From Fact-Checking to Journalistic Data Integration

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

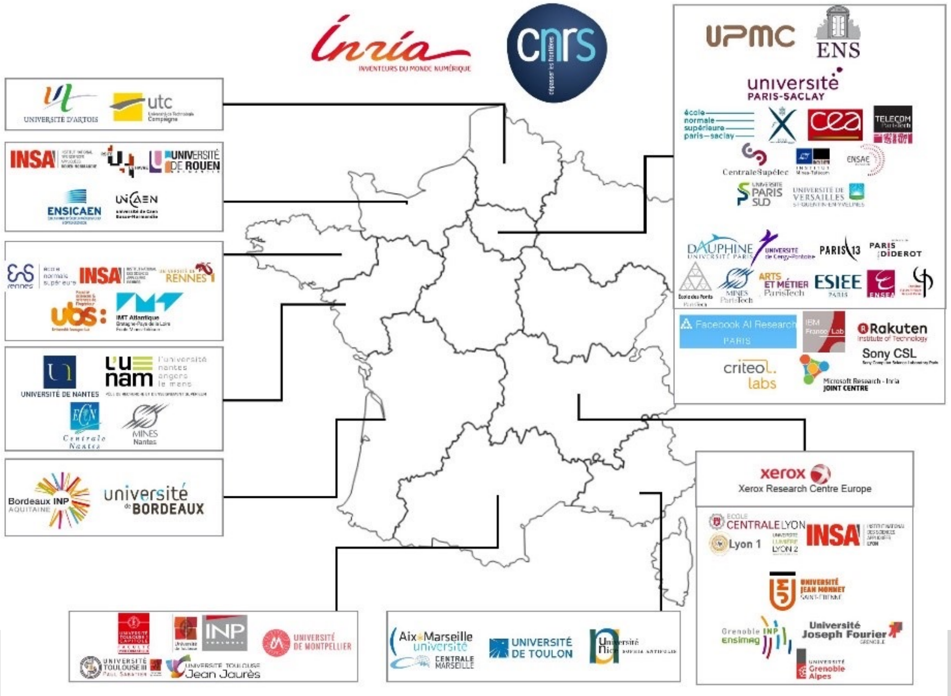
Outline

1. Group presentation
2. Motivation: why journalism?
3. Towards automatic checking of statistic claims
4. Integrating (very!) heterogeneous journalistic data
5. Related work and perspective

CEDAR team presentation

CEDAR: Inria and Ecole Polytechnique

Inria: French national research institute in Computer Science and Applied Mathematics, since 1976



Ecole Polytechnique: created in 1794 to train military engineers

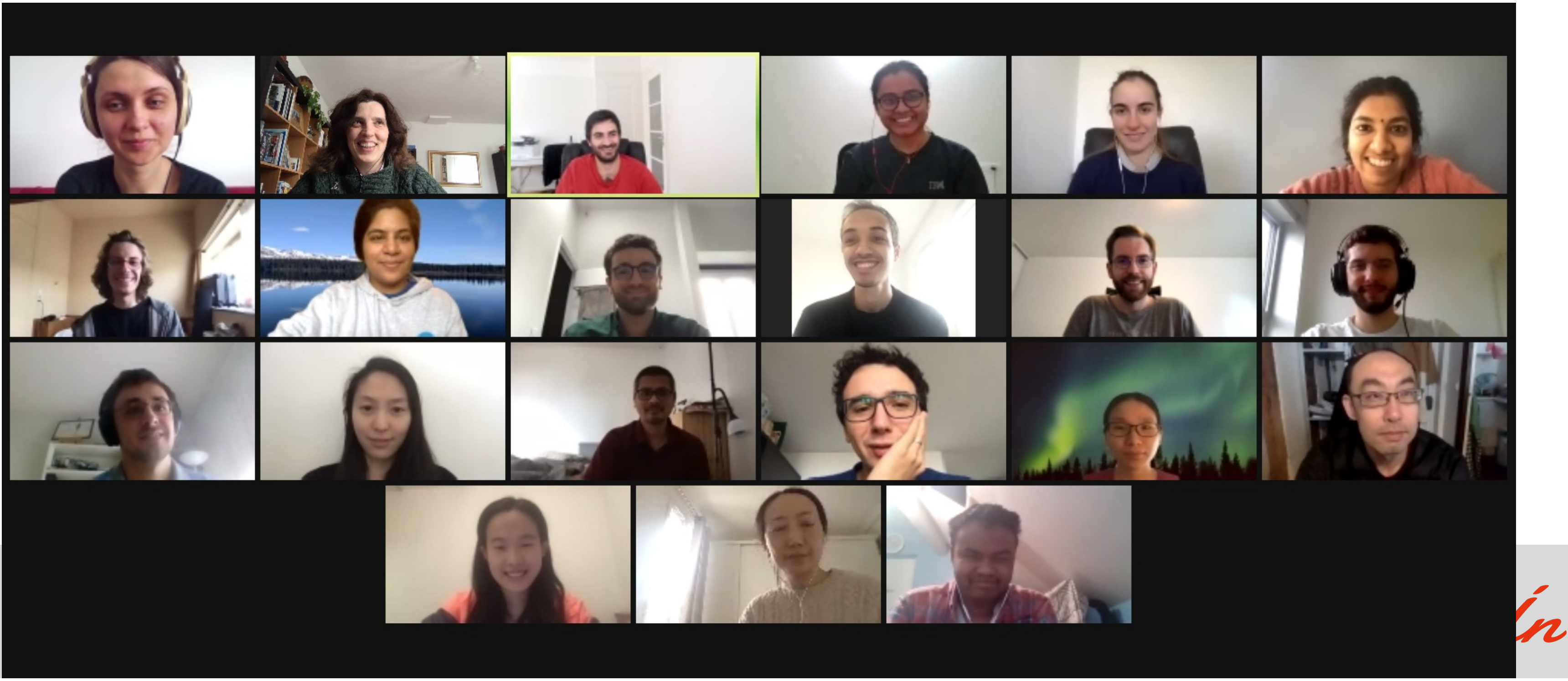
Among the professors: Ampère, Fourier, Monge, Laplace, Cauchy, Becquerel; 2018 Nobel prize in physics…

POUR LA PATRIE LES SCIENCES ET LA GLOIRE

CEDAR team: Inria and Polytechnique

Created in **2016** (I. Manolescu moved from U. Paris Sud, Yanlei Diao from U. Massachussets at Amherst)

Junior faculty: Angelos Anadiotis (EPFL), Oana Balalau (MPI Saarbrucken)



Motivation:
Why journalism?

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

ICIJ

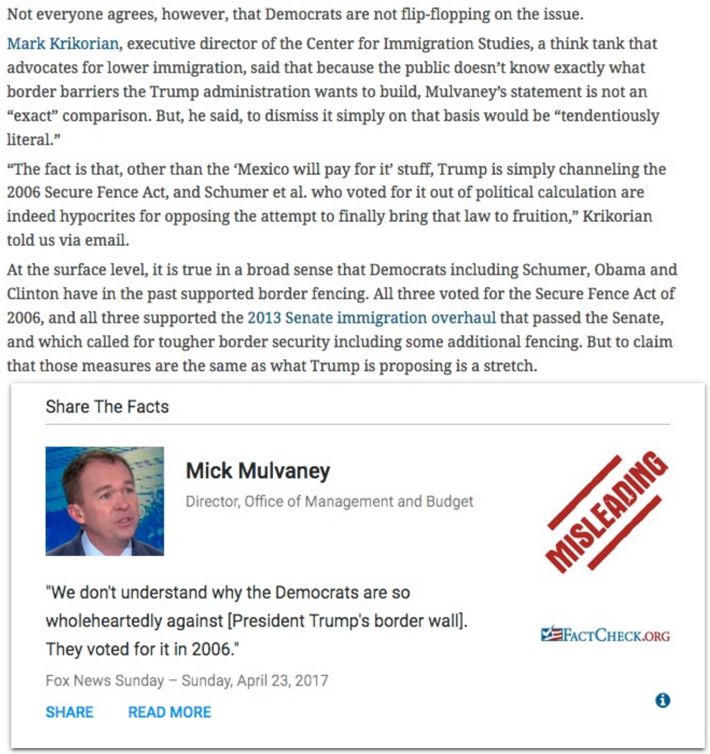
The International Consortium of Investigative Journalists

Home Projects Journalists About ICIJ Leak to us Donate

THE PANAMA PAPERS

Politicians, Criminals and the Rogue Industry That Hides Their Cash

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

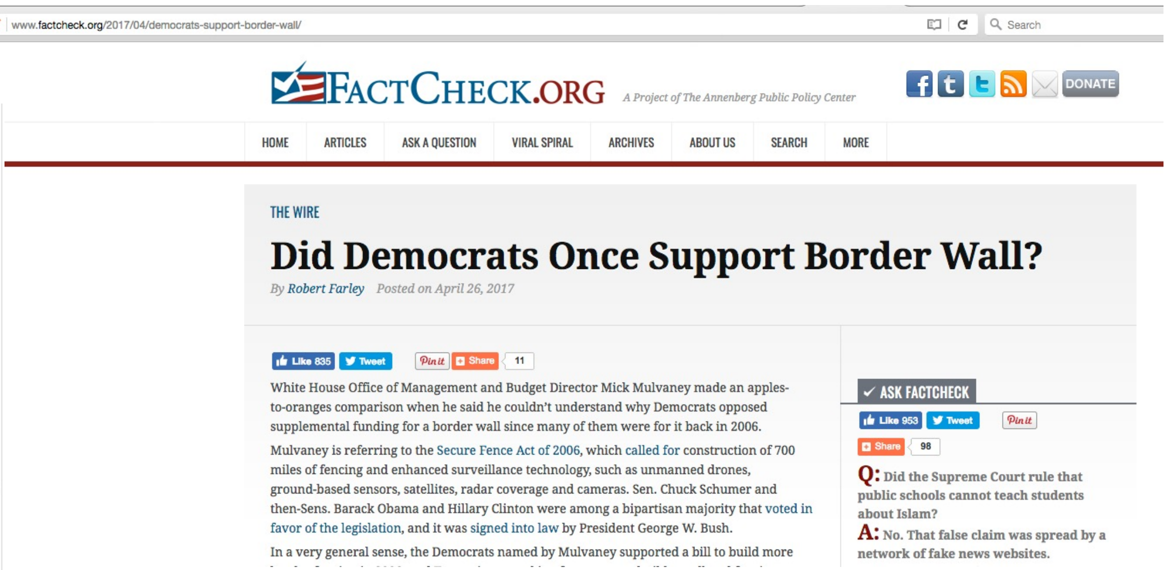
MISLEADING

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

Fox News Sunday – Sunday, April 23, 2017

FACTCHECK.ORG

SHARE READ MORE



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

FACTCHECK.ORG A Project of The Annenberg Public Policy Center

THE WIRE

Did Democrats Once Support Border Wall?

By Robert Farley Posted on April 26, 2017

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.

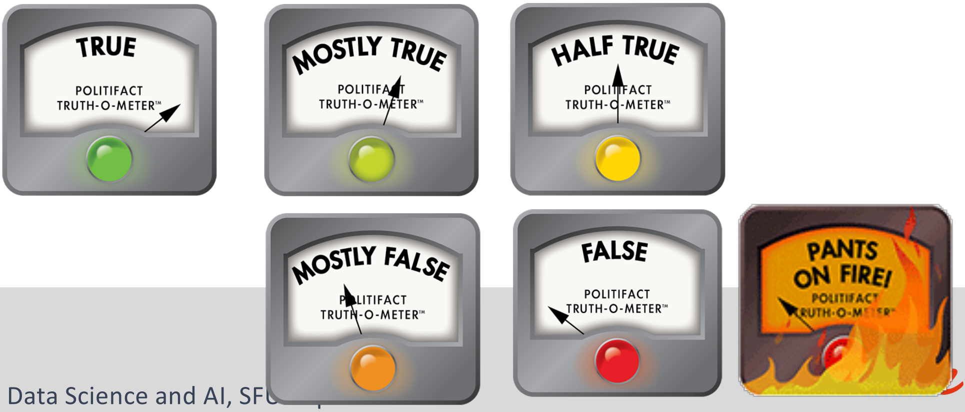
Mulvaney is referring to the Secure Fence Act of 2006, which called for construction of 700 miles of fencing and enhanced surveillance technology, such as unmanned drones, ground-based sensors, satellites, radar coverage and cameras. Sen. Chuck Schumer and then-Sens. Barack Obama and Hillary Clinton 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

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

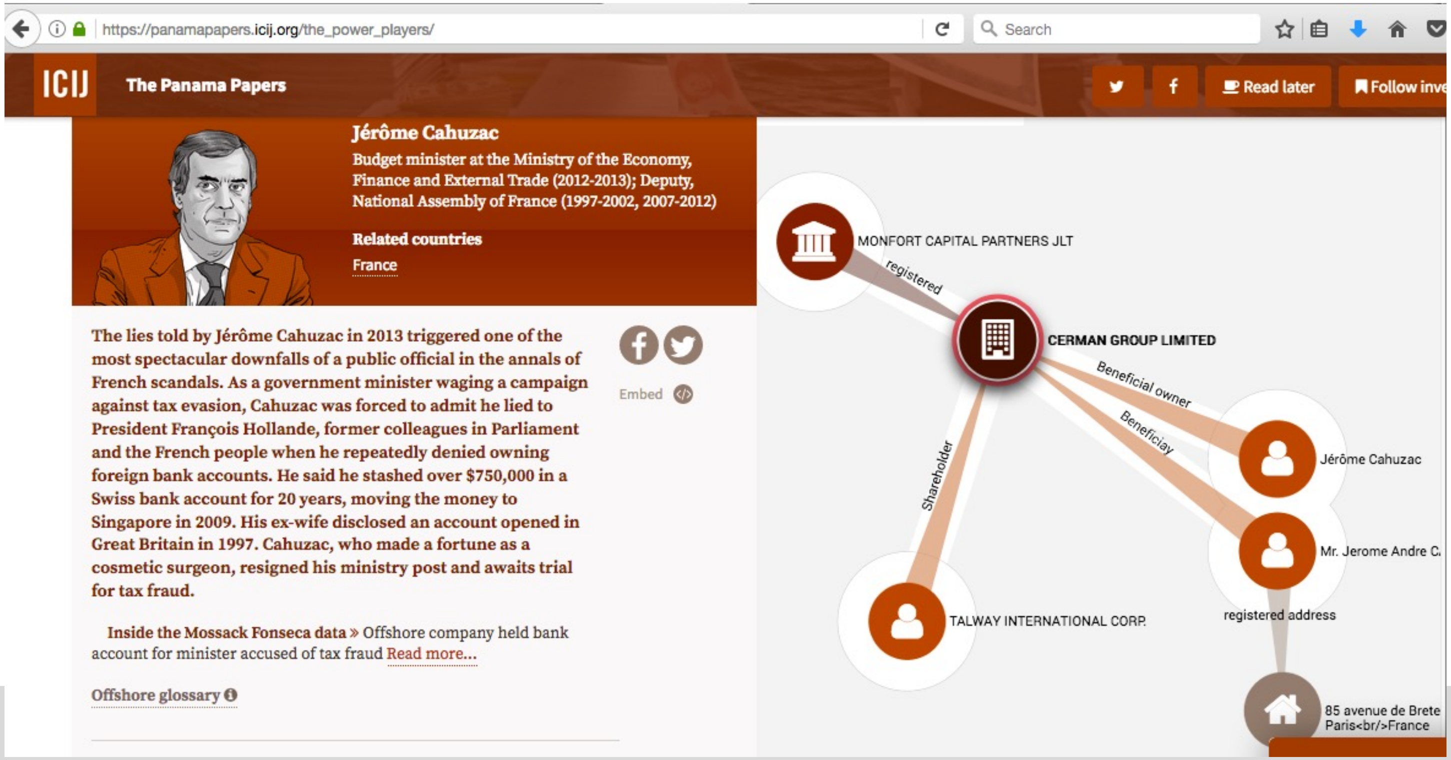
A: No. That false claim was spread by a network of fake news websites.



TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY TRUE
POLITIFACT TRUTH-O-METER™
HALF TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY FALSE
POLITIFACT TRUTH-O-METER™
FALSE
POLITIFACT TRUTH-O-METER™
PANTS ON FIRE!
POLITIFACT TRUTH-O-METER™

Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)



Jérôme Cahuzac

Budget minister at the Ministry of the Economy, Finance and External Trade (2012-2013); Deputy, National Assembly of France (1997-2002, 2007-2012)

Related countries

France

The lies told by Jérôme Cahuzac in 2013 triggered one of the most spectacular downfalls of a public official in the annals of French scandals. As a government minister waging a campaign against tax evasion, Cahuzac was forced to admit he lied to President François Hollande, former colleagues in Parliament and the French people when he repeatedly denied owning foreign bank accounts. He said he stashed over $750,000 in a Swiss bank account for 20 years, moving the money to Singapore in 2009. His ex-wife disclosed an account opened in Great Britain in 1997. Cahuzac, who made a fortune as a cosmetic surgeon, resigned his ministry post and awaits trial for tax fraud.

Inside the Mossack Fonseca data » Offshore company held bank account for minister accused of tax fraud Read more...

Offshore glossary

Projects and collaborations

Google Award (2015) with X. Tannier (U. Paris Sud)

ANR ContentCheck (2016-2020) with
Sorbonne Université, U. Lyon, U. Rennes 1,
Les Décodeurs (Le Monde) https://contentcheck.inria.fr

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

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

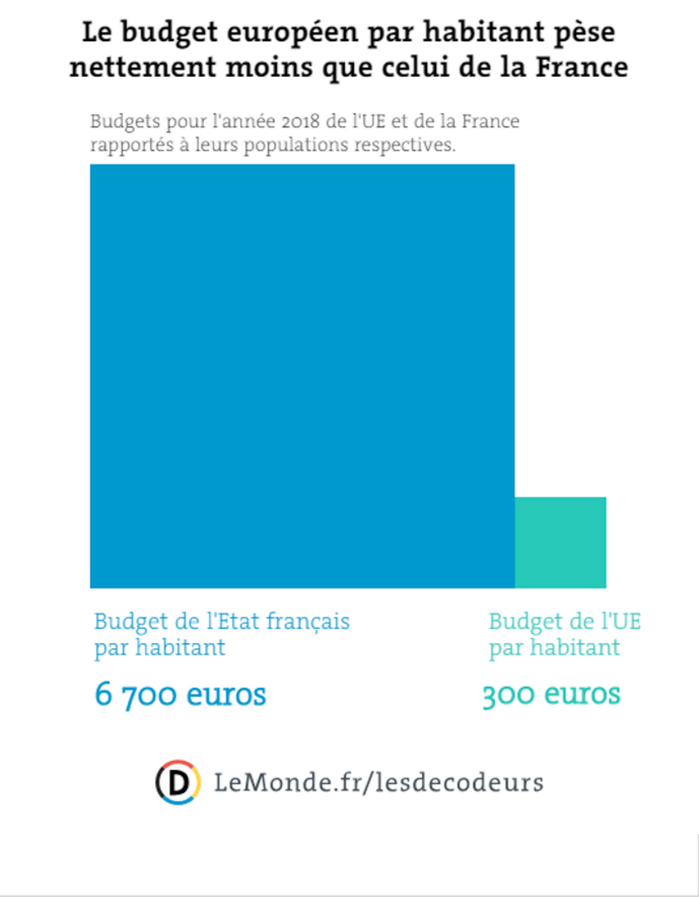
ANR SourcesSay AI Chair (2020-2023),
with Le Monde and WeDoData https://sourcessay.inria.fr

Towards automated fact-checking of statistic claims

Most common fact-checking scenarios

- ❑ "What is the value of metric X in space Y at time T"?
  - ❑ X=youth unemployment, Y=Germany, T=2018
  - ❑ X=illegal immigrants, Y=Italy, T=[2015-2018]
  - ❑ X=budget for research, Y=Canada, T=2020
- ❑ Comparisons
  - ❑ X1 against X2; Y1 against Y2; T1 against T2; temporal trend etc.

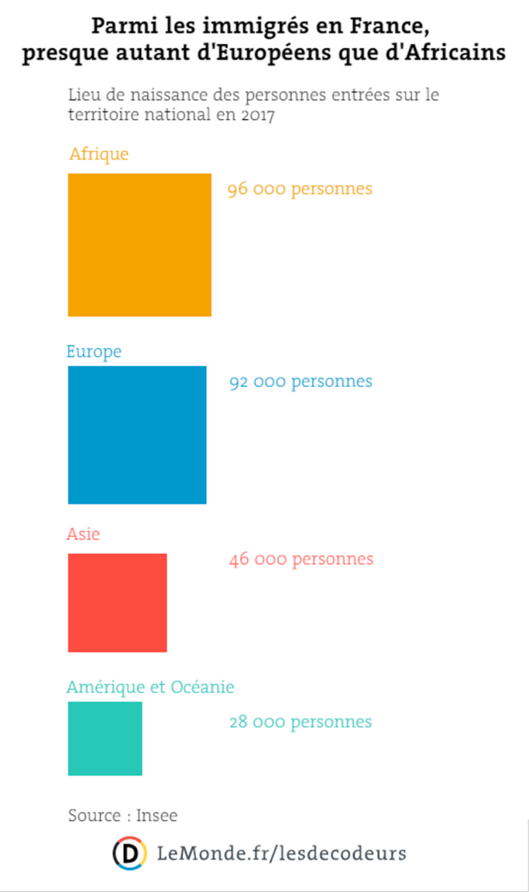
Most common fact-checking scenarios



Le budget européen par habitant pèse nettement moins que celui de la France

Budgets pour l'année 2018 de l'UE et de la France rapportés à leurs populations respectives.

LeMonde.fr/lesdecodeurs



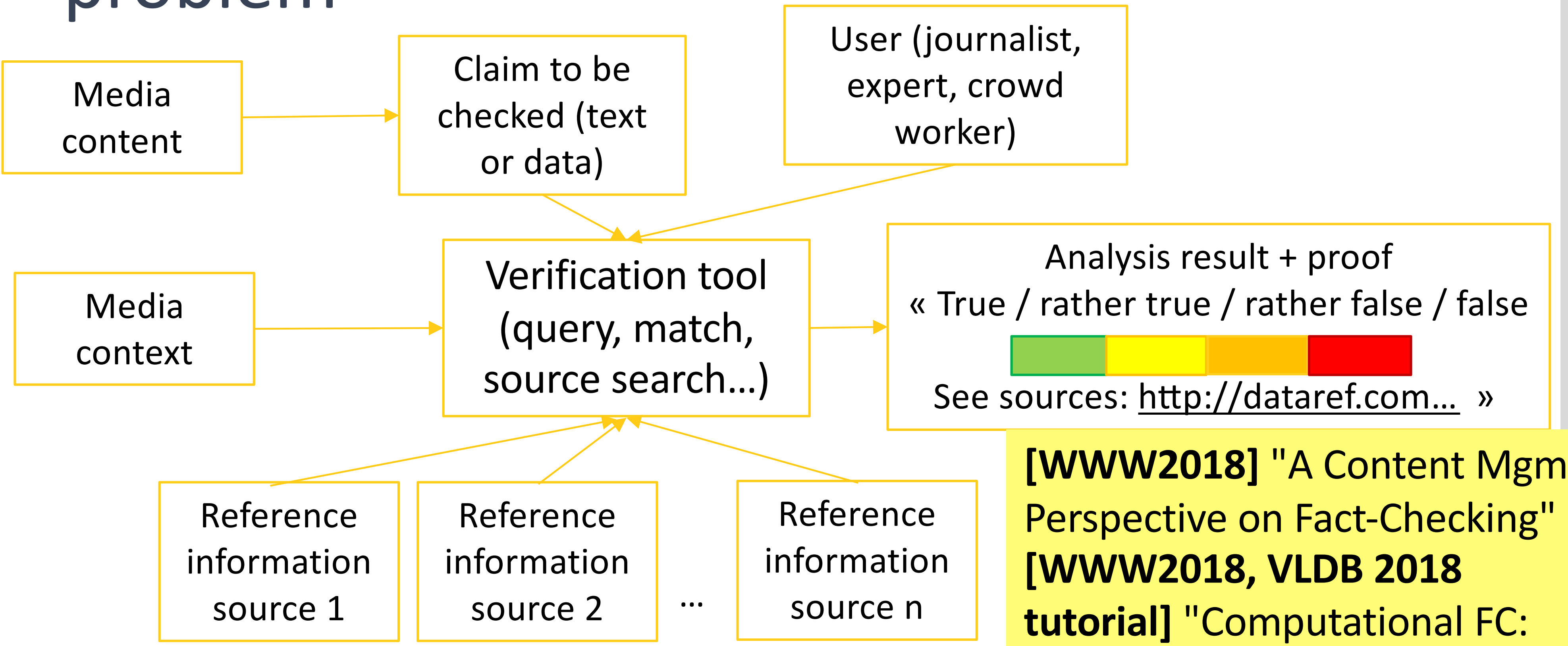
Parmi les immigrés en France, presque autant d'Européens que d'Africains

Lieu de naissance des personnes entrées sur le territoire national en 2017

Source : Insee

LeMonde.fr/lesdecodeurs

Fact-checking as a content management problem



Media content

Claim to be checked (text or data)

User (journalist, expert, crowd worker)

Media context

Verification tool (query, match, source search…)

Analysis result + proof

« True / rather true / rather false / false

🟩🟨🟧🟥

See sources: http://dataref.com… »

Reference information source 1

Reference information source 2

…

Reference information source n

[WWW2018] "A Content Mgmt. Perspective on Fact-Checking" **[WWW2018, VLDB 2018 tutorial]** "Computational FC: problems, state of the art, and perspectives"

Facilitating **statistical** fact-checking

INSEE: French national institute of statistics

- □ Publish valuable statistic datasets about economy, health, education etc. yearly or per quarter
- □ Web pages + statistic information as tabular files (mostly Excel)

UN, OCDE, IMF: SDMX databases

- □ Special Data Warehouse-style format for describing the data

Insee
Mesurer pour comprendre

sdmx
Statistical Data and Metadata eXchange

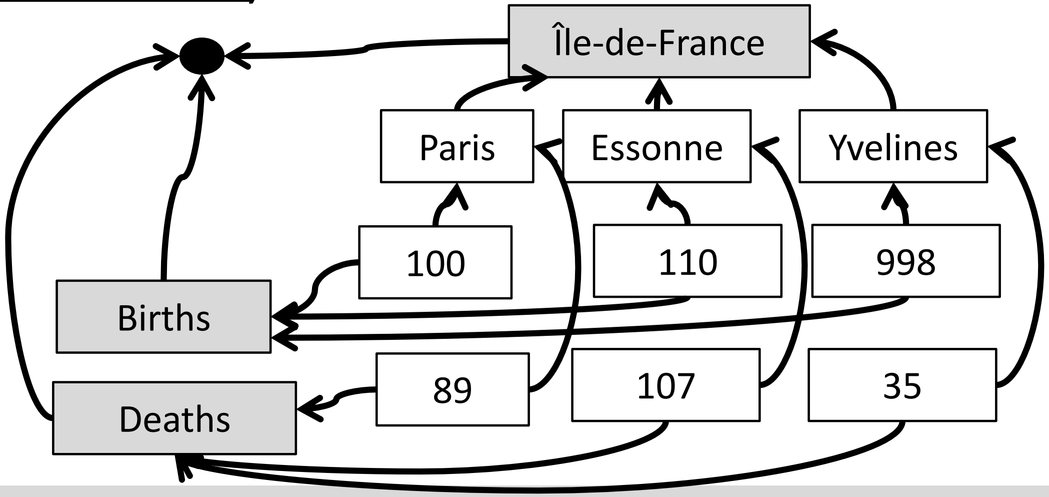
How to facilitate using them as reference data sources?

Fact-checking using INSEE data [SBD2017]

1. Crawled complete INSEE publication site nightly, gather HTML+XLS files (https://gitlab.inria.fr/cedar/insee-crawler)
2. Extract data from all statistic cells into RDF, preserving the connections between the cells (https://gitlab.inria.fr/cedar/excel-extractor)

Insee
Mesurer pour comprendre

	Ile-de-France		
	Paris	Essonne	Yvelines
Births	100	110	98
Deaths	89	107	35

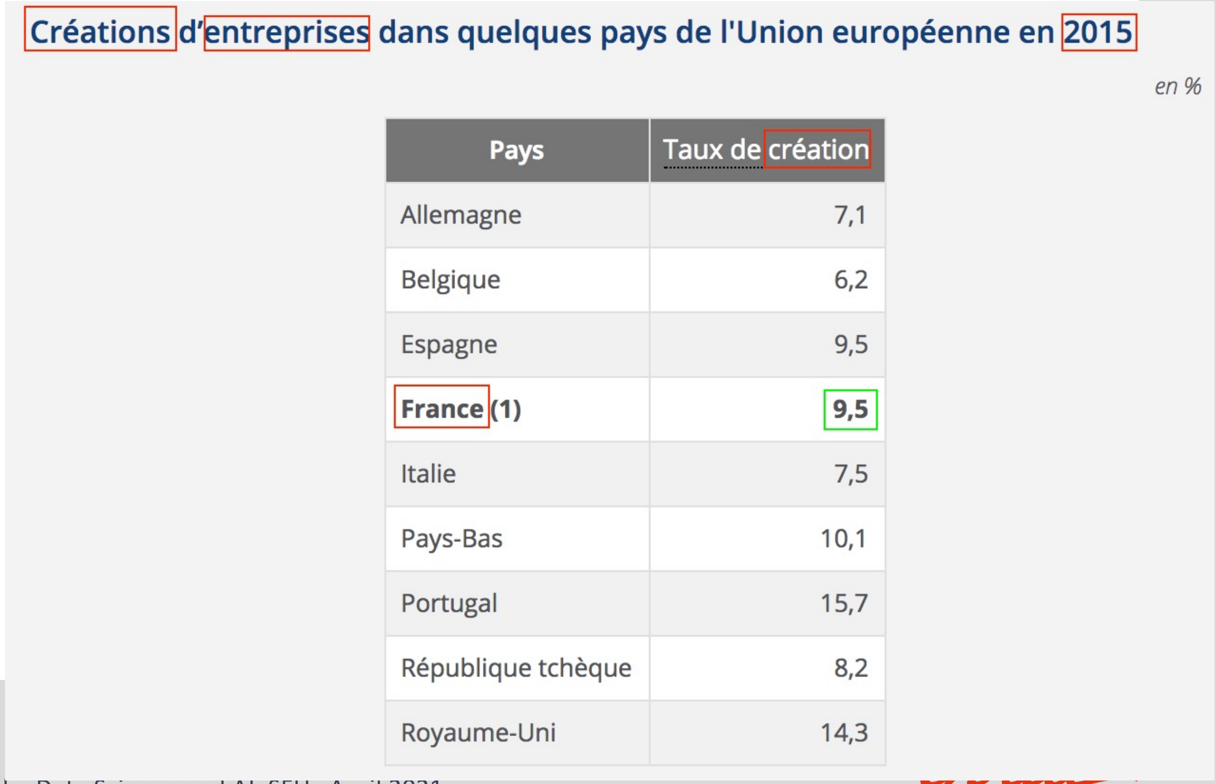


Fact-checking using INSEE data [WebDB2018]

3. **Keyword search** algorithm on resulting RDF graph which, given *“Créations d’entreprises en France en 2015”* , returns:

- ❑ Searches for line and column
- ❑ Returns **cell at their intersection**, if possible
- ❑ Otherwise, column or line
- ❑ Otherwise, spreadsheet
- ❑ Always with provenance (link to INSEE Web site)

Started similar project on SDMX [NLIWOD2020]



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

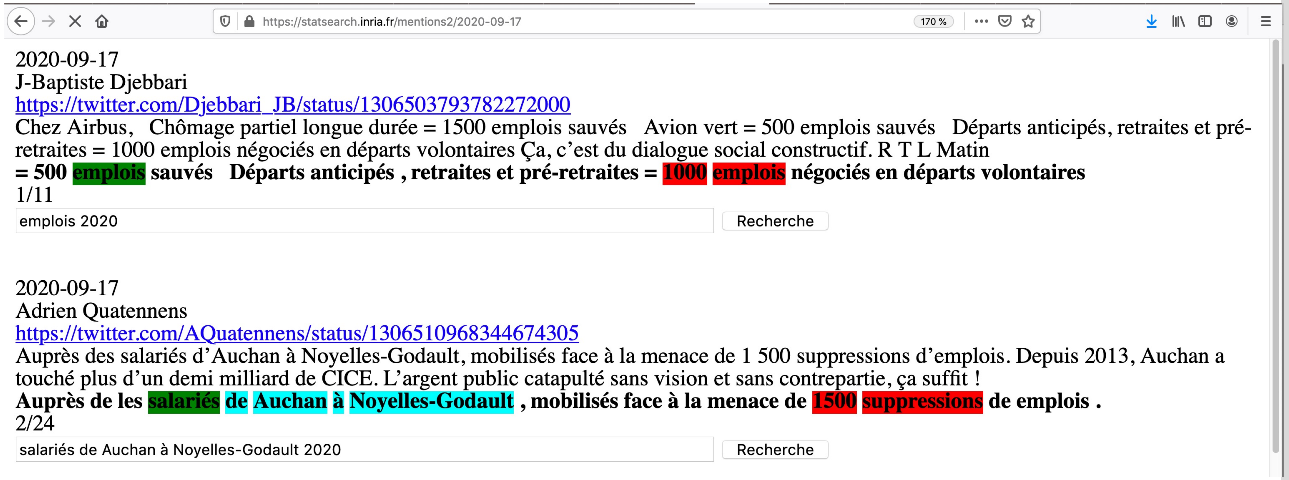
en %

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

Fact-checking using INSEE data [NLDB2019]

4. **Statistic claim extraction** algorithm which reads incoming tweets and identifies: Measure, Quantity, [Where], [When], [Other dimensions]

- ❑ Formulates keyword query with Measure, [Where], [When], [Other dimensions] for the INSEE search algoritm
- ❑ Leave it to the user to decide how to interpret the difference between claimed and found values



2020-09-17
J-Baptiste Djebbari
https://twitter.com/Djebbari_JB/status/1306503793782272000
Chez Airbus, Chômage partiel longue durée = 1500 emplois sauvés Avion vert = 500 emplois sauvés Départs anticipés, retraites et pré-retraites = 1000 emplois négociés en départs volontaires Ça, c'est du dialogue social constructif. R T L Matin
= 500 emplois sauvés Départs anticipés , retraites et pré-retraites = 1000 emplois négociés en départs volontaires
1/11

emplois 2020 | Recherche

2020-09-17
Adrien Quatennens
https://twitter.com/AQuatennens/status/1306510968344674305
Auprès des salariés d'Auchan à Noyelles-Godault, mobilisés face à la menace de 1 500 suppressions d'emplois. Depuis 2013, Auchan a touché plus d'un demi milliard de CICE. L'argent public catapulté sans vision et sans contrepartie, ça suffit !
Auprès de les salariés de Auchan à Noyelles-Godault , mobilisés face à la menace de 1500 suppressions de emplois .
2/24

salariés de Auchan à Noyelles-Godault 2020 | Recherche

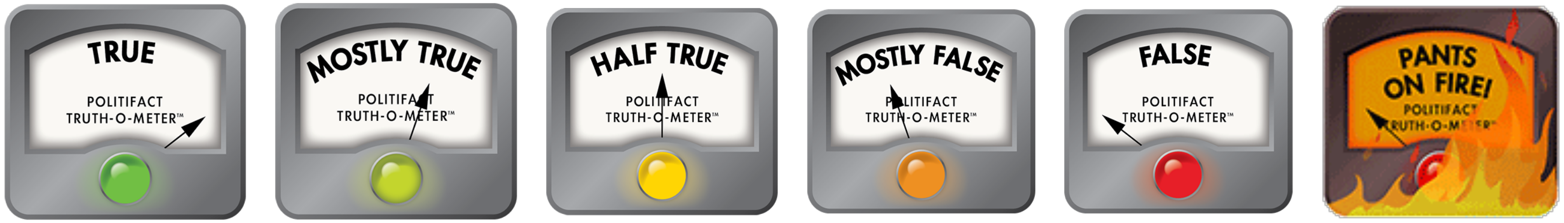
Integrating heterogeneous journalistic datasets

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

A 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


TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY TRUE
POLITIFACT TRUTH-O-METER™
HALF TRUE
POLITIFACT TRUTH-O-METER™
MOSTLY FALSE
POLITIFACT TRUTH-O-METER™
FALSE
POLITIFACT TRUTH-O-METER™
PANTS ON FIRE!
POLITIFACT TRUTH-O-METER™

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

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

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

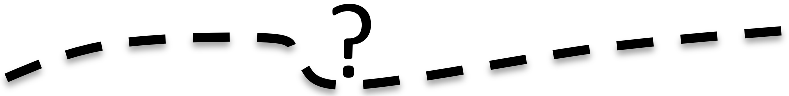
Online fact-checks: (semi)structured data sources (JSON, XML) listing

- 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

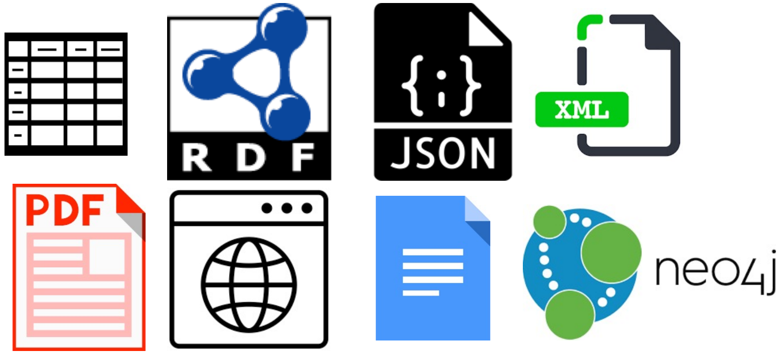
Question: When does a fake news post first cross into a supposedly legitimate community (e.g. members of the Parliament)?



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

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**

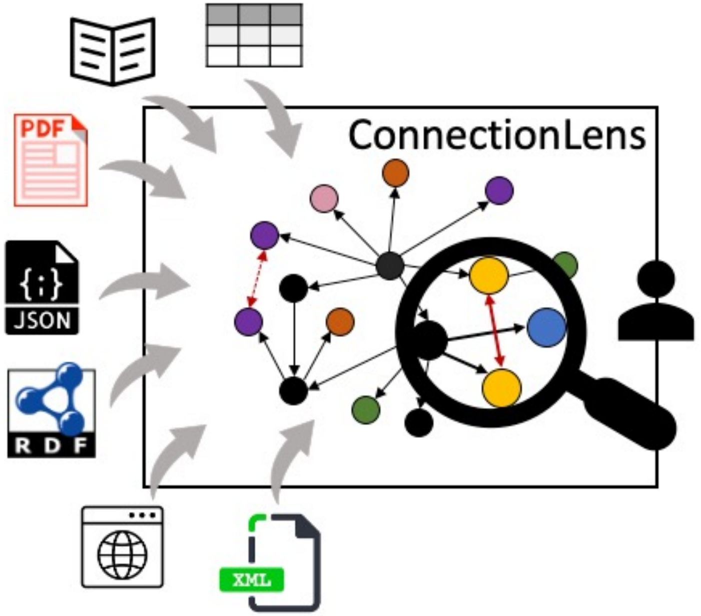
ConnectionLens: graph-based integration of heterogeneous data sources

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 [Chanial et al., 2018]

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 [Chanial et al., 2018]

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** →

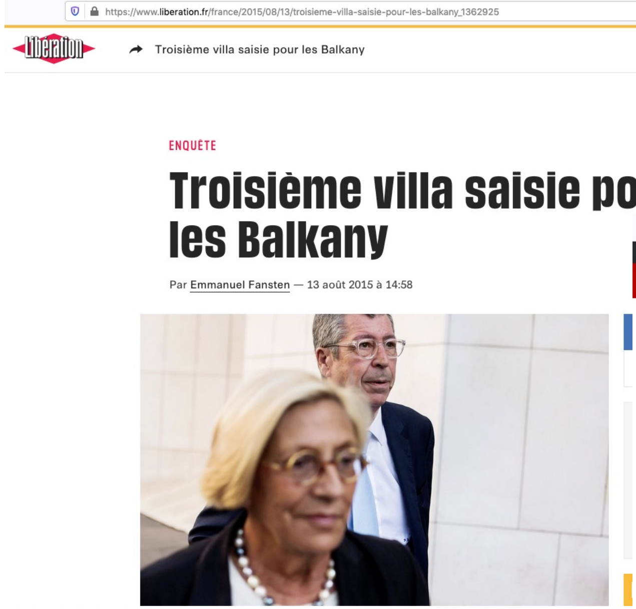
Quer d

❑ Cl

Rest of the talk based on state of the project 2+ years later:
https://arxiv.org/abs/2007.12488 [BDA 2020]
https://arxiv.org/abs/2009.04283 [BDA 2020]
https://arxiv.org/abs/2012.08830 [Invited to Elsevier Information Systems, under minor revision]
https://arxiv.org/abs/2102.04141

ConnectionLens graph construction

The Balkany and their African connections



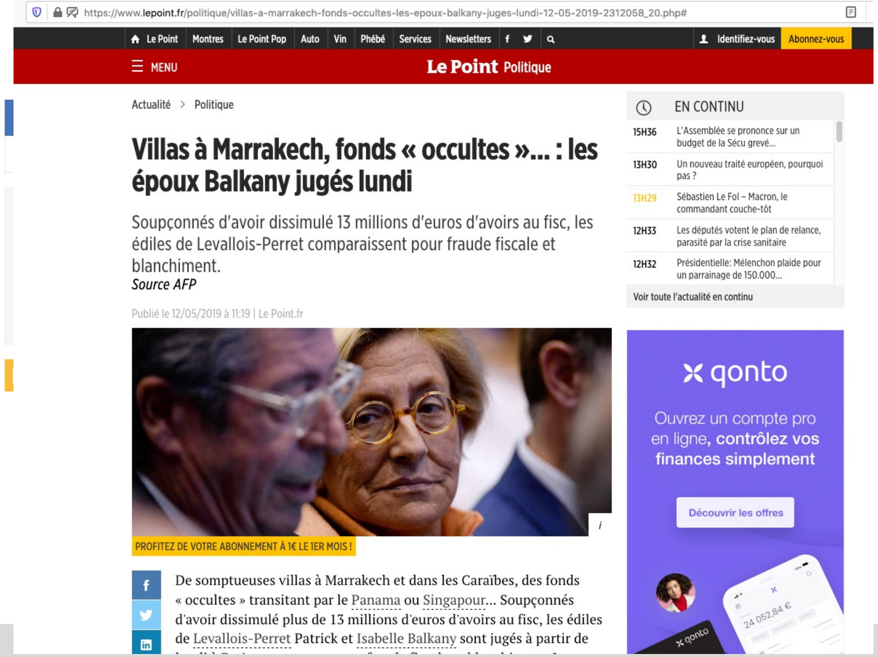
https://www.liberation.fr/france/2015/08/13/troisieme-villa-saisie-pour-les-balkany_1362925

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#

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 !

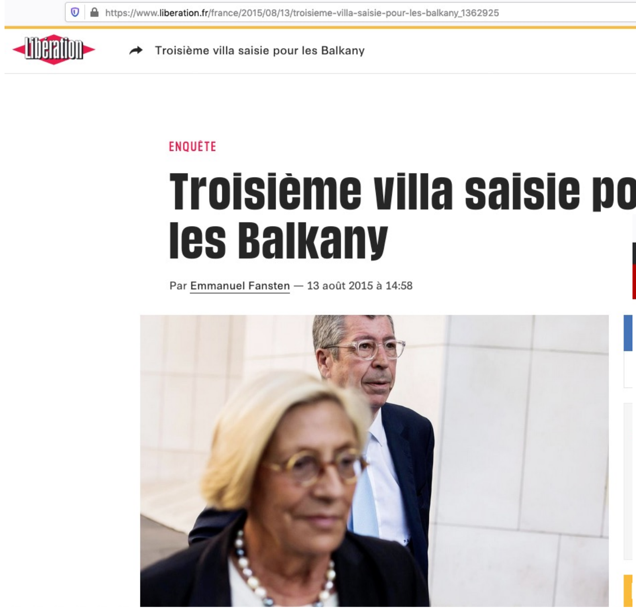
De somptueuses villas à Marrakech et dans les Caraïbes, des fonds « occultes » transitant par le Panama ou Singapour... Soupçonnés d'avoir dissimulé plus de 13 millions d'euros d'avoirs au fisc, les édiles de Levallois-Perret Patrick et Isabelle Balkany sont jugés à partir de

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



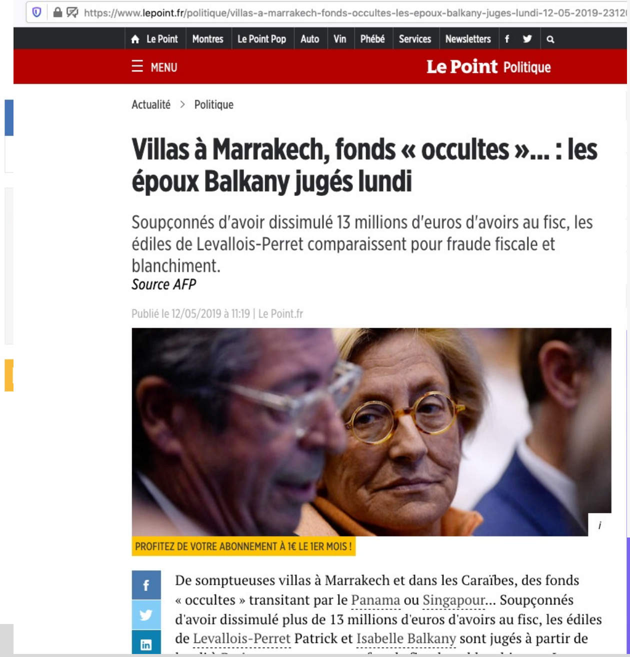
https://www.liberation.fr/france/2015/08/13/troisieme-villa-saisie-pour-les-balkany_1362925

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

Actualité > Politique

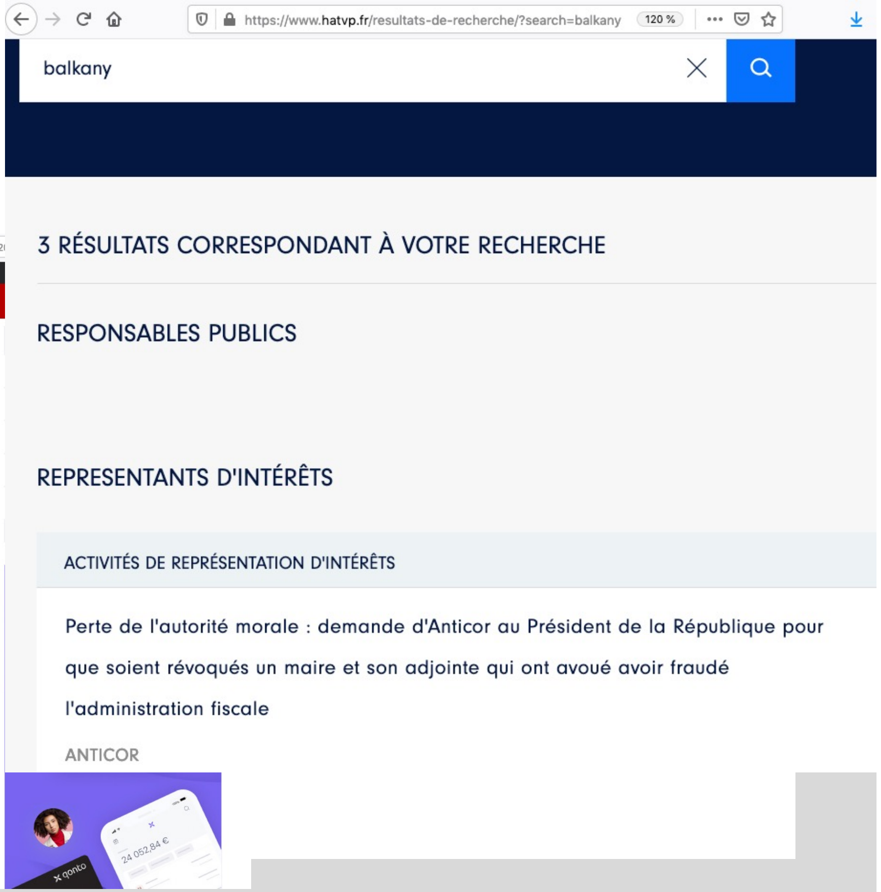
Villas à Marrakech, fonds « occultes »... : les époux Balkany jugés lundi

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

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

National Directory of Elected Officials (JSON)

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

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 .
}
```

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

National Directory of Elected Officials (JSON)

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

dbpedia.org (RDF)

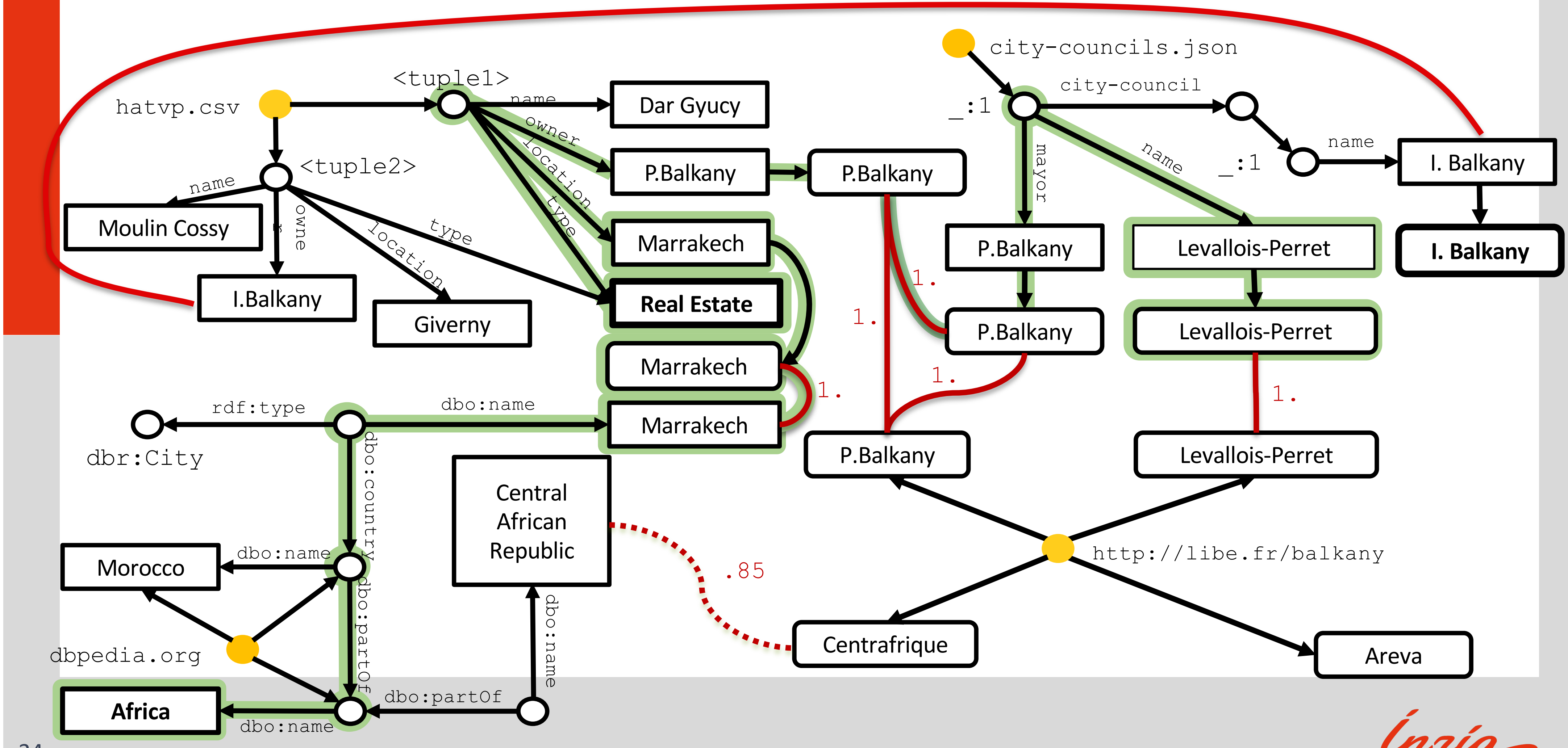
```
{
dbr:Marrakech
  dbr:name      "Marrakech"
  rdf:type      dbo:City ;
  dbo:country   dbr:Morrocco .
dbr:Morocco
  dbr:name      "Morocco"
  rdf:type      dbo:Country
  dbo:locatedIn dbr:Africa .
dbr:CentralAfricanRepublic
  dbr:name      "Central African Republic"
  dbo:locatedIn dbr:Africa .
}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

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. […]

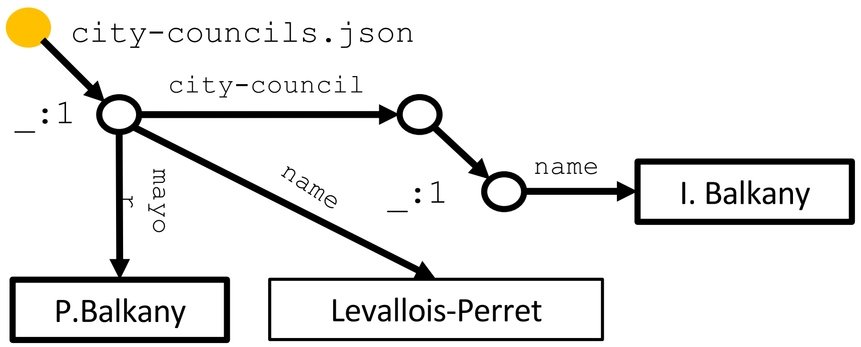
Idea: integrate all data sources into a heterogeneous graph



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



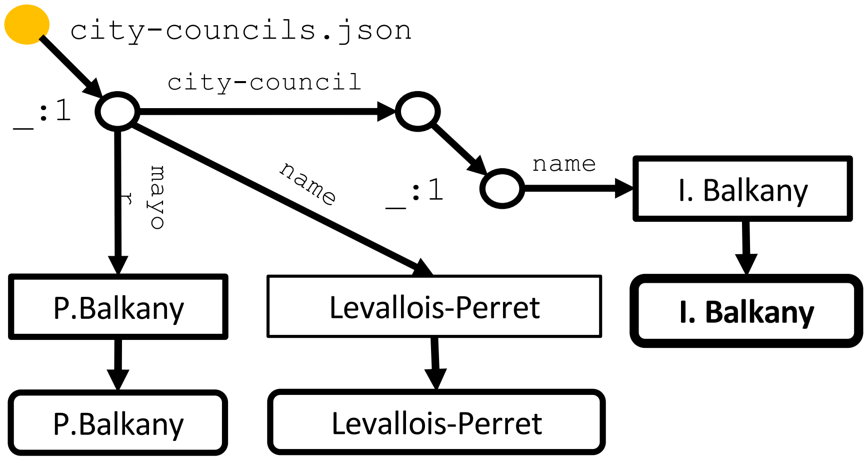
Graph construction stages

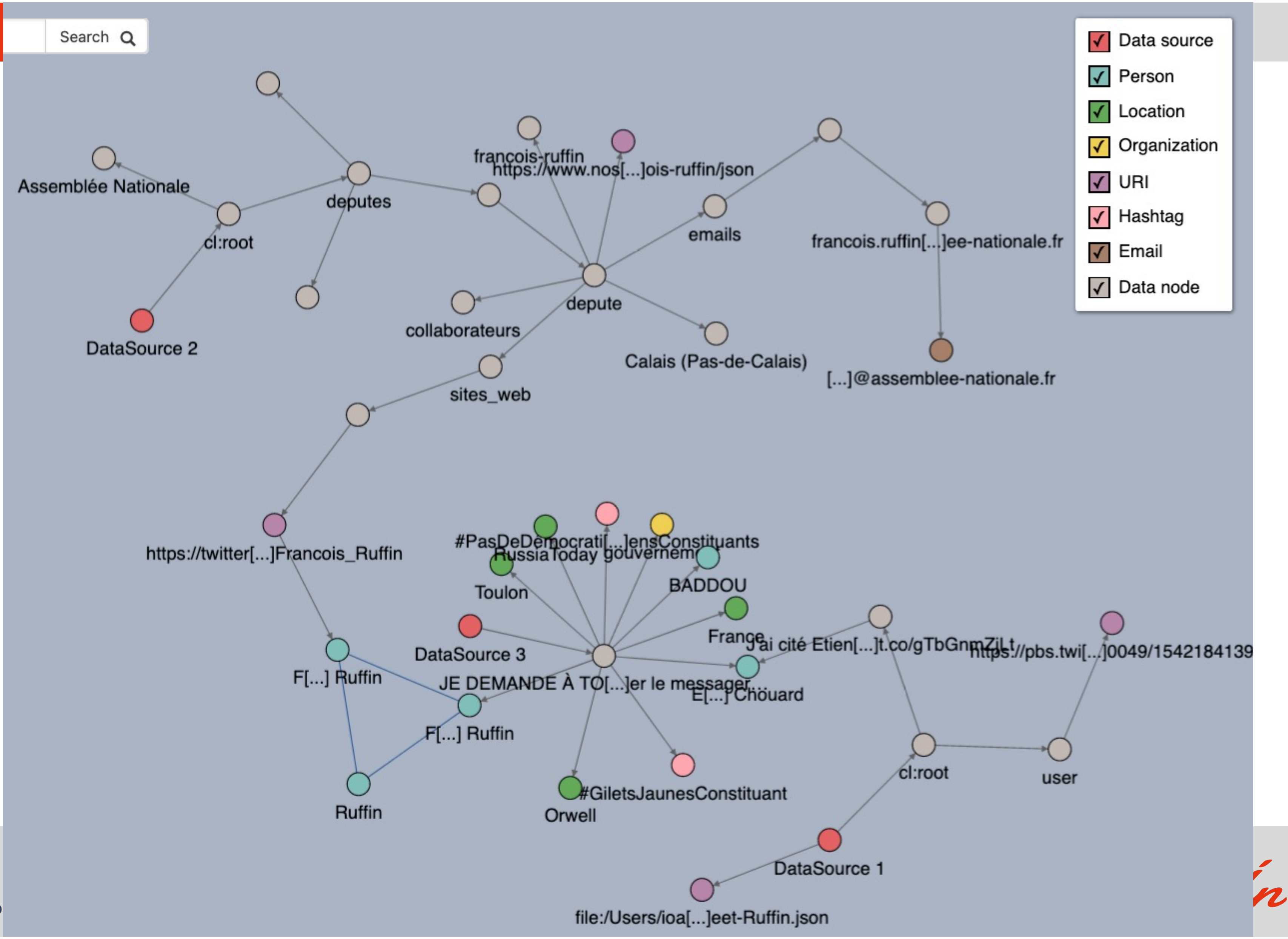
1. Primary node and edge construction

- ❑ 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
- ❑ [BDA2020] Developed and trained new entity extractor from French, based on Flair framework

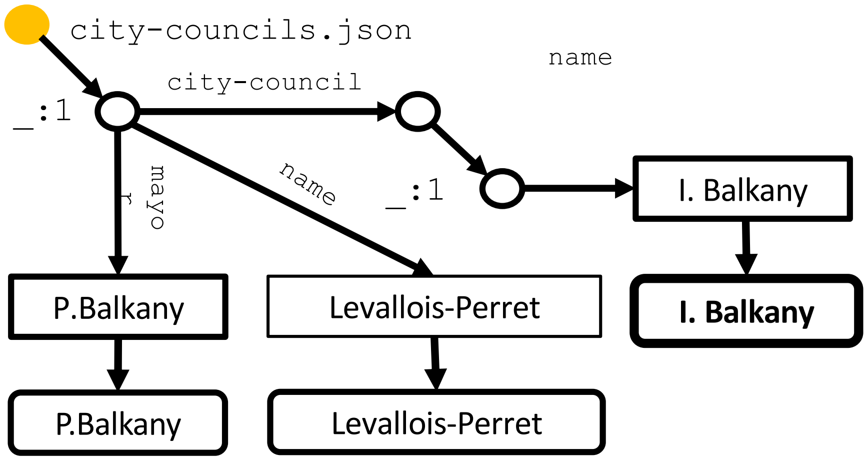



Search
☑ Data source
☑ Person
☑ Location
☑ Organization
☑ URI
☑ Hashtag
☑ Email
☑ Data node
Assemblée Nationale
cl:root
DataSource 2
deputes
francois-ruffin
https://www.nos[...]ois-ruffin/json
emails
francois.ruffin[...]ee-nationale.fr
depute
collaborateurs
Calais (Pas-de-Calais)
[...]@assemblee-nationale.fr
sites_web
https://twitter[...]Francois_Ruffin
#PasDeDemocrati[...]lensConstituants
RussiaToday gouvernem
Toulon
BADDOU
France
J'ai cité Etien[...]t.co/gTbGnmZjLt
https://pbs.twi[...]0049/1542184139
DataSource 3
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

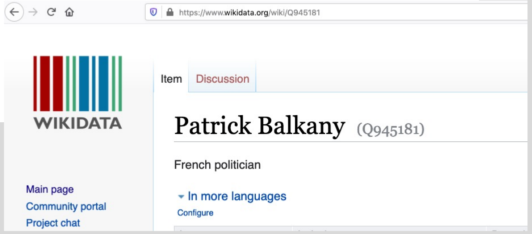
2. Entity extraction

- ❑ 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



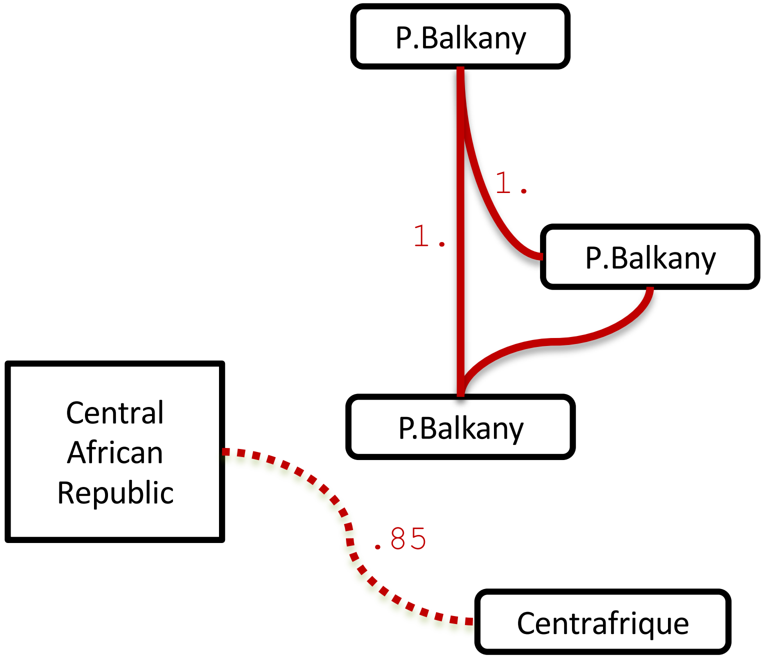
Patrick Balkany (Q945181)

French politician

Graph construction stages

4. Node matching

- ❑ To create sameAs edges:
  - ❑ Strong sameAs edges: equivalent nodes 1.
  - ❑ Weak sameAs edges: similar nodes .85
- ❑ Appropriate distance functions
- ❑ *New*: more normalization → better matching
- ❑ Remains quadratic at the core ☹ , so...



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

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.

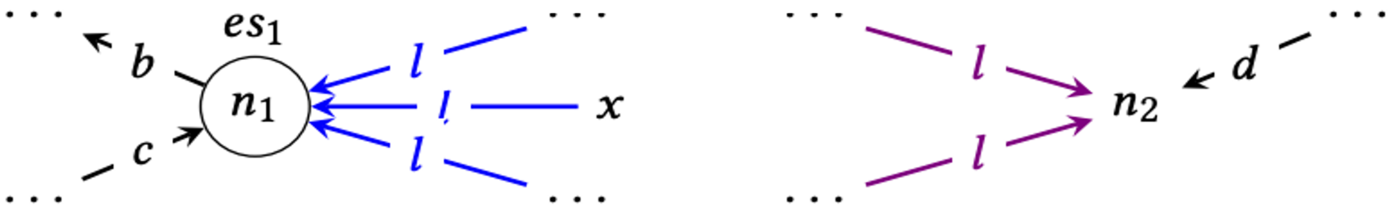
Search space and complexity

- ❑ 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.
  - ❑ Score function may be non-monotonous; no optimal sub-structure property
- ❑ **Large literature on kwd search in text, resp. structured data.**
  - ❑ 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

GAM (Grow and Aggressive Merge) Algorithm

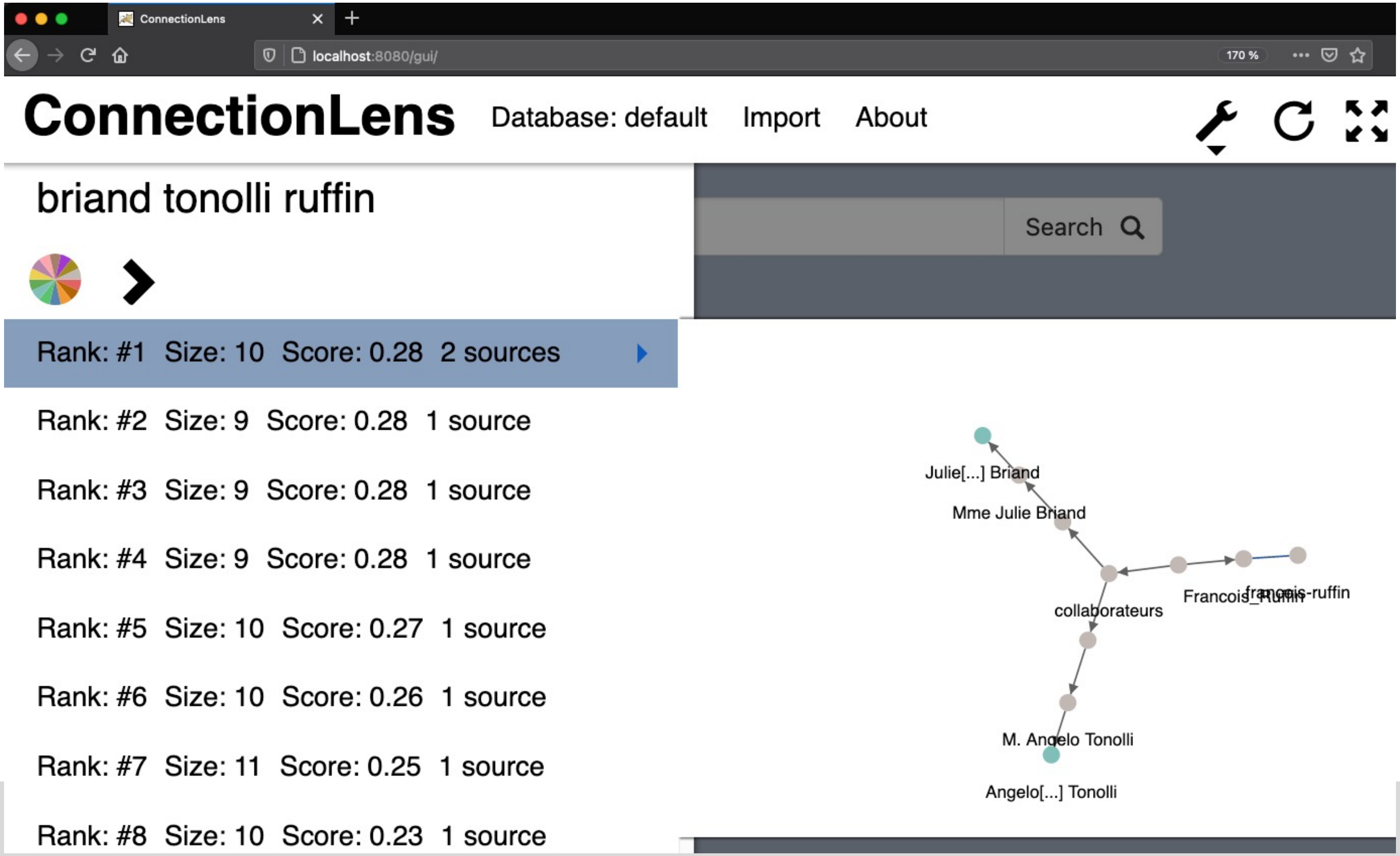
- Builds trees ”backward” from the keyword matches
- **GROW** adds an edge to the root of a tree, **MERGE** merges trees with the same root
- Exploration (GROW) order:
  1. Favor trees matching the largest number of query keywords
  2. To break ties, favor smaller trees
  3. To break second tie between (t1, e1), (t2, e2), we prefer the pair with the higher specificity edge.

The specificity of $e = n_1 \xrightarrow{l} n_2$ is: $s(e) = 2/(N^l_{n_1\rightarrow} + N^l_{\rightarrow n_2})$



Special measures to handle equivalence clusters efficiently

Sample query answers



briand tonolli ruffin

Rank: #1 Size: 10 Score: 0.28 2 sources

Rank: #2 Size: 9 Score: 0.28 1 source

Rank: #3 Size: 9 Score: 0.28 1 source

Rank: #4 Size: 9 Score: 0.28 1 source

Rank: #5 Size: 10 Score: 0.27 1 source

Rank: #6 Size: 10 Score: 0.26 1 source

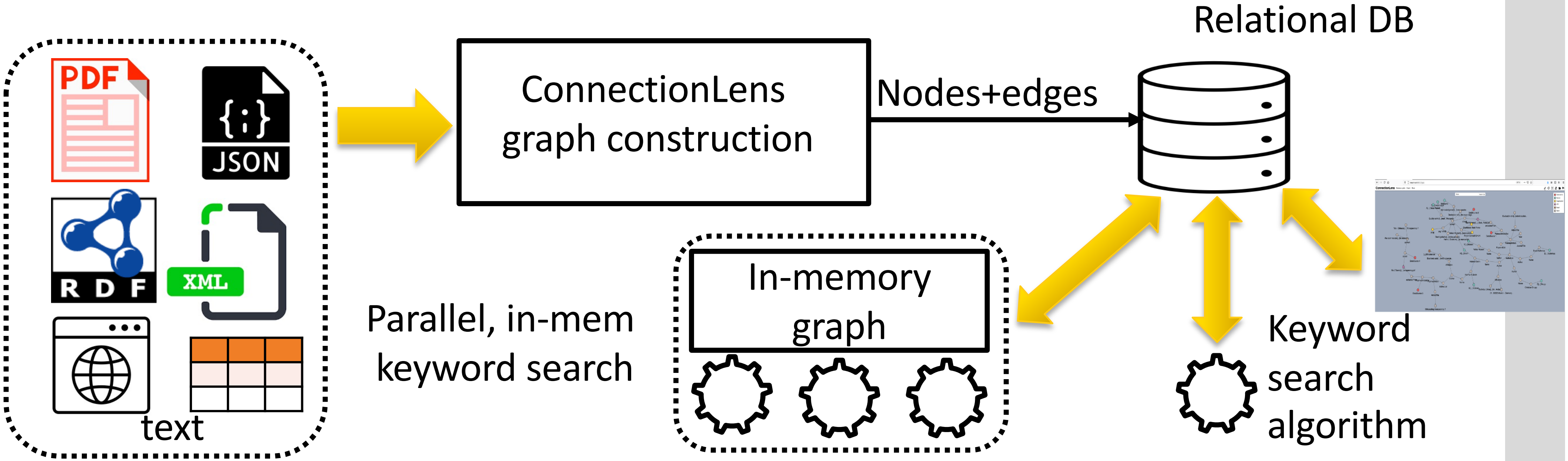
Rank: #7 Size: 11 Score: 0.25 1 source

Rank: #8 Size: 10 Score: 0.23 1 source

ConnectionLens architecture and performance

Implementation

- ❑ Java (220 classes/40K LOC), Python (25 classes/2700 LOC), JS + CSS
- ❑ Available online: https://gitlab.inria.fr/cedar/connectionlens



Implementation https://arxiv.org/abs/2012.08830

- ❑ 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 https://arxiv.org/abs/2012.08830

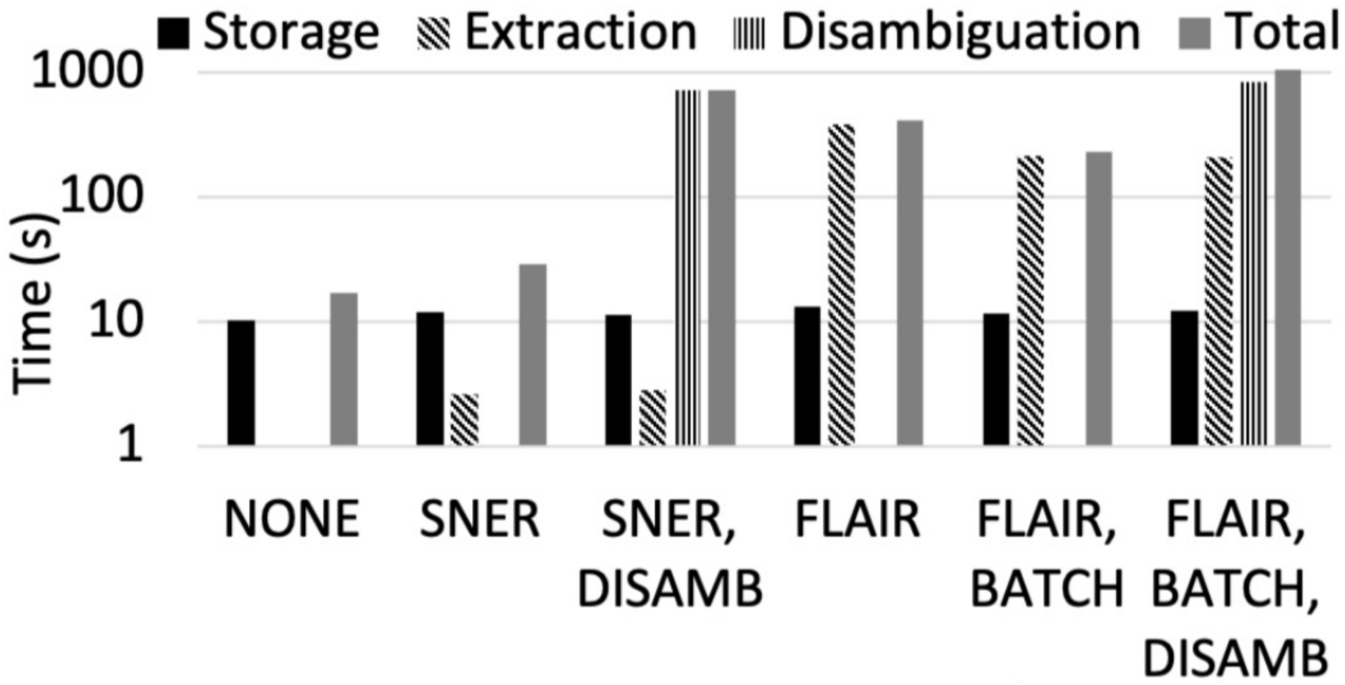


Figure 6: Graph construction time (seconds).

Graph creation performance: batch extraction

https://arxiv.org/abs/2012.08830

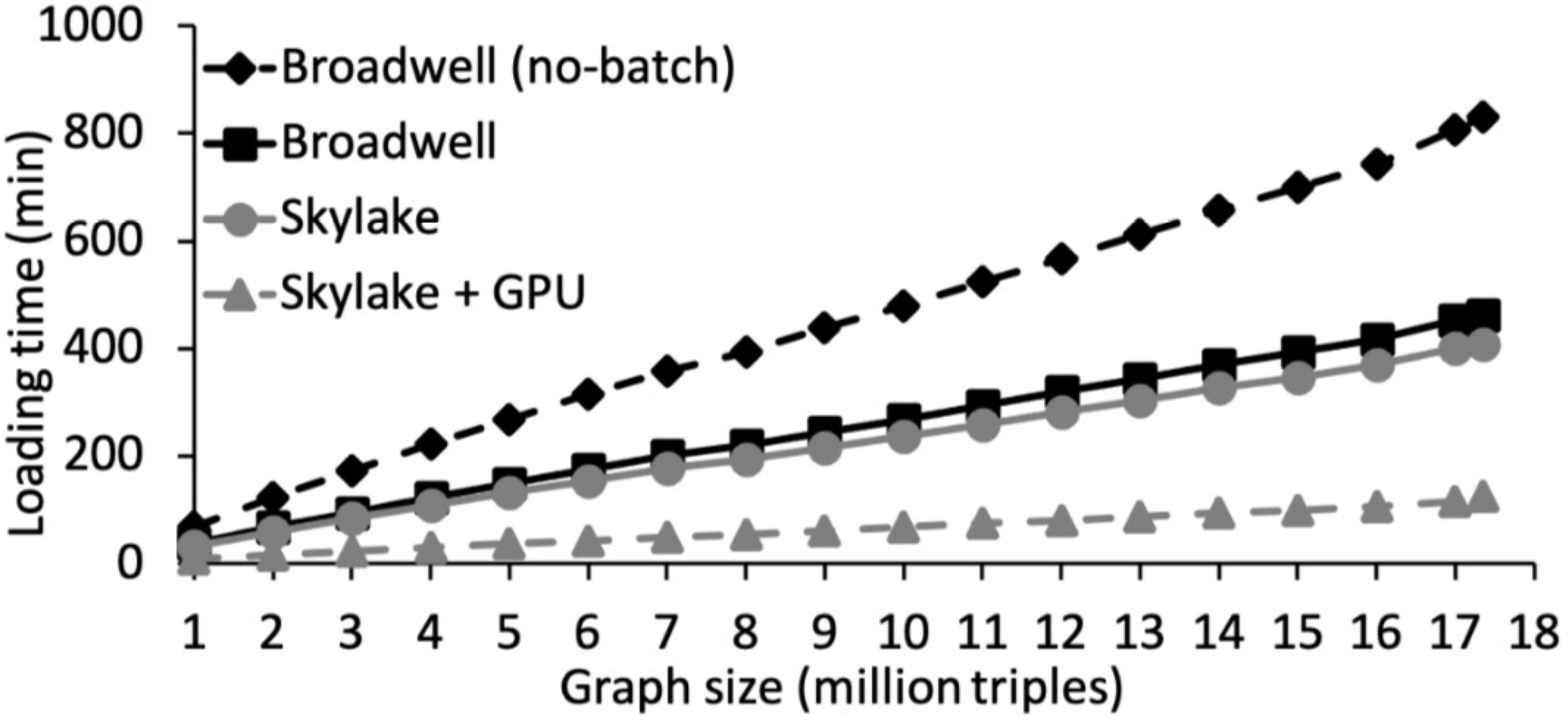
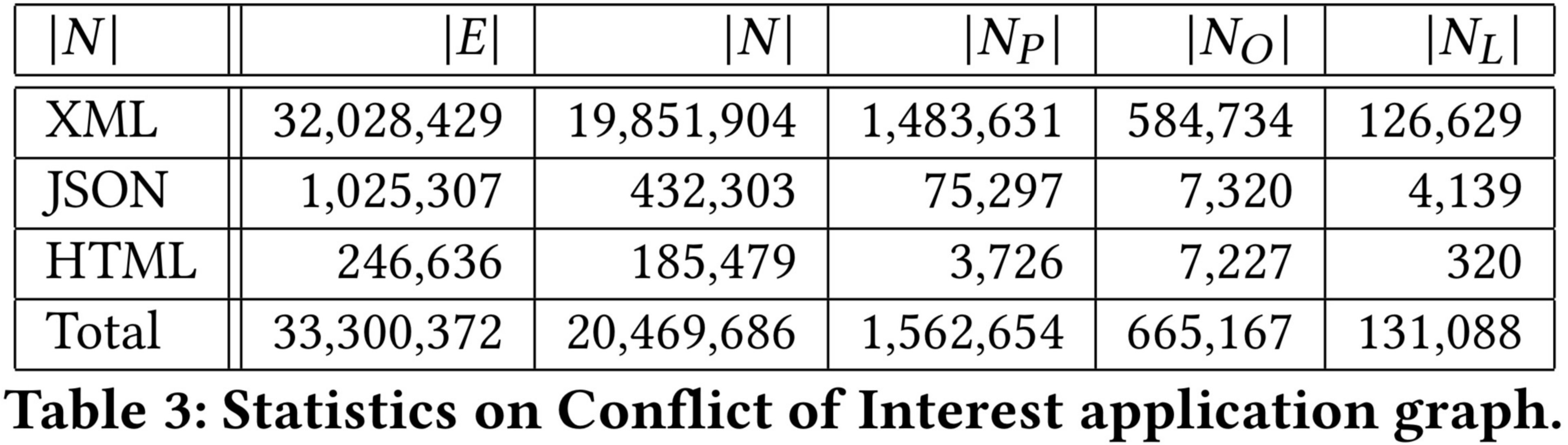


Figure 7: YAGO loading time (minutes) using Flair.

Application: conflicts of interest in the biomedical domain https://arxiv.org/abs/2102.04141

Collaboration with Stéphane Horel (Le Monde)

Data: XML, PDF→JSON, HTML



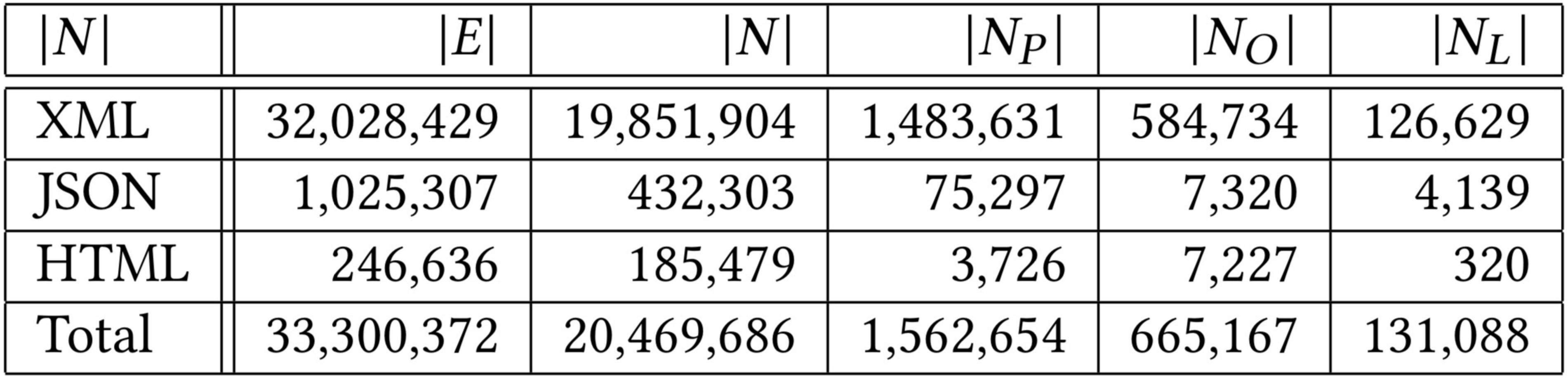
$\|N\|$	$\|E\|$	$\|N\|$	$\|N_P\|$	$\|N_O\|$	$\|N_L\|$
XML	32,028,429	19,851,904	1,483,631	584,734	126,629
JSON	1,025,307	432,303	75,297	7,320	4,139
HTML	246,636	185,479	3,726	7,227	320
Total	33,300,372	20,469,686	1,562,654	665,167	131,088

Table 3: Statistics on Conflict of Interest application graph.

Application: conflicts of interest in the biomedical domain https://arxiv.org/abs/2102.04141

Collaboration with Stéphane Horel (Le Monde)

Data: XML, PDF→JSON, HTML

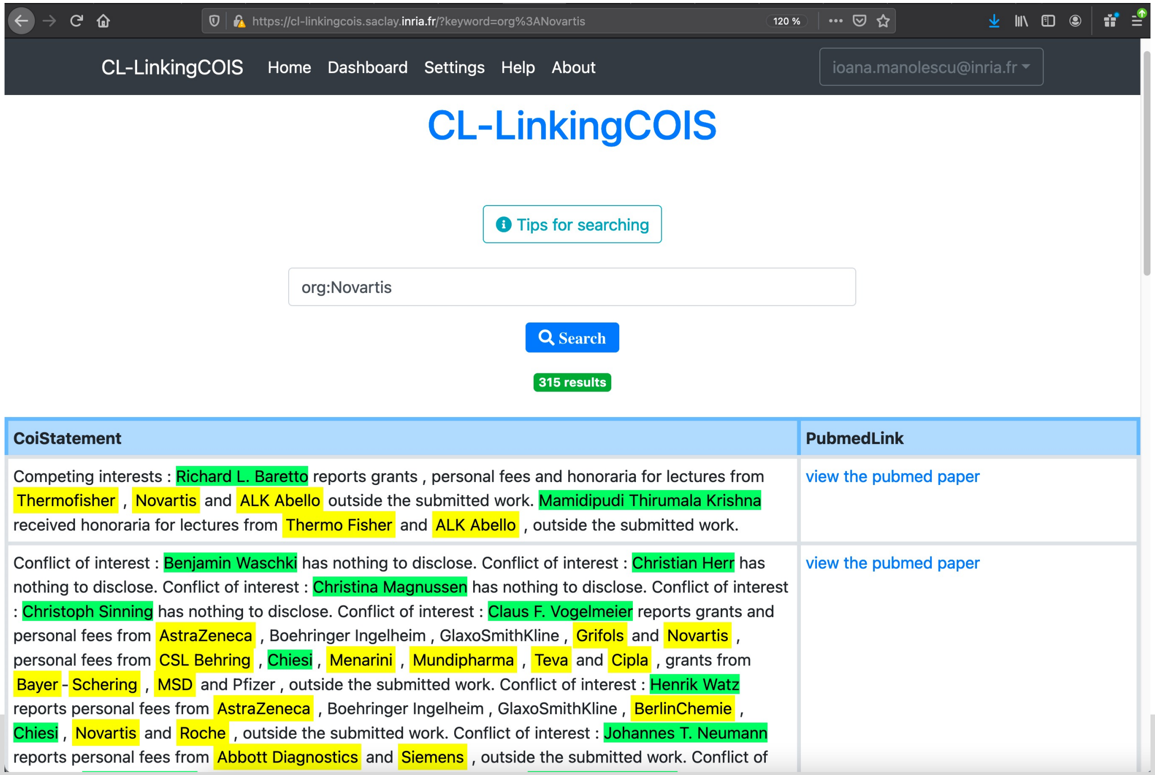


$\|N\|$	$\|E\|$	$\|N\|$	$\|N_P\|$	$\|N_O\|$	$\|N_L\|$
XML	32,028,429	19,851,904	1,483,631	584,734	126,629
JSON	1,025,307	432,303	75,297	7,320	4,139
HTML	246,636	185,479	3,726	7,227	320
Total	33,300,372	20,469,686	1,562,654	665,167	131,088

Table 3: Statistics on Conflict of Interest application graph.

Application: conflicts of interest in the biomedical domain https://arxiv.org/abs/2102.04141

Collaboration with Stéphane Horel (Le Monde)



CL-LinkingCOIS

org:Novartis

CoiStatement	PubmedLink
Competing interests : Richard L. Baretto reports grants , personal fees and honoraria for lectures from Thermofisher , Novartis and ALK Abello outside the submitted work. Mamidipudi Thirumala Krishna received honoraria for lectures from Thermo Fisher and ALK Abello , outside the submitted work.	view the pubmed paper
Conflict of interest : Benjamin Waschki has nothing to disclose. Conflict of interest : Christian Herr has nothing to disclose. Conflict of interest : Christina Magnussen has nothing to disclose. Conflict of interest : Christoph Sinning has nothing to disclose. Conflict of interest : Claus F. Vogelmeier reports grants and personal fees from AstraZeneca , Boehringer Ingelheim , GlaxoSmithKline , Grifols and Novartis , personal fees from CSL Behring , Chiesi , Menarini , Mundipharma , Teva and Cipla , grants from Bayer - Schering , MSD and Pfizer , outside the submitted work. Conflict of interest : Henrik Watz reports personal fees from AstraZeneca , Boehringer Ingelheim , GlaxoSmithKline , BerlinChemie , Chiesi , Novartis and Roche , outside the submitted work. Conflict of interest : Johannes T. Neumann reports personal fees from Abbott Diagnostics and Siemens , outside the submitted work. Conflict of	view the pubmed paper

Graph quality experiments
https://arxiv.org/abs/2012.08830

- ❑ 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

❑ Mediator previously developed [Bonaque et al., VLDB 2016] inappropriate due to complexity, lack of structure, and performance → graph warehouse

❑ Lack of structure forces reachability queries instead of join

  ❑ Price to pay for powerful integration

Data cleaning aspects: Similarity links require value or entity matching

❑ Avoid constructing structured objects ("clean tuples"): don't seem necessary

ConnectionLens in the scientific landscape (2)

Graph construction

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

Keyword search on structured data

- ❑ Previously studied for relational, graph, or XML databases
  - ❑ Typically assume structure/regularity in the graph
  - ❑ Exploit favorable properties of the score function
- ❑ First keyword search algorithm across heterogeneous data sources, 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

- ❑ Extending and improving the in-memory query processor (A. Anadiotis, F. Chimienti, IM)
- ❑ Relationship extraction (O. Balalau, M. Mohanty, IM)
- ❑ Natural language querying of the graph (O. Balalau, P. Upadhyay, IM + PhD in fall 2021)
- ❑ Improving the quality of graph linking (T. Bouganim, H. Galhardas, IM)
- ❑ Abstracting CL graphs (N. Barret, H. Galhardas, P. Upadhyay, IM)

- ❑ Applications:
  - ❑ Conflicts of Interests in the biomedical domain (w/ S. Horel and G. Fooks, Aston U., UK)
  - ❑ Mediacités (w/ WeDoData)

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/