



**HAL**  
open science

## Data integration for journalism: goals, tools, and architectures (Keynote)

Ioana Manolescu

► **To cite this version:**

Ioana Manolescu. Data integration for journalism: goals, tools, and architectures (Keynote). ii-WAS 2017 - 19th International Conference on Information Integration and Web-based Applications & Services, Dec 2017, Salzburg, Germany. pp.1-46. hal-01657152

**HAL Id: hal-01657152**

**<https://inria.hal.science/hal-01657152>**

Submitted on 8 Dec 2017

**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.

# Data integration for journalism: goals, tools, and architectures

Ioana Manolescu

Inria Saclay and LIX (CNRS and Ecole Polytechnique)

---

04/12/2017

*Inria*



# Plan

---

1. Motivation
2. Data journalism and fact-checking as content management problems
3. Data integration: architectures and solutions
  - ❑ Some works I have been involved to in this area (2013--)

**Joint work with:** D. Cao, B. Cautis, F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, J. Letelier, S. Ribeiro, X. Tannier, M. Thomazo, S. Zampetakis

**Thanks to:** S. Laurent, M. Vaudano, A. Senecat, M. Ferrer from Le Monde / Les Décodeurs



# Where I come from: Romania, 1989

---



# Where I come from: Romania, 1989

---



Ceaușescu re-elected  
at the 14th congress!



# Where I come from: Romania, 1989

---



Ceaușescu re-elected at the 14th congress!

He was in power since 1965.

# Where I come from: Romania, 1989

---



# Where I come from: Romania, 1989

---





# Where I come from: Romania, 1990

---





# Where I come from: Romania, 1990

---



1000 dead (approx)  
No one convicted.

# Democratic societies crucially need the press

---

To debate and express dissent



To analyze, confirm or refute public statements

Fact-checking (see next)

To expose and explain society functioning



# Democratic societies crucially need the press

---

To debate and express dissent



To analyze, confirm or refute public statements

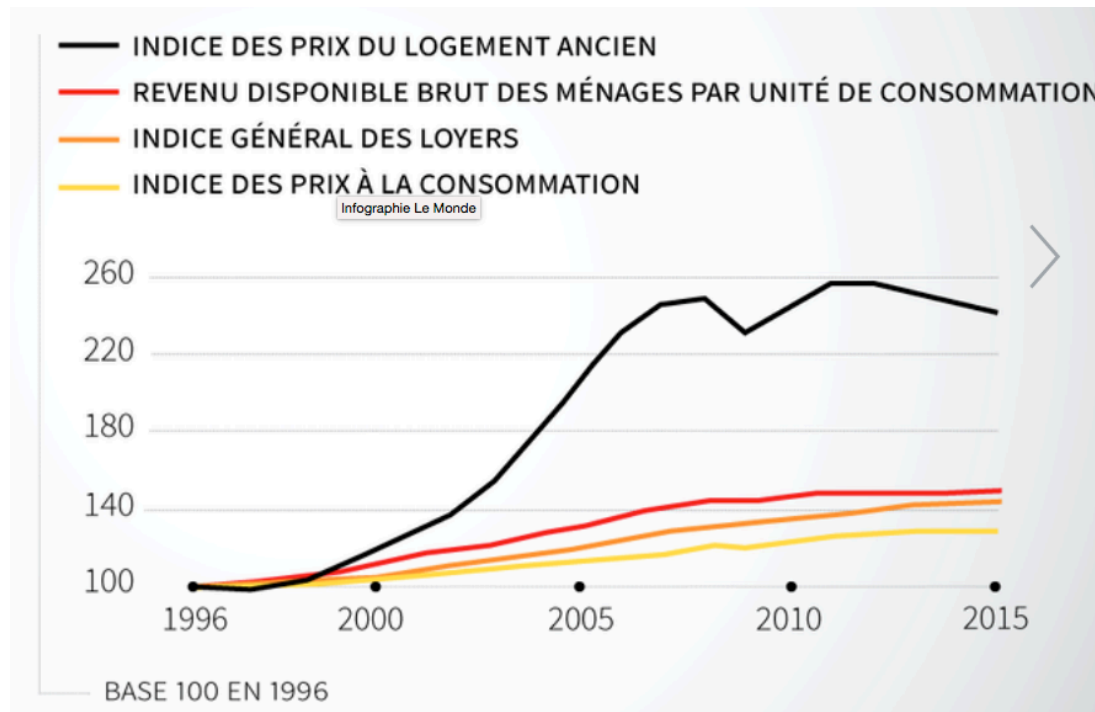
Fact-checking (see next)

To expose and explain society functioning



# Data journalism

Investigative journalism based on **complex and/or large data**



[http://abonnes.lemonde.fr/les-decodeurs/portfolio/2017/04/18/les-fractures-francaises-1-5-le-logement-les-raisons-de-la-crise\\_5112859\\_4355770.html](http://abonnes.lemonde.fr/les-decodeurs/portfolio/2017/04/18/les-fractures-francaises-1-5-le-logement-les-raisons-de-la-crise_5112859_4355770.html)



# Data journalism

## Panama Papers (International Consortium of Investigative Journalism, ICIJ)

The screenshot shows the ICIJ website interface for the Panama Papers. The browser address bar displays [https://panamapapers.icij.org/the\\_power\\_players/](https://panamapapers.icij.org/the_power_players/). The page title is "Panama Papers The Power Players". The navigation bar includes social media icons for Twitter and Facebook, and buttons for "Read later" and "Follow inves". The main content area features a filter menu with options: "All", "Africa", "Asia", "Europe", "Latin America & the Caribbean", and "Oceania". Below the filter menu, there are tabs for "Country leaders" and "Politicians/public officials", with "Filters Europe Country" visible. The main content displays a grid of 10 political figures, each with a stylized illustration and a brief description of their role.

Figure	Description
1	Advised former Greek prime minister
2	Former government minister in France
3	Iceland's minister of finance
4	Iceland's Interior Minister
5	Malta's minister of energy and health
6	Former member Hungarian Assembly
7	Deputy chief justice of Kenya's Supreme Court
8	Former Warsaw mayor, former member EU parliament
9	Baroness and lifetime member of U.K. Parliament
10	Member of British Parliament

# Data journalism

## Panama Papers (International Consortium of Investigative Journalism, ICIJ)

The screenshot shows a web browser displaying the ICIJ website. The page title is "The Panama Papers" and the URL is "https://panamapapers.icij.org/the\_power\_players/". The main content area features a profile for Jérôme Cahuzac, a French politician. To the right of the profile is a corporate structure diagram showing the relationships between several entities and individuals.

**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](#)

**Corporate Structure Diagram:**

- MONFORT CAPITAL PARTNERS JLT (registered) is connected to CERMAN GROUP LIMITED (Shareholder).
- TALWAY INTERNATIONAL CORP. (Shareholder) is connected to CERMAN GROUP LIMITED.
- CERMAN GROUP LIMITED (Beneficial owner) is connected to Jérôme Cahuzac (Beneficiary).
- CERMAN GROUP LIMITED (Beneficiary) is connected to Mr. Jerome Andre C. (Beneficiary).
- Mr. Jerome Andre C. is connected to a registered address: 85 avenue de Brete Paris-<br/>France.

# Fact-checking (since 1930 approx.)

---

**Fact-checking:** verification of facts mentioned **in media content**

- ❑ To protect media reputation and avoid legal action
- ❑ Verification supposes the existence of a **reference dataset**

“The day I became a fact-checker at The New Yorker,  
I received **one set of red pencils** [...] for underlining **passages on page proofs of articles that might contain checkable facts.** [...] confirmed **with the help of reference books** from the magazine’s library”



<http://www.nytimes.com/2010/08/22/magazine/22FOB-medium-t.html>

# Fact-checking (2012 – ongoing )

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

**FACTCHECK.ORG** A Project of The Annenberg Public Policy Center

HOME ARTICLES ASK A QUESTION VIRAL SPIRAL ARCHIVES ABOUT US SEARCH MORE

**THE WIRE**

## Did Democrats Once Support Border Wall?

By Robert Farley Posted on April 26, 2017

Like 835 Tweet Pin it Share 11

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**

Like 953 Tweet Pin it Share 98

**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.


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**

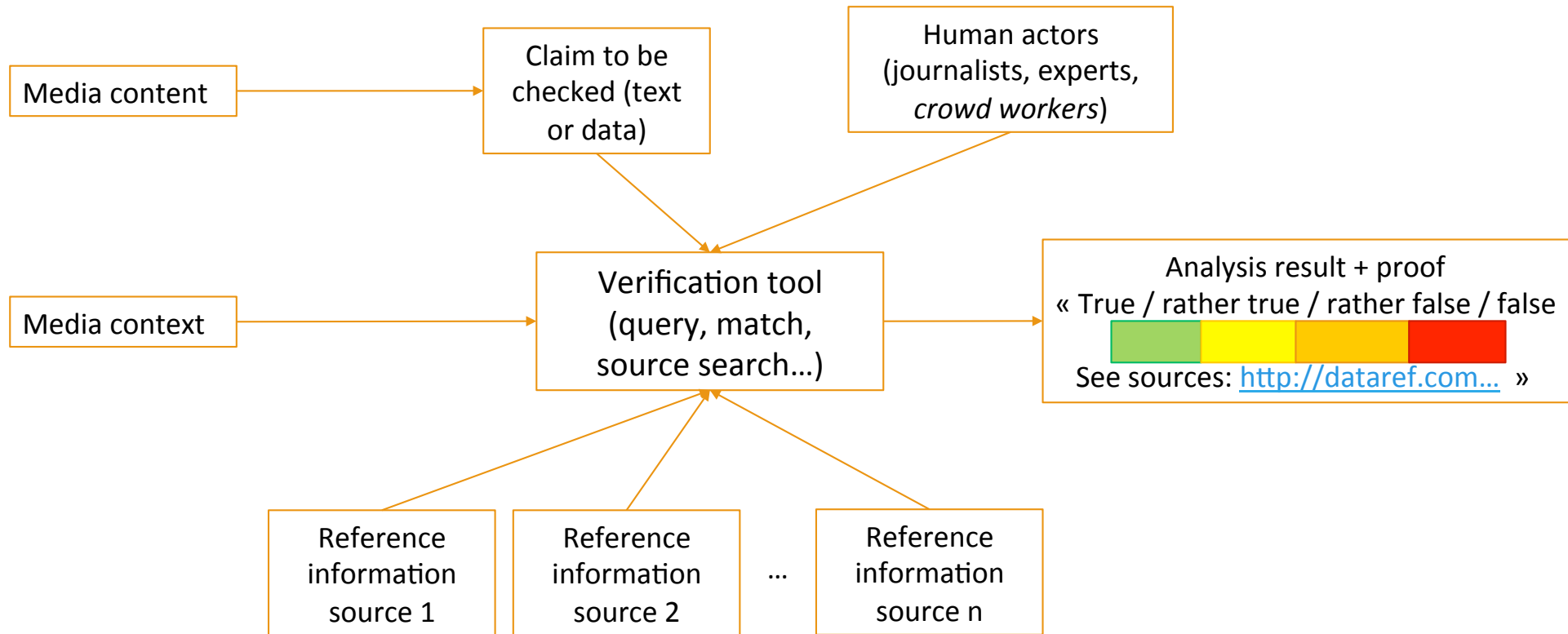
"We don't understand why the Democrats are so wholeheartedly against [President Trump's border wall]. They voted for it in 2006."

Fox News Sunday – Sunday, April 23, 2017

SHARE READ MORE

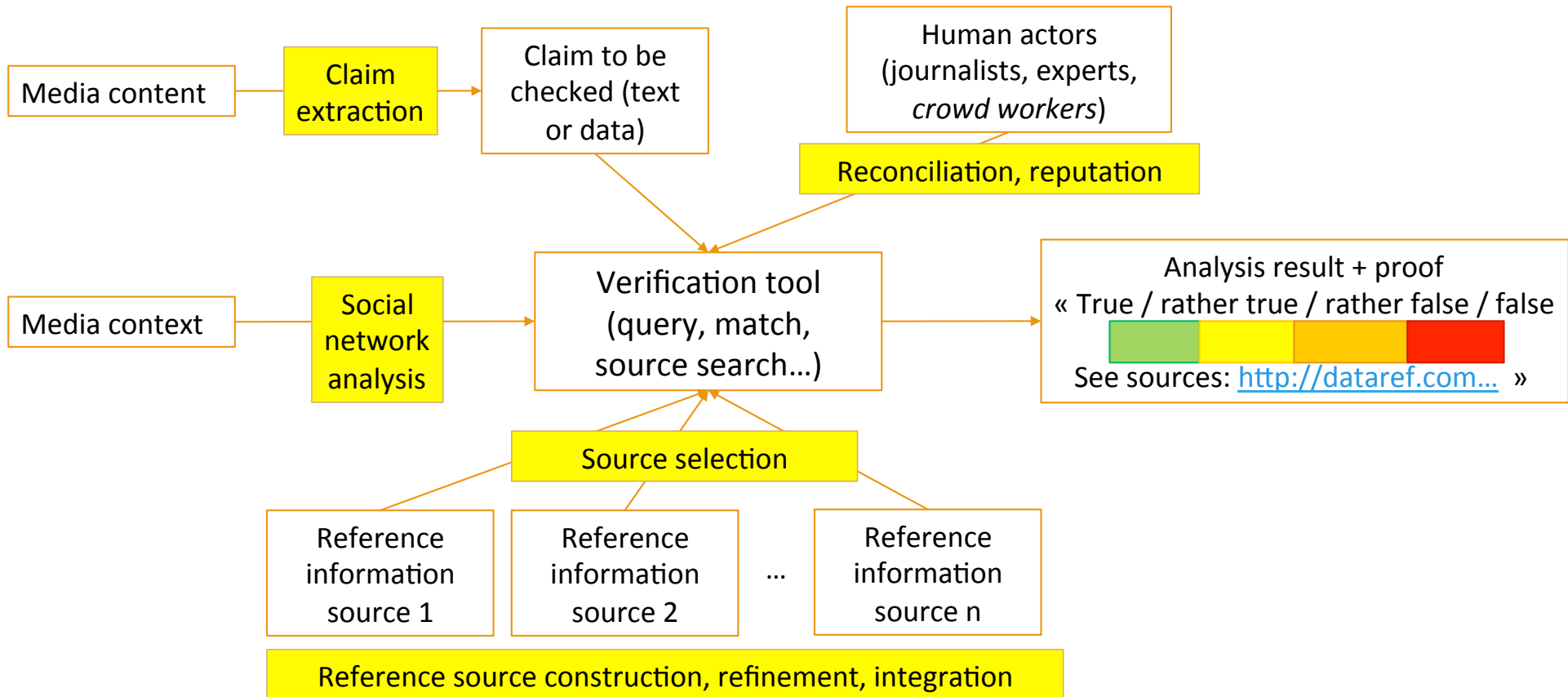
FACTCHECK.ORG

# Fact-checking is a content management problem

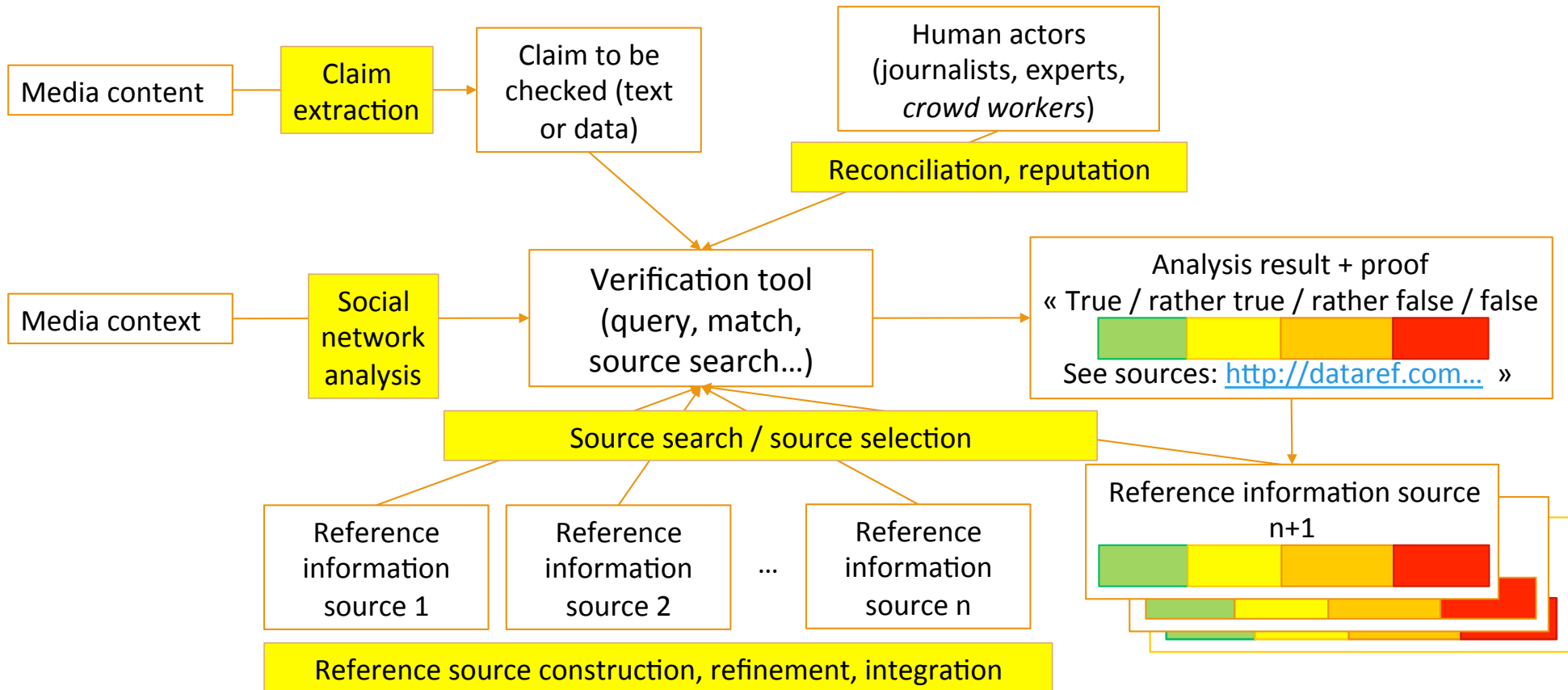




# Fact-checking is a content management problem



# Fact-checking is a content management problem



# It's not just checking

Most aspects of modern reality are complex

For the final audience, **explaining** is as important and useful as checking

Future is hard(er) to check!

abonnes.lermonde.fr/les-decodeurs/article/2017/05/05/une-sortie-de-l-euro-ferait-elle-exploser-la-dette-  
110% Search

Le Monde.fr ÉDITION ABONNÉS LE JOURNAL LES ARCHIVES LA CARTE VOS FAVORIS

INTERNATIONAL POLITIQUE SOCIÉTÉ ÉCO CULTURE IDÉES PLANÈTE SPORT SCIENCES PIXELS CAMPUS LE MAG

## LES DÉCODEURS

VENONS-EN AUX FAITS

LES DÉCODEURS Datavisualisation Vérification Nanographix Contexte Evasion fiscale Le blog du Décodex

### Une sortie de l'euro ferait-elle exploser la dette française ?

La mesure phare du Front national, à l'échéance repoussée par Marine Le Pen, est entourée de nombreuses zones d'incertitudes.

LE MONDE | 05.05.2017 à 11h51 • Mis à jour le 05.05.2017 à 12h15 |  
Par Maxime Vaudano

**Les décodeurs, mode d'emploi**

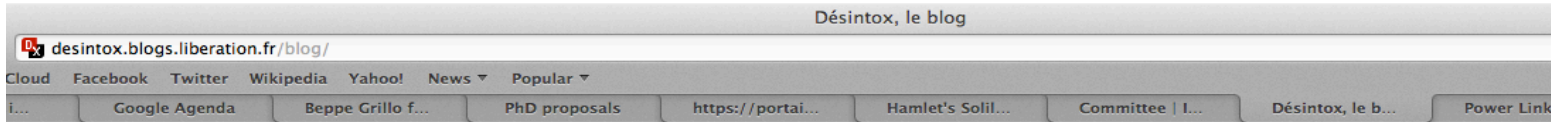
Les décodeurs du Monde.fr vérifient déclarations, assertions et rumeurs en tous genres ; ils mettent l'information en forme et la remettent dans son contexte; ils répondent à vos questions.

**LA CHARTE** Lire la charte >

ContentCheck  
ANR project  
(2016-2019)



# Libération *Désintox*: cost of saving Cyprus



DÉSINTOX le blog



mardi 23 avril 2013

0 Commentaires

Tweet 11

Like 34

MOTS CLÉS:

Florian Philippot  
Nicolas Dupont-Aignan

## Bobards en stock sur les plans de sauvetage européens

«Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de l'argent de l'UE, c'est-à-dire les Français, à des banques pourries à Chypre, on le donne ».

Nicolas Dupont-Aignan, le 25 mars sur France Bleu

«C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de dix milliards d'euros, dont deux milliards au titre de la France et des contribuables français».

Florian Philippot, le 25 mars sur France Info

«Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ? On a même pensé à une taxe sur les parcs d'attraction !»

Marine Le Pen, en septembre 2011, au moment du deuxième plan d'aide à la Grèce

### À PROPOS DE CE BLOG

Créé en 2008, Désintox est un observatoire des mensonges et des mots du discours politique.

Qui sommes nous?

Retrouvez-nous également du lundi au jeudi à 20h05 sur Arte dans l'émission 28 minutes.

### ALERTE EMAIL

Recevez des alertes Désintox par email

votre email

Je m'inscris gratuitement

### RECHERCHER

Q- OK

\$2-3 bn



\$10 bn



\$15 bn



# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not **given** but **lent**
- ❑ The European Mechanism of Stability will lend **\$9bn**
- ❑ Out of which France contributes 20% or a bit less than **\$2bn**



**INTOX** En période de rigueur budgétaire, les plan de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceintures. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de

*l'argent de l'UE , c'est-à-dire des Français, à des banques pourries à Chypre, on le donne», dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : «C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de 10 milliards d'euros, dont 2 milliards au titre de la France et des contribuables français». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce : «Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»*

**DÉSINTOX** Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?



# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not **given** but **lent**
- ❑ The European Mechanism of Stability will lend **\$9bn**
- ❑ Out of which France contributes 20% or a bit less than **\$2bn**

However, things are complicated because

- ❑ The initial contribution of France to the EMS (\$16 bn) **is** counted toward French public debt
- ❑ The remainder contribution (\$124) **is not**



## INTOX

En période de rigueur budgétaire, les plans de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceinture. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de

l'argent de l'UE, c'est-à-dire des Français, à des banques pourries à Chypre, on le donne», dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : «C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de 10 milliards d'euros, dont 2 milliards au titre de la France et des contribuables français». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce : «Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»

## DÉSINTOX

Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?

# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not
- ❑ The European Me
- ❑ Out of which France contributes 20% or a bit less than **\$2bn**

The initial statements are mostly false

However, things are complicated because

- ❑ The initial contribution of France to the EMS (\$16 bn) is counted toward French public debt
- ❑ The remainder contribution (\$124) is not



**INTOX** En période de rigueur budgétaire, les plan de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceintures. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de

l'argent de l'UE , c'est-à-dire des Français, à des banques pourries à Chypre, on le donne», dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : «C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de 10 milliards d'euros, dont 2 milliards au titre de la France et des contribuables français». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce : «Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»

**DÉSINTOX** Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?

# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not
- ❑ The European Me
- ❑ Out of which France contri than **\$2bn**

The initial statements are mostly false

The complete picture is complicated

However, things are complicated because

- ❑ The initial contribution of France to the EMS (\$16 bn) is counted toward French public debt
- ❑ The remainder contribution (\$124) is not



## INTOX

En période de rigueur budgétaire, les plan de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceintures. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de

UE, c'est-à-dire des Français, à des banques pourries à e donne», dénonçait Nicolas Dupont-Aignan. En version Florian Philippot, cela donne : «C'est nous qui allons encore verser de l'argent puisque le MES s'est engagé à hauteur de 10 milliards d'euros, dont 2 milliards au titre de la France et des contribuables français». Un propos qui fait écho -notamment- à celui entendu quelque deux ans avant, dans la bouche de Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce : «Comment peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»

## DÉSINTOX

Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?

# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not
- ❑ The European Me
- ❑ Out of which France contri than **\$2bn**

The initial statements are mostly false

The complete picture is complicated

However, things are complicated

- ❑ The initial contribution of France to counted toward French public debt
- ❑ The remainder contribution (\$124) is not

I would not have been able to give the answer



**INTOX** En période de rigueur budgétaire, les plans de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceinture. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de l'argent à Chypre, c'est-à-dire des Français, à des banques pourries à Chypre, ça donne», dénonçait Nicolas Dupont-Aignan. En version simplifiée : «C'est nous qui allons encore verser de l'argent engagé à hauteur de 10 milliards d'euros, à Chypre, France et des contribuables français». Un exemple - à celui entendu quelque deux ans plus tôt - de la politique de Nicolas Dupont-Aignan et Florian Le Pen, au sujet cette fois du plan d'aide à Chypre : «Peut-on imposer aux Français, aux classes populaires et moyennes, aux petites entreprises, de nouvelles taxes, des taxes sur les sodas, une hausse de la CSG de 550 millions d'euros, et d'un autre côté alimenter de 15 milliards d'euros supplémentaires en Grèce l'incendie de la zone euro ?»

**DÉSINTOX** Quel impact ont donc les plans de sauvetage successifs de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?

# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not given
- ❑ The European Mechanism
- ❑ Out of which France

The initial statements are mostly false

However, things are complicated

- ❑ the initial contribution of France toward French public debt
- ❑ the remainder contribution (\$124) is not

The complete picture is complicated

I would not have been able to give the answer

Yet I vote!



**INTOX** En période de rigueur budgétaire, les plan de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceintures. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de l'argent à l'étranger, ça change. C'est la logique de l'UE, c'est-à-dire des Français, à des banques pourries à Chypre», dénonçait Nicolas Dupont-Aignan. En version française : «C'est nous qui allons encore verser de l'argent engagé à hauteur de 10 milliards d'euros, à l'étranger, en France et des contribuables français». Un économiste, à celui entendu quelque deux ans auparavant par le député Le Pen, au sujet cette fois du plan d'aide à la Grèce : «On va-t-il imposer aux Français, aux classes moyennes, de nouvelles taxes, des milliards supplémentaires en Grèce».

**DESINTOX** ... de l'Euro? Et le dernier en date, en direction de Chypre, va-t-il contraindre les Français à allonger 2 milliards d'euros?



# Saving Cyprus: how much does it cost?

The article goes on saying:

- ❑ The money is not given
- ❑ The European Mechanism
- ❑ Out of which France

The initial statements are mostly false

However, things are complicated

- ❑ the initial contribution of France toward French public debt
- ❑ the remainder contribution (\$124) is not

The complete picture is complicated

I would not have been able to give the answer

Yet I vote!

Don't even mention EU politics (commission, parliament...)



**INTOX** En période de rigueur budgétaire, les plans de sauvegarde successifs offrent depuis trois ans un boulevard aux détracteurs de l'euro et de l'Europe. Nicolas Dupont-Aignan et Florian Philippot ont ainsi, de concert, dénoncé récemment les milliards déversés à Chypre au moment où les Français se serrent la ceinture. «Les Français vont devoir donner 2 à 3 milliards d'euros pour des banques à Chypre. D'un côté on supprime les infirmières, on surtaxe les PME en France [...] mais quand il s'agit de donner de l'argent à Chypre, c'est-à-dire des Français, à des banques pourries à Chypre, ça ne se passe pas comme ça», dénonçait Nicolas Dupont-Aignan. En version anglaise, il a écrit : «C'est nous qui allons encore verser de l'argent engagé à hauteur de 10 milliards d'euros, à Chypre, à France et des contribuables français». Un jour, il faut arrêter ça. À celui entendu quelque deux ans plus tôt par Marine Le Pen, au sujet cette fois du plan d'aide à la Grèce, on peut-on imposer aux Français, aux classes moyennes, de nouvelles taxes, des milliards supplémentaires en Grèce

# Democracy needs data integration!

---

1. Democratic societies crucially need the press
2. The press needs to report on an increasingly complex reality...
3. ... much of which is documented in digital datasources to which journalists have (or can wrangle) access
4. **They need help finding and exploiting this data!**

## State of the industry

2014, Financial Times reporter: "the data is in my home on a hard disk"

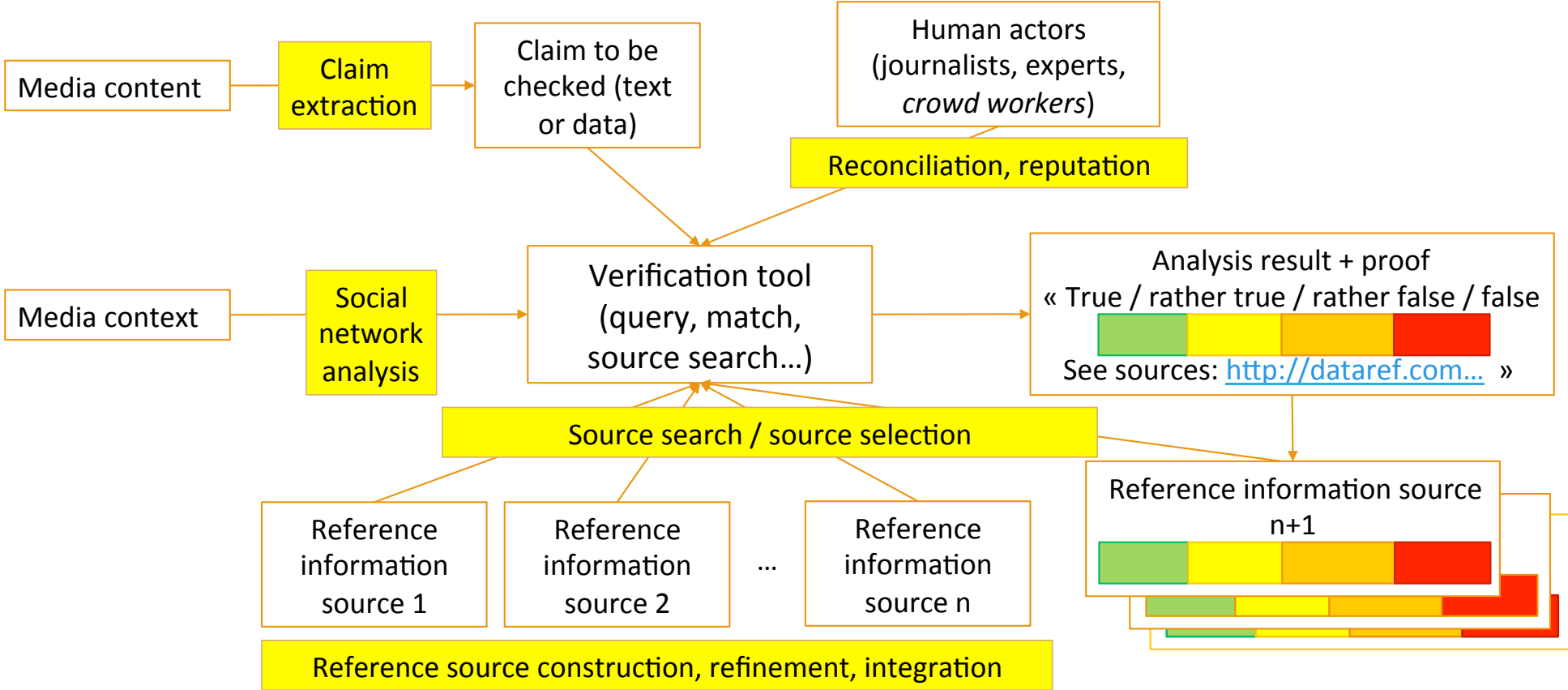
2016, Le Monde *old-school* journalist: "*the data is in my paper+pencil storage system*"

2014, Le Monde, Les Décodeurs : Google docs, starting to use shared folders

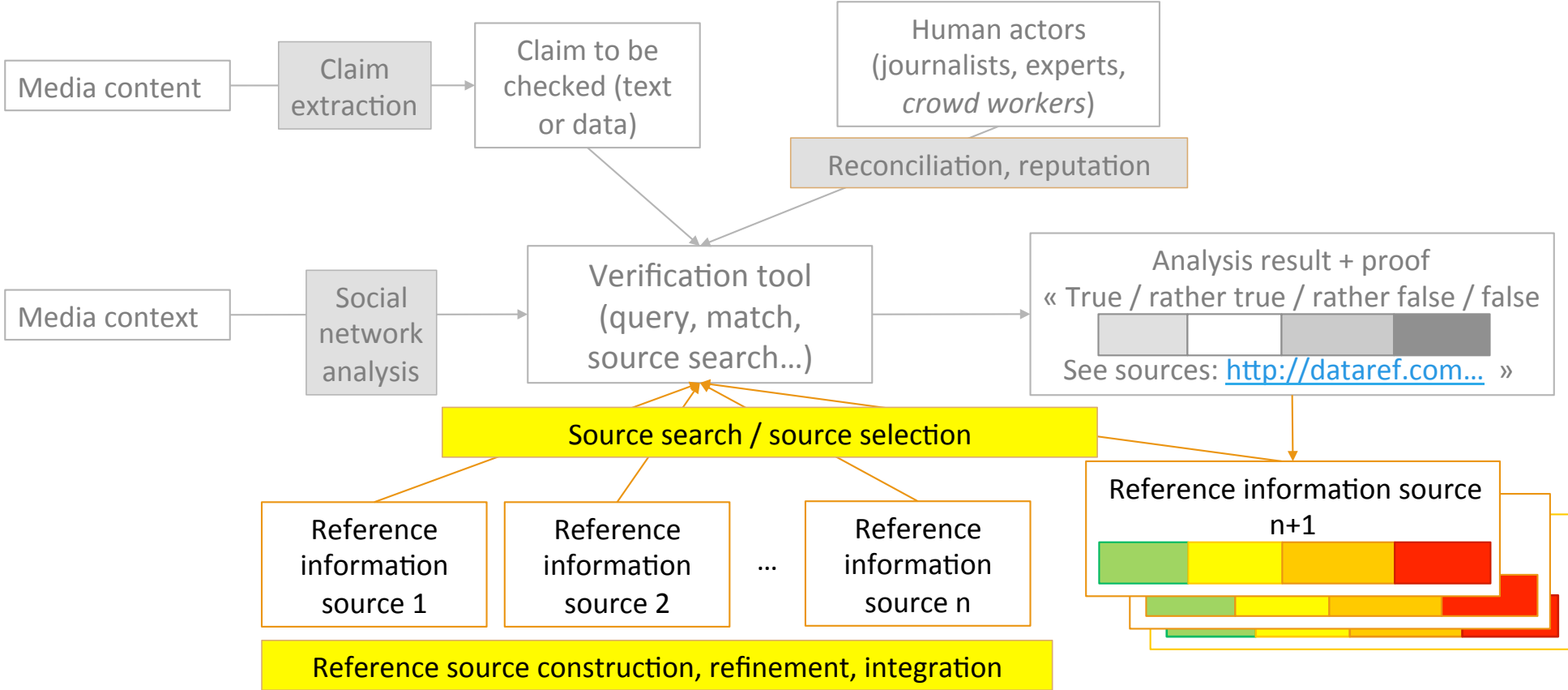
2017, Le Monde, Les Décodeurs: fact-checking browser plug-in, open source, open data

Slowly getting there...

# Back to the content management problem



# Back to the content management problem



# What data journalism needs

---

- ❑ **Access** to data → any model (text, relational, graph, image, video, JSON...)
- ❑ **Trust** in the data: serious journalists do not use data they don't trust
  - ❑ Akin to trusting their sources (e.g. AFP "censorship")
  - ❑ A claim needs to be backed by data easy to trace to a trusted source, to convince the reader
- ❑ **Exploratory** tools: journalists don't always know what they will find...
  - ❑ ... but they have a hunch
- ❑ Understanding of the data / application domain

## What data journalism doesn't have

- ❑ Time
- ❑ Much money



# Source identification, integration and selection

---

# Data integration architectures

---

Goal: turn N information sources into 1 system

Architectures:

## 1. Data warehouse

- ❑ Ingest all sources in a single system (Extract-Transform-Load)
- ❑ Star product for relational database vendors
- ❑ Schema design, maintenance, active rules...
- ❑ Commerce, banking, telecom...

## 2. Mediator systems (revisited as: polystores)

## 3. Dataspaces

## 4. Data lakes...

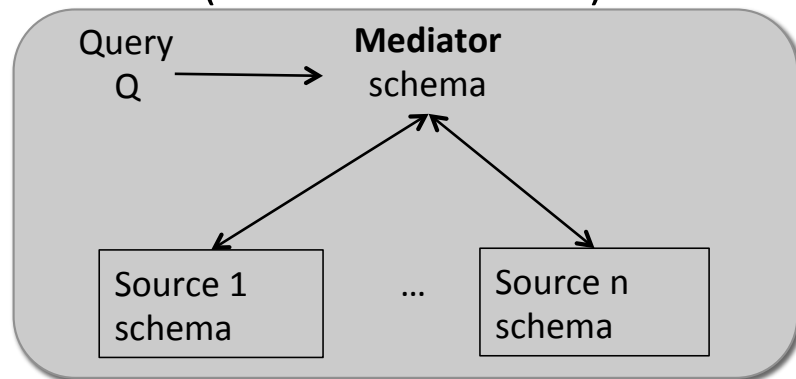
# Mediator systems (1980's—2000's)

A set of **data sources**, of the same or different data model, query language; source schemas

A **mediator** data model and mediator schema, used for queries

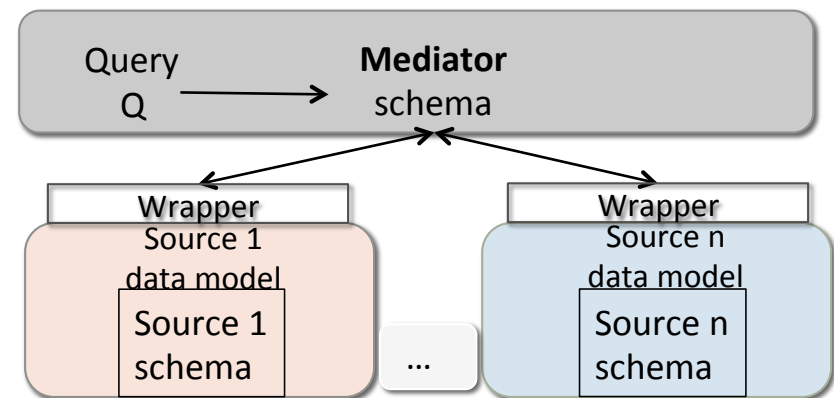
Independent sources; mediator-driven integration

Common data model  
(sources+mediator)



or

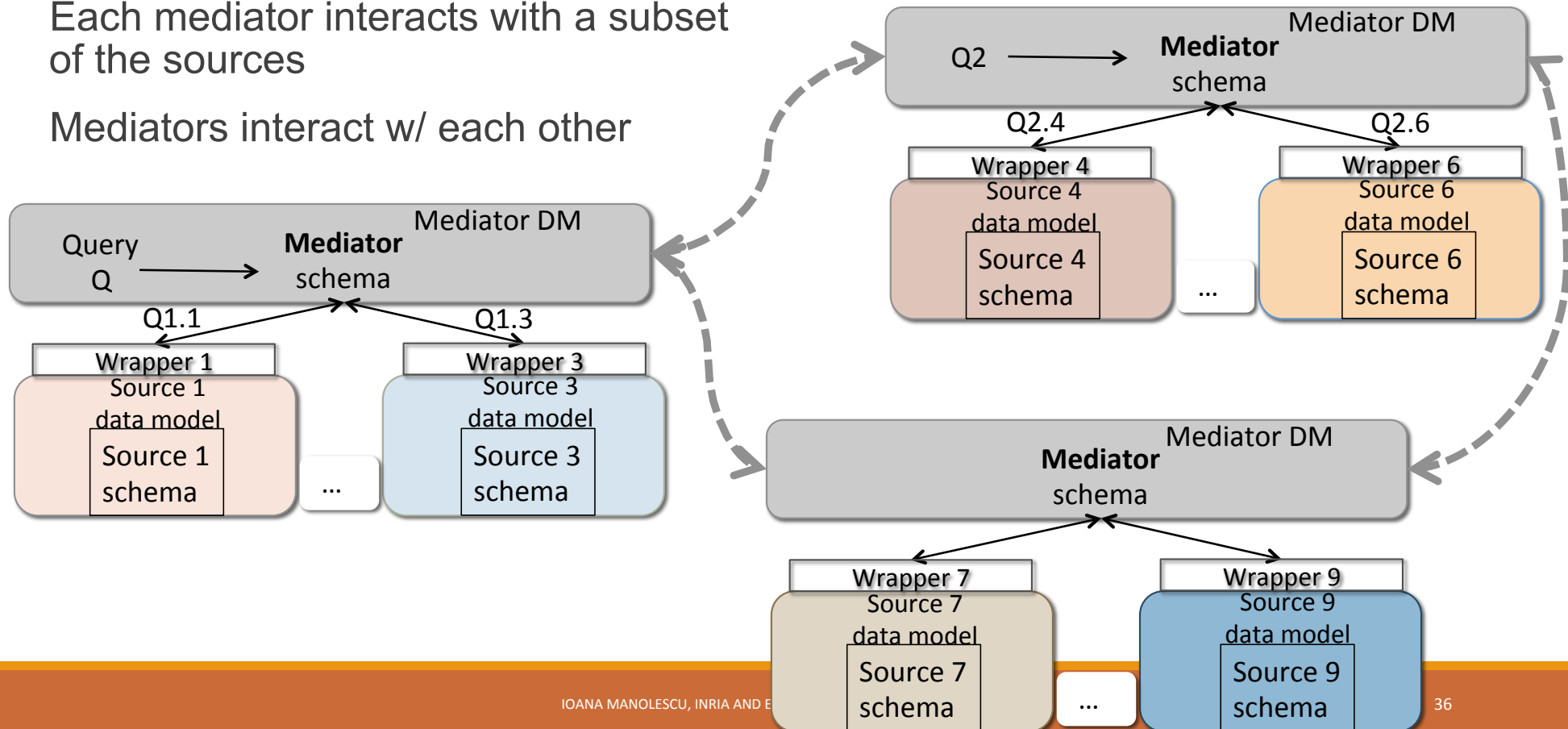
Mediator data model



# Mediator systems (1980's—2000's)

Each mediator interacts with a subset of the sources

Mediators interact w/ each other



# Mediator systems (1980's—2000's)

---

*Mappings* allow to relate global schema and source schemas

**Global-as-view:** mediator (global) schema expressed as a query over the source schemas

- + simplicity of query answering (= view unfolding)
- mediator schema may need to change when  $\pm$  sources

**Local-as-view:** source (local) schemas expressed as queries over the global schema

- + sources can be added to / removed from the integration system independently
- complexity of query answering (= rewriting queries using views)

**Global-local-as-view:** query over global schema  $\approx$  query over local schema



# Mediators revisited: polystores (2010's —)

---

Polystores: access to multiple stores that natively support different data models

Similar to mediator systems, but:

- ❑ Without a global schema
- ❑ A source can be a massively distributed system, e.g. a Hadoop or Spark cluster

Global-as-view style: Elmore, Duggan, Stonebraker, Balazinska et al. "A Demonstration of the BigDAWG Polystore System" (demo), PVLDB 2015

Local-as-view style: Bugiotti, Bursztyn, Deutsch, Ileana, Manolescu: "Flexible Hybrid Stores: Constraint-Based Rewriting to the Rescue" (demo), ICDE 2016

More recent work from EPFL (A. Ailamaki)

# Dataspaces

---

Franklin, Halevy, Maier: "From Databases to Dataspaces: A New Abstraction for Information Management". SIGMOD Record 34, 4 (Dec. 2005)

Dataspace = a large number of **diverse, interrelated, heterogeneous data sources**

Does not assume all semantic relationships between sources are known and have been specified

For: Personal information management, Scientific data management, ...

**Uniform access to data: keyword search**

- ❑ In parallel over different data sources

Dong, Halevy: "Indexing Dataspaces", SIGMOD 2007

# Automated source selection: FactMinder

Goasdoué, Karanasos, Katsis, Leblay, Manolescu, Zampetakis:  
"Fact-checking and analyzing the Web", SIGMOD 2013

Browser plug-in

Bringing up rich context for a Web page, from document and knowledge bases

Source search, contextualization

Supported by template queries over XML documents and RDF graphs

« Second screen »

The screenshot displays the FactMinder browser plug-in interface. The main content area shows a news article titled "Bill Clinton apporte son soutien à Barack Obama". The sidebar on the right is divided into several sections:

- Concepts:** A list of entities with their dbpedia:Person or dbpedia:Place identifiers, including Bill Clinton, Barack Obama, Des Moines, Iowa, Hillary Clinton, Michelle Obama, Bruce Springsteen, and Pennsylvanie.
- Facts & figures:** A section titled "Curriculum" with a photo of Bill Clinton and biographical details: Born William Jefferson, August 19, 1946, Hope, Arkansas, U.S., Political party Democratic Party, Spouse(s) Hillary Rodham.
- Quotes:** Two quotes with links: "Launching the Africa Regional Media Hub in Johannesburg <http://bit.ly/SSjCt>" and "President Obama speaking LIVE for the last time before the election <http://bit.ly/PSfky>".
- Related stories:** A list of related news items with links, such as "Ohio, Floride, ... être compliqué <http://lemonde.fr/...829254.html>" and "Des indicateurs ... d'Obama <http://lemonde.fr/...dobama/>".
- Sources:** A section for the source of the article, "Romney... agressifs <http://lemonde.fr/...>".

# Data lakes

---

"The closest thing to dataspace that actually happened / imperfect realization" (A. Halevy, 2017)

Reality of enterprise systems:

- ❑ Glut of databases accumulating over time
  - ❑ None is ever discarded
  - ❑ NoSQL and massively parallel systems added to previous ones...
  - ❑ Most frequent scenario: SQL + Hadoop/Spark clusters (RDBMSs still rock!)
- Accordingly: no semantic integration; schema on load; algorithms for schema and join discovery...

D. Deng, R. C. Fernandez, Z. Abedjan, S. Wang, M. Stonebraker et al. "The Data Civilizer system" In CIDR, 2017

<https://www.ibm.com/analytics/us/en/data-management/data-lake/>

<https://azure.microsoft.com/en-us/solutions/data-lake/>

# Toward usable data integration tools for journalists

Bonaque, Cao, Cautis, Goasdoué, Letelier, Manolescu, Mendoza, Ribeiro, Tannier, Thomazo:  
"Tatooine: a lightweight data integration architecture for data journalism" (demo), VLDB2016

So many data sources, so little time!

~ Data lake

« Can we **group tweets of politicians by political current** and analyze their **most frequent topics**? »

« Can we **classify articles** by their main **concept** and do a **tag cloud**? »

Can we **industrialize** answering such requests?





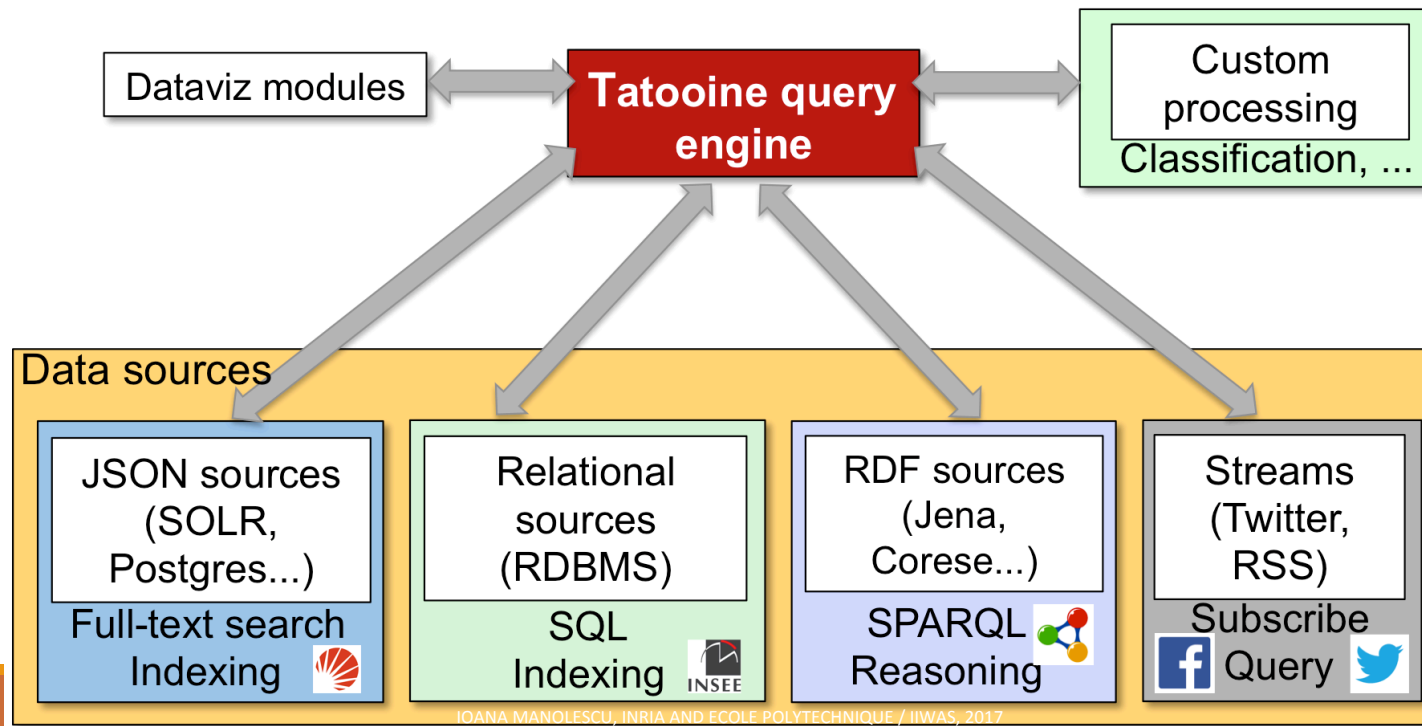
# Content management for data journalism

Bonaque, Cao, Cautis, Goasdoué, Letelier, Manolescu, Mendoza, Ribeiro, Tannier, Thomazou  
"Tatooine: a lightweight data integration architecture for data journalism" (demo), VLDB2017

So many data sources, so little time!

Data journalism; reference source enrichment

**Demo**



# Conclusion

---

**Data integration:** desirable goal/vision, still fundamentally hard

Maybe the human understanding is the bottleneck?

Too much data?

Journalism needs to make sense of increasingly complex data

And trust it

Journalism remains by the humans for the humans (despite automated sport reports)

Essential for a free society, which we hope to keep

# Much more to do

---

Detecting fake news based on the content, the social context, the language

- ❑ "Michelle was caught cheating with Eric Holder – Obama is FURIOUS!!!"

Statistic / ML / IR / graph (social) analysis approaches

Publishing content with proofs or provenance

- ❑ Restricted to well-behaved content.. not those we worry about

Building fact bases: crowd-sourcing, knowledge base construction, extraction

Formalizing explanations and proofs; argumentation theory

Calling Bullshit in the Age of Big Data course at U. Washington:

<https://www.youtube.com/playlist?list=PLPnZfvKID1Sje5jWxt-4CSZD7bUI4gSPS>

# Merci / questions?

---