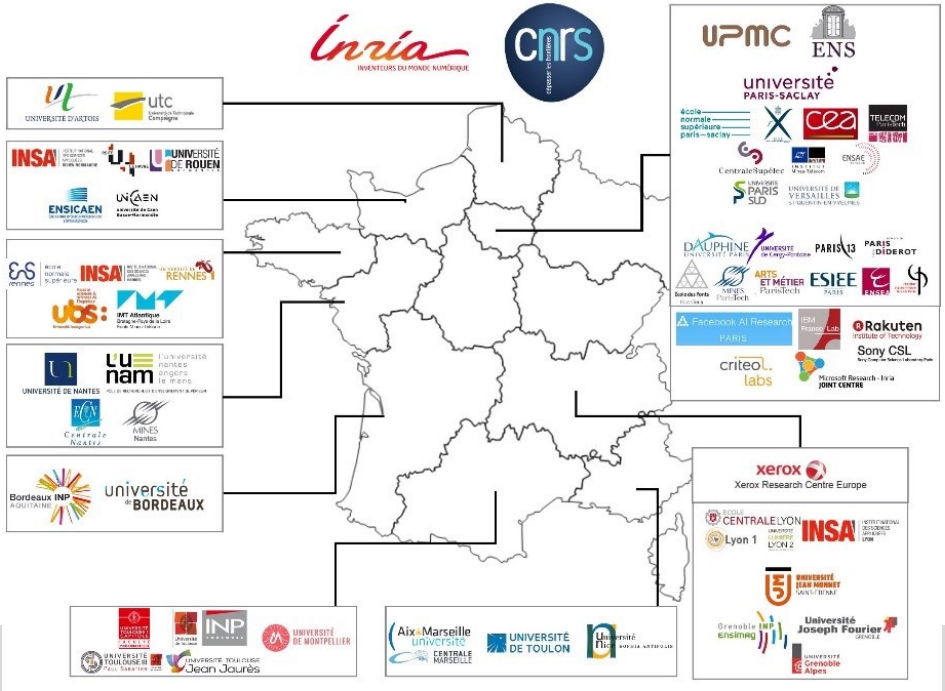
Intégration de données hétérogènes et detection de liens/conflits d'intérêt Système ConnectionLens

Ioana Manolescu
CEDAR team
Inria Saclay-Île-de-France, Institut Polytechnique de Paris
@ioanamanol @cedarinrialix

Equipe CEDAR, Inria et Ecole polytechnique

Inria: Institut national de recherche en informatique et automatique, depuis 1976



Ecole Polytechnique, depuis 1794

Plan

- **Notre domaine** : la gestion de données
- **Projet de recherche** (2018--) : intégration de données très hétérogènes sans schéma préétabli, notamment pour le journalisme
- **Application** : conflits d'intérêt dans le domaine biomédical
- **Démo : ConnectionStudio** sur des données HATVP

Plus d'informations:
https://connectionstudio.inria.fr
https://teams.inria.fr/cedar

La gestion des données: de l'art à l'industrie

Tout code informatique manipule des données

- Entrées, sorties, structures auxiliaires

Les données sont volumineuses, importantes →
Code dédié à la gestion des données (créer / modifier / trouver / effacer)

- Un code pour pour les données sur les contribuables…
- Un code pour les données sur les étudiants / cours / propriétés immobiliaires etc.
  → Effort multiplié pour écrire, maintenir, améliorer ces codes

Problème résolu depuis 1970: les systèmes de gestion des bases de données !

La gestion des données: de l'art à l'industrie

Les **systèmes** (logiciels) **de gestion des bases de données (SGBD):**

- Encapsulent des fonctionnalités nécessaires pour la gestion efficace des données
- Fournissent aux utilisateurs des interfaces simples pour mettre en oeuvre de multiples applications de gestion de données sans jamais devoir coder.
  - 1 SGBD, N applications: 1. contribuables; 2. étudiants; 3. propriétés immobiliaires…

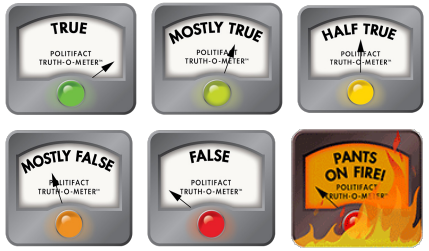
SGBDs omniprésents dans toute société moderne: comptes bancaires, paiements CB, badges d'accès, toute réservation de billet, système de santé, impôts, etc. Industrie de MD$/an

Exige que les données soient relationelles (dans des tableaux, qui peuvent se référencer)

Fausses nouvelles et leur propagation sur Twitter (1)

Vérifications (fact-checks) mises en ligne: des documents (non pas tableaux) contenant

- Lien vers une affirmation (media, réseau social etc.), auteur
- Analyse, conclusion, auteur, date, institution



Parmi les premiers corpus de vérifications publiés:
https://www.lemonde.fr/webservice/decodex/updates

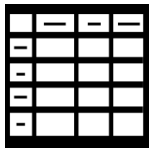
Puis **ClaimReview** (format structuré), par Google
https://www.claimreviewproject.com/

Fausses nouvelles et leur propagation sur Twitter (2)

Vérifications (fact-checks) mises en ligne: des documents (non pas tableaux)

- Lien vers une affirmation (media, réseau social etc.), auteur
- Analyse, conclusion, auteur, date, institution

Base des Décodeurs (Excel): figures publiques (Parlement, Gouvernement, etc.)



- Nom, prénom, ID twitter, poste, parti

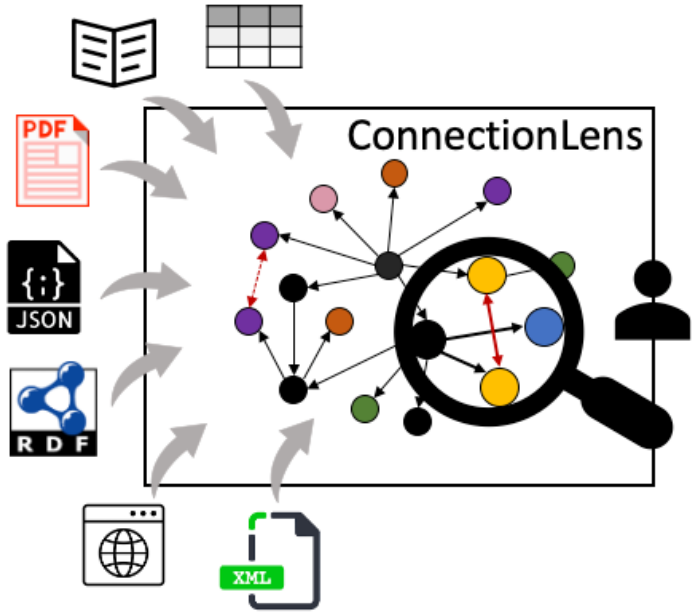
Question: Quand est-ce qu'une fausse info est propagée la première fois dans une certaine communauté (p.ex., Parlement)?

?

- Trouver des tweets connectés à un auteur de faux, et à une personnalité politique
- Toute structure possible pour les deux connections: écrit/rediffuse/aime/suit…

ConnectionLens: intégration de sources de données hétérogènes dans des graphes (réseaux) de donnés

https://team.inria.fr/cedar/connectionlens/

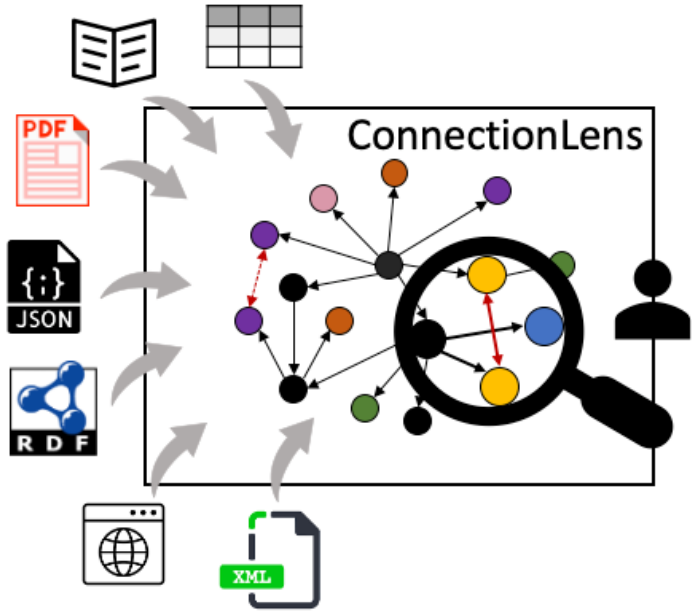


ConnectionLens: intégration de sources de données hétérogènes dans des graphes (réseaux) de donnés

https://team.inria.fr/cedar/connectionlens/

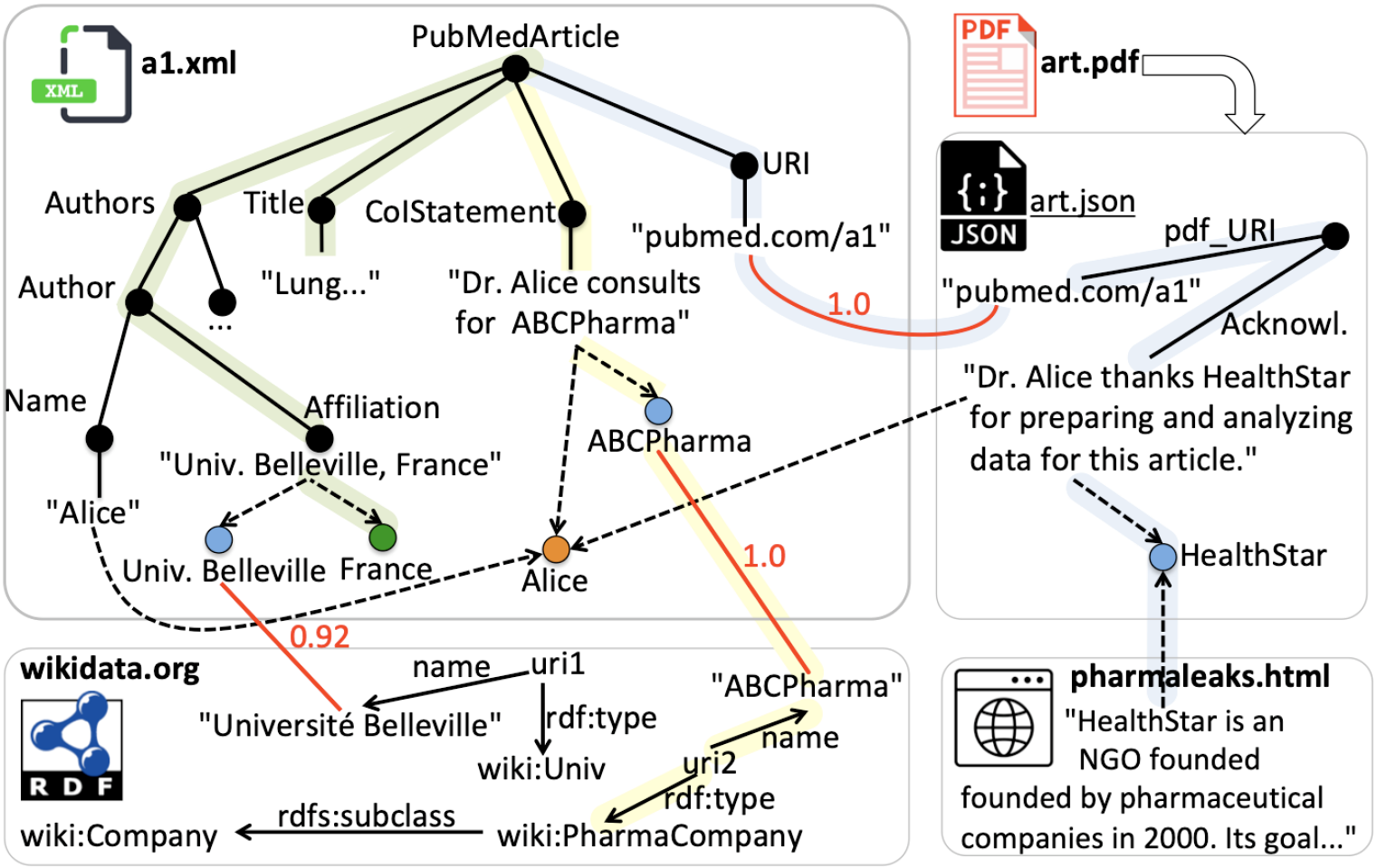
Données :

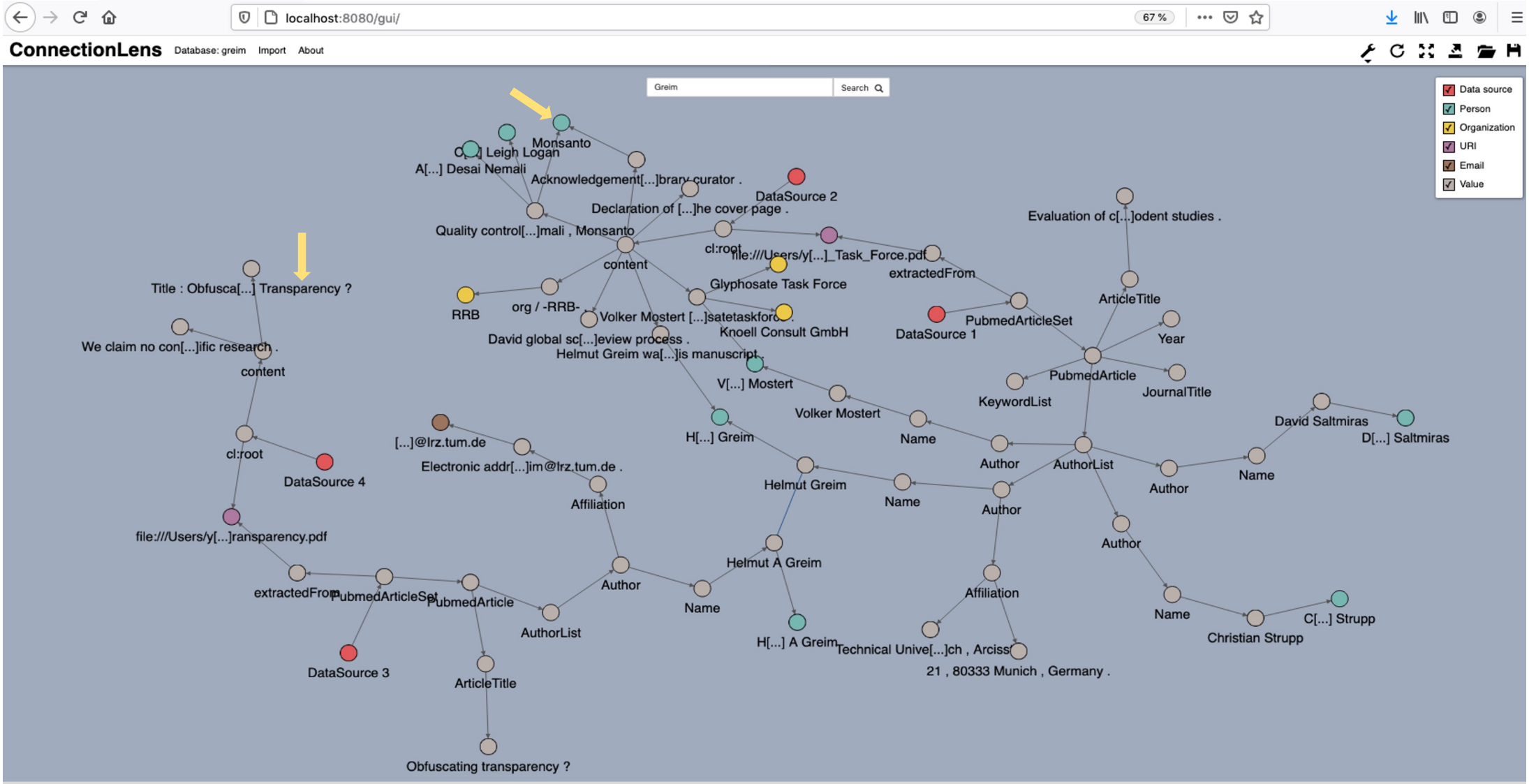
- Pages Web
- Documents (PDF, Office, …)
- Données (CSV, RDF, JSON, XML…) notamment Open Data, réseaux sociaux
- Bases de données structurées…



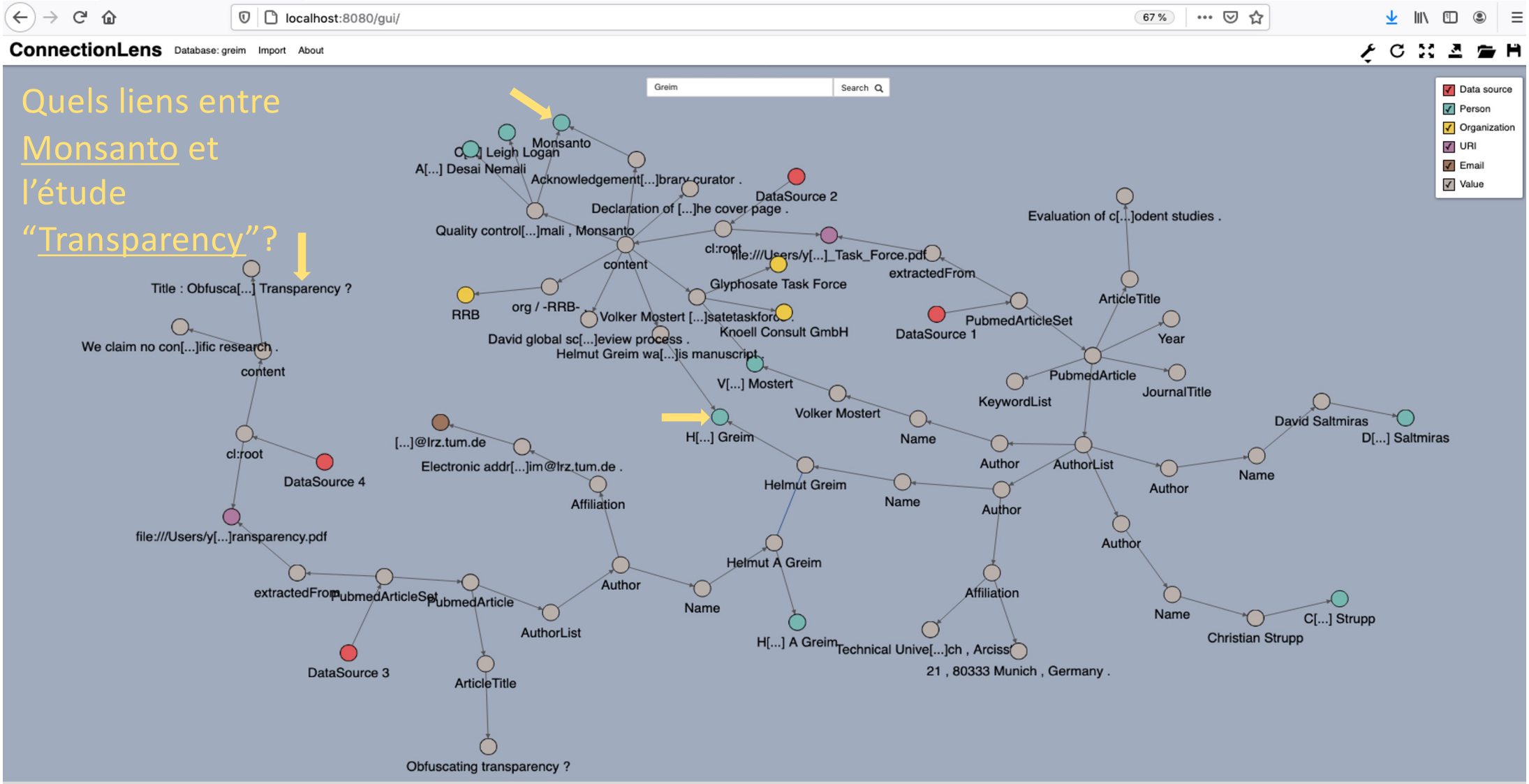
Enrichies avec:

- Des entités nommées extraites automatiquement depuis les données, à l'aide de modèles entraînées (IA)
- Personnes, lieux, organisations, emails, hashtags, dates…
- **Liens** entre des entités qui sont probablement les mêmes malgré des différentes notations


a1.xml
XML
PubMedArticle
Authors
Title
CoIStatement
URI
"pubmed.com/a1"
Author
...
"Lung..."
"Dr. Alice consults for ABCPharma"
1.0
Name
Affiliation
"Univ. Belleville, France"
ABCPharma
"Alice"
Univ. Belleville
France
Alice
1.0
0.92
PDF
art.pdf
art.json
JSON
pdf_URI
"pubmed.com/a1"
Acknowl.
"Dr. Alice thanks HealthStar for preparing and analyzing data for this article."
HealthStar
wikidata.org
R D F
name
uri1
"Université Belleville"
rdf:type
wiki:Univ
"ABCPharma"
name
uri2
rdf:type
rdfs:subclass
wiki:Company
wiki:PharmaCompany
pharmaleaks.html
"HealthStar is an NGO founded founded by pharmaceutical companies in 2000. Its goal..."



Merci à S. Horel (Le Monde)



Merci à S. Horel (Le Monde)

Questions / démo

ConnectionLens: https://team.inria.fr/cedar/connectionlens/

ConnectionStudio: https://connectionstudio.inria.fr

CEDAR team: https://teams.inria.fr/cedar