SourcesSay: Intelligent Analysis and Interconnexion of Heterogeneous Data in Digital Arenas

AI Chair project, ANR & DGA

Ioana Manolescu

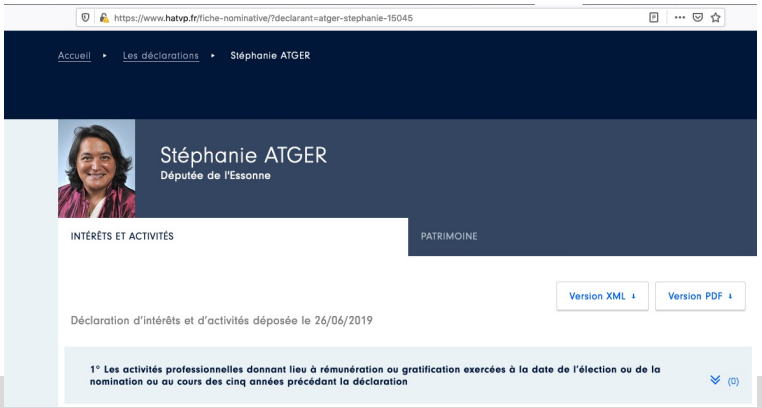
Inria and Institut Polytechnique de Paris

Motivation

Data production has been democratized: unprecedented data generation rates by humans, software, and (equipped) physical objects

Numerous opportunities to **add value by integrating data from several sources**. Examples from data journalism:

- Follow **official communication** by politicians together with their **social media presence**, **laws** they promote, and their **conflicts of interest**



Data journalism problem: working with heterogeneous data

Digital data sources are **heterogeneous**

- For Open Data, W3C standard advocates RDF. Yet...
- **INSEE**: some RDF, lots of Excel and HTML**; NosDéputés.fr**: JSON, XML
- **HATVP** (Haute Autorité pour la Transparence de la Vie Publique): CSV, XML
- **EFSA** (European Food Safety Administration): PDF

Different format, organization, structure, value representation convention...

Application: analyzing a fake news **arena**

An **arena** consists of a set of **entities** (users, organizations etc.) and **contents** they author, share, or are mentioned in

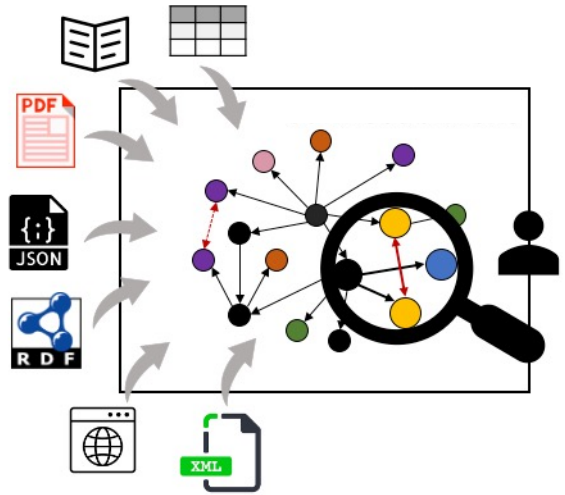
Fake news arena:

- content: HTML, JSON, text or PDF (pages, articles, posts, tweets)
- publishers and distributors, e.g., in relational or JSON;
- fact-checks (usually semi-structured, XML or JSON)

Given a new content with their authors, re-distributors, links etc.

What can we say about the trustworthiness of the content and its environment?

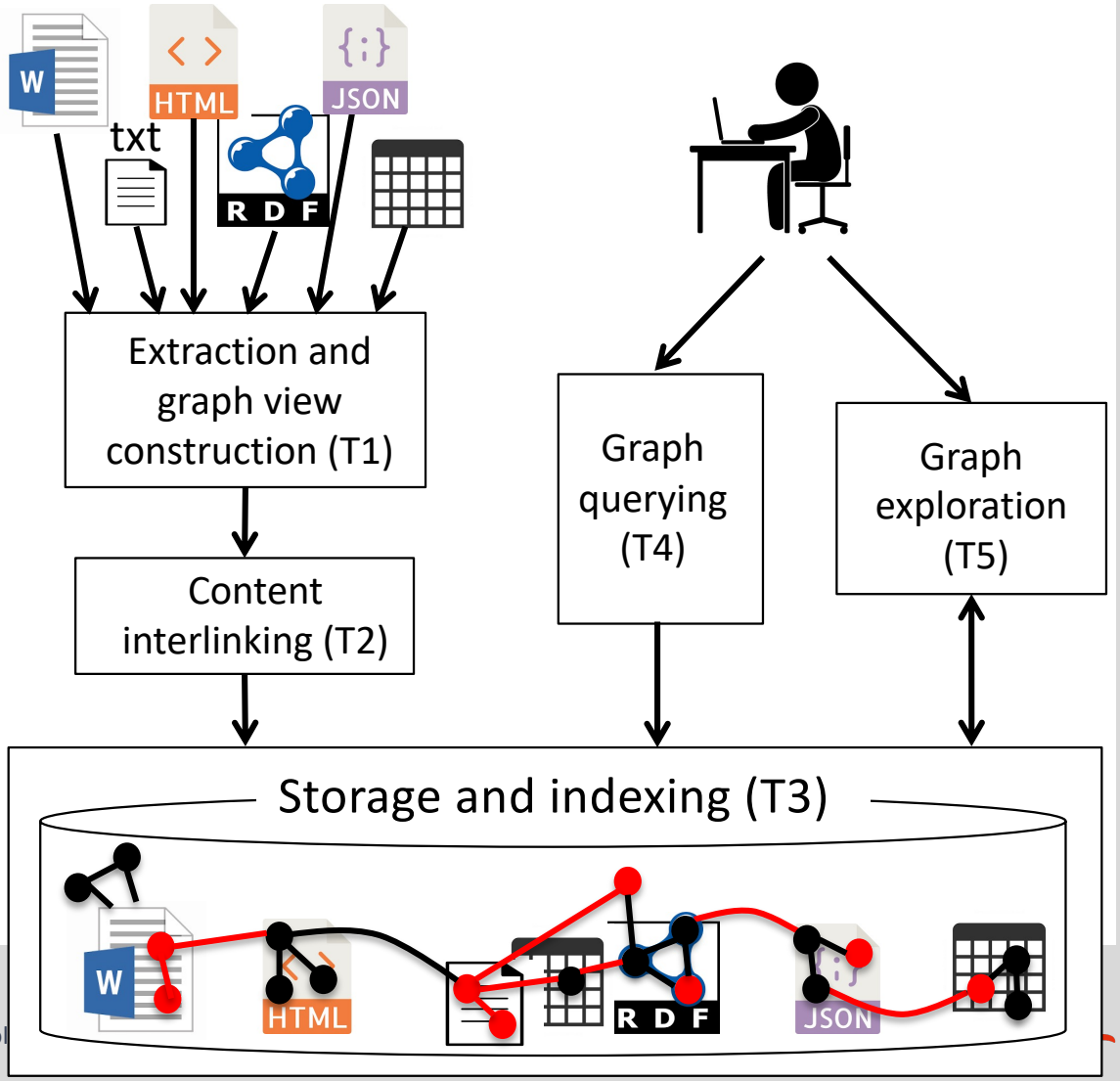
SourcesSay vision



- **Integrate** and **interpret** heterogeneous data from digital arenas as graphs
- **Sourcing** precisely every info
- Novel graph **query** and **exploration**
- Precision, efficiency, friendliness to non-technical users

SourcesSay Architecture

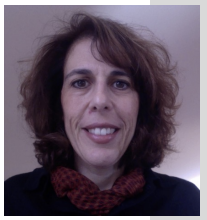
Unifying technical hypothesis:
integrate data in a **graph**



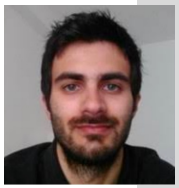
Challenges

How to **enrich and interconnect sources**? Collab. O. Balalau (CEDAR)
H. Galhardas (U. Lisbon)

- Entity and relationship extraction
- Node/entity matching, disambiguation w/r knowledge base

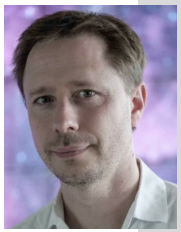


How to **efficiently store large volumes of heterogeneous content, and the connections** extracted from them? Collab. A. Anadiotis (CEDAR)



- Disk-based and || novel in-memory querying engine

How to **efficiently and flexibly query the data** using keywords or NL?
How to know when an answer is **interesting**? Collab. O. Balalau (CEDAR)



- Heterogeneous graph embeddings

How to **explore and interact with** the graphs? Collab E. Pietriga (ILDA)

Applications and collaborations

- Non-funded partners bring applications:
  Le Monde (S. Horel, European Press Prize « Investigative Reporting Award »),
  WeDoData
- Inria AI engineer (2019-2022)
- DIM RSFI PhD (2020-2022) with WeDoData
- Joint work with: 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

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

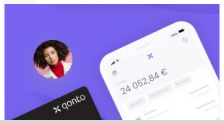
RESPONSABLES PUBLICS

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 are the Balkany 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)

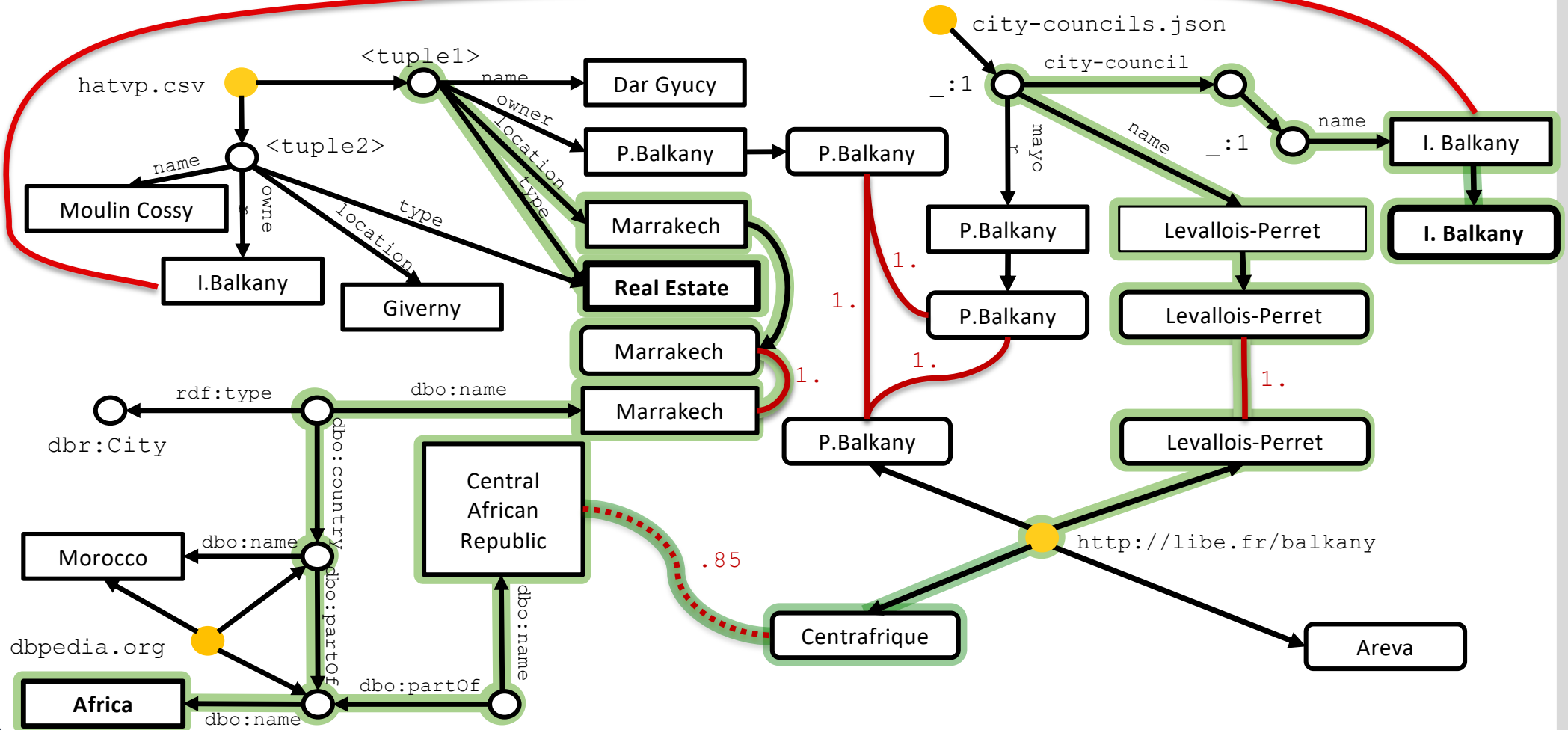
```
[{
    name: “Levallois-Perret”,
    mayor: “P. Balkany”,
    city-council: [
      {name: ”I. Balkany”},
        …
    ]
}, …]
```

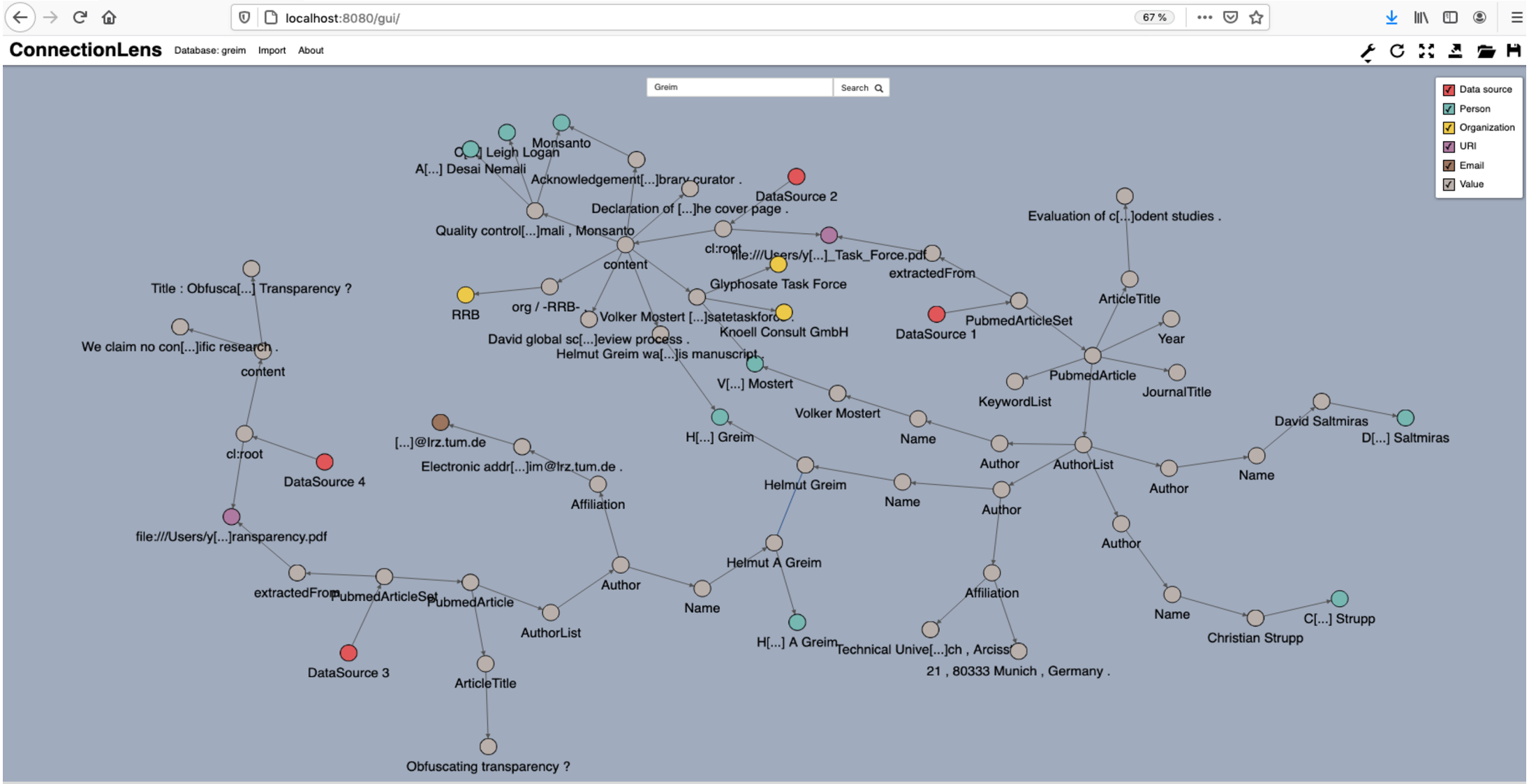
Libération – Nov. 13, 2014 (Text)

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




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
content
cl:root
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
Year
PubmedArticleSet
DataSource 1
PubmedArticle
JournalTitle
KeywordList
Title : Obfusca[...] Transparency ?
We claim no con[...]ific research .
content
cl:root
DataSource 4
V[...] Mostert
Volker Mostert
H[...] Greim
Name
Author
AuthorList
David Saltmiras
D[...] Saltmiras
[...]@lrz.tum.de
Electronic addr[...]im@lrz.tum.de .
Affiliation
Helmut Greim
Name
Author
Author
Name
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 ?

Publications and visibility

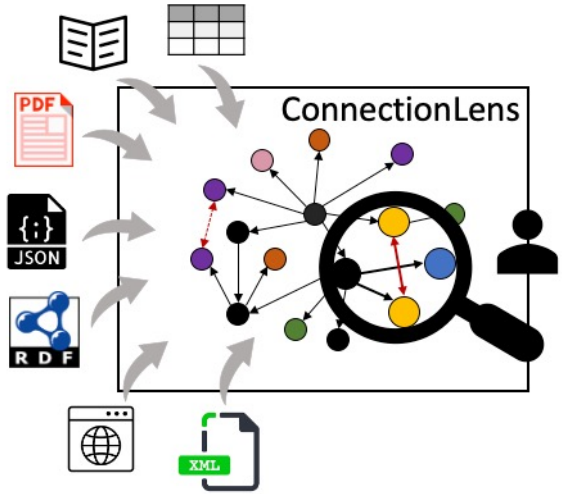
Publications

- « Graph integration of structured, semistructured and unstructured data for data journalism » (BDA 2020)
- "Graph-based keyword search in heterogeneous data sources » (BDA 2020)
- « Graph integration of structured, semistructured and unstructured data for data journalism », accepted for publication in Elsevier Journal of Information Systems, 2021
- « Empowering Investigative Journalism with Graph-Based Heterogeneous Data Management », invited in special issue of IEEE TKDE, 2021

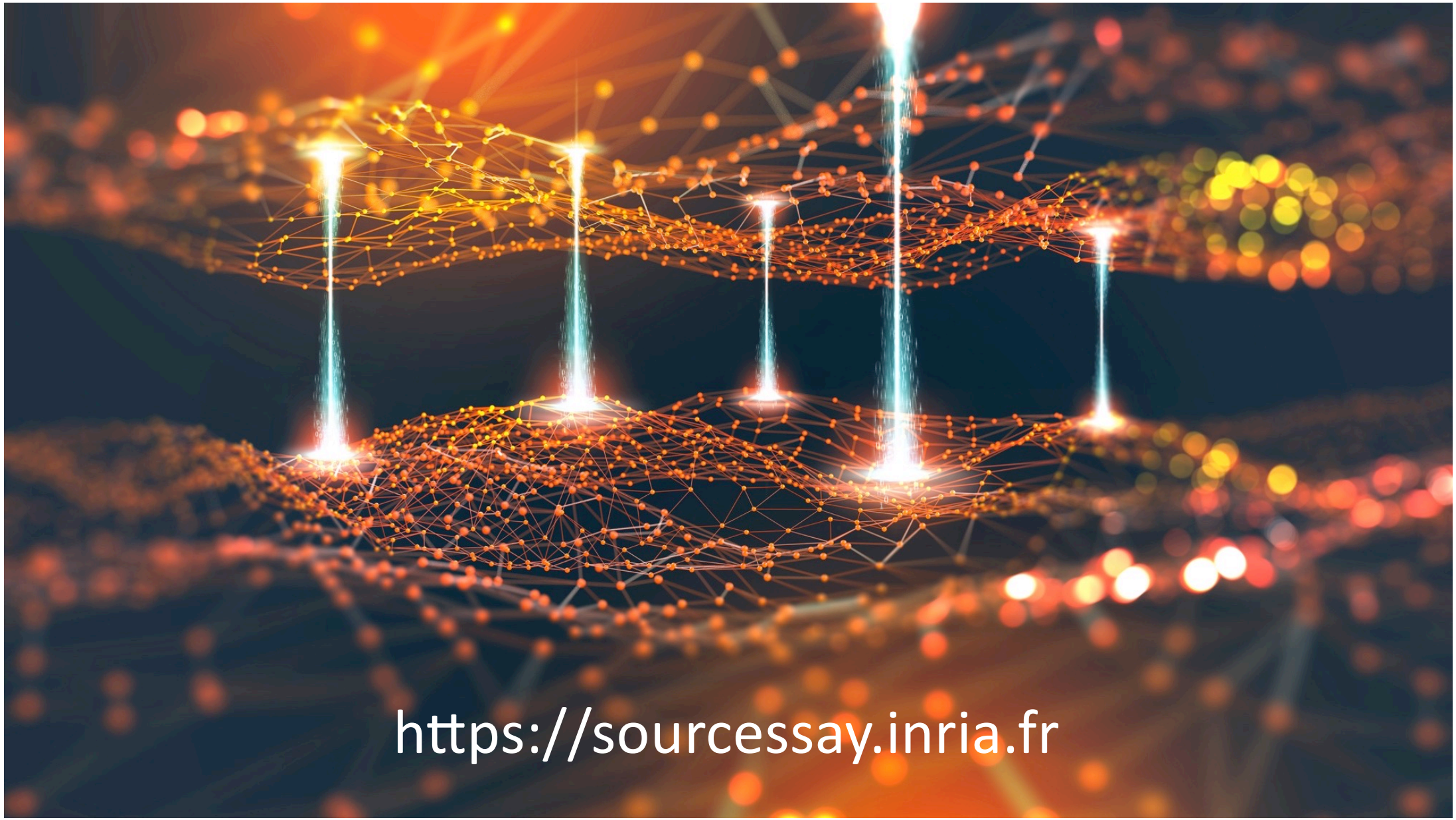
Invited keynotes at ADBIS 2020, DATA 2020, KnoD 2021? DOLAP 2021, ICFCA 2021; lab talks IRISA, LIG, Simon Fraser U. (Canada)
Presented SourcesSay in Data Journalism session at DataHarvest 2020

Implementation

- Java (220 classes/40K LOC), Python (25 classes/2700 LOC)
- Available online: https://gitlab.inria.fr/cedar/connectionlens
- Graph creation time mostly **linear in the size of the data**
- Costlier operations involve ML (disambiguation, extraction)



Data model	$\|E\|$	$\|N\|$	$\|N_P\|$	$\|N_O\|$	$\|N_L\|$
XML	35,318,110	22,204,487	1,561,352	718,434	147,256
JSON	2,800,959	998,013	133,794	147,431	9,822
HTML	232,675	174,849	5,144	4,479	581
Total	38,351,744	23,377,349	1,700,290	870,344	157,659


https://sourcessay.inria.fr

Application: analyzing a fake news ecosystem

Fact-checking: verification of public statements in the (social) media

Collaboration since 2014 with:

Le Monde

LES DÉCODEURS
VENONS-EN AUX FAITS

Les Décodeurs publish as Open Data
their classification of 1300 web sites in:

{ **rather reliable**; **satirical**; **has published fakes**; **agregateur (re-check)** }

https://www.lemonde.fr/webservice/decodex/updates
https://toolbox.google.com/factcheck/

Google Fact Check Tools

Inria

Other digital arenas

- **Scientific and general-audience publications** on a topic
  - Particle air pollution, controversial drugs, a company's products...
- **Journalistic investigations**
  - Tax evasion (Panama Papers): relational database + PDF documents
  - Mongering doubt on tobacco effects or global warming