Integrating (very) heterogeneous data sources: a structured and an unstructured perspective

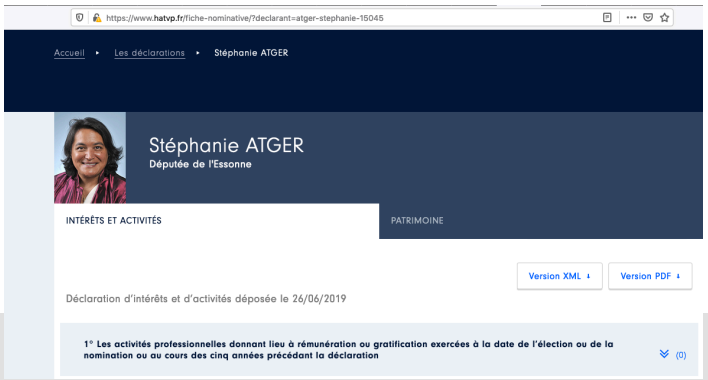
Ioana Manolescu
Inria and Institut Polytechnique de Paris

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**



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

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!

Integrating heterogeneous data

Data is heterogeneous in: its format, organization, structure, value representation convention...

- Different data producers which often do not collaborate/are not aware of each other

The talk will cover two projects:

1. **Estocada: view-based data integration in a polystore**

- Mediator-style: each source stores a separate piece of the data

2. **ConnectionLens: keyword search across heterogeneous data sources**

- Building a warehouse of heterogeneous data into a graph

1

Estocada: views across multiple data models in a polystore system

Joint work with: R. Al-Otaibi, A. Deutsch (UCSD);
D. Bursztyn, S. Zampetakis, F. Bugiotti, I. Ileana (Inria);
B. Cautis (U. Paris-Saclay)

Landscape of data management systems

"One-Size-fits-None":

None of these systems is a universal one-size-fits-all

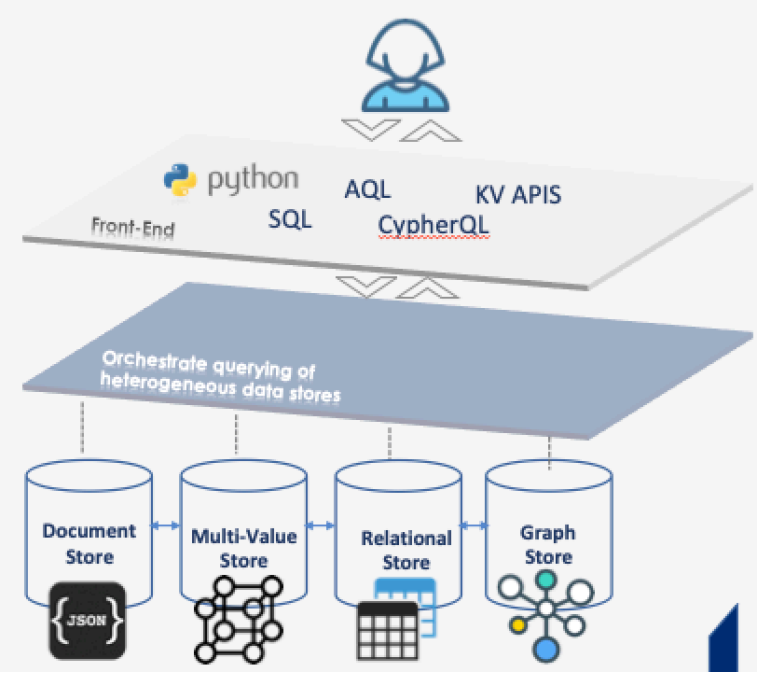
Specific systems have **excellent support for certain tasks** and **perform very poorly on others**

Idea: why not use each system for what it does best?

Emergence of polystore systems

Polystore design principles:

- Use different DBMSs side by side
- Place each type of data in a dedicated native store
- Users directly access the data using different query languages



Shortcomings and missed opportunities in existing polystore systems

Depending on where data is stored, operations needed on it may be **impossible** (e.g., joins in document stores) or **very poorly supported**

Possible **data redundancy**

- The same data may be accessible with different performance from distinct stores... yet the polystore doesn't know!

No way to take advantage of **previous evaluation efforts**

Estocada: materialize in a data store views with data from other store(s)

To improve the performance of operations on that data

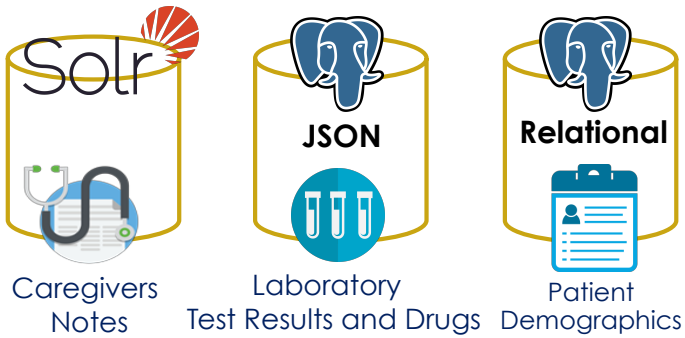
- Data may not be "born" in the best store for the application needs on it

1. Define and materialize the views.

2. Query the data "as it was in its original store"

3. The system will transparently select the best views to use in order to answer the query

Example: MIMIC medical scenario

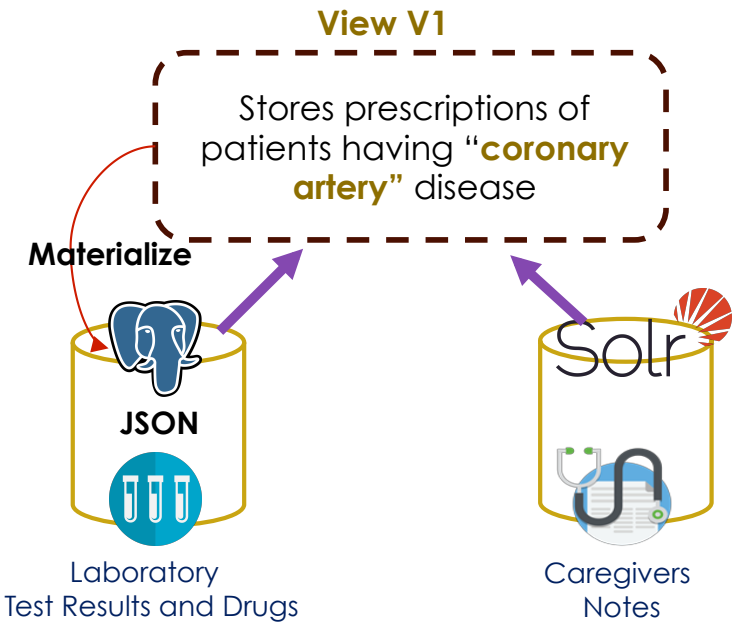


MIMIC-III Dataset

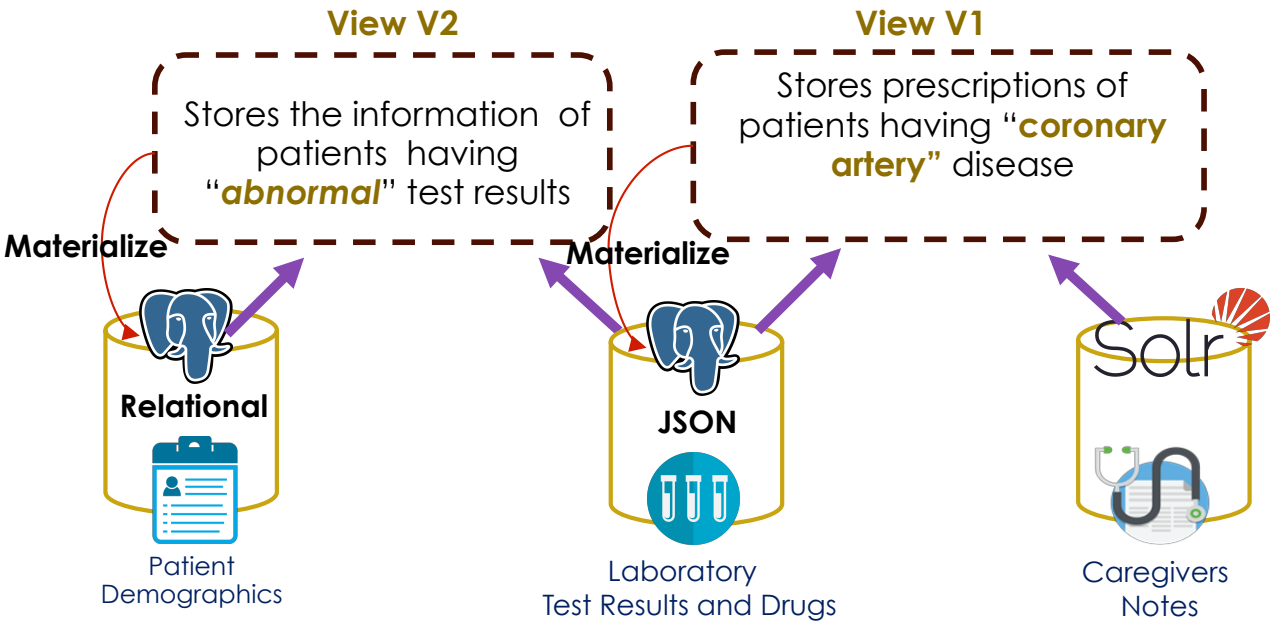
Q1: "For ***'female'*** patients who have been diagnosed with '***coronary artery***' disease and '***abnormal'*** lab test results, ***find*** date/time of their hospital admission and the drugs of types '***addictive'*** prescribed to them."

User Query

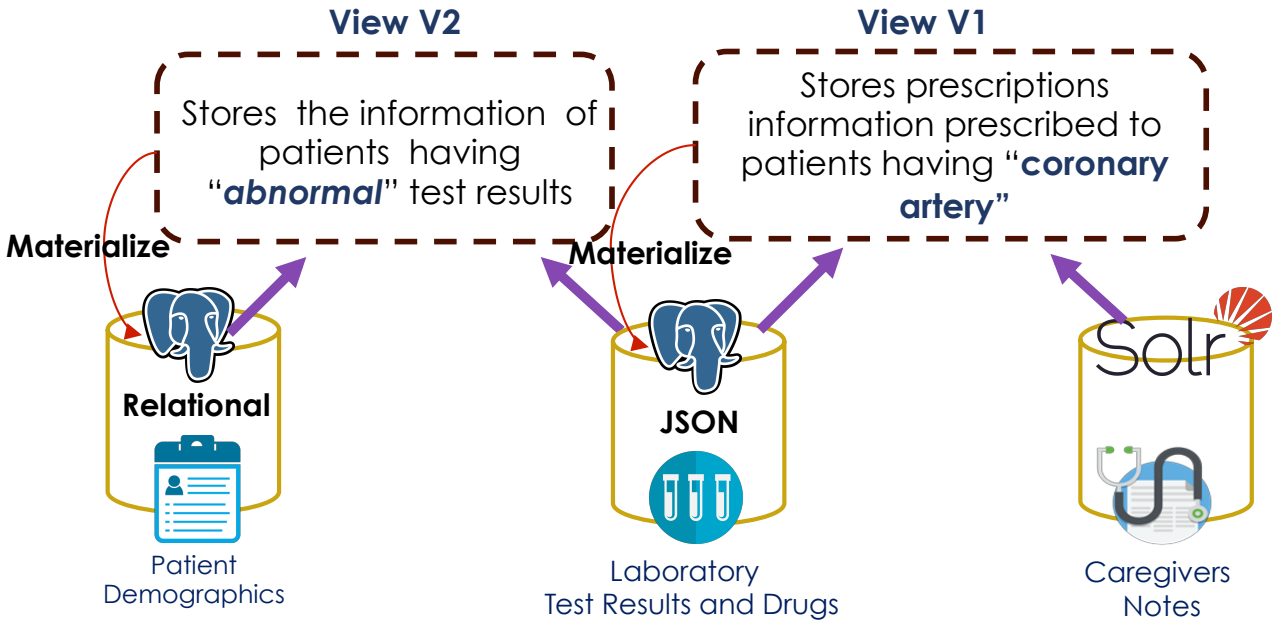
Example: MIMIC medical scenario

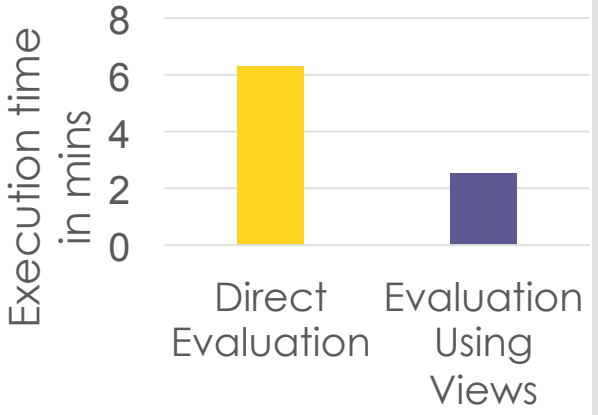


Example: MIMIC medical scenario



Example: MIMIC medical scenario





Cross-data model view-based rewriting in Estocada

1. **Translate view definitions** into a set of **virtual relations + constraints**
2. Describe each **data model** as a set of **virtual relations + constraints**
3. **Translate query** into a **relational query** over the data model relations
4. **Rewrite** the relational query using the available views under the known integrity constraints (Provenance-Aware Chase and Back Chase [Ileana et al., SIGMOD 2014] + improvemens)
5. **Decode** (translate back) the **relational query rewriting** thus obtained, into **source queries + operations to be applied in the mediator**

Estocada publications

- Rana Alotaibi, Bogdan Cautis, Alin Deutsch, Moustafa Latrache, Ioana Manolescu, Yifei Yang: "ESTOCADA: Towards Scalable Polystore Systems" **(demonstration), PVLDB 2020** *Extension to matrix computations!* Video online: https://www.youtube.com/watch?v=IieAQBfoE6o
- Rana Alotaibi, Damian Bursztyn, Alin Deutsch, Ioana Manolescu, Stamatis Zampetakis: "Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue". **SIGMOD Conference 2019**: 1660-1677
- Francesca Bugiotti, Damian Bursztyn, Alin Deutsch, Ioana Manolescu, Stamatis Zampetakis: "Flexible hybrid stores: Constraint-based rewriting to the rescue". **ICDE 2016**: 1394-1397
- Francesca Bugiotti, Damian Bursztyn, Alin Deutsch, Ioana Ileana, Ioana Manolescu: "Invisible Glue: Scalable Self-Tunning Multi-Stores". **CIDR 2015**

2

ConnectionLens: keyword search across heterogeneous data sources

Joint work with: A. Anadiotis, I. Burger, C. Chanial, J. Feitz, M.-H. Le Nguyen (Ecole Polytechnique); O. Balalau, Y. Haddad, T. Merabti, E. Pietriga, Y. Youssef (Inria); C. Conceiçao, H. Galhardas (U. Lisbon); J. Leblay (AIST Tokyo)

Motivation

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

Collaboration since 2014-2015 with:

Le Monde

LES DÉCODEURS
VENONS-EN AUX FAITS

Les Décodeurs publish as Open Data their classification of approximately 1300 web sites in 4 categories:

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

https://www.lemonde.fr/webservice/decodex/updates

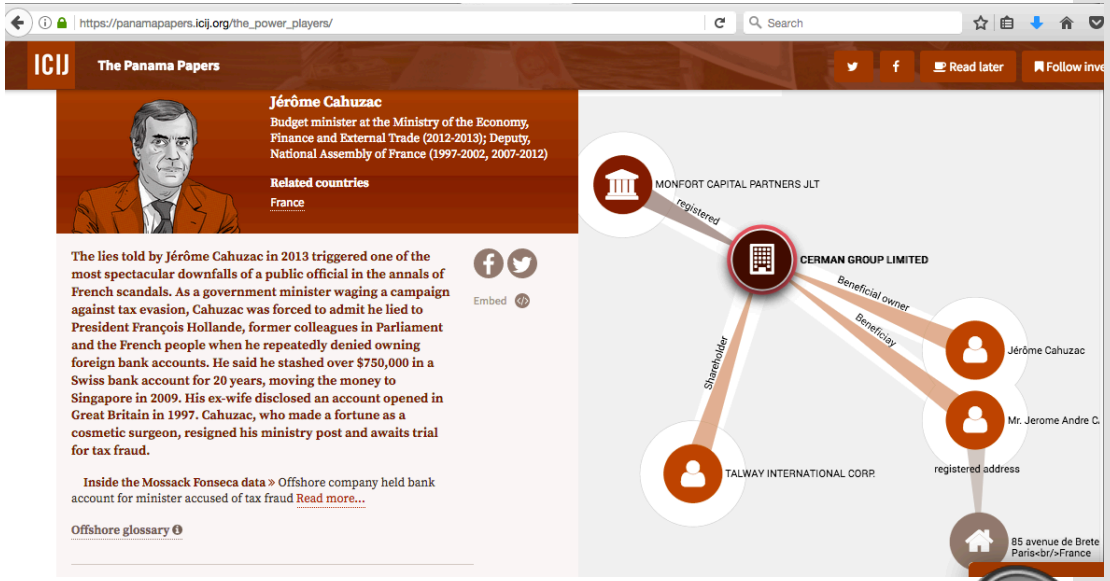
Heterogeneous data integration for journalism (1)

Panama Papers (International Consortium of Investigative Journalism, ICIJ) integrated:

- a **relational** database +
- a set of **PDF** documents

into a Neo4J graph +
graph queries

How does this generalize?



ConnectionLens approach: turning heterogeneous data into a graph

- **Relational** or **CSV** data: each node is a tuple, fk's are edges among them
- **RDF** graphs: direct mapping
- **JSON**, **XML** documents: trees
- **Text**: split in successive fragments
- **PDF**: transformed in JSON + bidimensional tables encoded in RDF

+ Entity extraction from any text field or snippet

Common entities lead to **connections across datasets**

Heterogeneous data integration for journalism (2)

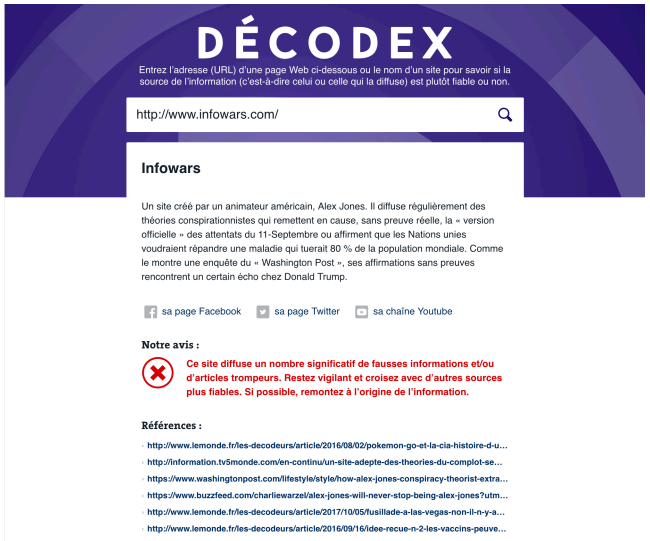
Décodex: browser plug-in from Le Monde

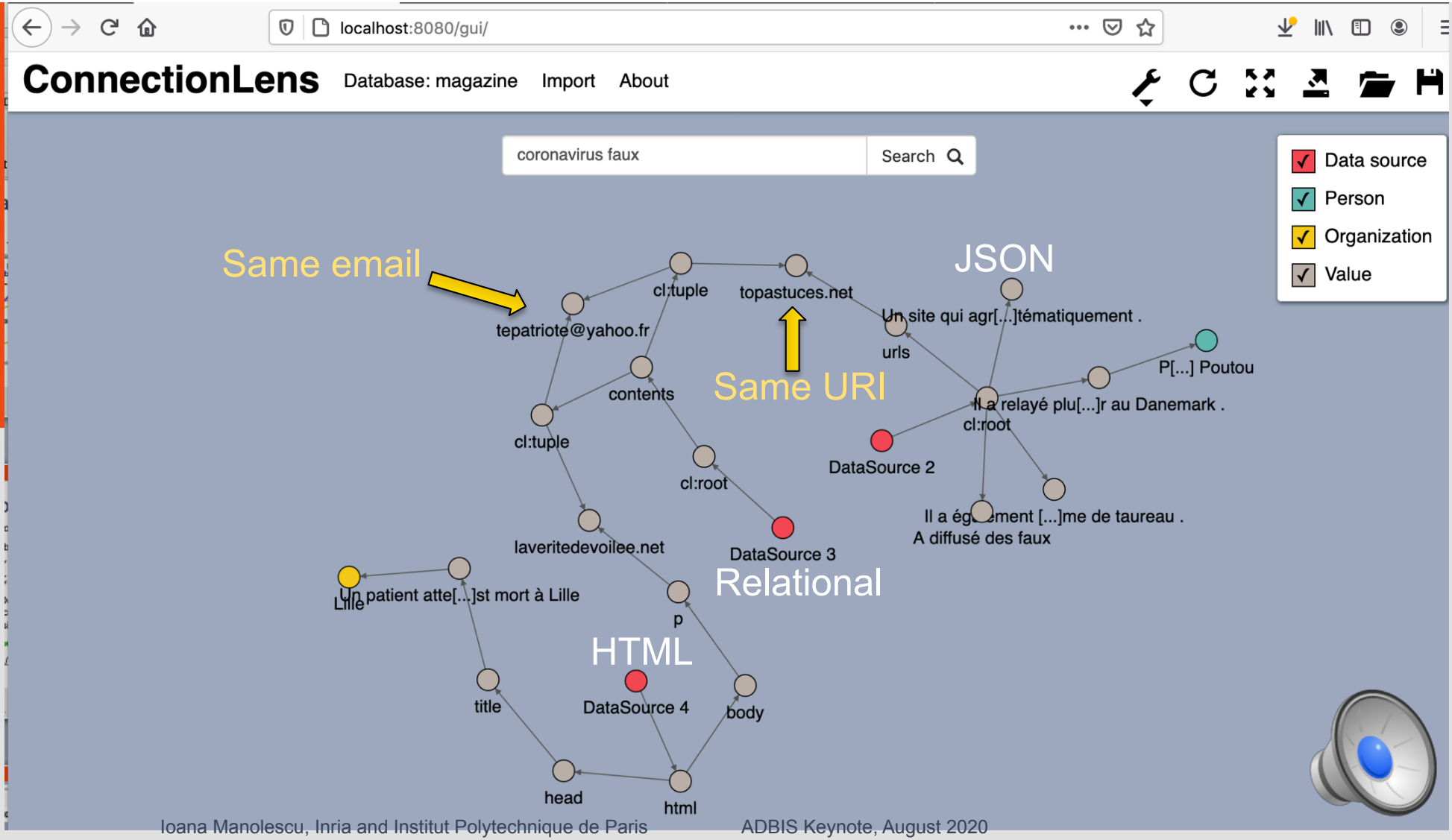
- When visiting a **Web page**
- Given a classification of sites in

{ **rather reliable**; **satirical**;
has published fakes; **agregateur (re-check)** }
based on previous manual effort

- Shows the category of the site to which the page belongs (if the site is known)

What if the site is not in the Décodex base?




localhost:8080/gui/
ConnectionLens
Database: magazine
Import
About
coronavirus faux
Search
Data source
Person
Organization
Value
Same email
Same URI
JSON
Relational
HTML
cl:tuple
topastuces.net
tepatriote@yahoo.fr
Un site qui agr[...]tématiquement .
urls
P[...] Poutou
contents
Il a relayé plu[...]r au Danemark .
cl:root
cl:tuple
DataSource 2
cl:root
Il a ég[...]ment [...]me de taureau .
A diffusé des faux
laveritedevoilee.net
DataSource 3
Un patient atte[...]st mort à Lille
Lille
p
title
DataSource 4
body
head
html

ConnectionLens publications

- Camille Chanial, Rédouane Dziri, Helena Galhardas, Julien Leblay, Minh-Huong Le Nguyen, Ioana Manolescu.
  ConnectionLens: Finding Connections Across Heterogeneous Data Sources, PVLDB 2018 (demonstration)
- "Graph-based keyword search in heterogeneous data sources" by Angelos Christos Anadiotis, Mhd Yamen Haddad and Ioana Manolescu, BDA 2020 (informal publication)
- Oana Balalau, Catarina Conceição, Helena Galhardas, Ioana Manolescu, Tayeb Merabti, Jingmao You, Youssr Youssef.
  Graph integration of structured, semistructured and unstructured data for data journalism, BDA 2020 (informal publication)
- Irène Burger, Ioana Manolescu, Emmanuel Pietriga, Fabian Suchanek.
  Toward Visual Interactive Exploration of Heterogeneous Graphs, SEAData Workshop (w/ EDBT), 2020

Wrap-up

Data integration: fascinating problem with multiple applications

Polystores (e.g., BigDawg, Myria, Tatooine etc.): enterprise data management with several systems already in place

- Estocada: materialized views to be established in one system based on data from one or more other systems → performance savings

Data journalism scenarios: many heterogeneous sources and no data management system already in place

- ConnectionLens: integrating very heterogeneous data into a graph + querying the graph through keywords

Perspective

Data integration is an old area with many open questions

- Gio Wiederhold's work since 1988 https://dblp.org/pid/w/GioWiederhold.html

Recent architectures focus on **structured databases**:

- **Polystores** (MIT, Microsoft, U. Washington): query the data using native languages (or combinations thereof)
- **Ontology-based data access** (U. Rome, U. Bolzano, Inria): use ontologies as conceptual schemas and map sources to the ontologies

How to extend these to **humanities/literature/non-business topics**?

- J. Darmont's: Data Lakes for digital humanities...

AI Chair SourcesSay (2020-2024, hiring!) https://project.inria.fr/sourcessay