AuthorList, PubmedArticle, CoiStatement, and CoiStatement#val. The results list authors A. Coates and A. Nishiyama, with associated CoiStatement values such as 'CP', 'Helperby Therapeutics', and 'Japan Society for the promotion of Science'. The interface also includes navigation controls like 'COLONNES', 'FILTRES', 'DENSITÉ', 'EXPORTER', and 'ETENDRE LE TEXTE'. At the bottom, a breadcrumb trail shows the current path: Affiliation#val > Affiliation > Author > AuthorList > PubmedArticle > CoiStatement > CoiStatement#val (252 chemins).

Name#val	Name	Author	AuthorList	PubmedArticle	CoiStatement	CoiStatement#val
A. Coates	Name	Author	AuthorList	PubmedArticle	CoiStatement	CP
A. Coates	Name	Author	AuthorList	PubmedArticle	CoiStatement	Helperby Therapeutics
A. Nagano	Name	Author	AuthorList	PubmedArticle	CoiStatement	Japan Society for the promotion of Science
A. Nishiyama	Name	Author	AuthorList	PubmedArticle	CoiStatement	Grant-in-Aid for Scientific Research B of Japan Society for the Promotion of Science
A. Nishiyama	Name	Author	AuthorList	PubmedArticle	CoiStatement	Grant-in-Aid for Scientific Research C of Japan Society for the Promotion of Science

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	<code>exact:Airbus</code>	Un noeud exactement labellisé Airbus
Un mot	<code>Airbus / "Airbus"</code>	Un noeud qui contient : <ul style="list-style-type: none">- Airbus (non-sensible à la casse)- "Airbus" (sensible)
Exploration par mot-clé	<code>explore Capgemini</code>	Les noeuds labellisés Capgemini ainsi que leurs voisins
Exactement plusieurs mots	<code>exact:CREDIT+AGRICOLE</code>	Les noeuds labellisés exactement et uniquement CREDIT AGRICOLE
Plusieurs mots	<code>"Airbus Engie"</code>	Les noeuds qui contiennent Airbus et Engie (dans n'importe quel ordre)
Exploration par type	<code>type:Person</code>	Tous les noeuds de type Personne

Requête 2 : recherche par mots-clés

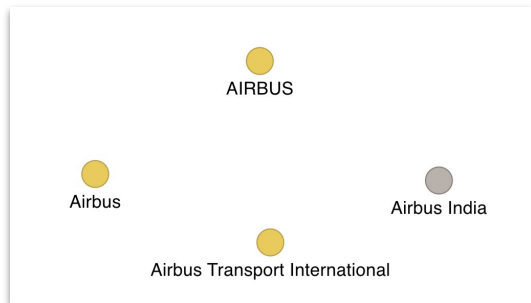
Information du Noeud ×

 Airbus

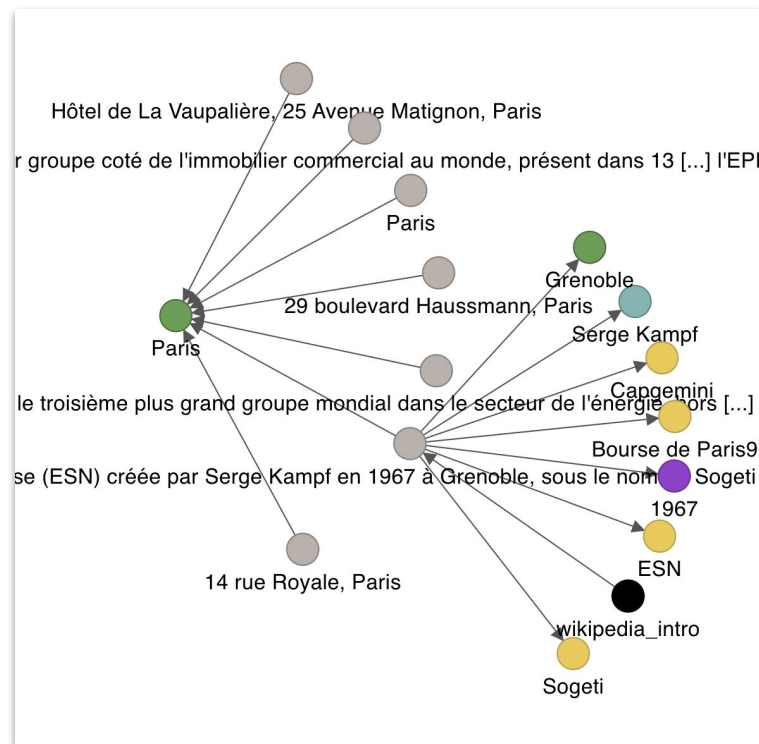
Noeud **Airbus** de type **Organisation** extrait depuis le fichier **Cac_40_with_id_actionnaires_and_filiales.csv**

Identifiant interne: 2073

Un noeud et ses informations

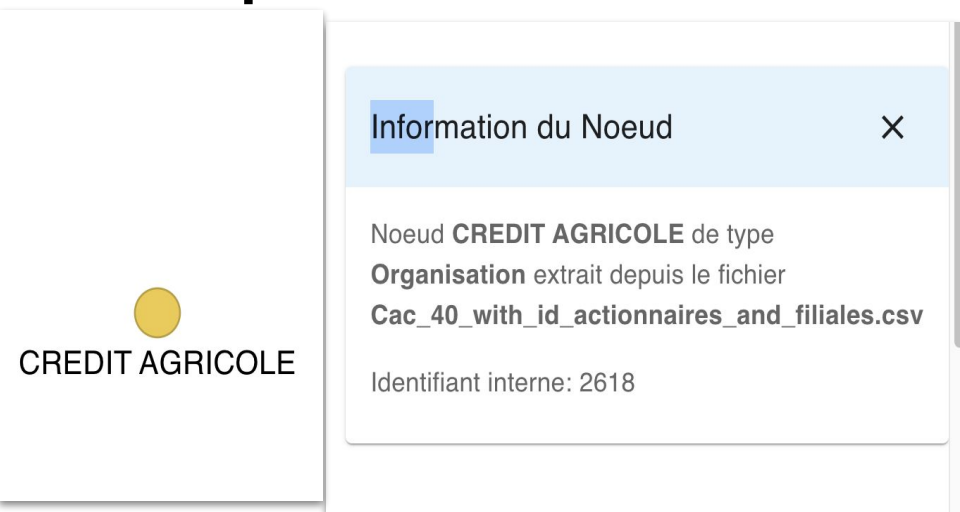


Les noeuds "Airbus"



Exploration manuelle

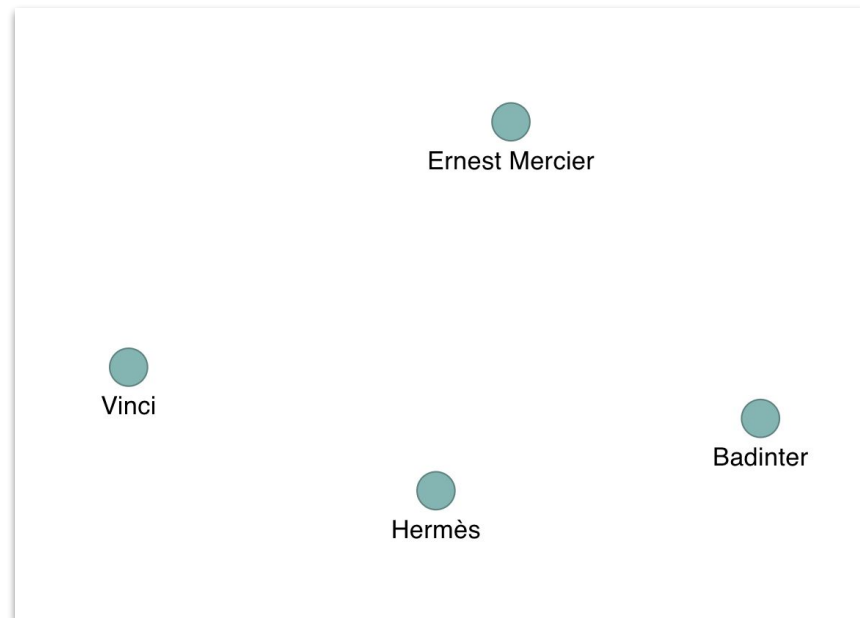
Requête 2 : recherche par mots-clés



The image shows a user interface element for a node. On the left, there is a yellow circle icon above the text "CREDIT AGRICOLE". To the right, a white popup window with a light blue header is open. The header contains the text "Information du Noeud" and a close button (X). The main content of the popup reads: "Noeud **CREDIT AGRICOLE** de type **Organisation** extrait depuis le fichier **Cac_40_with_id_actionnaires_and_filiales.csv**" and "Identifiant interne: 2618".

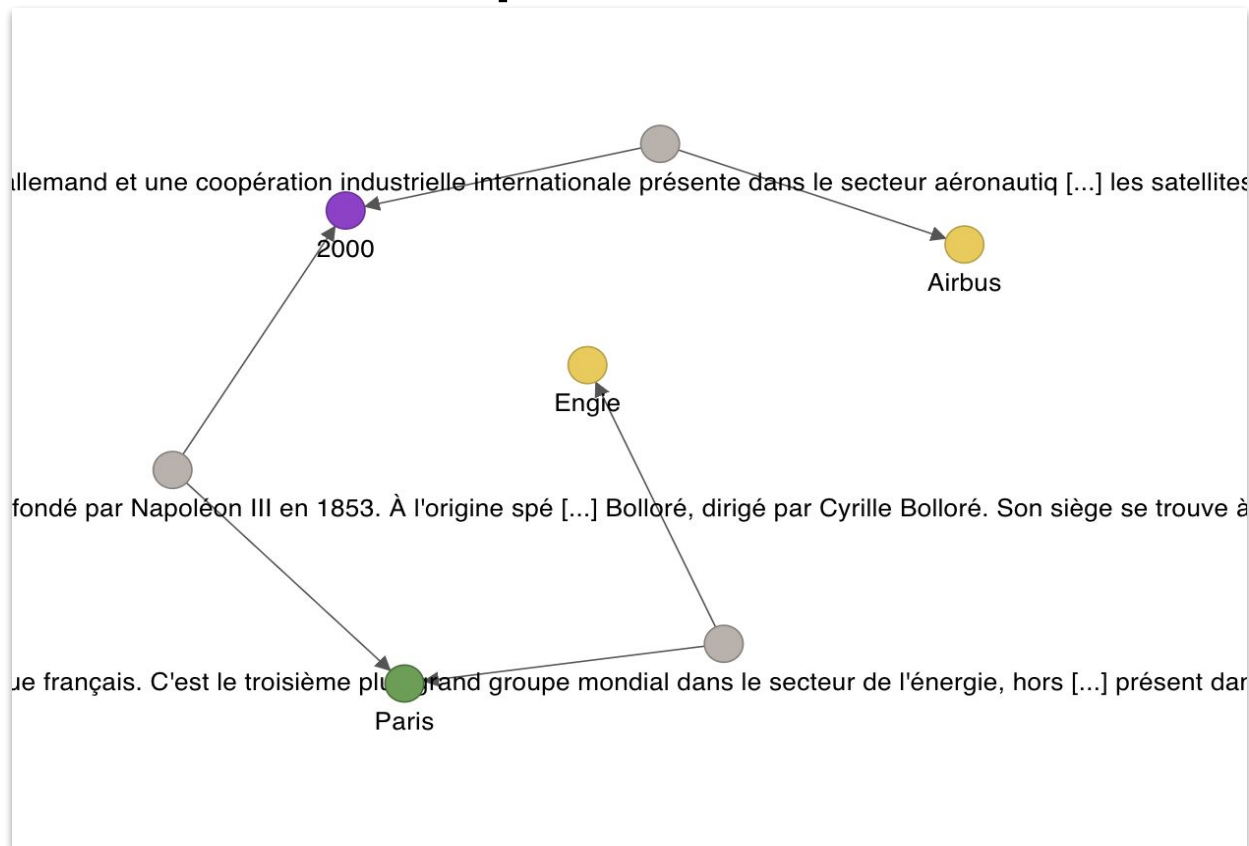
Un noeud et ses informations

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

The screenshot displays the Connection Studio interface. At the top, the browser address bar shows 'localhost:3000/cl_cac40/view'. The main header is blue with 'Connection Studio' and 'Vue des données' on the left, and 'Projet: Cac40' on the right. Below the header, there are three input fields for file selection, path selection, and a path field. To the right of these fields are three buttons: 'Afficher la requête', 'EVALUER LA REQUÊTE', and 'SAUVER LES MODIFICATIONS'. Below the input fields, there is a table with two columns: 'var0' and 'var1'. The table contains 11 rows of data, including values like '1', '14', '27', '40', '53', '66', '79', '92', '105', and '118' in the first column, and company names like 'air liquide', 'airbus', 'alstom', 'arcelormittal', 'axa', 'bnp paribas', 'bouygues', 'cappemini', 'carrefour', and 'crédit agricole' in the second column.

Explorer **Connection Studio** Vue des données Projet: Cac40

Sélectionner un fichier
Cac_40_with_id.csv

Sélectionner un chemin
row.nom_entreprise.#val

Chemin 1
row.nom_entreprise.#val

Afficher la requête **EVALUER LA REQUÊTE** SAUVER LES MODIFICATIONS

COLONNES FILTERS DENSITÉ EXPORTER

var0	var1
1	air liquide
14	airbus
27	alstom
40	arcelormittal
53	axa
66	bnp paribas
79	bouygues
92	cappemini
105	carrefour
118	crédit agricole

Requête 3 : Requêtes utilisateur

The screenshot shows the Connection Studio interface. At the top, there's a navigation bar with 'Explorer', 'Connection Studio', and 'Vue des données'. The project name 'Projet: Cac40' is visible on the right. Below the navigation bar, there are several configuration fields for a query:

- 'Sélectionner un fichier': Cac_40_with_id.csv
- 'Sélectionner un chemin': row.nom_entreprise.#val.extract:o
- 'Chemin 1': row.description_wikipedia.#val
- 'Chemin 2': row.description_wikipedia.#val.extract:o
- 'Chemin 3': row.nom_entreprise.#val.extract:o

Buttons for 'Afficher la requête', 'EVALUER LA REQUÊTE', and 'SAUVER LES MODIFICATIONS' are present. Below the configuration, there are settings for 'Variable de départ' and 'Variable d'arrivée' for each path, along with 'Jointure' options (Requis, Optionnel).

The bottom part of the image shows a table view with columns: ligne, texte, entreprise, and ligne2. The first row has the value 66 in the 'ligne' column and a detailed text description in the 'texte' column. The 'entreprise' column contains 'bnp paribas' and the 'ligne2' column contains '66'.

COLONNES	FILTRES	DENSITÉ	EXPORTER
ligne	texte	entreprise	ligne2
66	bnp paribas est une banque commerciale française elle est la première banque européenne par son activité et sa rentabilité avec 3 080 milliards \$ d'actifs et le 8e groupe bancaire international présent dans 65 pays il est coté au premier marché d'euronext paris et fait partie de l'indice cac 40 au 31 décembre 2022 le bénéfice net part du groupe s'élève à 10 2 milliards d'euros en hausse de 7 5 % par rapport à 202110 avec 193 000 employés en février 2023 la banque est organisée selon trois grands domaines d'activités : services bancaires pour particuliers et pour commerçants services d'investissement et de protection services bancaires pour entreprises et institutions le groupe est issu de la fusion en mai 2000 entre la banque nationale de paris banque née en 1965 de la fusion de l'ancienne banque nationale de crédit et du comptoir national d'escompte de paris et de la banque paribas établissement né au cours du xixe siècle	bnp paribas	66

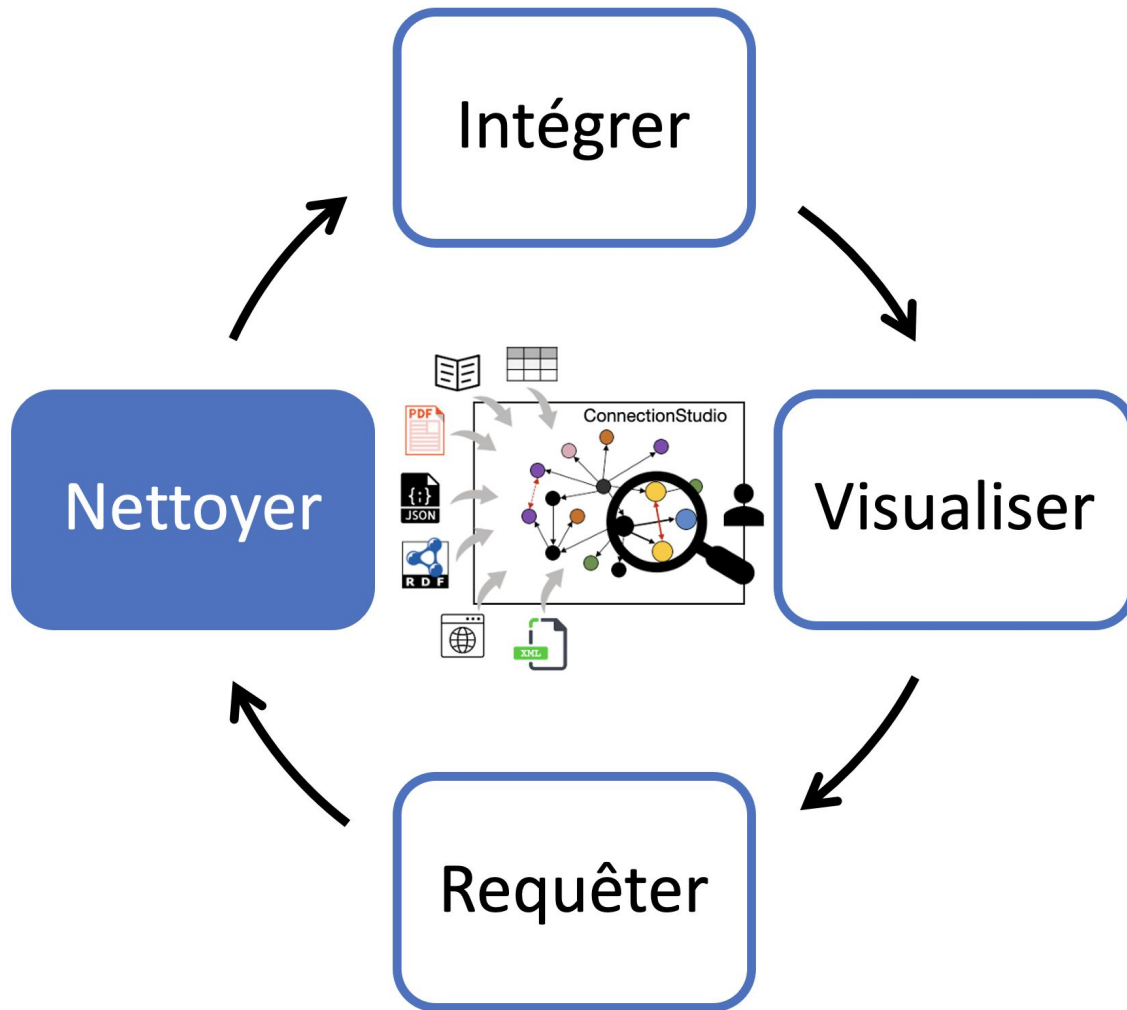
Requête 3 : Requêtes utilisateur

The screenshot displays the Connection Studio interface for a project named 'Cac40'. The main view is titled 'Vue des données'. The interface includes a file selection dropdown set to 'Cac_40_with_id.csv' and a path selection dropdown set to 'row.nom_entreprise.#val.extract:o'. Below these are three path configuration rows:

- Chemin 1: row.description_wikipedia.#val, with 'ligne' as the start variable and 'texte' as the end variable.
- Chemin 2: row.description_wikipedia.#val.extract:o, with 'ligne' as the start variable and 'entreprise' as the end variable. The join type is 'Requis'.
- Chemin 3: row.nom_entreprise.#val.extract:o, with 'ligne2' as the start variable and 'entreprise' as the end variable. The join type is 'Requis'.

Buttons for 'Cacher la requête', 'EVALUER LA REQUÊTE', and 'SAUVER LES MODIFICATIONS' are visible. The bottom section shows the generated SQL query:

```
Requête générée
WITH path0
AS (
  SELECT n1.id
  AS ligne, n3.normallabel
  AS texte
  FROM nodes n1, nodes n2, nodes n3, edges e1, edges e2
  WHERE n1.label='row'
  AND n1.type=7
  AND n2.label='description_wikipedia'
  AND n2.type=7
  AND e1.source=n1.id
  AND e1.target=n2.id
  AND e2.source=n2.id
  AND e2.target=n3.id) , path1
AS (
  SELECT n1.id
  AS ligne, n4.normallabel
  AS entreprise
  FROM nodes n1, nodes n2, nodes n3, nodes n4, edges e1, edges e2, edges e3
  WHERE n1.label='row'
  AND n1.type=7
```



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

The screenshot shows a web interface for file selection and path management. At the top, there is a dropdown menu labeled 'Sélectionner un fichier' with the selected file 'Cac_40_with_id.csv'. Below it is a text input field labeled 'Sélectionner un chemin'. The main area displays a list of paths under the heading 'row'. The paths are organized into groups, with some groups highlighted in grey. The paths include:

- row.description_wikipedia
- row.description_wikipedia.#val
- row.description_wikipedia.#val.extract:d
- row.description_wikipedia.#val.extract:f
- row.description_wikipedia.#val.extract:l
- row.description_wikipedia.#val.extract:o
- row.description_wikipedia.#val.extract:p
- row.nom_entreprise
- row.nom_entreprise.#val
- row.nom_entreprise.#val.extract:l
- row.nom_entreprise.#val.extract:o
- row.nom_entreprise.#val.extract:p
- row.numero
- row.numero.#val

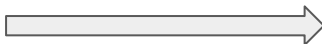
Nettoyage : politiques d'extraction

Sélectionner un fichier
Cac_40_with_id.csv

Sélectionner un chemin

```
row
row.description_wikipedia
row.description_wikipedia.#val
row.description_wikipedia.#val.extract:d
row.description_wikipedia.#val.extract:f
row.description_wikipedia.#val.extract:l
row.description_wikipedia.#val.extract:o
row.description_wikipedia.#val.extract:p
row.nom_entreprise
row.nom_entreprise.#val
row.nom_entreprise.#val.extract:l
row.nom_entreprise.#val.extract:o
row.nom_entreprise.#val.extract:p
row.numero
row.numero.#val
```

Politique d'extraction
row.nom_entreprise.#val
Organization



Sélectionner un fichier
Cac_40_with_id.csv

Sélectionner un chemin

```
row
row.description_wikipedia
row.description_wikipedia.#val
row.description_wikipedia.#val.extract:d
row.description_wikipedia.#val.extract:f
row.description_wikipedia.#val.extract:l
row.description_wikipedia.#val.extract:o
row.description_wikipedia.#val.extract:p
row.nom_entreprise
row.nom_entreprise.#val
row.nom_entreprise.#val.extract:o
row.numero
row.numero.#val
```

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

The screenshot shows the Connection Studio interface. At the top, there's a blue header with "Explorer", "Connection Studio", "Vue des données", and "Projet: Hatvp Cac". Below the header, there are three input fields: "Sélectionner un fichier" (Cac40.csv), "Sélectionner un chemin" (row.description_wikipedia.#val.extract:o), and "Chemin 1" (row.description_wikipedia.#val.extract:o). To the right of the second field is a button "Afficher la requête". Further right are two buttons: "EVALUER LA REQUÊTE" and "SAUVER LES MODIFICATIONS". Below these fields is a table with columns "COLONNES", "FILTRES", "DENSITÉ", and "EXPORTER". The table contains data rows with two columns: "var5" and "var6".

var5	var6
5186881	absolut
5186842	agache sca
5186595	air liquide
5186595	air liquide al
5186608	airbus
5186621	alcatel alsthom

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

The screenshot shows the Connection Studio web interface. The browser address bar indicates the URL is localhost:3000/cl_hatvp_cac/view. The page title is "Connection Studio" and the view is "Vue des données". The project name is "Projet: Hatvp Cac".

Configuration fields:

- Chemin 1: `declaration.general.declarant.nom#val`
- Variable de départ: `decla`
- Variable d'arrivée: `nom`
- Chemin 2: `declaration.general.mandat.label#val`
- Variable de départ: `decla`
- Variable d'arrivée: `typeMandat`
- Jointure: Requis Optionnel

Table headers: COLONNES, FILTRES, DENSITÉ, EXPORTER

decla	nom	typemandat
237676	abbassia hakem	elu local ou membre d'un établissement public de coopération intercommunale
1836220	abdlatif ammar	elu local ou membre d'un établissement public de coopération intercommunale
3156530	abdallah hassani	député ou sénateur
2987788	abdel madjid sadi	elu local ou membre d'un établissement public de coopération intercommunale
558126	abdelaziz hamida	elu local ou membre d'un établissement public de coopération intercommunale
4894366	abdelghani ghezali	elu local ou membre d'un établissement public de coopération intercommunale

Scénario : HATVP & CAC40

The screenshot shows a web browser window with the address bar at localhost:3000/cl_hatvp_cac/view. The application header is blue and contains the text 'Explorer Connection Studio Vue des données' and 'Projet: Hatvp Cac'. Below the header, there are search filters for 'declaration.generale.mandat.la0en#var', 'decia', and 'typemandat'. There are radio buttons for 'Requis' (selected) and 'Optionnel'. The main content area displays a table with columns 'type de mandat' and 'nbpersonnes'. The table lists various types of mandates and their corresponding counts. At the bottom right, there is a pagination control showing 'Lignes par page : 20' and '1-8 sur 8'.

type de mandat	nbpersonnes
elu local ou membre d'un établissement public de coopération intercommunale	4470
député ou sénateur	911
député européen	77
titulaire d'une fonction soumise uniquement au dépôt d'une déclaration d'intérêts	46
membre du gouvernement	42
membre ou dirigeant d'une autorité administrative indépendante	13
dirigeant du secteur public (sociétés publiques offices publics de l'habitat epic de l'etat)	2
membre d'un organe chargé de la déontologie parlementaire	1

Scénario : HATVP & CAC40

The screenshot displays the Connection Studio interface. The top navigation bar includes 'Explorer', 'Connection Studio', 'Vue des données', and 'Projet: Hatvp Cac'. The main area is divided into three mapping sections:

- Chemin 1:** Maps `row.nom_entreprise.#val.extract.o` to `ligneNum` (Variable de départ) and `societe` (Variable d'arrivée).
- Chemin 2:** Maps `declaration.participationFinanciereDto.items.item.nomSociete#val.extract.o` to `declaNum` (Variable de départ) and `societe` (Variable d'arrivée). The jointure is set to **Requis**.
- Chemin 3:** Maps `declaration.participationFinanciereDto.items.item.nombreParts#val` to `declaNum` (Variable de départ) and `nbParts` (Variable d'arrivée). The jointure is set to **Optionnel**.

Below the mappings is a table with the following columns: COLONNES, FILTRES, DENSITÉ, and EXPORTER. The table data is as follows:

lignenum	societe	declanum	nbparts ↓
5186647	axa	556656	998
5186725	danone	556656	998
5186751	engie	556656	998
5186933	saint gobain	556656	998
5186946	sanofi	556656	998
5186972	societe generale	556656	998
5187000	vivendi	556656	998

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

