# Data integration research for data journalism


## Ioana Manolescu

**Inria and Institut Polytechnique de Paris**

htttps://pages.saclay.inria.fr/ioana.manolescu,
@ioanamanol

# Who we are

**Computer scientists** (research staff, faculty, engineer)

- Employed by **public organizations** (Inria, FR national research institute in CS, and Ecole Polytechnique, FR engineering school), U. Lisbon
- Many students (M1, M2, PhD)

Research in **data management**

Since 2012, inspired by **data journalism** and **computational fact-checking**

ACM SIGMOD 2013

**Fact Checking and Analyzing the Web**

François Goasdoué[1] Konstantinos Karanasos[2]* Yannis Katsis[3]
Julien Leblay[1] Ioana Manolescu[1] Stamatis Zampetakis[1]

[1]OAK team, Inria Saclay & LRI, Université Paris-Sud, Orsay, France
[2]IBM Almaden Research Center, San Jose, CA
[3]UCSD Database group & WebDam project, Inria Saclay, San Diego, CA

firstname.lastname@inria.fr

**ABSTRACT**

*Fact checking* and *data journalism* are currently strong trends. The sheer amount of data at hand makes it difficult even for trained professionals to spot biased, outdated or simply incorrect information. We propose to demonstrate FactMinder, a fact checking and analysis assistance application. SIGMOD attendees will be able to analyze documents using FactMinder and experience how background knowledge and open data repositories help build insightful overviews of current topics.

of traffic on asthma cases) can comb the Web for bits of information, connect, interpret, annotate and re-share them. Such data gathering and fact checking have come at the core of "data journalism", pioneered, e.g., in Europe by The Guardian[2] and growing through efforts such as FactCheck[3], Politifact[4], and similar French sites[5].

*Fact checking and analysis* (*FCA*, for short), viewed as the process of analyzing a piece of information, crossing it with existing knowledge, verifying its accuracy and possibly enriching it with nuances, comments and connections to reputable sources, has an inherent part of human effort, thus

# Our collaboration with Le Monde

**Google Award in Computational Journalism** (2014-2015), with U. Paris Sud, mostly data viz

**ANR ContentCheck**: Models, Algorithms and Tools for Data Journalism and Journalistic Fact-Checking (2015-2020), https://contentcheck.inria.fr
700 K€, w/ U. Paris Sud, U. Rennes, U. Lyon, and Les Décodeurs (fact-checking team from Le Monde):
Samuel Laurent, Maxime Ferrer, Adrien Sénécat

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

LES DÉCODEURS
VENONS-EN AUX FAITS

**AI Chair SourcesSay**: Intelligent Data Analysis and Interconnection in Digital Arenas (2020-2024), https://sourcessay.inria.fr
600 K€; Le Monde (Stéphane Horel) and WeDoData (Karine Bastien) as non-funded, supporting partners



Inria

# Improving access to digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text

**Créations d'entreprises dans quelques pays de l'Union européenne en 2015**

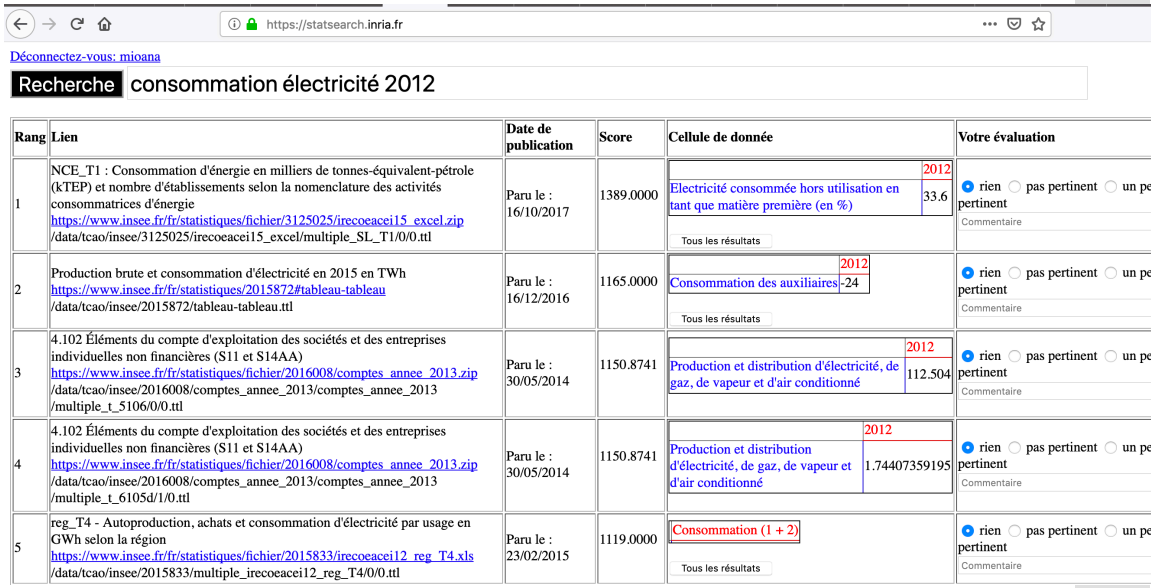
*en %*

| Pays | Taux de création |
|---|---|
| Allemagne | 7,1 |
| Belgique | 6,2 |
| Espagne | 9,5 |
| **France (1)** | **9,5** |
| Italie | 7,5 |
| Pays-Bas | 10,1 |
| Portugal | 15,7 |
| République tchèque | 8,2 |
| Royaume-Uni | 14,3 |

# Improving access to digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text

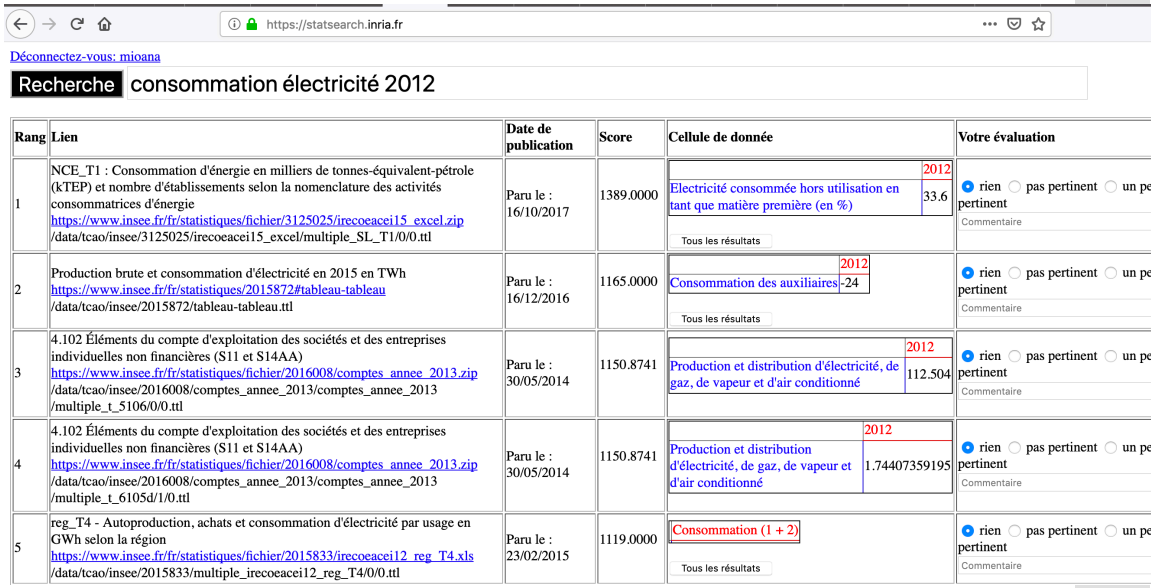



| Rang | Lien | Date de publication | Score | Cellule de donnée | Votre évaluation |
|---|---|---|---|---|---|
| 1 | NCE_T1 : Consommation d'énergie en milliers de tonnes-équivalent-pétrole (kTEP) et nombre d'établissements selon la nomenclature des activités consommatrices d'énergie https://www.insee.fr/fr/statistiques/fichier/3125025/irecoeacei15_excel.zip /data/tcao/insee/3125025/irecoeacei15_excel/multiple_SL_T1/0/0.ttl | Paru le : 16/10/2017 | 1389.0000 | 2012<br>Electricité consommée hors utilisation en tant que matière première (en %) 33.6<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |
| 2 | Production brute et consommation d'électricité en 2015 en TWh https://www.insee.fr/fr/statistiques/2015872#tableau-tableau /data/tcao/insee/2015872/tableau-tableau.ttl | Paru le : 16/12/2016 | 1165.0000 | 2012<br>Consommation des auxiliaires -24<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |
| 3 | 4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_5106/0/0.ttl | Paru le : 30/05/2014 | 1150.8741 | 2012<br>Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 112.504 | rien pas pertinent un pe pertinent<br>Commentaire |
| 4 | 4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_6105d/1/0.ttl | Paru le : 30/05/2014 | 1150.8741 | 2012<br>Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 1.74407359195 | rien pas pertinent un pe pertinent<br>Commentaire |
| 5 | reg_T4 - Autoproduction, achats et consommation d'électricité par usage en GWh selon la région https://www.insee.fr/fr/statistiques/fichier/2015833/irecoeacei12_reg_T4.xls /data/tcao/insee/2015833/multiple_irecoeacei12_reg_T4/0/0.ttl | Paru le : 23/02/2015 | 1119.0000 | Consommation (1 + 2)<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |

# Improving access to digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text



https://statsearch.inria.fr

Déconnectez-vous: mioana

Recherche consommation électricité 2012

| Rang | Lien | Date de publication | Score | Cellule de donnée | Votre évaluation |
|---|---|---|---|---|---|
| 1 | NCE_T1 : Consommation d'énergie en milliers de tonnes-équivalent-pétrole (kTEP) et nombre d'établissements selon la nomenclature des activités consommatrices d'énergie https://www.insee.fr/fr/statistiques/fichier/3125025/irecoeacei15_excel.zip /data/tcao/insee/3125025/irecoeacei15_excel/multiple_SL_T1/0/0.ttl | Paru le : 16/10/2017 | 1389.0000 | 2012<br>Electricité consommée hors utilisation en tant que matière première (en %) 33.6<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |
| 2 | Production brute et consommation d'électricité en 2015 en TWh https://www.insee.fr/fr/statistiques/2015872#tableau-tableau /data/tcao/insee/2015872/tableau-tableau.ttl | Paru le : 16/12/2016 | 1165.0000 | 2012<br>Consommation des auxiliaires -24<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |
| 3 | 4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_5106/0/0.ttl | Paru le : 30/05/2014 | 1150.8741 | 2012<br>Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 112.504 | rien pas pertinent un pe pertinent<br>Commentaire |
| 4 | 4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_6105d/1/0.ttl | Paru le : 30/05/2014 | 1150.8741 | 2012<br>Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 1.74407359195 | rien pas pertinent un pe pertinent<br>Commentaire |
| 5 | reg_T4 - Autoproduction, achats et consommation d'électricité par usage en GWh selon la région https://www.insee.fr/fr/statistiques/fichier/2015833/irecoeacei12_reg_T4.xls /data/tcao/insee/2015833/multiple_irecoeacei12_reg_T4/0/0.ttl | Paru le : 23/02/2015 | 1119.0000 | Consommation (1 + 2)<br>Tous les résultats | rien pas pertinent un pe pertinent<br>Commentaire |

Follow-up: question answering on SDMX databases

# A platform for integrating heterogeneous data sources

Data sources produced by independent, non-coordinating authors

Data in: PDF, HTML, text, CSV, JSON, XML, XLS, RDF, relational databases...

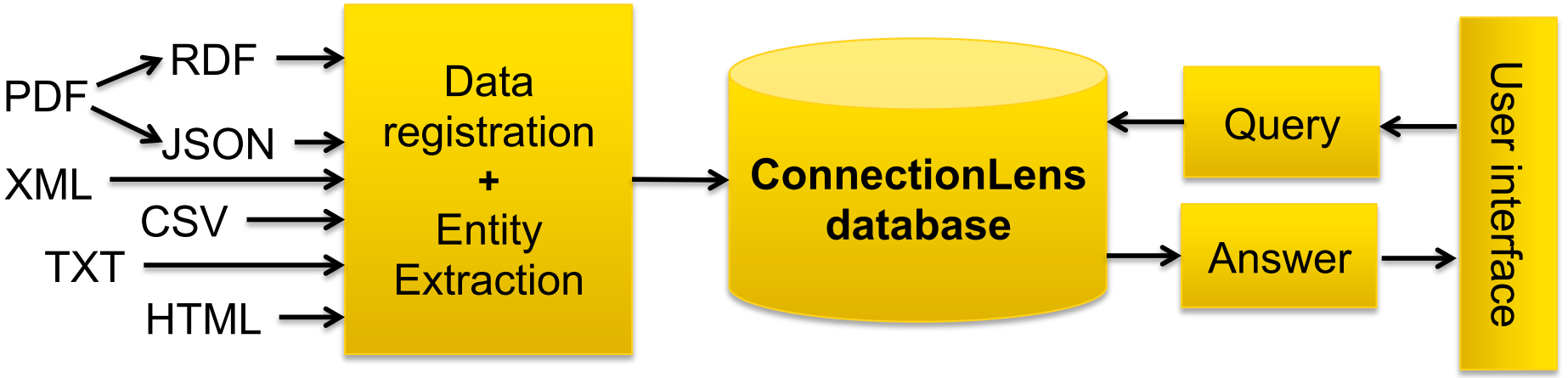
- Most **structured**: relational > CSV > JSON, XML > RDF > XLS
- Least **structured**: PDF < text < HTML

Already there: one or more systems/tools for each data format

- Need to learn the language, possibly convert the data

How to **find answers across different data sources?**

# The ConnectionLens platform




+ provenance

**Entities**: Person, Location, Organization, URI, email, hashtag

1. PostgreSQL
2. In-memory DB

# Using ConnectionLens

1. **Select data sources** (possibly w/ provenance)
   We also work on some *crawlers* to acquire sources
2. **Ingest the sources** in a ConnectionLens database
3. [Optional] Copy & host the database on a different server
4. Specifies **queries** (questions):
   a. as set of keywords: CL finds connections between them
   b. "known queries", e.g., find all organizations from which Dr. Helmut Greim has received funding

# Building a CoI database in ConnectionLens

**Data sources** signaled by SH:

- **XML** publication records from PubMed
- **PDF** articles (for now: those in open access from PubMed)
- Web sites (= sets of **HTML** sources) where scientific experts and organizations are described

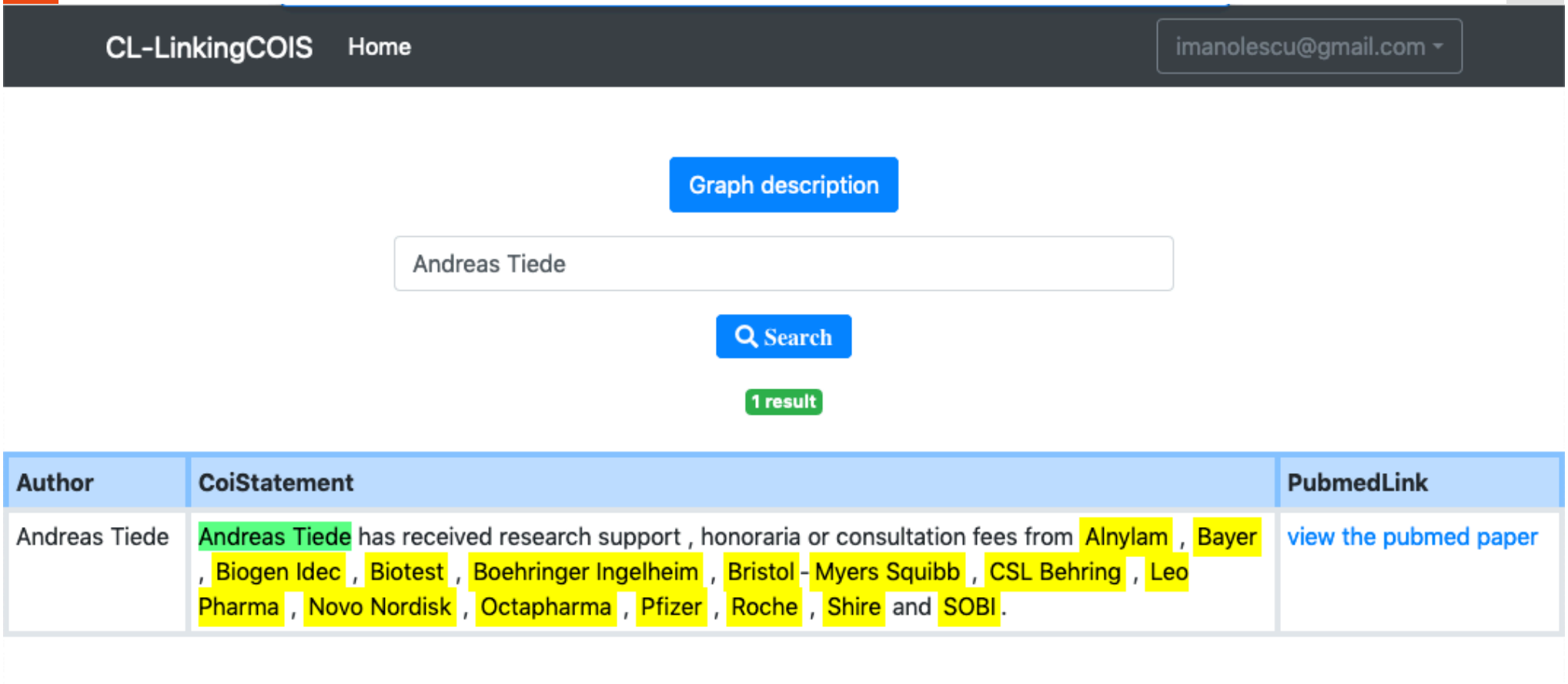
**Custom query interface:**

aut:Eva Brand org:Takeda

Search

| Author | CoiStatement | PubmedLink |
|---|---|---|
| Eva Brand | Disclosures Malte Lenders and Eva Brand received speaker fees and research grants from Takeda, Sanofi Genzyme, and Amicus Therapeutics. None of the funding companies had a role in writing or submission of this manuscript. | view the pubmed paper |

# Building a CoI database in ConnectionLens




| Author | CoiStatement | PubmedLink |
|---|---|---|
| Andreas Tiede | Andreas Tiede has received research support , honoraria or consultation fees from Alnylam , Bayer , Biogen Idec , Biotest , Boehringer Ingelheim , Bristol - Myers Squibb , CSL Behring , Leo Pharma , Novo Nordisk , Octapharma , Pfizer , Roche , Shire and SOBI . | view the pubmed paper |

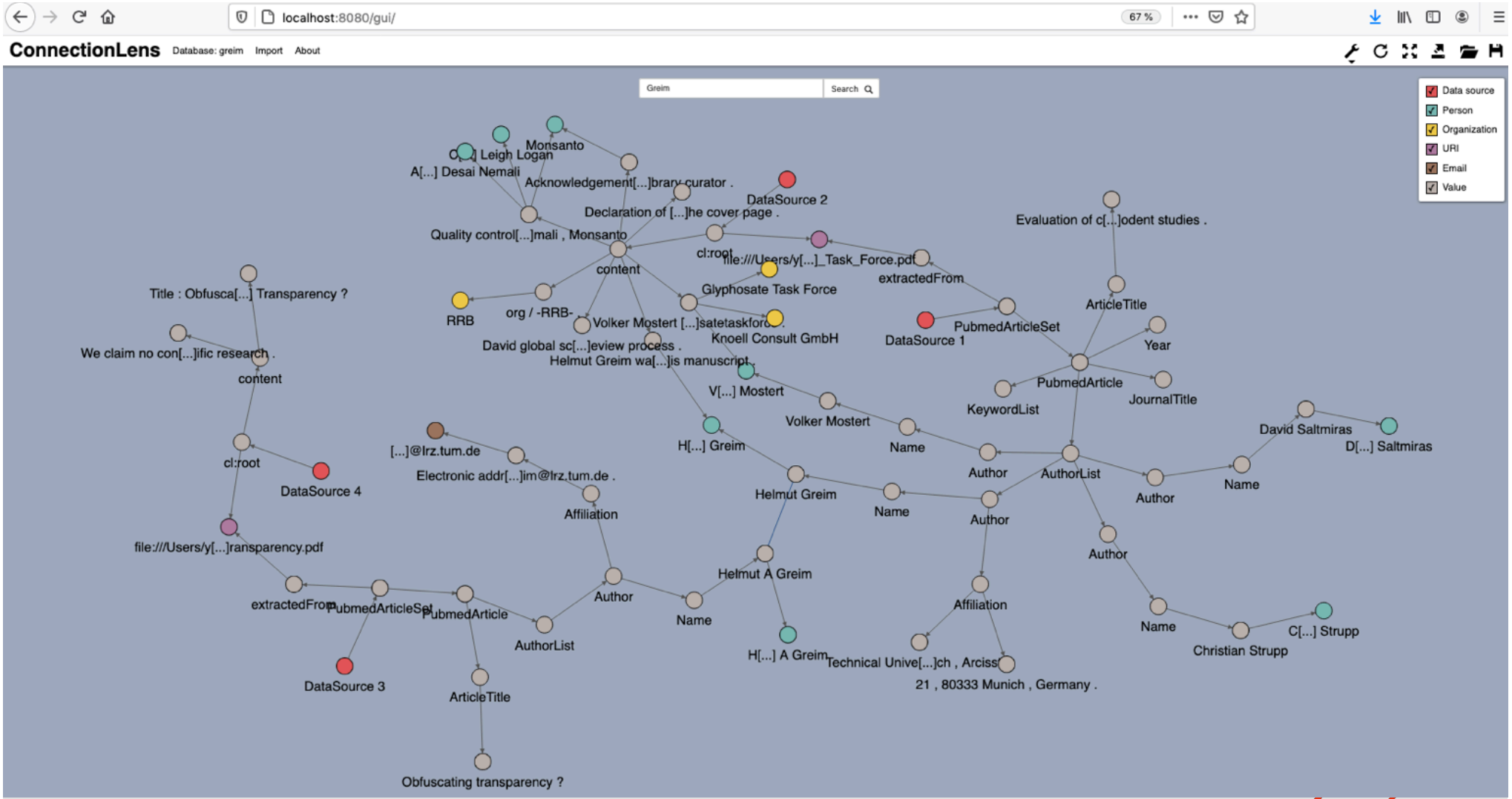
# Useful links

**ContentCheck (ANR, 2015-2020)**
https://contentcheck.inria.fr

**ConnectionLens**: https://gitlab.inria.fr/cedar/connectionlens

Most recent/detailed **ConnectionLens papers**:
https://arxiv.org/abs/2007.12488, https://arxiv.org/abs/2009.04283

**SourcesSay (ANR + DGA, 2020-2024)**
https://sourcessay.inria.fr


localhost:8080/gui/
67 %
ConnectionLens
Database: greim
Import
About
Greim
Search
Data source
Person
Organization
URI
Email
Value
Monsanto
Leigh Logan
A[...] Desai Nemali
Acknowledgement[...]brary curator .
DataSource 2
Declaration of [...]he cover page .
Quality control[...]mali , Monsanto
cl:root
file:///Users/y[...]_Task_Force.pdf
extractedFrom
content
Glyphosate Task Force
Title : Obfusca[...] Transparency ?
RRB
org / -RRB-
Volker Mostert [...]satetaskforce .
Knoell Consult GmbH
David global sc[...]eview process .
Helmut Greim wa[...]is manuscript .
DataSource 1
PubmedArticleSet
Evaluation of c[...]odent studies .
ArticleTitle
Year
We claim no con[...]ific research .
content
V[...] Mostert
Volker Mostert
PubmedArticle
JournalTitle
KeywordList
David Saltmiras
D[...] Saltmiras
[...]@lrz.tum.de
H[...] Greim
Name
Author
AuthorList
Author
Name
cl:root
DataSource 4
Electronic addr[...]im@lrz.tum.de .
Affiliation
Helmut Greim
Name
Author
file:///Users/y[...]ransparency.pdf
Author
Helmut A Greim
extractedFrom
PubmedArticleSet
PubmedArticle
Author
Name
Affiliation
Name
C[...] Strupp
AuthorList
H[...] A Greim
Technical Unive[...]ch , Arciss
Christian Strupp
DataSource 3
ArticleTitle
21 , 80333 Munich , Germany .
Obfuscating transparency ?