Teasing journalistic findings out of heterogeneous sources: a data/AI journey

Ioana Manolescu

Teasing journalistic findings out of heterogeneous sources: a data/AI journey

Ioana Manolescu
Inria Saclay-Île-de-France, Institut Polytechnique de Paris

Outline

1. Motivation: why journalism?
2. Data integration problems raised by data journalism
3. Graph-based data integration in ConnectionLens
4. Related work and perspectives

Motivation:
Why journalism?

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, confim or refute public statements

Fact-checking

(Data) journalism

- ❑ To expose and explain society functioning

Fact-checking

Not everyone agrees, however, that Democrats are not flip-flopping on the issue.

Mark Krikorian, executive director of the Center for Immigration Studies, a think tank that advocates for lower immigration, said that because the public doesn't know exactly what border barriers the Trump administration wants to build, Mulvaney's statement is not an "exact" comparison. But, he said, to dismiss it simply on that basis would be "tendentiously literal."

"The fact is that, other than the 'Mexico will pay for it' stuff, Trump is simply channeling the 2006 Secure Fence Act, and Schumer et al. who voted for it out of political calculation are indeed hypocrites for opposing the attempt to finally bring that law to fruition," Krikorian told us via email.

At the surface level, it is true in a broad sense that Democrats including Schumer, Obama and Clinton have in the past supported border fencing. All three voted for the Secure Fence Act of 2006, and all three supported the 2013 Senate immigration overhaul that passed the Senate, and which called for tougher border security including some additional fencing. But to claim that those measures are the same as what Trump is proposing is a stretch.

Share The Facts

Mick Mulvaney

Director, Office of Management and Budget

MISLEADING

"We don't understand why the Democrats are so wholeheartedly against [President Trump's border wall]. They voted for it in 2006."

FACTCHECK.ORG

Fox News Sunday – Sunday, April 23, 2017

SHARE READ MORE

www.factcheck.org/2017/04/democrats-support-border-wall/

FACTCHECK.ORG A Project of The Annenberg Public Policy Center

DONATE

HOME | ARTICLES | ASK A QUESTION | VIRAL SPIRAL | ARCHIVES | ABOUT US | SEARCH | MORE

THE WIRE

Did Democrats Once Support Border Wall?

By Robert Farley Posted on April 26, 2017

Like 835 Tweet Pin it Share 11

White House Office of Management and Budget Director Mick Mulvaney made an apples-to-oranges comparison when he said he couldn't understand why Democrats opposed supplemental funding for a border wall since many of them were for it back in 2006.

Mulvaney is referring to the Secure Fence Act of 2006, which called for construction of 700 miles of fencing and enhanced surveillance technology, such as unmanned drones, ground-based sensors, satellites, radar coverage and cameras. Sen. Chuck Schumer and then-Sens. Barack Obama and Hillary Clinton were among a bipartisan majority that voted in favor of the legislation, and it was signed into law by President George W. Bush.

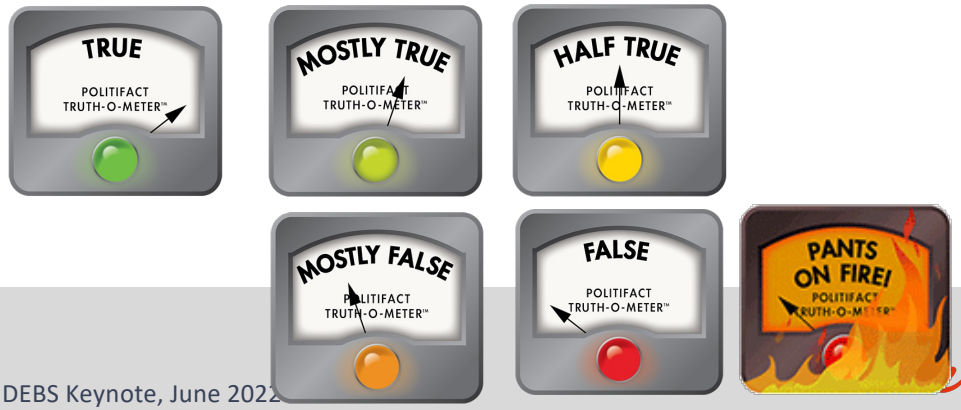
In a very general sense, the Democrats named by Mulvaney supported a bill to build more

ASK FACTCHECK

Like 953 Tweet Pin it Share 98

Q: Did the Supreme Court rule that public schools cannot teach students about Islam?

A: No. That false claim was spread by a network of fake news websites.



Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)

Projects and collaborations

Google Award (2015) with U. Paris Sud

ANR ContentCheck (2016-2020) with
Les Décodeurs (Le Monde) https://contentcheck.inria.fr

Inria Associated Team WebClaimExplain (2017-2019),
with AIST Japan

Collaboration with H. Galhardas (University of Lisbon),
A. Anadiotis, O. Balalau, E. Pietriga (Inria, Polytechnique)

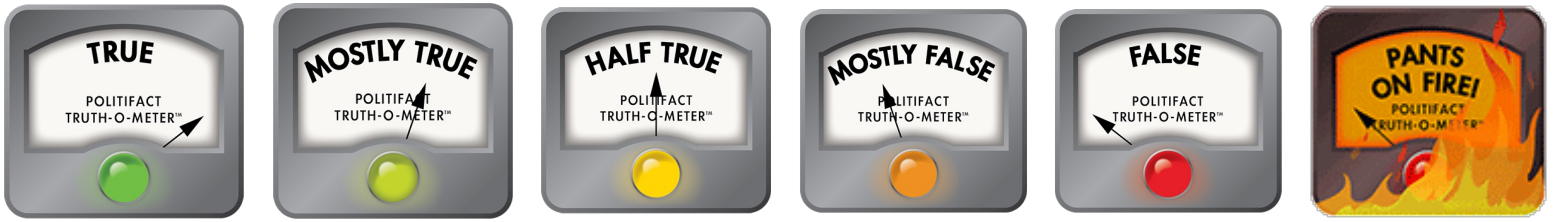
ANR SourcesSay AI Chair (2020-2023),
with Le Monde and WeDoData https://sourcessay.inria.fr

Fact-checking for **RadioFrance** (2022-2023)

Project discussed with Les Décodeurs: fake new detection and propagation on Twitter

Online fact-checks: (semi)structured data sources (JSON, XML) listing

- Link to claim (media, social network etc.), **claim author**
- **Fact-check**, containing: analysis (details), final assessment, fc author, date, institution



Among the first published: https://www.lemonde.fr/webservice/decodex/updates

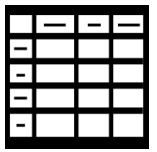
Years later: **ClaimReview** by Google and others (https://www.claimreviewproject.com/)

Project discussed with Les Décodeurs: fake new detection and propagation on Twitter

Online fact-checks: (semi)structured data sources (JSON, XML) listing

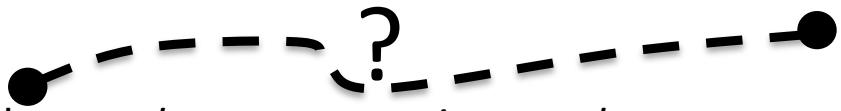
- Link to claim (media, social network etc.), **claim author**
- **Fact-check**, containing: analysis (details), final assessment, fc author, date

Décodeurs' database of French public figures (Excel)



- First name, last name, Twitter ID, position, political party when known

Question: When does a fake news post first cross into a community, e.g., members of the Parliament?



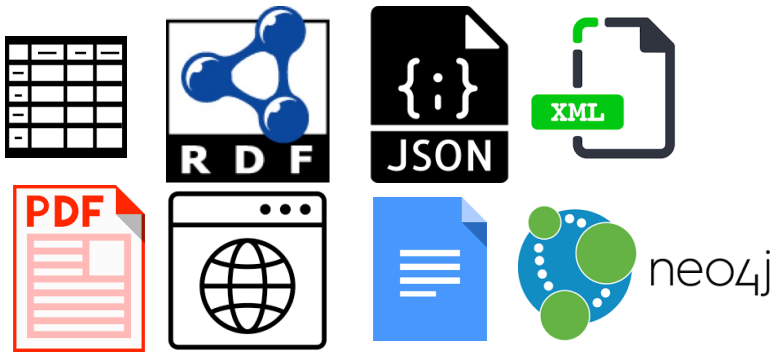
- Looking for tweets *connected to* a fake news author, *and to a community member*; arbitrary paths (chains of author/likes/retweets/inParty/…)

Integrating heterogeneous journalistic datasets with ConnectionLens

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

Graph-based integration of heterogeneous data sources

- The sources are not RDF. They can be (semi)structured, or unstructured (text).
- The sources may be very dynamic (projects started and abandoned as per news cycle and data availability).



- There is no schema. Data producers often uncollaborative.
- For most journalists, databases do not come naturally, and IT support is limited. They know keyword-based search…

→ **Integrate heterogeneous sources within a graph, query w/ keywords**

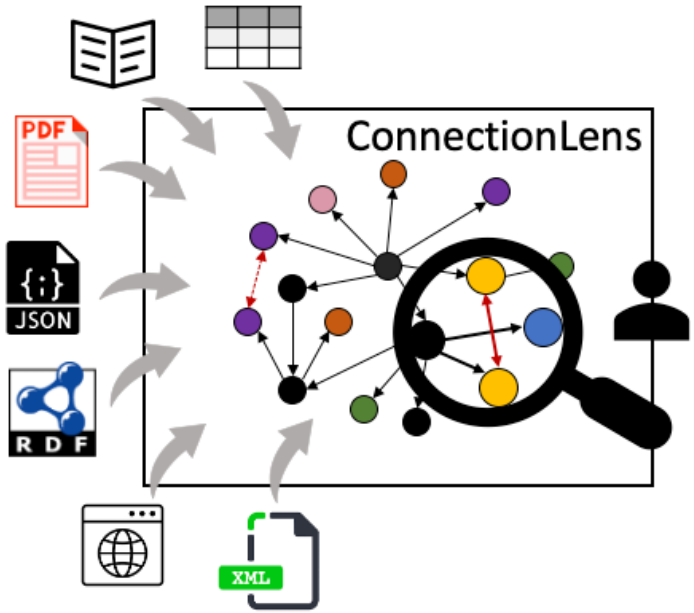
ConnectionLens: graph-based integration of heterogeneous data sources

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

Joint work with: J. Leblay (AIST Japan),
H. Galhardas and C. Conceiçao (U. Portugal),
A. Anadiotis, O. Balalau, N. Barret, T. Bouganim,
F. Chimienti, M.-Y. Haddad, T. Merabti,
P. Upadhyay (CEDAR) + interns

S.Horel (Le Monde, European Press Prize
"Investigative Reporting Award 2018")

Ongoing work in ANR/DGA AI Chair SourcesSay (https://sourcessay.inria.fr), DIM RFSI



ConnectionLens principles [VLDBdemo2018]

Integrate **any kind of data** into a **graph**

Extract entities from any text node (regardless of the model of the data source where the text comes from)

- ❑ Same entity in two different text nodes = link among the text nodes (*densification* of the graph)

The graph is **heterogeneous** and **irregular** →

Query it through **keywords**: find trees that connect 1 node matching each kwd

- ❑ Closely related to the Group Steiner Tree Problem (GSTP)

ConnectionLens principles [VLDBdemo2018]

Integrate **any kind of data** into a **graph**

Extract entities from any text node (regardless of the model of the data source where the text comes from)

- ❑ Same entity in two different text nodes = link among the text nodes (*densification* of the graph)

The graph is **heterogeneous** and **irregular** →

Query it through **keywords**: find trees that connect 1 node matching each kwd

- ❑ Closely related to the Group Steiner Tree Problem (GSTP)

Rest of the talk based on current state of the project:
[CIKMdemo2021, Elsevier InfSys 2022, IEEE DataEngBull 2021, ongoing works]

ConnectionLens graph construction

The Balkany and their African connections

https://www.liberation.fr/france/2015/08/13/troisieme-villa-saisie-pour-les-balkany_1362925

Libération — Troisième villa saisie pour les Balkany

ENQUÊTE

Troisième villa saisie pour les Balkany

Par Emmanuel Fansten — 13 août 2015 à 14:58

https://www.lepoint.fr/politique/villas-a-marrakech-fonds-occultes-les-epoux-balkany-juges-lundi-12-05-2019-2312058_20.php#

Le Point | Montres | Le Point Pop | Auto | Vin | Phébé | Services | Newsletters | Identifiez-vous | Abonnez-vous

MENU — Le Point Politique

Actualité > Politique

Villas à Marrakech, fonds « occultes »... : les époux Balkany jugés lundi

Soupçonnés d'avoir dissimulé 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret comparaissent pour fraude fiscale et blanchiment.
Source AFP

Publié le 12/05/2019 à 11:19 | Le Point.fr

PROFITEZ DE VOTRE ABONNEMENT À 1€ LE 1ER MOIS !

De somptueuses villas à Marrakech et dans les Caraïbes, des fonds « occultes » transitant par le Panama ou Singapour... Soupçonnés d'avoir dissimulé plus de 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret Patrick et Isabelle Balkany sont jugés à partir de

EN CONTINU

15H36 L'Assemblée se prononce sur un budget de la Sécu grevé...
13H30 Un nouveau traité européen, pourquoi pas ?
13H29 Sébastien Le Fol – Macron, le commandant couche-tôt
12H33 Les députés votent le plan de relance, parasité par la crise sanitaire
12H32 Présidentielle: Mélenchon plaide pour un parrainage de 150.000...

Voir toute l'actualité en continu

The Balkany and their African connections

https://www.liberation.fr/france/2015/08/13/troisieme-villa-saisie-pour-les-balkany_1362925

Libération

Troisième villa saisie pour les Balkany

ENQUÊTE

Troisième villa saisie pour les Balkany

Par Emmanuel Fansten — 13 août 2015 à 14:58

https://www.lepoint.fr/politique/villas-a-marrakech-fonds-occultes-les-epoux-balkany-juges-lundi-12-05-2019-2312

Le Point | Montres | Le Point Pop | Auto | Vin | Phébé | Services | Newsletters

MENU — Le Point Politique

Actualité > Politique

Villas à Marrakech, fonds « occultes »... : les époux Balkany jugés lundi

Soupçonnés d'avoir dissimulé 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret comparaissent pour fraude fiscale et blanchiment.
Source AFP

Publié le 12/05/2019 à 11:19 | Le Point.fr

PROFITEZ DE VOTRE ABONNEMENT À 1€ LE 1ER MOIS !

De somptueuses villas à Marrakech et dans les Caraïbes, des fonds « occultes » transitant par le Panama ou Singapour... Soupçonnés d'avoir dissimulé plus de 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret Patrick et Isabelle Balkany sont jugés à partir de

https://www.hatvp.fr/resultats-de-recherche/?search=balkany 120 %

balkany

3 RÉSULTATS CORRESPONDANT À VOTRE RECHERCHE

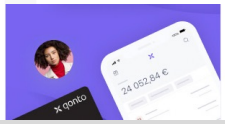
RESPONSABLES PUBLICS

REPRESENTANTS D'INTÉRÊTS

ACTIVITÉS DE REPRÉSENTATION D'INTÉRÊTS

Perte de l'autorité morale : demande d'Anticor au Président de la République pour que soient révoqués un maire et son adjointe qui ont avoué avoir fraudé l'administration fiscale

ANTICOR



The Balkanys and their African connections

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

The Balkanys and their African connections

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

National Directory of Elected Officials (JSON)

```
[{
    name: "Levallois-Perret",
    mayor: "P. Balkany",
    city-council: [
      {name: "I. Balkany"},
        …
    ]
}, …]
```

The Balkanys and their African connections

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

dbpedia.org (RDF)

```
{
dbr:Marrakech
  dbr:name       ”Marrakech”
  rdf:type       dbo:City ;
  dbo:country    dbr:Morrocco .
dbr:Morocco
  dbr:name       ”Morocco”
  rdf:type       dbo:Country
  dbo:locatedIn  dbr:Africa .
dbr:CentralAfricanRepublic
  dbr:name       ”Central African Republic”
  dbo:locatedIn  dbr:Africa .
}
```

National Directory of Elected Officials (JSON)

```
[{
    name: “Levallois-Perret”,
    mayor: “P. Balkany”,
    city-council: [
      {name: ”I. Balkany”},
        …
    ]
}, …]
```

The Balkanys and their African connections

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

dbpedia.org (RDF)

```
{
dbr:Marrakech
  dbr:name       ”Marrakech”
  rdf:type       dbo:City ;
  dbo:country    dbr:Morrocco .
dbr:Morocco
  dbr:name       ”Morocco”
  rdf:type       dbo:Country
  dbo:locatedIn  dbr:Africa .
dbr:CentralAfricanRepublic
  dbr:name       ”Central African Republic”
  dbo:locatedIn  dbr:Africa .
}
```

National Directory of Elected Officials (JSON)

```
[{
    name: “Levallois-Perret”,
    mayor: “P. Balkany”,
    city-council: [
      {name: ”I. Balkany”},
        …
    ]
}, …]
```

Libération – Nov. 13, 2014 (Text)

Balkany mineur de fonds

L’élu de Levallois-Perret est soupçonné d’avoir touché 5 millions de dollars de commission en 2009 grâce à son rôle d’intermédiaire entre Areva et la Centrafrique dans le dossier Uramin. […]

How is Levallois-Perret connected to Africa and "real estate"?

Public officials transparency high authority (CSV)

Name	Owner	Location	Type
Dar Gyucy	P. Balkany	Marrakech	Real Estate
Moulin Cossy	I. Balkany	Giverny	Real Estate

dbpedia.org (RDF)

```
{
dbr:Marrakech
  dbr:name       ”Marrakech”
  rdf:type       dbo:City ;
  dbo:country    dbr:Morrocco .
dbr:Morocco
  dbr:name       ”Morocco”
  rdf:type       dbo:Country
  dbo:locatedIn  dbr:Africa .
dbr:CentralAfricanRepublic
  dbr:name       ”Central African Republic”
  dbo:locatedIn  dbr:Africa .
}
```

National Directory of Elected Officials (JSON)

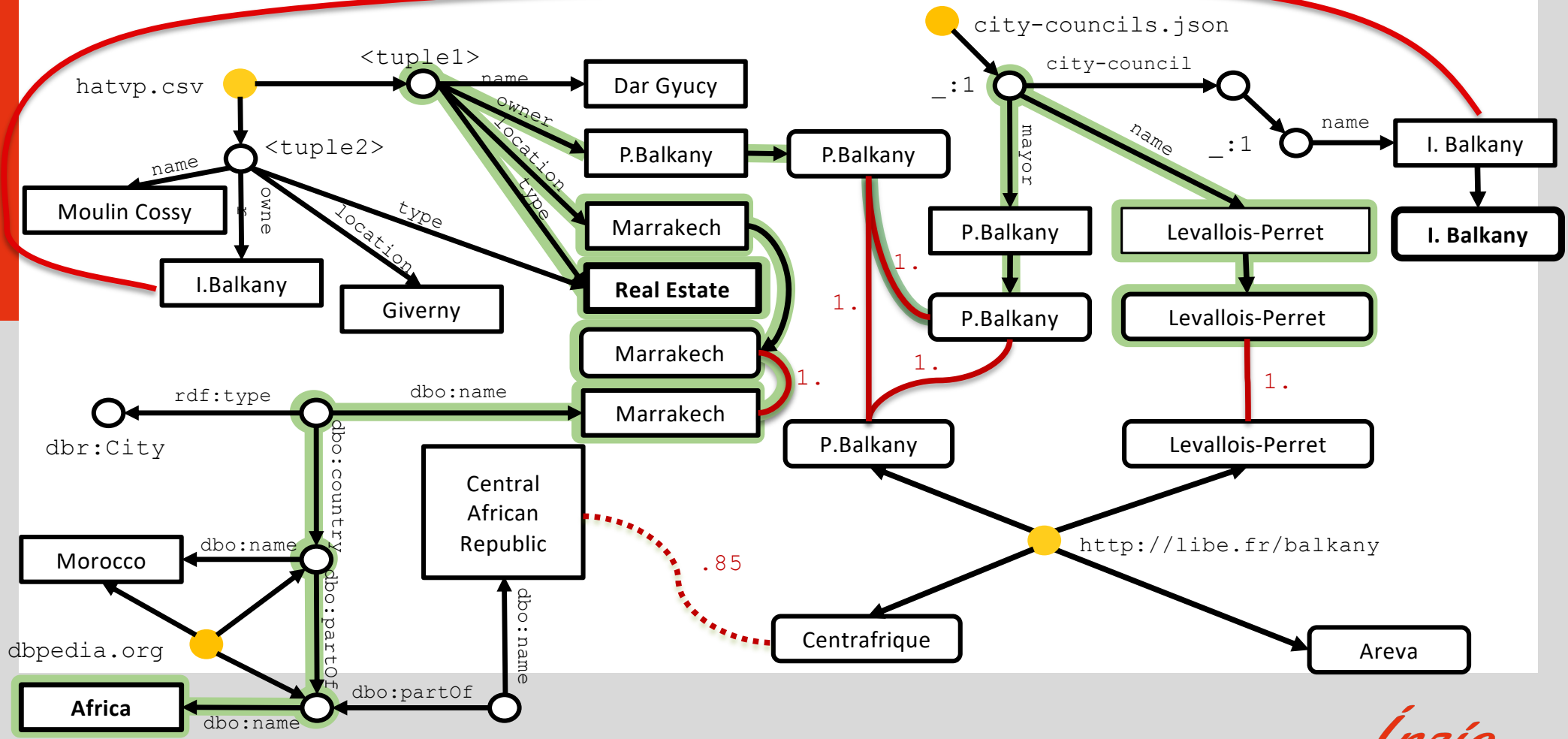
```
[{
    name: “Levallois-Perret”,
    mayor: “P. Balkany”,
    city-council: [
      {name: ”I. Balkany”},
        …
    ]
}, …]
```

Libération – Nov. 13, 2014 (Text)

Balkany mineur de fonds

L’élu de Levallois-Perret est soupçonné d’avoir touché 5 millions de dollars de commission en 2009 grâce à son rôle d’intermédiaire entre Areva et la Centrafrique dans le dossier Uramin. […]

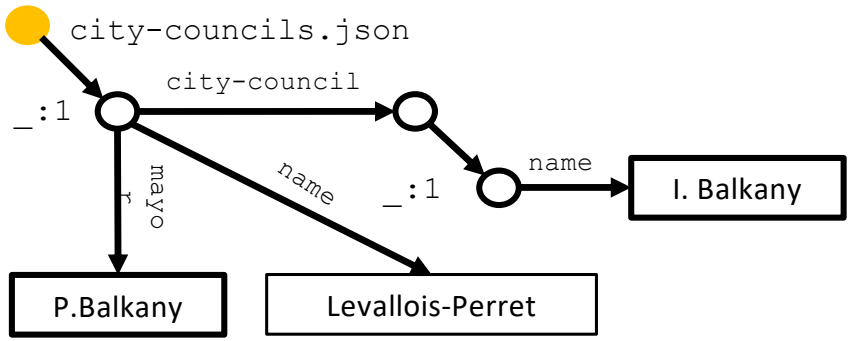
Idea: integrate all data sources into a heterogeneous graph



Graph construction stages

1. Primary node and edge construction

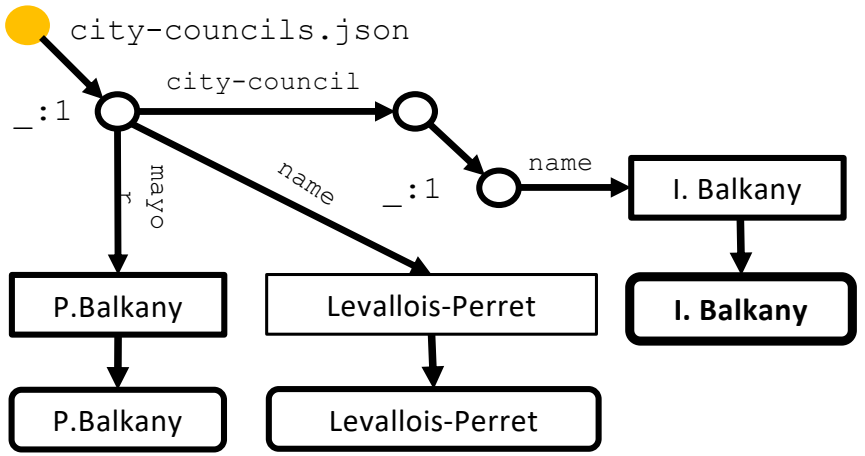
- Direct for XML, JSON, RDF, HTML
- 1 relational tuple=1 node;
  primary keys-foreign keys as links
- Convert information from PDF into:
  - JSON for text content
  - RDF describing tables



Graph construction stages

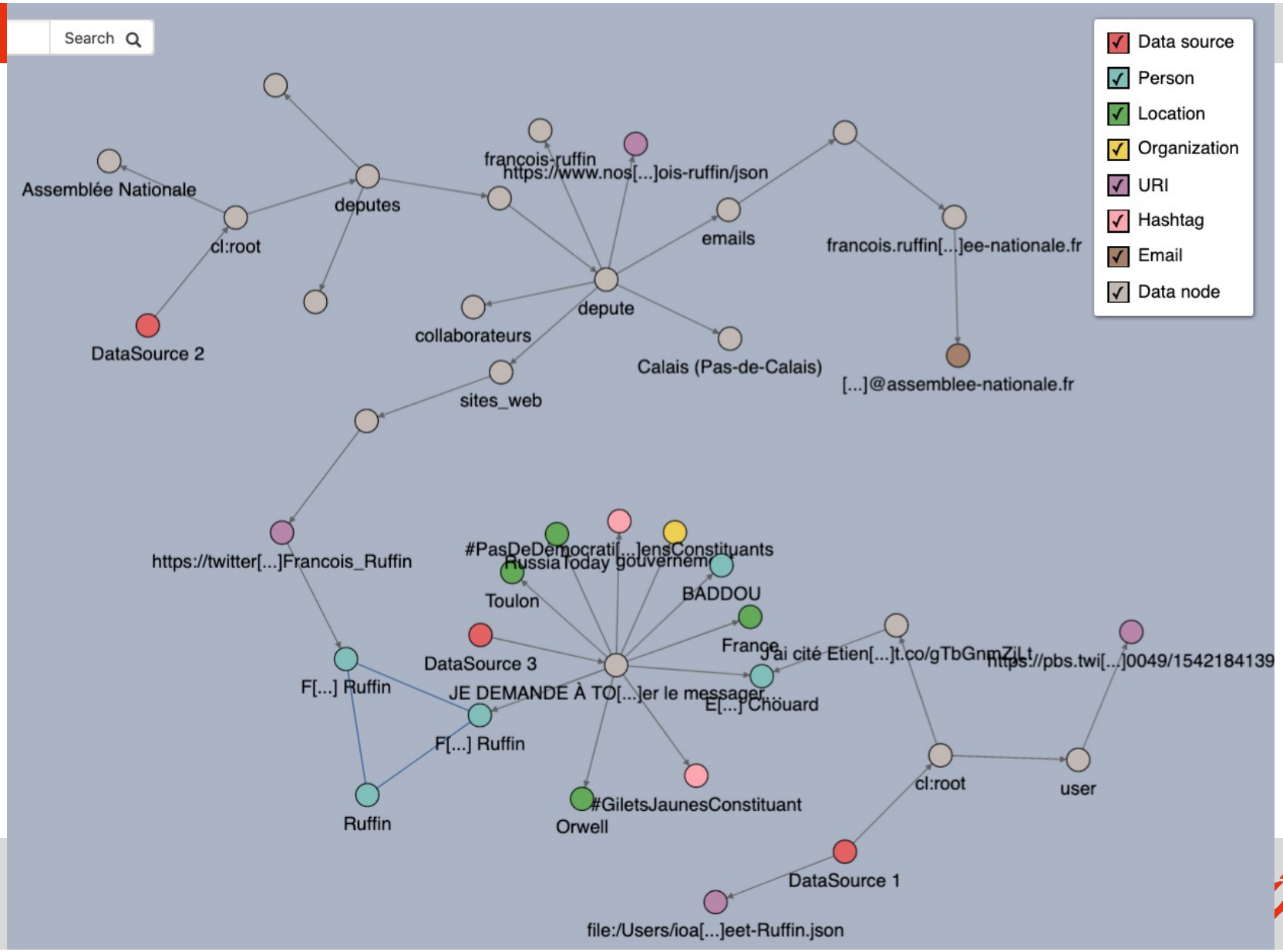
1. Primary node and edge construction

- Direct for XML, JSON, RDF, HTML
- 1 relational tuple=1 node; PK-FKs as links
- [Optional] segment text documents
- Extract information from PDF into: (a) JSON, and (b) RDF describing tables



2. Entity extraction

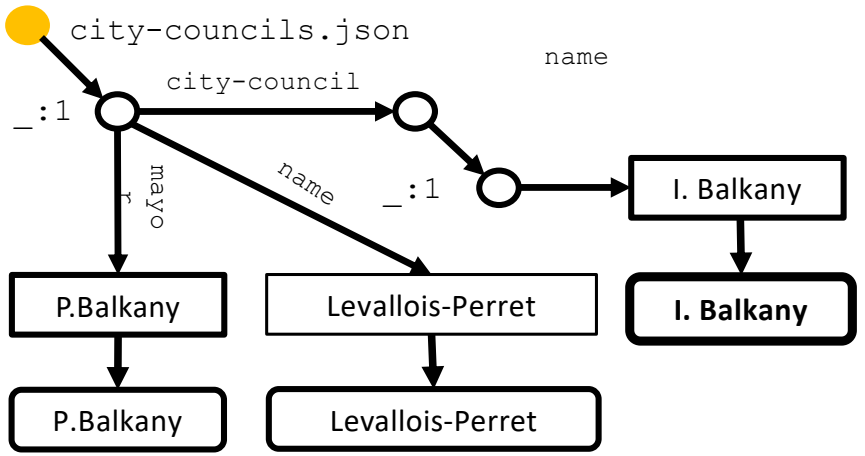
- From all text nodes of all the sources: **entity node** child of text node
- [VLDB2018]: based on Stanford NER
- [InfSys2021] Developed and trained new entity extractor FR/EN, based on Flair framework
- Date extraction based on HeidelTime


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
Ruffin
#GiletsJaunesConstituant
Orwell
cl:root
user
DataSource 1
file:/Users/ioa[...]eet-Ruffin.json

Graph construction stages

2. Entity extraction

- From all text nodes of all the sources: **entity node** child of text node
- [VLDB2018]: based on Stanford NER
- [BDA2020] Developed and trained new entity extractor from French, based on Flair framework



3. Entity disambiguation

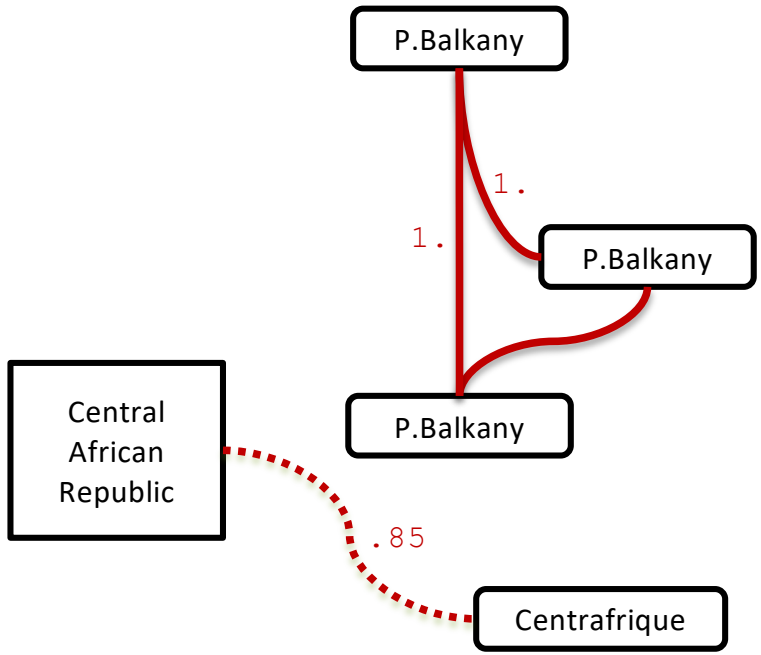
- For each recognized entity, e.g., "Hollande" the place or the person?
- Built novel disambiguation pipeline for French, based on Ambiverse framework
  - Based on knowledge bases (WikiData, YAGO) and Wikipedia
  - Helpful on well-known entities

Graph construction stages

4. Node matching

- To create sameAs edges:
  - Strong sameAs edges: equivalent nodes 1.
  - Weak sameAs edges: similar nodes .85
- Label normalization, distance functions
- Remains quadratic at the core ☹ , so...

Node factorization (heuristic): create only one node per label per document (or per graph)



More complex entity matching/cleaning pipelines for: Person, Organization, Location

ConnectionLens graph querying

Querying problem statement

- Given the graph G = (N, E) built out of the datasets D and a query Q= {w1, ...,wm}, return the **k highest-score minimal answer trees**.
- An answer tree is a set of edges which (*i*) form a tree (*ii*) contain at least one node whose label matches each keyword wi.
- We are interested in **minimal answer trees**, that is:
  - Removing an edge from the tree should make it lack some keyword(s).
  - If a keyword matches more than one nodes in the answer tree, then all these matching nodes must be equivalent.

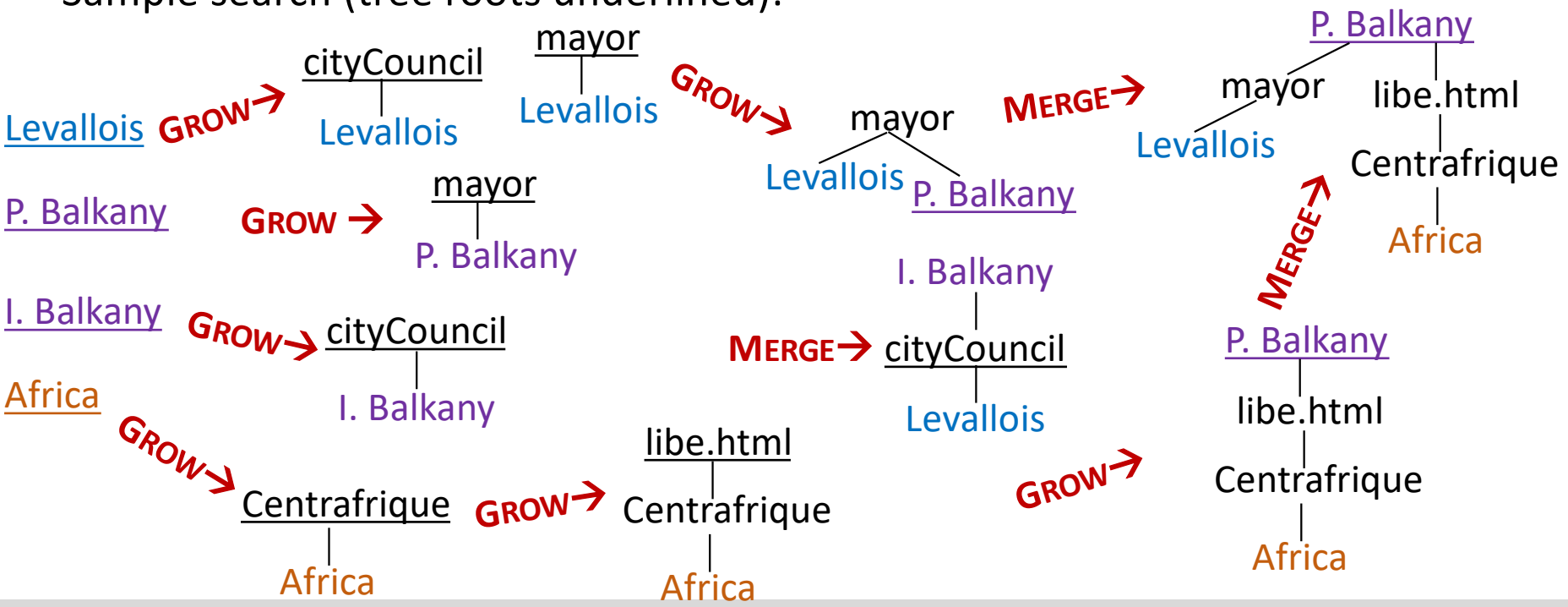
Search space and complexity

- Problem related to the **(Group) Steiner Tree Problem**
  - Given graph G, and nodes n1, . . . , nm, the Steiner Tree Problem (STP) requires the smallest tree in G that connects all the nodes. Known NP-hard problem in |G|
  - Group STP: start with m groups of nodes
- **Differences with our problem:**
  - Each edge can be taken in both directions: exponential increase in search space size
  - We need the k smallest-cost trees, not just one.
  - Support arbitrary score functions
- **Large literature (30+) on kwd search in data graphs.**
  - Differ in search space and/or make limitative assumptions on score
- **Our approach: enumerate solutions until time-out or max number of solutions reached.**
  - Return best k solutions found, for given score function

GAM (Grow and Aggressive Merge) Algorithm

- Builds trees exploring around the keyword matches
- GROW adds an edge to the root of a tree, MERGE merges trees with the same root

Sample search (tree roots underlined):



GAM (Grow and Aggressive Merge) Algorithm

- Builds trees exploring around the keyword matches
- GROW adds an edge to the root of a tree, MERGE merges trees with the same root

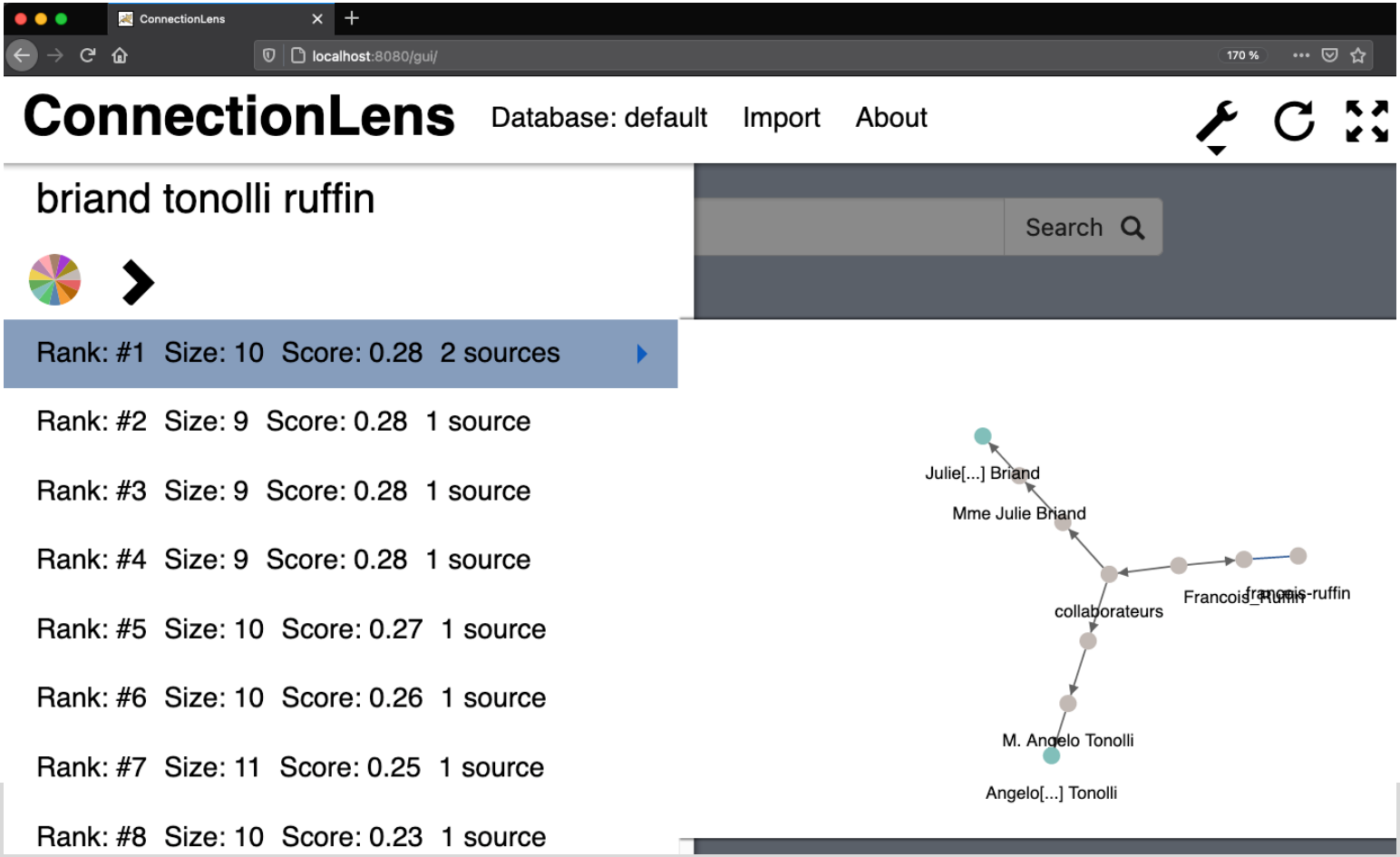
Every solution is a **conection (edge set):** the root is not meaningful to users

Ideally, we should develop each edge set only once... But:

- Grow and Merge depend on the tree root → Many edge sets with the same root
- Also, there are many Grow-Merge combinations that produce the same rooted tree!

Optimizations and careful pruning to keep exploration highly efficient (ongoing)

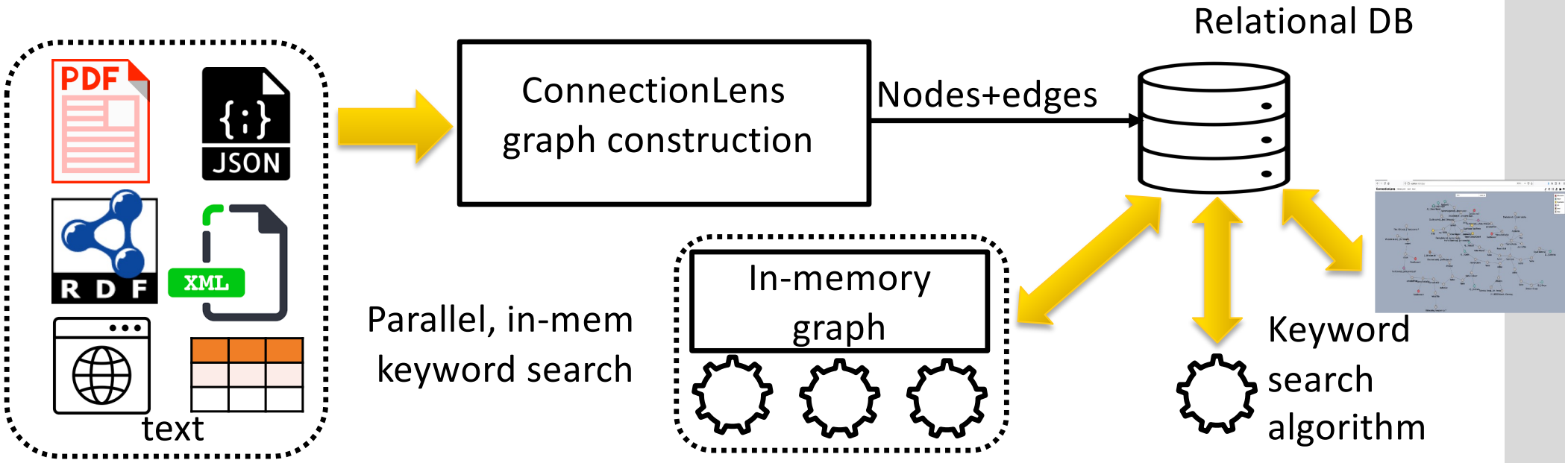
Sample query answers



ConnectionLens architecture and performance

Implementation

- Java (230 classes/46K LOC), Python (24 classes/2800 LOC), JS + CSS
- Available online: https://gitlab.inria.fr/cedar/connectionlens



Implementation [InfSys2022]

- Graph creation time mostly **linear in the size of the data**
- Costliest operations involve ML (disambiguation, extraction)
  - **Batch extraction:** 20x speed-up on GPU, 2x speed-up on regular server
  - **Extraction policies** replace or avoid extraction in some parts of the data

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

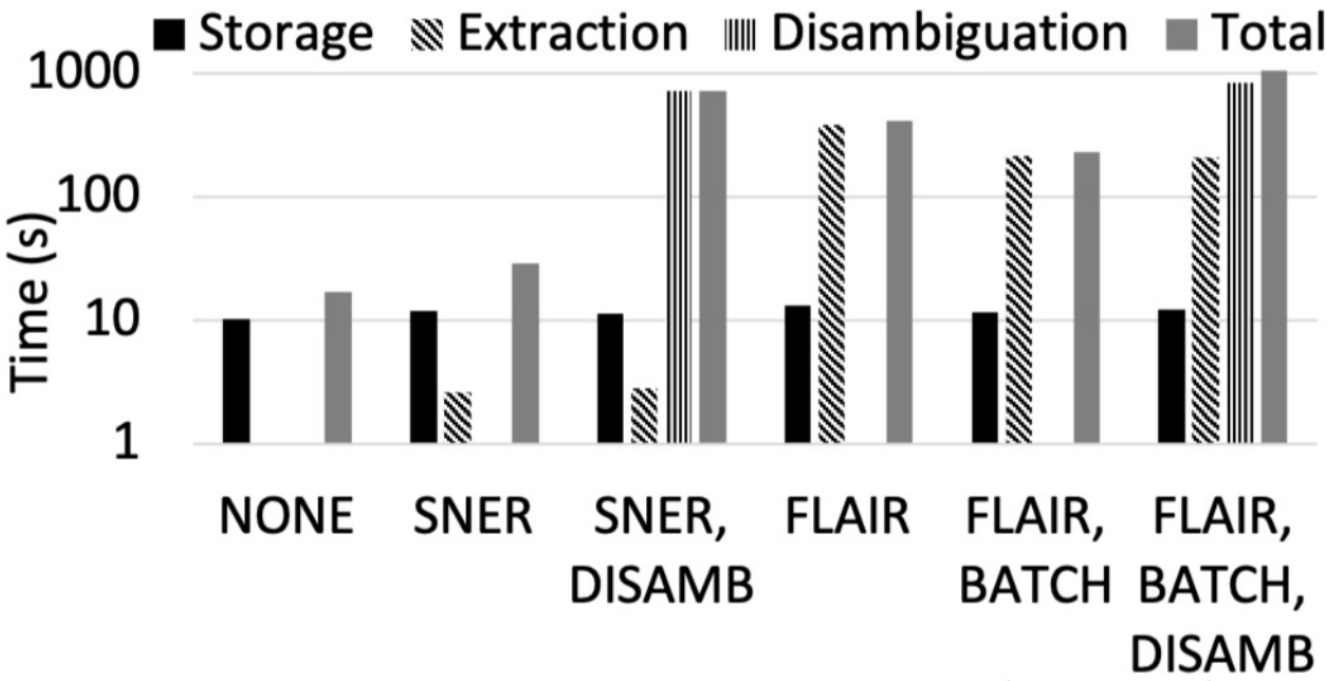


Figure 6: Graph construction time (seconds).

Graph creation performance: batch extraction [InfSys2022]

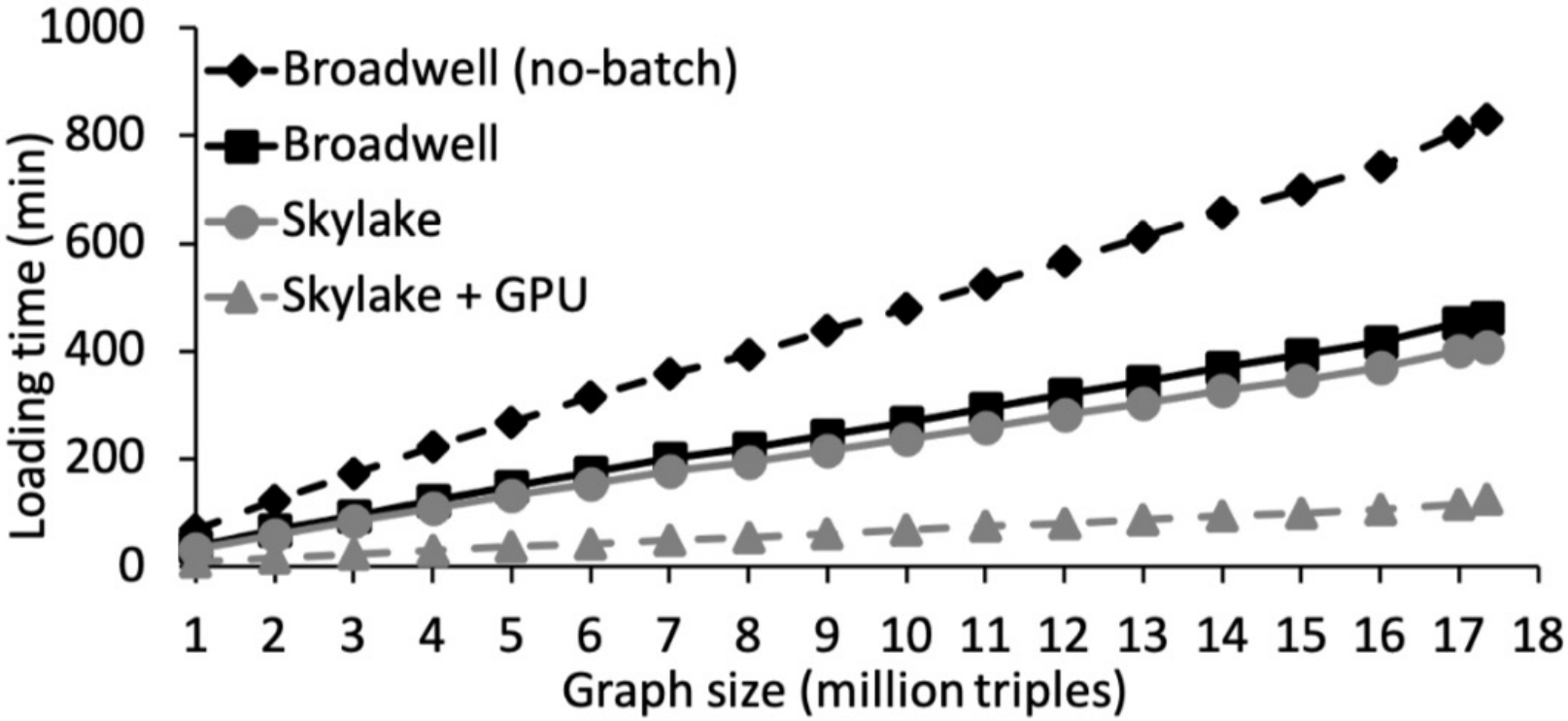


Figure 7: YAGO loading time (minutes) using Flair.

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

Collaboration with Stéphane Horel (Le Monde)

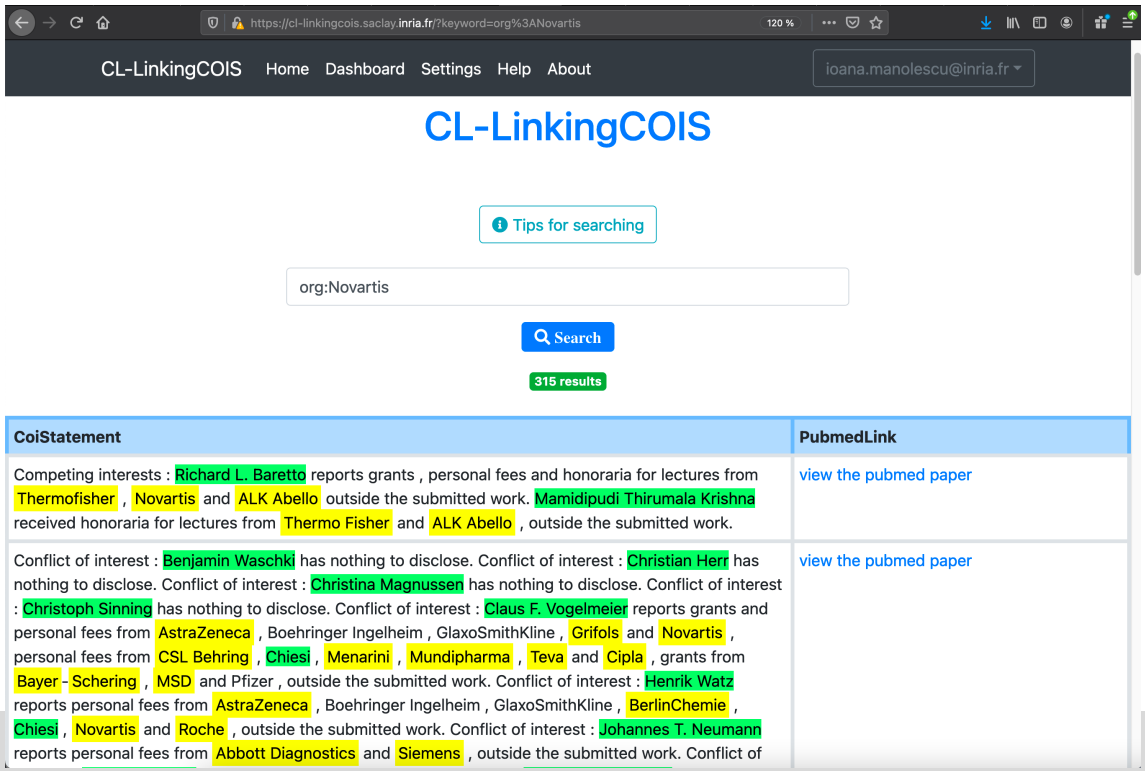
Data: XML, PDF→JSON, HTML

$\|N\|$	$\|E\|$	$\|N\|$	$\|N_P\|$	$\|N_O\|$	$\|N_L\|$
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.

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

Collaboration with Stéphane Horel (Le Monde)



Graph quality experiments [InfSys2022]

- PDF extraction accuracy: 63%
- F1 score for entity extraction from French:
  - Flair stacked forward and backward embeddings with French fastText embeddings: 73%
  - Spacy: 63%
  - StanfordNER: 45%
- F1 score of disambiguation: 86%

ConnectionLens in the scientific landscape

ConnectionLens in the scientific landscape (1)

Data integration for structured, semistructured and unstructured data

- "Ad-hoc" (combinations of sources to be unioned, joined, or chained)
- No schema, ontologies, queries known in advance
- Reachability queries instead of join: price to pay for powerful integration
- Comparable work by R. S. Roy, G. Weikum: GST querying of RDF graph enriched on-the-fly with content extracted from Web sources
  - Web info may be missing or not trusted

Data cleaning: Similarity links require value or entity matching

- Use reference data sources, heuristics, interactive cleaning (under integrated)

ConnectionLens in the scientific landscape (2)

Graph construction

- Users of **entity extraction** modules, trained a model for French

Keyword search on structured data

- Intensely studied for relational, graph, or XML databases: DISCOVER, Banks & Banks2, DPBF, ObjectRank, BLINKS, Spark, … QGPT (WWW 2021)
  - Some assume regularity in the graph and translate to SQL or SPARQL
  - Consider more limited search spaces (edge direction; one tree per root; …)
  - Exploit favorable properties of fixed score function; establish approximations…
- We focus on efficient keyword search for arbitrary graphs (up to time-out) w/o assumptions on score, w/o sub-optimal structure prop., w/ bidirectional search
- In-memory graph store and parallel query processor (200x speed-up)

Ongoing work

- Learning to predict extraction success ~ detecting disguised nulls (BDA 2021, invited TLDKS 2022)
- Integrating keyword search into structured graph querying (with A. Anadiotis, M. Mohanty)
- Relationship extraction based on OpenIE (with O. Balalau, P. Upadhyay)
- Natural language graph querying (with O. Balalau, K. Zhang)
- Improving the quality of graph linking (with T. Bouganim, H. Galhardas)
- Abstracting CL graphs (with N. Barret, P. Upadhyay) https://team.inria.fr/cedar/projects/abstra/
- Applications:
  - Conflicts of Interests in the biomedical domain (w/ S. Horel and G. Fooks, Aston U., UK)
  - France's business x political sphere (w/ Radio France Investigation)

Why data journalism?

Because I grew up in a dictatorship, and I value free press

Because journalists are threatened and killed still today in Europe

Daphne Galizia, 1964-2017

Jan Kuciak, 1990-2018

Because the press' economic model is threatened by IT giants

Because this industry is currently underserved by IT – and we could really make an impact!

Thank you

Questions?

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