All your digital data sources in one place: the SourcesSay ANR/DGA project

https://sourcessay.inria.fr

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

Why should computer scientists focus on tools for journalists?

Bad memories: Romania, 1989

Bad memories: Romania, 1989

Ceauşescu re-elected at the 14th congress!

Bad memories: Romania, 1989

Ceauşescu re-elected at the 14th congress!

He had held power since 1965.

Bad memories: Romania, 1985

1990: things got better

... kind of

1000 dead (approx.)
No one convicted.

Democratic societies crucially need the press

- ❑ To debate and express dissent

Socialist Romania, 1984

- ❑ To analyze, confirm or refute public statements

Fact-checking

(Data) journalism

- ❑ To expose and explain society functioning

Democratic societies crucially need the press

- ❑ To debate and express dissent

Socialist Romania, 1984

- ❑ To analyze, confirm or refute public statements

Fact-checking

Open-source intelligence

- ❑ To expose and explain society functioning

bellingcat

Information, misinformation, disinformation and data

How do we analyze statements?

A statement is made

How do we analyze statements?

A statement is made

What should we think of this?

How do we analyze statements?

A statement is made

What should we think of this?

Information?
Misinformation?
Dysinformation?

How do we analyze statements?

A statement is made

What should we think of this?

Information?
Misinformation?
Dysinformation?

Evaluate
source authority
(past statements…)

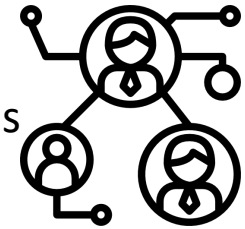
How do we analyze statements?

A statement is made

What should we think of this?

Information?
Misinformation?
Dysinformation?

Evaluate source connections



Evaluate source authority (past statements…)

How do we analyze statements?

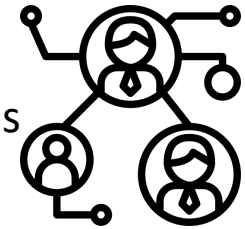
A statement is made

Consult reference data

What should we think of this?

Information?
Misinformation?
Dysinformation?

Evaluate source connections



Evaluate source authority (past statements…)

How do we analyze statements?

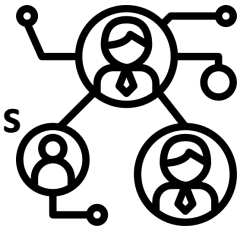
A statement is made

Consult **reference data**

What should we think of this?

Information?
Misinformation?
Dysinformation?

Evaluate **source connections**



Evaluate **source authority (past statements…)**

How do we analyze statements?

A statement is made

Information?
Misinformation?
Dysinformation?

source connections

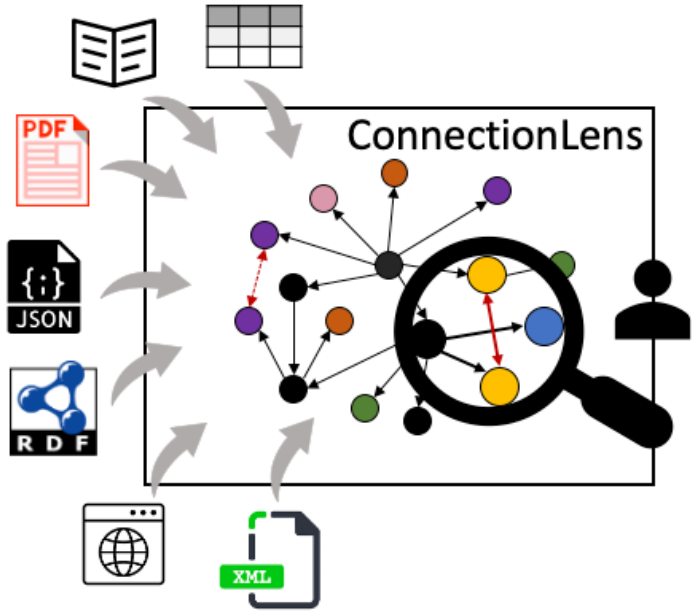
source past statements

fact-checks

reference data

ConnectionLens: graph-based integration of heterogeneous data sources

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

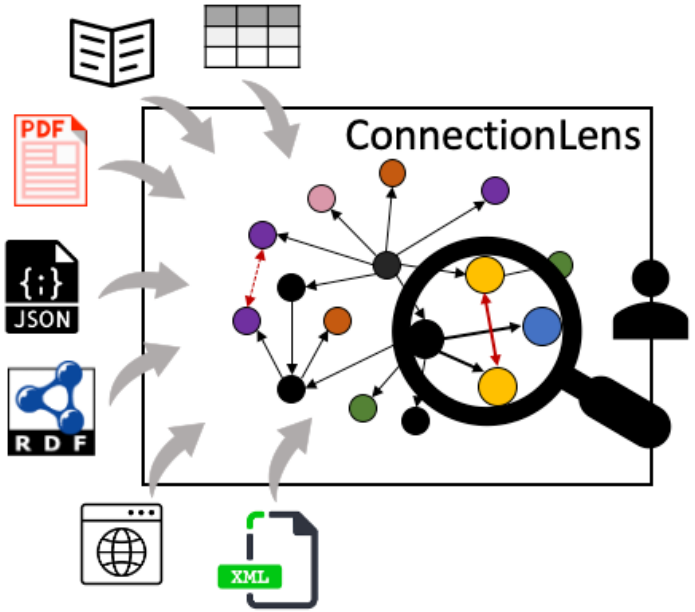


ConnectionLens: graph-based integration of heterogeneous data sources

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

Data ingestion:

- Web pages
- Documents
- Data:
  - Social media content
  - (Semi)structured data: CSV, RDF, XML, JSON…

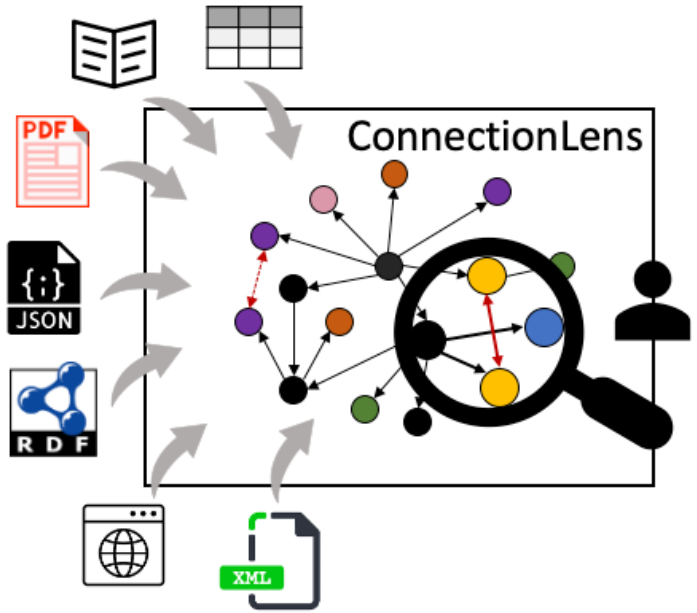


ConnectionLens: graph-based integration of heterogeneous data sources

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

Data ingestion:

- Web pages
- Documents
- Data:
  - Social media content
  - (Semi)structured data: CSV, RDF, XML, JSON…



Data enrichment:

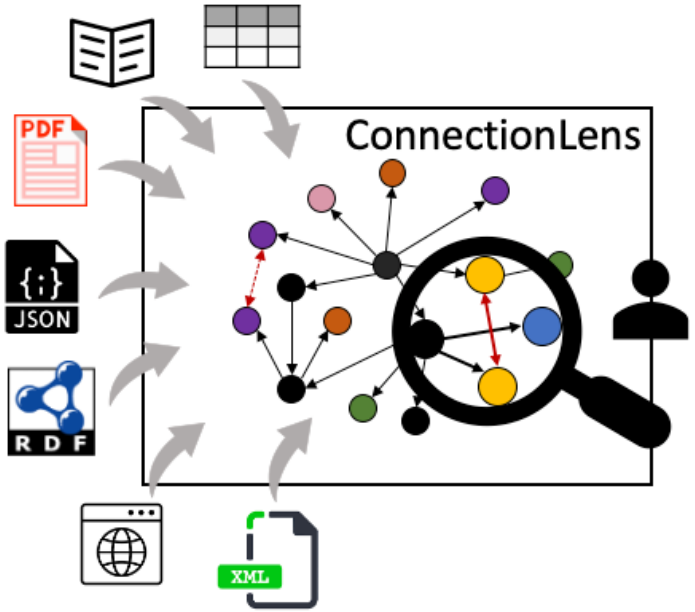
- Extraction of **entities** from text (people, places, dates, organizations, emails, URIs, …)
- Extraction of **relationships** from text

ConnectionLens: graph-based integration of heterogeneous data sources

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

Data ingestion:

- Web pages
- Documents
- Data:
  - Social media content
  - (Semi)structured data: CSV, RDF, XML, JSON…



Data enrichment:

- Extraction of **entities** from text (people, places, dates, organizations, emails, URIs, …)
- Extraction of **relationships** from text
- Disambiguation of entity names based on context, e.g., « Hollande »

ConnectionLens: graph-based integration of heterogeneous data sources

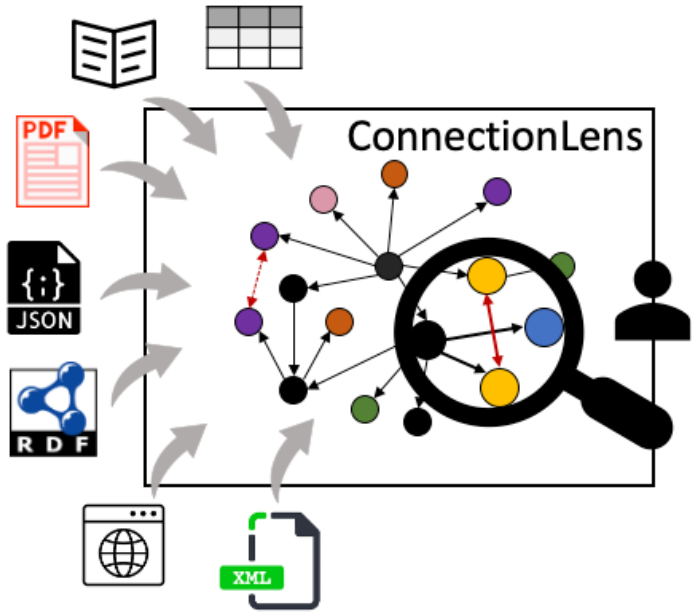
https://team.inria.fr/cedar/connectionlens/

Data ingestion:

- ❑ Web pages
- ❑ Documents
- ❑ Data

Data enrichment:

- ❑ Entity extraction
- ❑ Relationship extraction
- ❑ Disambiguation



Data exploitation (1):

- ❑ I don't know much about « data formats ». What does this dataset contain?

ConnectionLens: graph-based integration of heterogeneous data sources

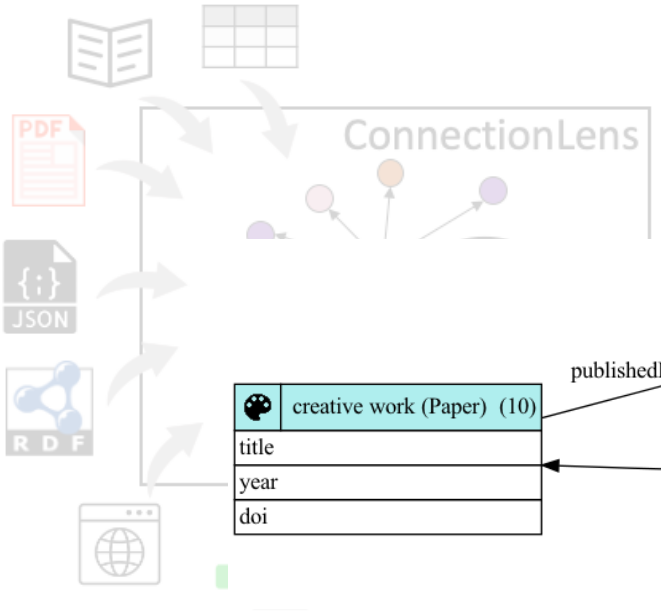
https://team.inria.fr/cedar/connectionlens/

Data ingestion:

- ❑ Web pages
- ❑ Documents
- ❑ Data

Data enrichment:

- ❑ Entity extraction
- ❑ Relationship extraction
- ❑ Disambiguation



Data exploitation (1):

- ❑ I don't know much about « data formats ». What does this dataset contain?

ConnectionLens: graph-based integration of heterogeneous data sources

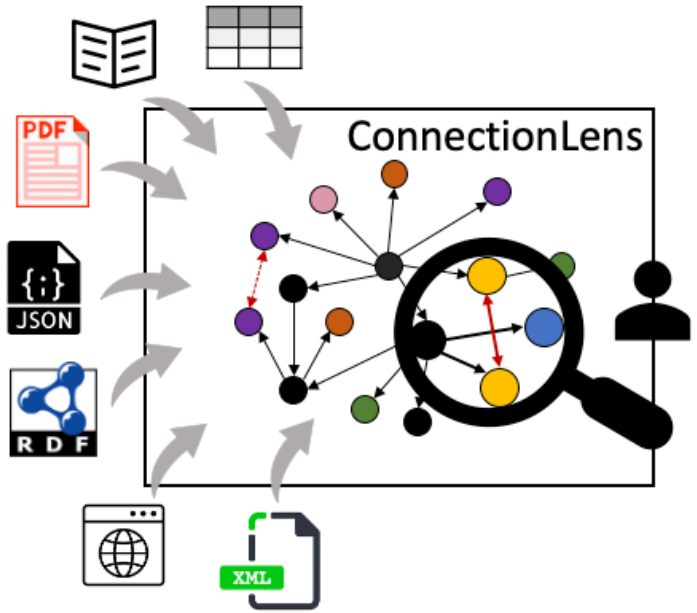
https://team.inria.fr/cedar/connectionlens/

Data ingestion:

- ❑ Web pages
- ❑ Documents
- ❑ Data

Data enrichment:

- ❑ Entity extraction
- ❑ Relationship extraction
- ❑ Disambiguation



Data exploitation (1):

- ❑ I don't know much about « data formats ». What does this dataset contain?
- ❑ Entity search, statistics, co-occurrence
  - ❑ « Which organizations (dates, people, …) are found in connection with A.Kohler? »
- ❑ Connections between entities
  - ❑ « How is A.Kohler connected to the STX and MSC companies? »

ConnectionLens: graph-based integration of heterogeneous data sources

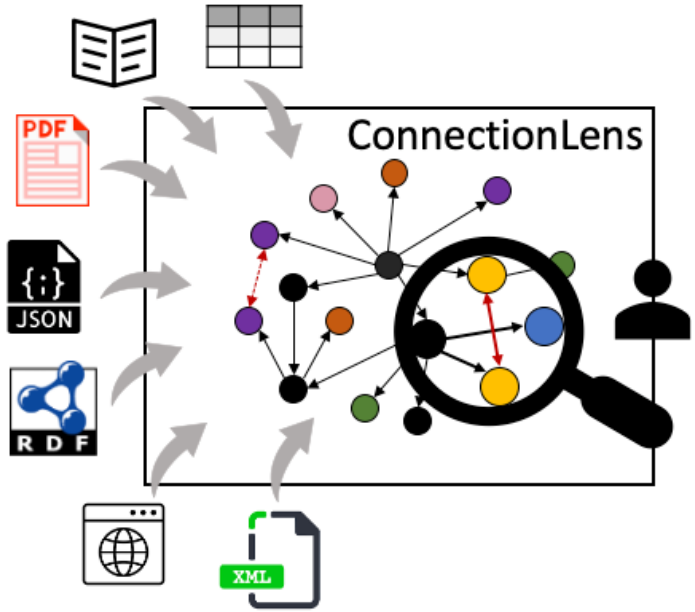
https://team.inria.fr/cedar/connectionlens/

Data ingestion:

- ❑ Web pages
- ❑ Documents
- ❑ Data

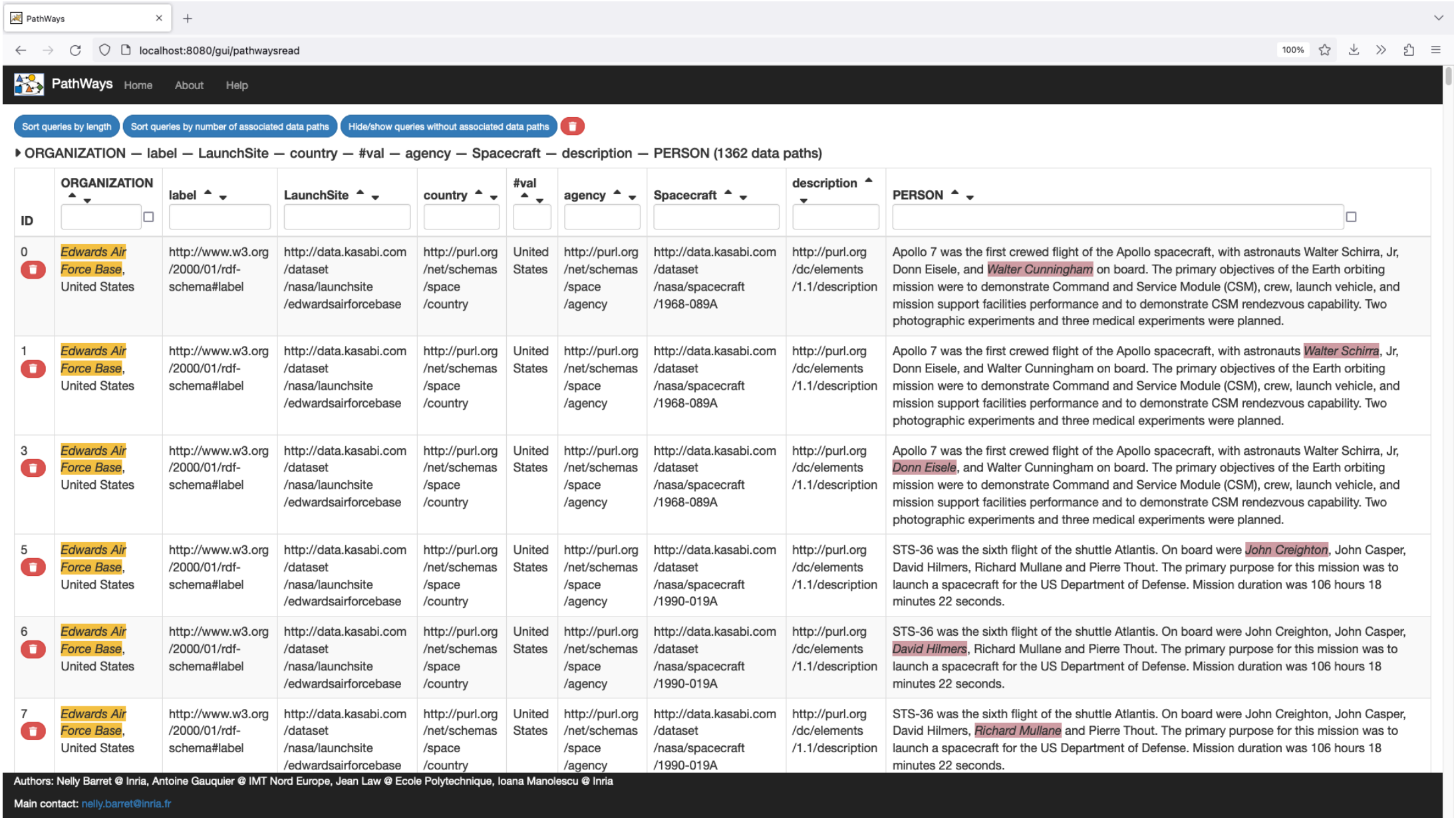
Data enrichment:

- ❑ Entity extraction
- ❑ Relationship extraction
- ❑ Disambiguation



Data exploitation (2):

- ❑ What are the connections between people and organizations in this dataset?
  - ❑ Which people (organizations)?
  - ❑ Which connections? (« worked for », « funded by »…)


PathWays
localhost:8080/gui/pathwaysread
100%
PathWays Home About Help
Sort queries by length
Sort queries by number of associated data paths
Hide/show queries without associated data paths
ORGANIZATION — label — LaunchSite — country — #val — agency — Spacecraft — description — PERSON (1362 data paths)
ID
ORGANIZATION
label
LaunchSite
country
#val
agency
Spacecraft
description
PERSON
0
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1968-089A
http://purl.org/dc/elements/1.1/description
Apollo 7 was the first crewed flight of the Apollo spacecraft, with astronauts Walter Schirra, Jr, Donn Eisele, and Walter Cunningham on board. The primary objectives of the Earth orbiting mission were to demonstrate Command and Service Module (CSM), crew, launch vehicle, and mission support facilities performance and to demonstrate CSM rendezvous capability. Two photographic experiments and three medical experiments were planned.
1
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1968-089A
http://purl.org/dc/elements/1.1/description
Apollo 7 was the first crewed flight of the Apollo spacecraft, with astronauts Walter Schirra, Jr, Donn Eisele, and Walter Cunningham on board. The primary objectives of the Earth orbiting mission were to demonstrate Command and Service Module (CSM), crew, launch vehicle, and mission support facilities performance and to demonstrate CSM rendezvous capability. Two photographic experiments and three medical experiments were planned.
3
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1968-089A
http://purl.org/dc/elements/1.1/description
Apollo 7 was the first crewed flight of the Apollo spacecraft, with astronauts Walter Schirra, Jr, Donn Eisele, and Walter Cunningham on board. The primary objectives of the Earth orbiting mission were to demonstrate Command and Service Module (CSM), crew, launch vehicle, and mission support facilities performance and to demonstrate CSM rendezvous capability. Two photographic experiments and three medical experiments were planned.
5
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1990-019A
http://purl.org/dc/elements/1.1/description
STS-36 was the sixth flight of the shuttle Atlantis. On board were John Creighton, John Casper, David Hilmers, Richard Mullane and Pierre Thout. The primary purpose for this mission was to launch a spacecraft for the US Department of Defense. Mission duration was 106 hours 18 minutes 22 seconds.
6
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1990-019A
http://purl.org/dc/elements/1.1/description
STS-36 was the sixth flight of the shuttle Atlantis. On board were John Creighton, John Casper, David Hilmers, Richard Mullane and Pierre Thout. The primary purpose for this mission was to launch a spacecraft for the US Department of Defense. Mission duration was 106 hours 18 minutes 22 seconds.
7
Edwards Air Force Base, United States
http://www.w3.org/2000/01/rdf-schema#label
http://data.kasabi.com/dataset/nasa/launchsite/edwardsairforcebase
http://purl.org/net/schemas/space/country
United States
http://purl.org/net/schemas/space/agency
http://data.kasabi.com/dataset/nasa/spacecraft/1990-019A
http://purl.org/dc/elements/1.1/description
STS-36 was the sixth flight of the shuttle Atlantis. On board were John Creighton, John Casper, David Hilmers, Richard Mullane and Pierre Thout. The primary purpose for this mission was to launch a spacecraft for the US Department of Defense. Mission duration was 106 hours 18 minutes 22 seconds.
Authors: Nelly Barret @ Inria, Antoine Gauquier @ IMT Nord Europe, Jean Law @ Ecole Polytechnique, Ioana Manolescu @ Inria
Main contact: nelly.barret@inria.fr

ConnectionLens: graph-based integration of heterogeneous data sources

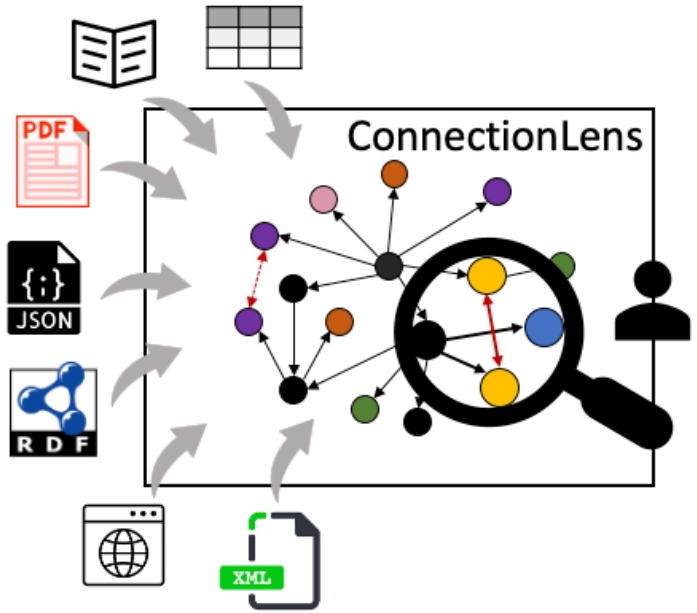
https://team.inria.fr/cedar/connectionlens/

Data ingestion:

- ❑ Web pages
- ❑ Documents
- ❑ Data

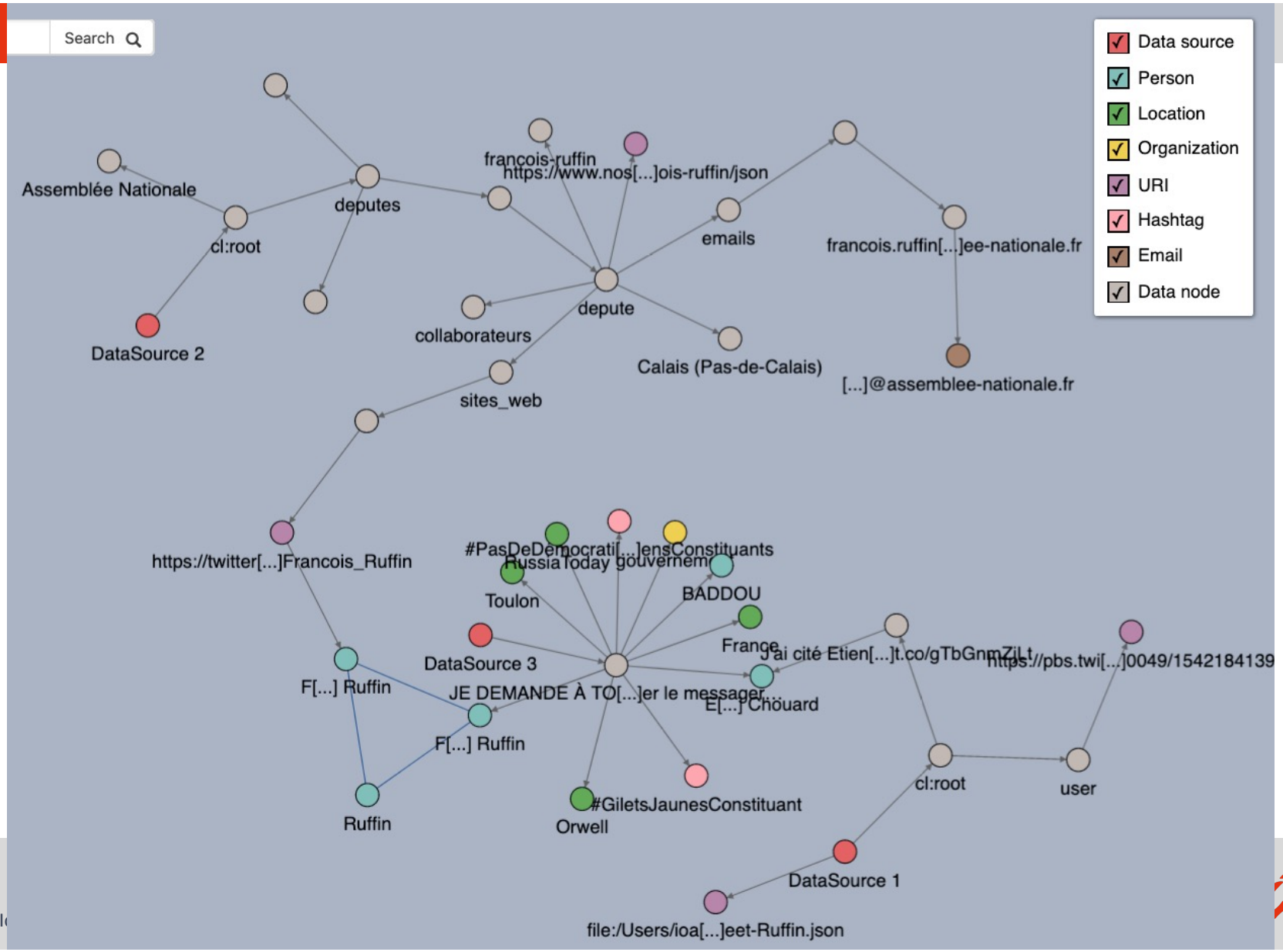
Data enrichment:

- ❑ Entity extraction
- ❑ Relationship extraction
- ❑ Disambiguation

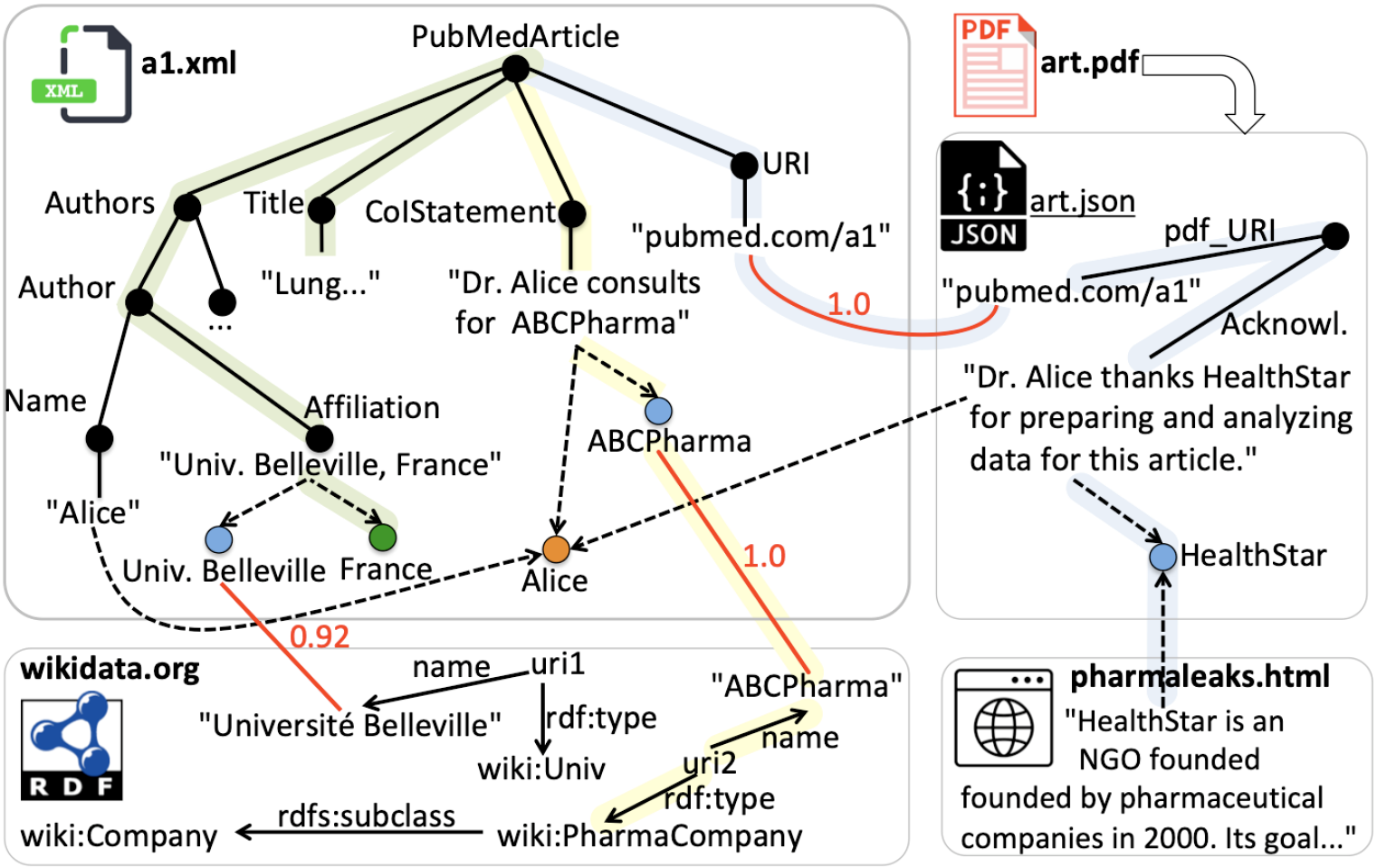


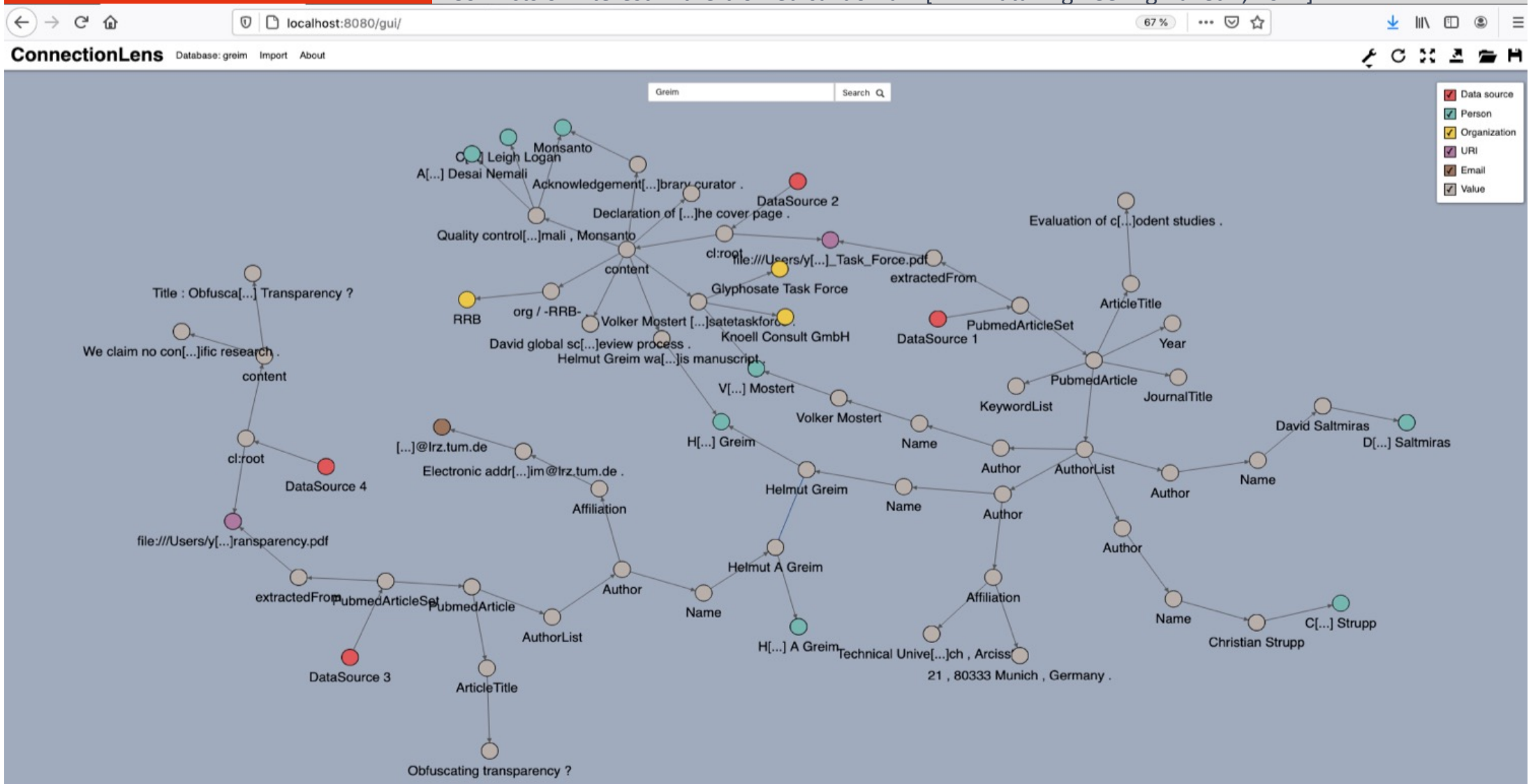
Data exploitation (2):

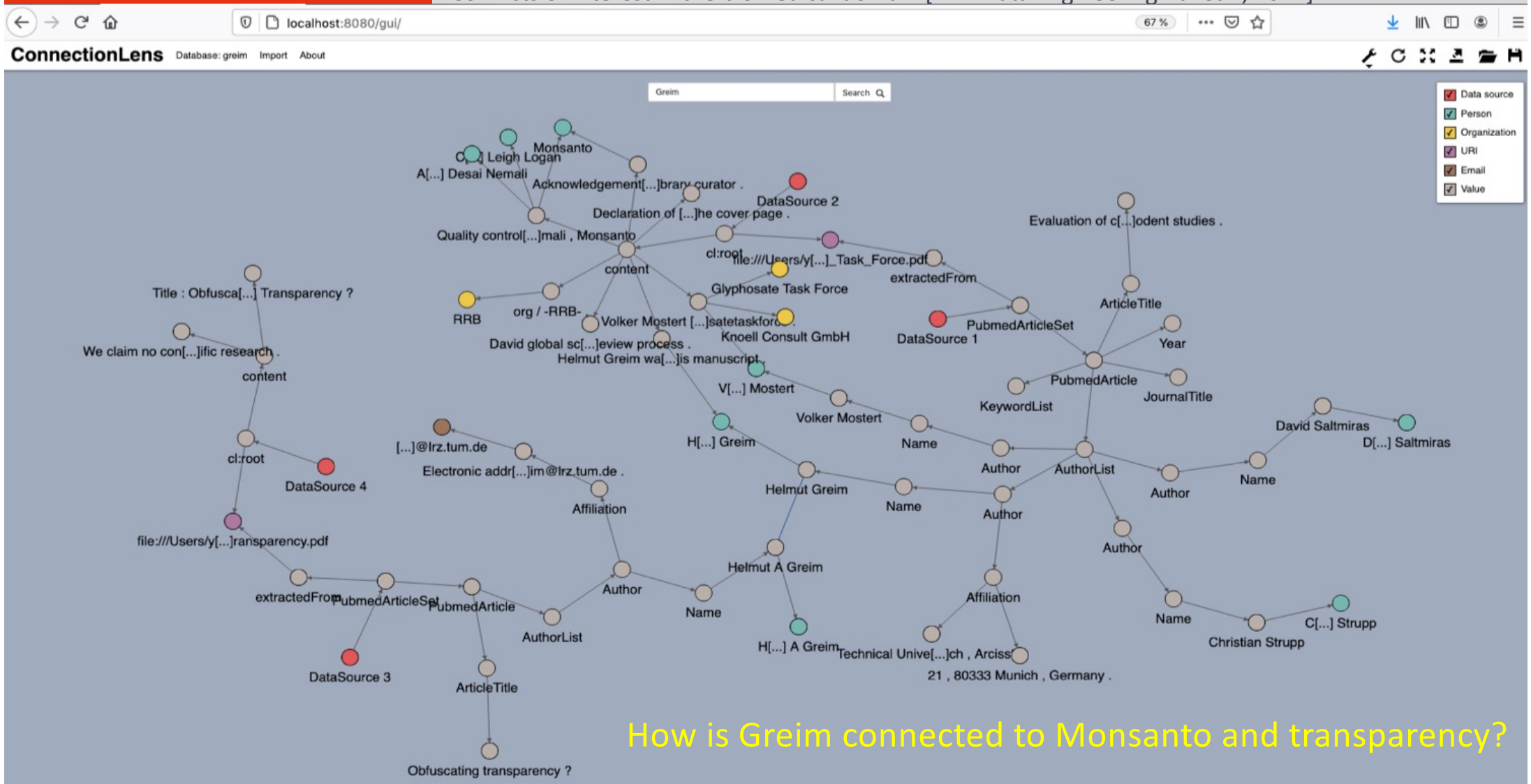
- ❑ What are the connections between people and organizations in this dataset?
- ❑ Interactive graph exploration
  - ❑ « What should I know about A.Kohler? »
- ❑ Structured and semistructured querying (requires database skills)


Search
Data source
Person
Location
Organization
URI
Hashtag
Email
Data node
Assemblée Nationale
deputes
francois-ruffin
https://www.nos[...]ois-ruffin/json
cl:root
emails
francois.ruffin[...]ee-nationale.fr
depute
collaborateurs
DataSource 2
Calais (Pas-de-Calais)
[...]@assemblee-nationale.fr
sites_web
https://twitter[...]Francois_Ruffin
#PasDeDemocrati[...]ensConstituants
RussiaToday gouvernem
Toulon
BADDOU
France
DataSource 3
J'ai cité Etien[...]t.co/gTbGnmZiLt
https://pbs.twi[...]0049/1542184139
F[...] Ruffin
JE DEMANDE À TO[...]er le messager...
E[...] Chouard
F[...] Ruffin
cl:root
user
Ruffin
#GiletsJaunesConstituant
Orwell
DataSource 1
file:/Users/ioa[...]eet-Ruffin.json

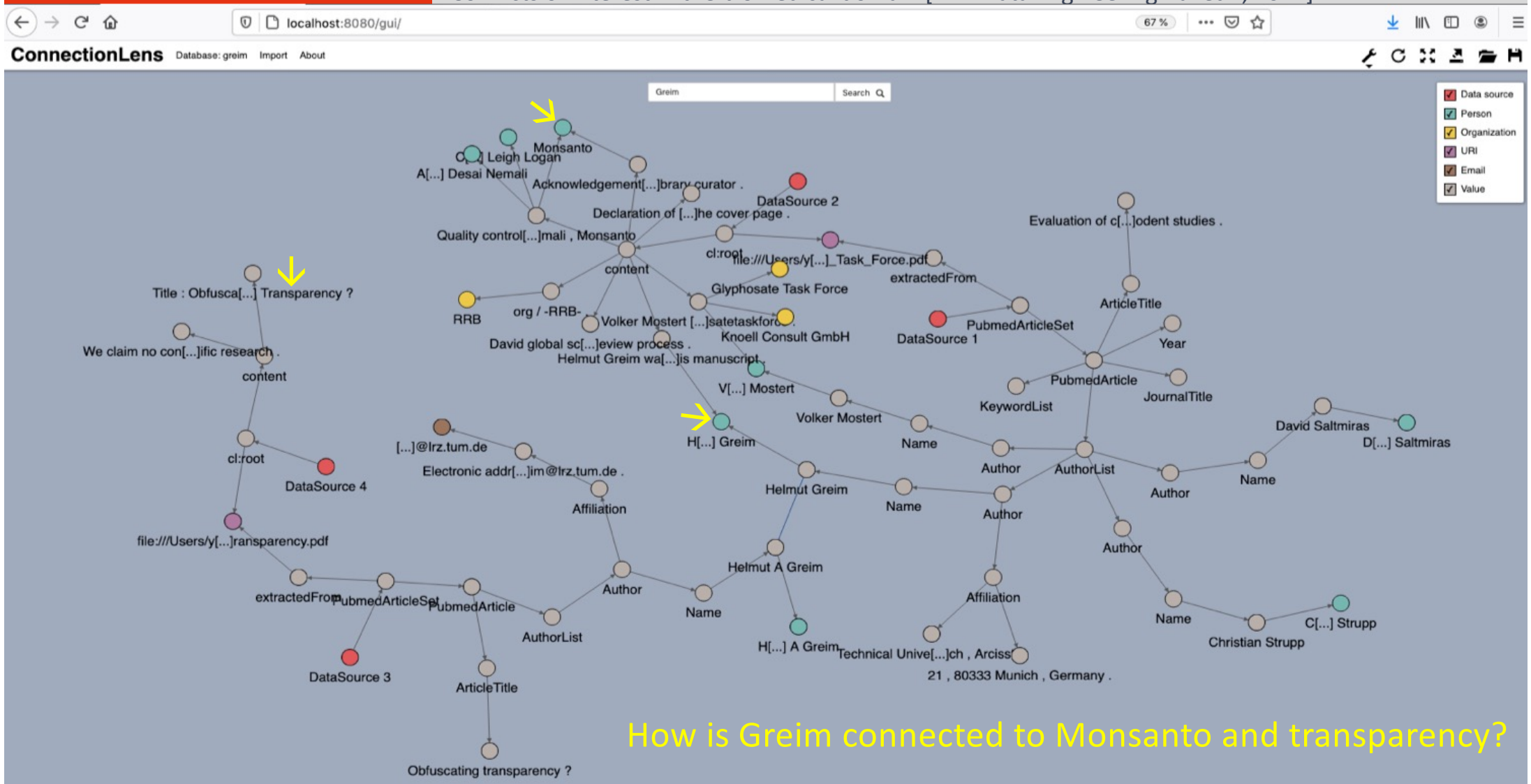
Application: conflicts of interest in the biomedical domain


a1.xml
XML
PubMedArticle
URI
Authors
Title
CoIStatement
"pubmed.com/a1"
Author
...
"Lung..."
"Dr. Alice consults for ABCPharma"
1.0
Name
Affiliation
ABCPharma
"Univ. Belleville, France"
"Alice"
Univ. Belleville
France
Alice
1.0
0.92
PDF
art.pdf
{:}
JSON
art.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..."


localhost:8080/gui/
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
content
file:///Users/y[...]_Task_Force.pdf
extractedFrom
Glyphosate Task Force
RRB
org / -RRB-
Volker Mostert [...]satetaskforc...
Knoell Consult GmbH
David global sc[...]eview process .
Helmut Greim wa[...]is manuscript .
Evaluation of c[...]odent studies .
ArticleTitle
PubmedArticleSet
DataSource 1
Year
PubmedArticle
KeywordList
JournalTitle
Title : Obfusca[...] Transparency ?
We claim no con[...]ific research .
content
V[...] Mostert
Volker Mostert
Name
Author
AuthorList
David Saltmiras
D[...] Saltmiras
Name
Author
cl:root
DataSource 4
[...]@lrz.tum.de
Electronic addr[...]im@lrz.tum.de .
Affiliation
H[...] Greim
Helmut Greim
Name
Author
file:///Users/y[...]ransparency.pdf
Helmut A Greim
Author
Author
extractedFrom
PubmedArticleSet
PubmedArticle
AuthorList
Name
Affiliation
Name
C[...] Strupp
Christian Strupp
H[...] A Greim
Technical Unive[...]ch , Arciss
21 , 80333 Munich , Germany .
DataSource 3
ArticleTitle
Obfuscating transparency ?

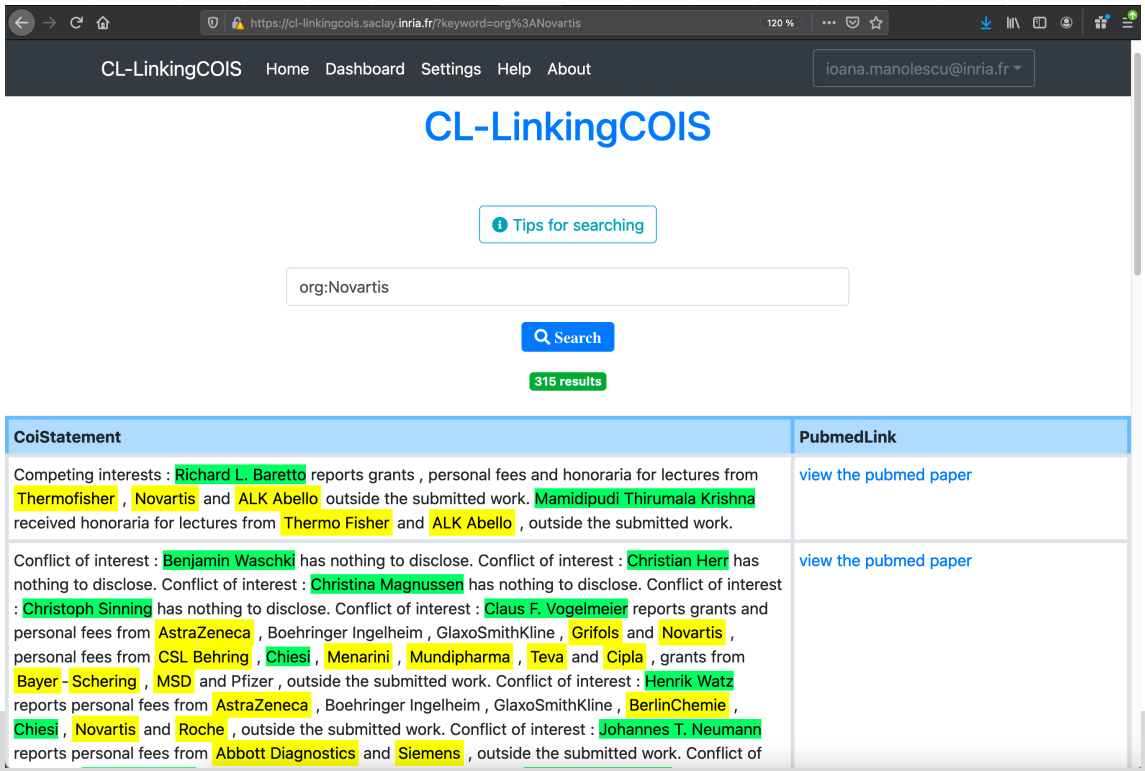


How is Greim connected to Monsanto and transparency?


localhost:8080/gui/
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
content
file:///Users/y[...]_Task_Force.pdf
extractedFrom
Glyphosate Task Force
RRB
org / -RRB-
Volker Mostert [...]satetaskforc...
Knoell Consult GmbH
David global sc[...]eview process .
Helmut Greim wa[...]is manuscript .
V[...] Mostert
Volker Mostert
H[...] Greim
Name
Author
AuthorList
Helmut Greim
Helmut A Greim
H[...] A Greim
Affiliation
Technical Unive[...]ch , Arciss
21 , 80333 Munich , Germany .
Evaluation of c[...]odent studies .
ArticleTitle
Year
PubmedArticleSet
DataSource 1
PubmedArticle
JournalTitle
KeywordList
David Saltmiras
D[...] Saltmiras
Name
Author
Author
Author
Name
Christian Strupp
C[...] Strupp
Title : Obfusca[...] Transparency ?
We claim no con[...]ific research .
content
cl:root
DataSource 4
file:///Users/y[...]ransparency.pdf
[...]@lrz.tum.de
Electronic addr[...]im@lrz.tum.de .
Affiliation
extractedFrom
PubmedArticleSet
PubmedArticle
AuthorList
Author
Name
DataSource 3
ArticleTitle
Obfuscating transparency ?
How is Greim connected to Monsanto and transparency?

Application: conflicts of interest in the biomedical domain [IEEE DataEngBull 2021, CIKM2021]

Collaboration with Stéphane Horel (Le Monde)



Application: conflicts of interest in the biomedical domain [IEEE DataEngBull 2021, CIKM2021]

Collaboration with Stéphane Horel (Le Monde)

Data: XML, PDF→JSON, HTML

$\|N\|$	$\|E\|$	$\|N\|$	$\|N_P\|$	$\|N_O\|$	$\|N_L\|$
XML	32,028,429	19,851,904	1,483,631	584,734	126,629
JSON	1,025,307	432,303	75,297	7,320	4,139
HTML	246,636	185,479	3,726	7,227	320
Total	33,300,372	20,469,686	1,562,654	665,167	131,088

Table 3: Statistics on Conflict of Interest application graph.

Graph creation performance: storage, extraction, disambiguation [InfSys2022]

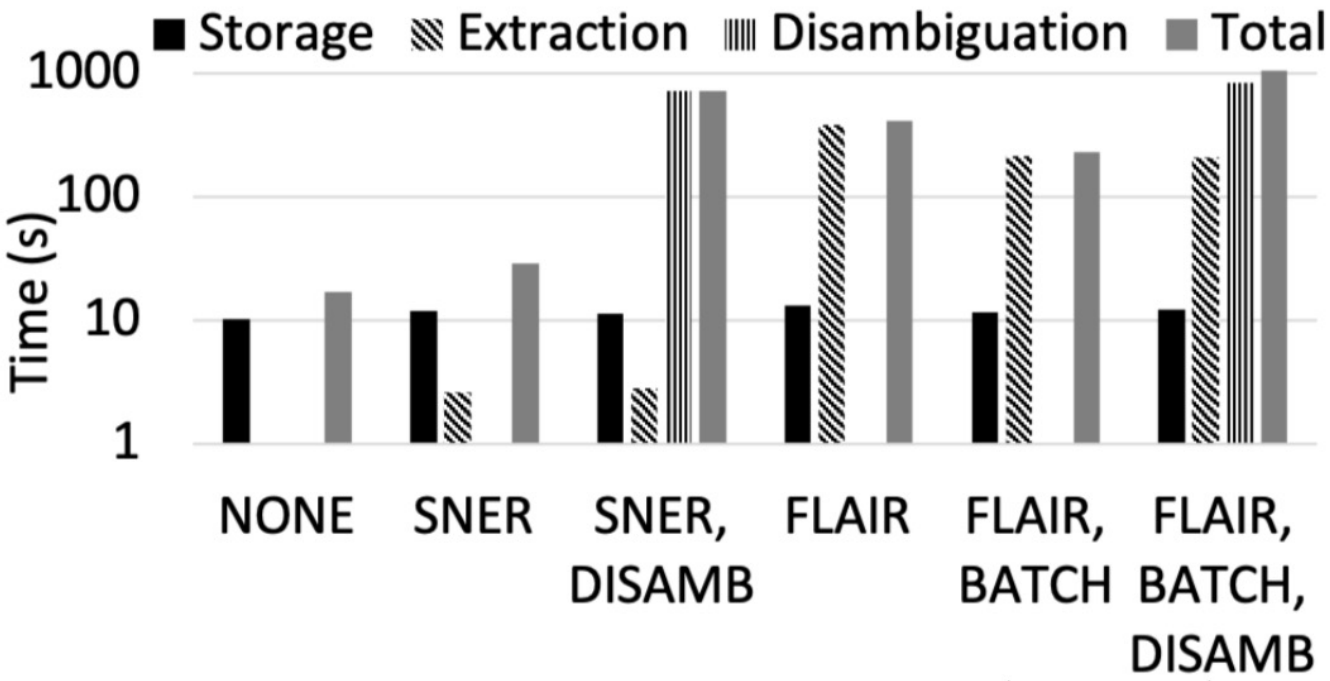


Figure 6: Graph construction time (seconds).

Graph creation performance: batch extraction [InfSys2022]

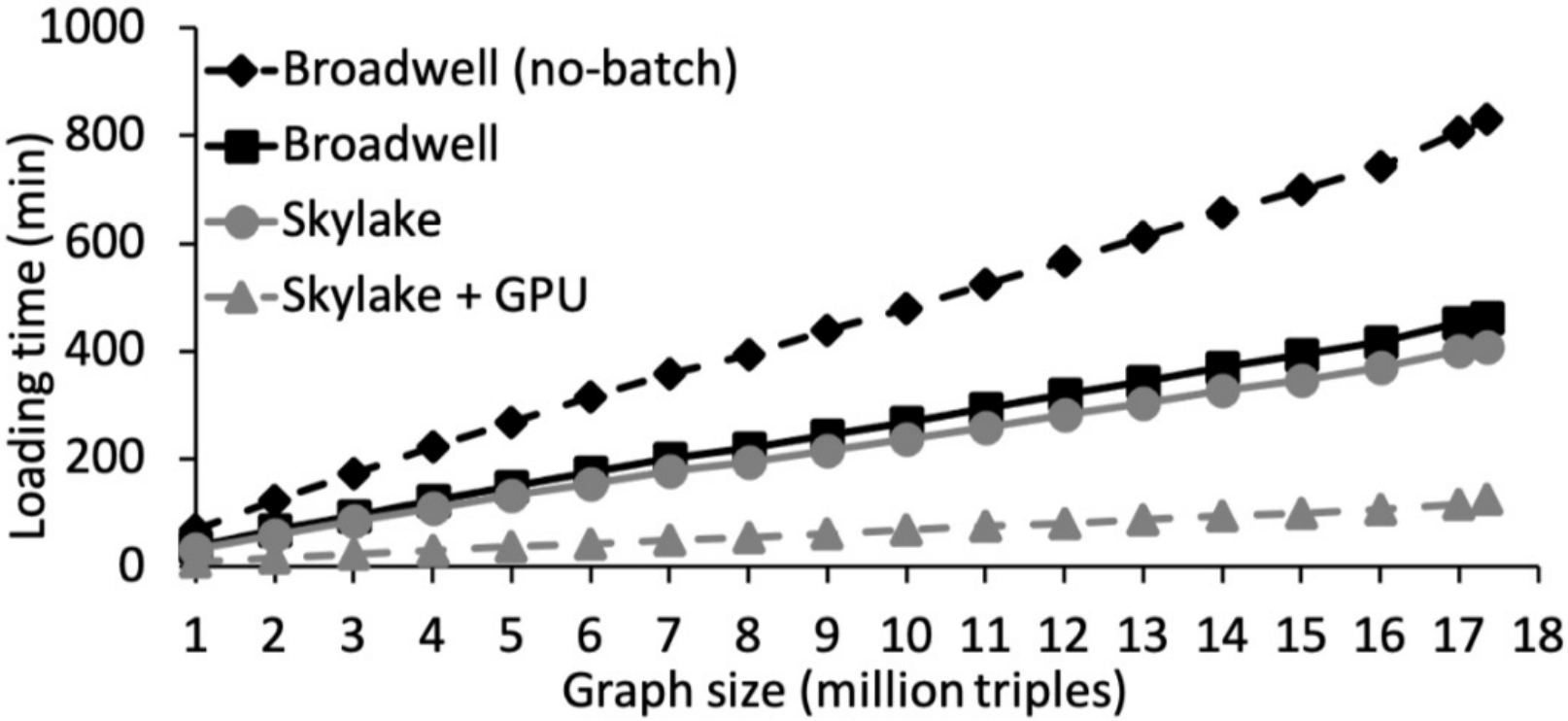


Figure 7: YAGO loading time (minutes) using Flair.

Credits and references https://sourcessay.inria.fr

ConnectionLens core architecture: A. Anadiotis, O. Balalau, H. Galhardas, J. Leblay, I. Manolescu, T. Merabti, Y. Haddad, J. You, Y. Youssef [VLDB18, InfSys22, CIKM21demo]

Dataset abstraction: N. Barret, I. Manolescu, P. Upadhyay [CIKM22demo, BDA2022, 2 under submission]

Entity path finding: N. Barret, A. Gauquier, J. Law, I. Manolescu [1 under submission]

Graph exploration: T. Bouganim, I. Manolescu, E. Pietriga [work in progress]

Structure and unstructured querying: A. Anadiotis, Y. Haddad, I. Manolescu, M. Mohanty [IEEE DEBull21, ICDE23, 1 under submission]

Biomedical Conflicts of Interest application: A. Anadiotis, O. Balalau, T. Bouganim, G. Fooks, H. Galhardas, S. Horel, I. Manolescu, T. Mills, C. Pettineo, P. Upadhyay [IEEE DEBull21, EACL23, 1 under preparation]

Looking forward

Perspectives

SourcesSay: 2-3 more years to go

- ❑ Finalizing abstraction + finding interesting entity paths (PhD N. Barret)
- ❑ Exploratory querying (PhD T. Bouganim)

Inria Exploratory Action JoDaIA (“Data and AI for Journalism”) (2023-2026)

- ❑ One axis: identifying and developing more applications based on ConnectionLens
- ❑ *Open to exploring more proposals!*

ELIAS EU project (coord. U. Trento, with O. Goga): Development of learning algorithms to protect and secure democracy

Thank you

Questions?

SourcesSay: https://sourcessay.inria.fr

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