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

Ioana Manolescu

## To cite this version:

Ioana Manolescu. Teasing journalistic findings out of heterogeneous sources: a data/AI journey: (invited keynote). DEBS 2022: The 16th ACM International Conference on Distributed and Eventbased Systems, Jun 2022, Copenhagen, Denmark. pp.1-1, 10.1145/3524860.3544406 . hal-03945733

HAL Id: hal-03945733
https://inria.hal.science/hal-03945733
Submitted on 18 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

Cirs


## 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



## Bad memories: Romania, 1989



Ceauşescu re-elected at the 14th congress!
ack-
He had held power since 1965.

## Bad memories: Romania, 1985



## 1990: things got better



## ... kind of



## Democratic societies crucially need the press

To debate and express dissent


Socialist Romania, 1984
To analyze, confim or refute public statements
Fact-checking

To expose and explain society functioning

## Fact-checking



FастСнеск.org
ATE』
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


"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
SHARE READ MORE
Mick Mulvaney
Director, Office of Management and Budget

Home artcles aska question viral spral archues about us search more

THE WIRE
Did Democrats Once Support Border Wall?

## 

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 . 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, satelites, radar coverage and cameras. Sen. Chuck Schumer and then-Sens. Barakictarna and wa
favor of the legislation, and it was signed into law by President George W. Bush favor very sense the bus

$\checkmark$ ASK FAGTOHEOK
 0
Q: Did the Supreme Court rule that public schools cannot teach students abut 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）
（ ）（1） $\mathrm{nttps} / /$／panamapapers．icij．org／the＿power＿players／
c $Q$ Search
自自 令
ICIJ The Panama Papers

## Jérôme Cahuzac

Budget minister at the Ministry of the Economy， inance 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 6

－
ad later $\quad$ Follow inve

## 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

LES DÉCODEURS
VENONS-EN AUX FAITS
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)

## fe Monde



## 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

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


PDF


There is no schema. Data producers often uncollaborative.
$\square$ 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)
$\square$ Same entity in two different text nodes = link among the text nodes (densification of the graph)

The graph is heterogeneous and irregular $\rightarrow$
Query it through keywords: find trees that connect 1 node matching each kwd
$\square$ 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 $\rightarrow$
Query it through keywords: find trees that connect 1 node matching each kwd
$\square$ 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



## The Balkany and their African connections



## 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"\},
]
\}, ...]

## Sample integration scenario

## 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)
i
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"},
```

            ...
    $\left.\begin{array}{c} \\ \} \\ \\ \hline\end{array} . ..\right]$
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"},
```

            ...
    ]
$\}, \ldots$
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



## ConnectionLens graph construction

## Graph construction stages

1. Primary node and edge construction
$\square$ Direct for XML, JSON, RDF, HTML
$\square 1$ relational tuple=1 node; primary keys-foreign keys as links
$\square$ Convert information from PDF into:
$\square$ JSON for text content
$\square$ RDF describing tables


## ConnectionLens graph construction

## Graph construction stages

## 1. Primary node and edge construction

$\square$ Direct for XML, JSON, RDF, HTML
$\square 1$ relational tuple=1 node; PK-FKs as links
$\square$ [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
$\square$ [InfSys2021] Developed and trained new entity extractor FR/EN, based on Flair framework
$\square$ Date extraction based on HeideITime



## ConnectionLens graph construction

## Graph construction stages

## 2. Entity extraction

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

## 3. Entity disambiguation


$\square$ For each recognized entity, e.g., "Hollande" the place or the person?
$\square$ Built novel disambiguation pipeline for French, based on Ambiverse frameworkBased on knowledge bases (WikiData, YAGO) and WikipediaHelpful on well-known entities

## Graph construction stages

4. Node matchingTo create sameAs edges:
$\square$ Strong sameAs edges: equivalent nodes 1.
$\square$ Weak sameAs edges: similar nodes .85
$\square$ Label normalization, distance functions
$\square$ Remains quadratic at the core $:$, so...
Node factorization (heuristic): create only


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

## ConnectionLens graph querying

## Querying problem statement

$\square$ Given the graph $G=(N, E)$ built out of the datasets $D$ and a query $Q=\{w 1, \ldots, w m\}$, return the $\mathbf{k}$ highest-score minimal answer trees.
$\square$ 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.
$\square$ 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

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

## GAM (Grow and Aggressive Merge) Algorithm

$\square$ Builds trees exploring around the keyword matches
$\square$ 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

$\square$ Builds trees exploring around the keyword matches
$\square$ 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:
$\square$ Grow and Merge depend on the tree root $\rightarrow$ 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



Rank: \#1 Size: 10 Score: 0.282 sources
Rank: \#2 Size: 9 Score: 0.281 source
Rank: \#3 Size: 9 Score: 0.281 source
Rank: \#4 Size: 9 Score: 0.281 source
Rank: \#5 Size: 10 Score: 0.271 source
Rank: \#6 Size: 10 Score: 0.261 source
Rank: \#7 Size: 11 Score: 0.251 source
Rank: \#8 Size: 10 Score: 0.231 source

## ConnectionLens architecture and performance

## Implementation

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



## Implementation [InfSys2022]

G Graph creation time mostly linear in the size of the data
$\square$ Costliest operations involve ML (disambiguation, extraction)
$\square$ Batch extraction: 20x speed-up on GPU, 2x speed-up on regular server
$\square$ 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 $\rightarrow$ JSON, HTML

| $\|N\|$ | $\|E\|$ | $\|N\|$ | $\left\|N_{P}\right\|$ | $\left\|N_{O}\right\|$ | $\left\|N_{L}\right\|$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| 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)


CoiStatement
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.

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 AstraZeneca, Boehringer ingelhim, GlaxoSmithKine, Grifols and Novartis , Mensorin , 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

| PubmedLink |
| :--- |
| view the pubmed paper |
| view the pubmed paper |

## Graph quality experiments [InfSys2022]

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


## ConnectionLens in the scientific landscape

## ConnectionLens in the scientific landscape (1)

Data integration for structured, semistructured and unstructured data
$\square$ "Ad-hoc" (combinations of sources to be unioned, joined, or chained)
$\square$ No schema, ontologies, queries known in advance
$\square$ 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
$\square$ Web info may be missing or not trusted
Data cleaning: Similarity links require value or entity matching
$\square$ Use reference data sources, heuristics, interactive cleaning (under integrated)

## ConnectionLens in the scientific landscape (2)

Graph construction
$\square$ Users of entity extraction modules, trained a model for French

## Keyword search on structured data

$\square$ 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 SPARQLConsider more limited search spaces (edge direction; one tree per root; ...)Exploit favorable properties of fixed score function; establish approximations...
$\square$ 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

$\square$ Learning to predict extraction success ~ detecting disguised nulls (BDA 2021, invited TLDKS 2022)
$\square$ Integrating keyword search into structured graph querying (with A. Anadiotis, M. Mohanty)
$\square$ Relationship extraction based on OpenIE (with O. Balalau, P. Upadhyay)
I Natural language graph querying (with O. Balalau, K. Zhang)
$\square$ Improving the quality of graph linking (with T. Bouganim, H. Galhardas)
$\square$ Abstracting CL graphs (with N. Barret, P. Upadhyay) https://team.inria.fr/cedar/projects/abstra/

- Applications:
$\square$ Conflicts of Interests in the biomedical domain (w/ S. Horel and G. Fooks, Aston U., UK)
$\square$ 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!

Questions?

## Thank you

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

