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

AI Chair project, ANR & DGA

Ioana Manolescu

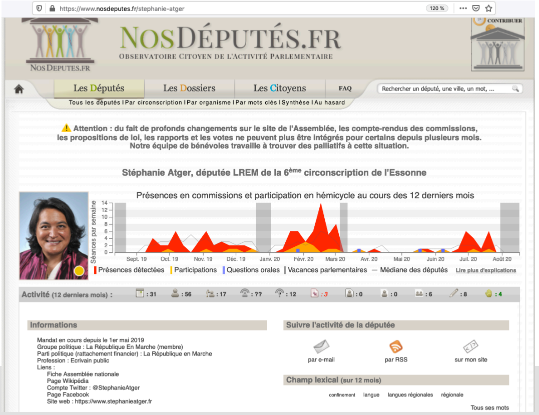
Inria and Institut Polytechnique de Paris

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**


https://www.nosdeputes.fr/stephanie-atger
120 %
NOSDÉPUTÉS.FR
OBSERVATOIRE CITOYEN DE L'ACTIVITÉ PARLEMENTAIRE
NOSDEPUTES.FR
CONTRIBUER
Les Députés
Les Dossiers
Les Citoyens
FAQ
Rechercher un député, une ville, un mot, ...
Tous les députés | Par circonscription | Par organisme | Par mots clés | Synthèse | Au hasard

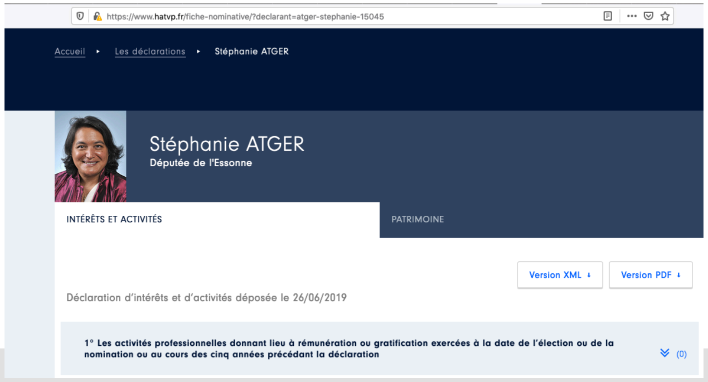
⚠ Attention : du fait de profonds changements sur le site de l'Assemblée, les compte-rendus des commissions, les propositions de loi, les rapports et les votes ne peuvent plus être intégrés pour certains depuis plusieurs mois. Notre équipe de bénévoles travaille à trouver des palliatifs à cette situation.

Stéphanie Atger, députée LREM de la 6ème circonscription de l'Essonne

Présences en commissions et participation en hémicycle au cours des 12 derniers mois

Informations

Mandat en cours depuis le 1er mai 2019
Groupe politique : La République En Marche (membre)
Parti politique (rattachement financier) : La République en Marche
Profession : Ecrivain public
Liens :
Fiche Assemblée nationale
Page Wikipédia
Compte Twitter : @StephanieAtger
Page Facebook
Site web : https://www.stephanieatger.fr



Stéphanie ATGER
Députée de l'Essonne

Déclaration d'intérêts et d'activités déposée le 26/06/2019

1° Les activités professionnelles donnant lieu à rémunération ou gratification exercées à la date de l'élection ou de la nomination ou au cours des cinq années précédant la déclaration

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

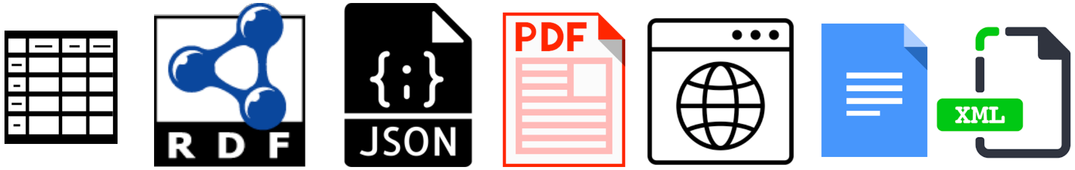
Jan Kuciak, 1990-2018

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!

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 ecosystem

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

Collaboration since 2014 with:

Le Monde

LES DÉCODEURS
VENONS-EN AUX FAITS

☑ Active (149)
□ Inactive (63)

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

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:

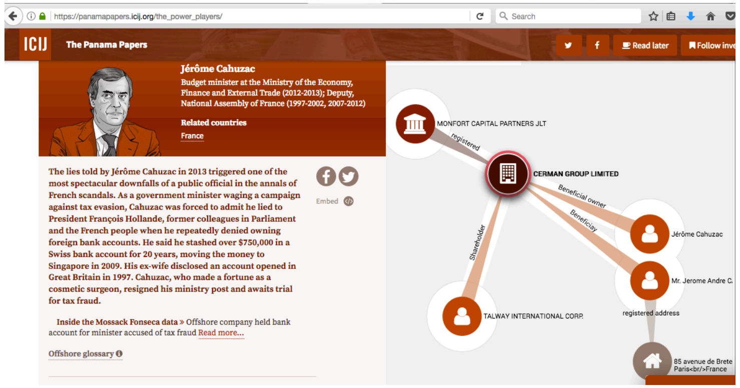
- content: HTML, JSON, text or PDF (pages, articles, posts, tweets)
- publishers and distributors, e.g., in relational or JSON;
- fact-checks (usually semi-structured, XML or JSON)

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

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

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



https://panamapapers.icij.org/the_power_players/

ICIJ The Panama Papers

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)

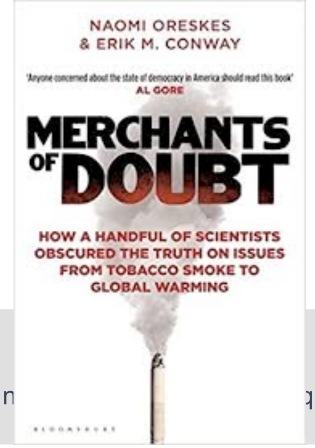
Related countries

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...

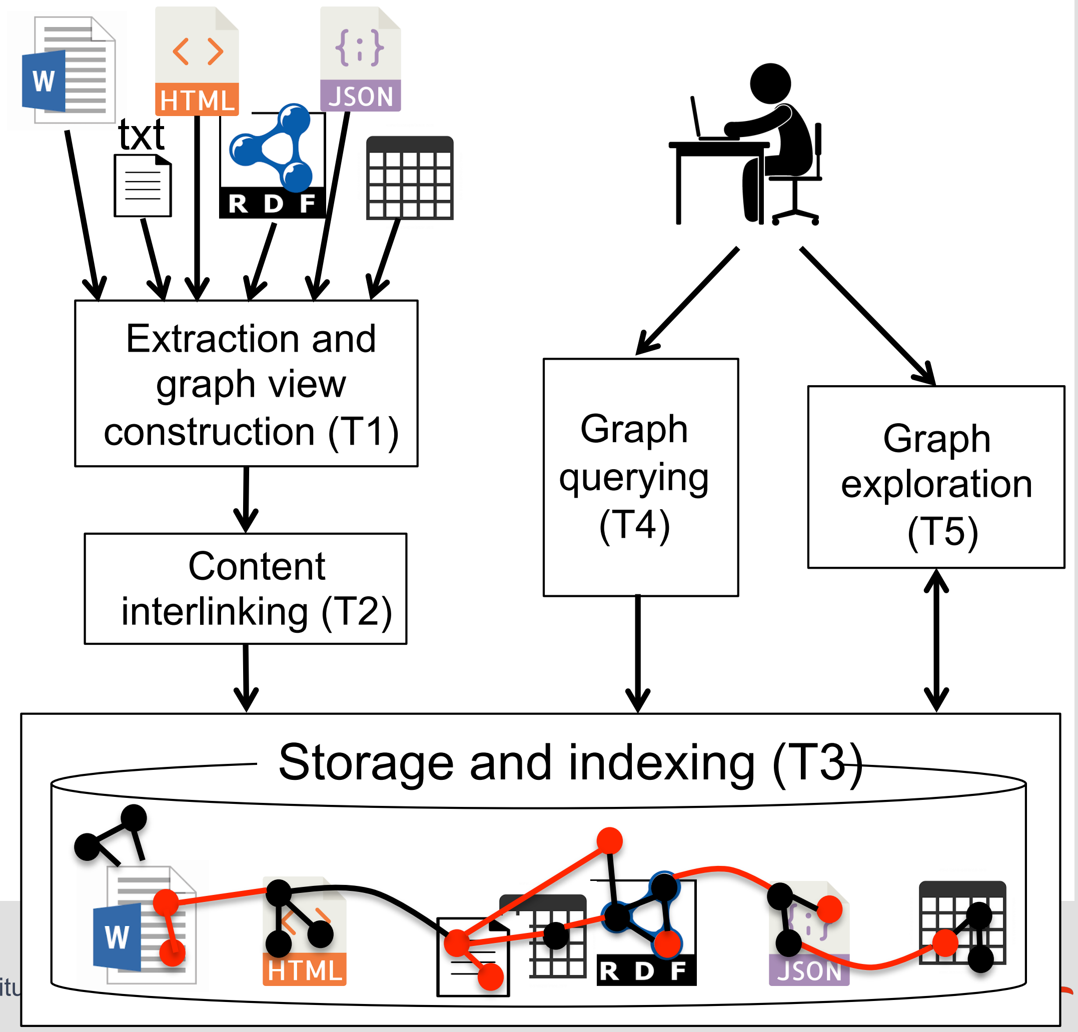
Offshore glossary



NAOMI ORESKES
& ERIK M. CONWAY
'Anyone concerned about the state of democracy in America should read this book'
AL GORE
MERCHANTS OF DOUBT
HOW A HANDFUL OF SCIENTISTS OBSCURED THE TRUTH ON ISSUES FROM TOBACCO SMOKE TO GLOBAL WARMING
BLOOMSBURY

SourcesSay Architecture

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



Extraction and graph view construction (T1)

Content interlinking (T2)

Storage and indexing (T3)

Graph querying (T4)

Graph exploration (T5)

Challenges

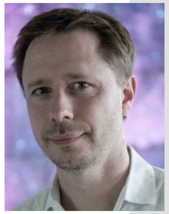
How to **enrich and interconnect sources**? Collab. O. Balalau (CEDAR)
H. Galhardas (U. Lisbon)

- Entity and relationship extraction through NLP and AI
- 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)

- Learn usage patterns, devise novel data processing engines

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



- Learn to rank

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

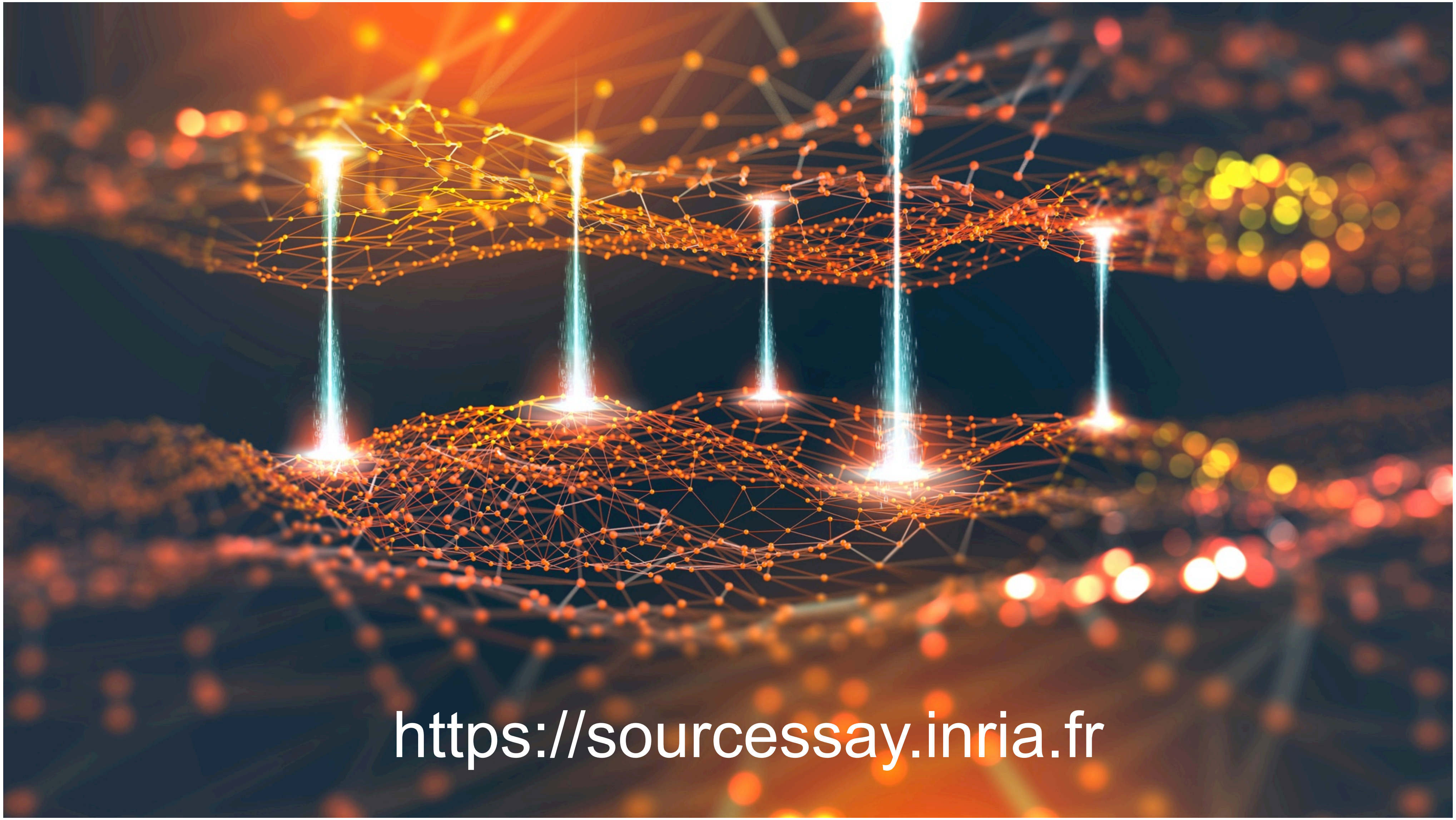
Context and collaborations

- Non-funded partners bring applications: Le Monde, WeDoData

- Inria AI engineer (2019-2022): Tayeb Merabti
- DIM RSFI PhD (2020-2022) with WeDoData
- Work started in seven 2020 internships: I. Burger, F. Chimienti, J. Feitz, Y. Haddad, J. You, Y. Youssef, X. Zhang

The vision

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



https://sourcessay.inria.fr