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

Ioana Manolescu

Data integration for journalism: goals, tools, and architectures

Ioana Manolescu

Inria Saclay and LIX (CNRS and Ecole Polytechnique)

04/12/2017

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

Ceauşescu re-elected
at the 14th congress!

Where I come from: Romania, 1989

Ceauşescu 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

Democratic societies crucially need the press

To debate and express dissent

To analyze, confim 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

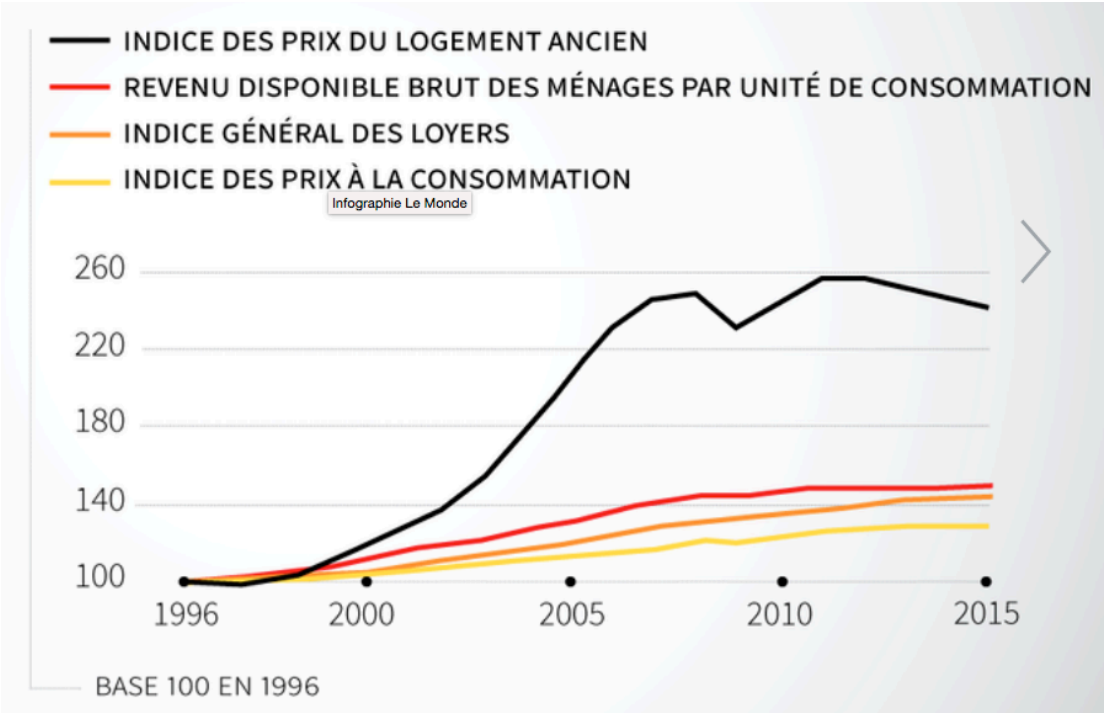
To analyze, confim or refute public statements

Fact-checking (see next)

To expose and explain society functioning

Data journalism

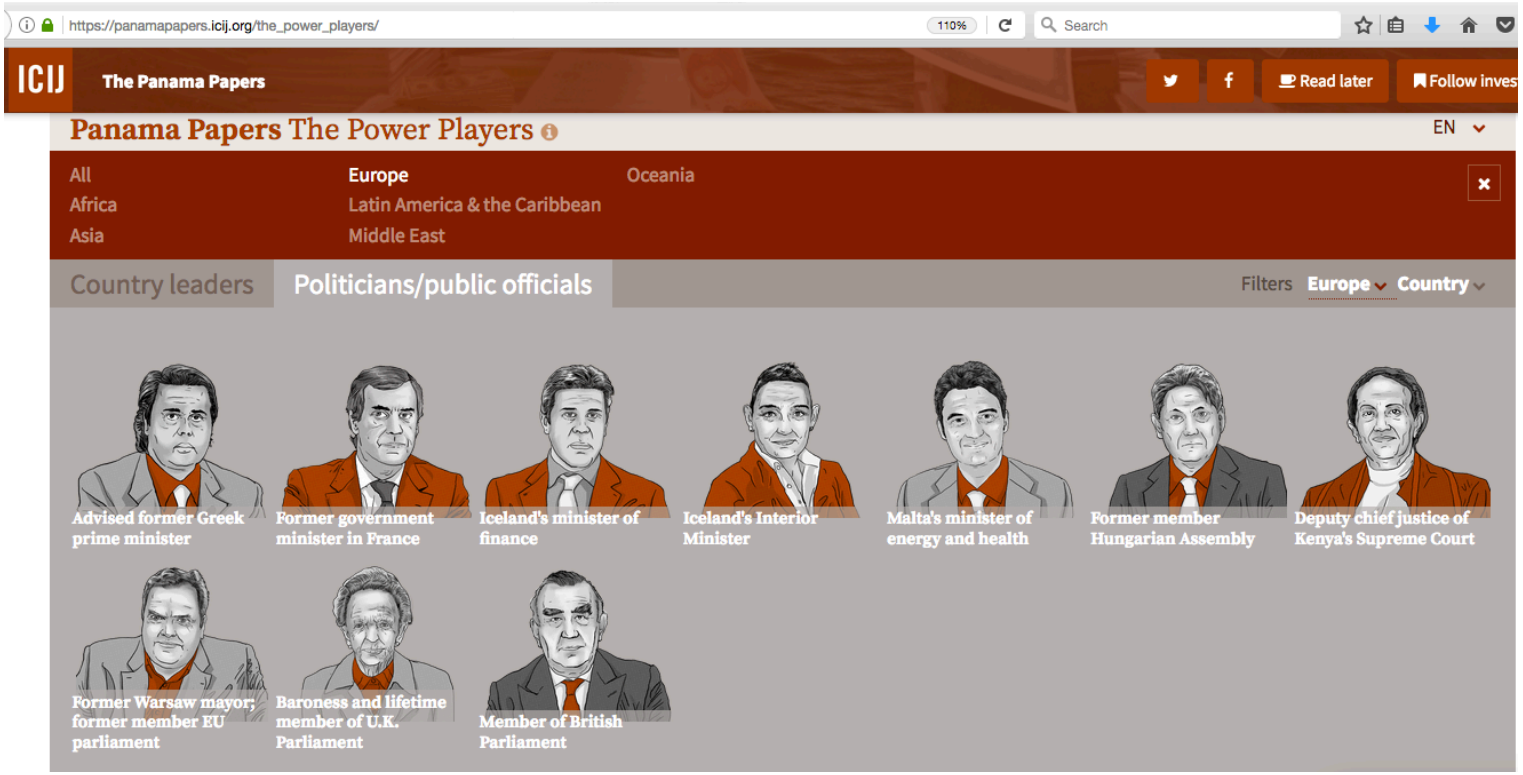
Investigative journalism based on **complex and/or large data**



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

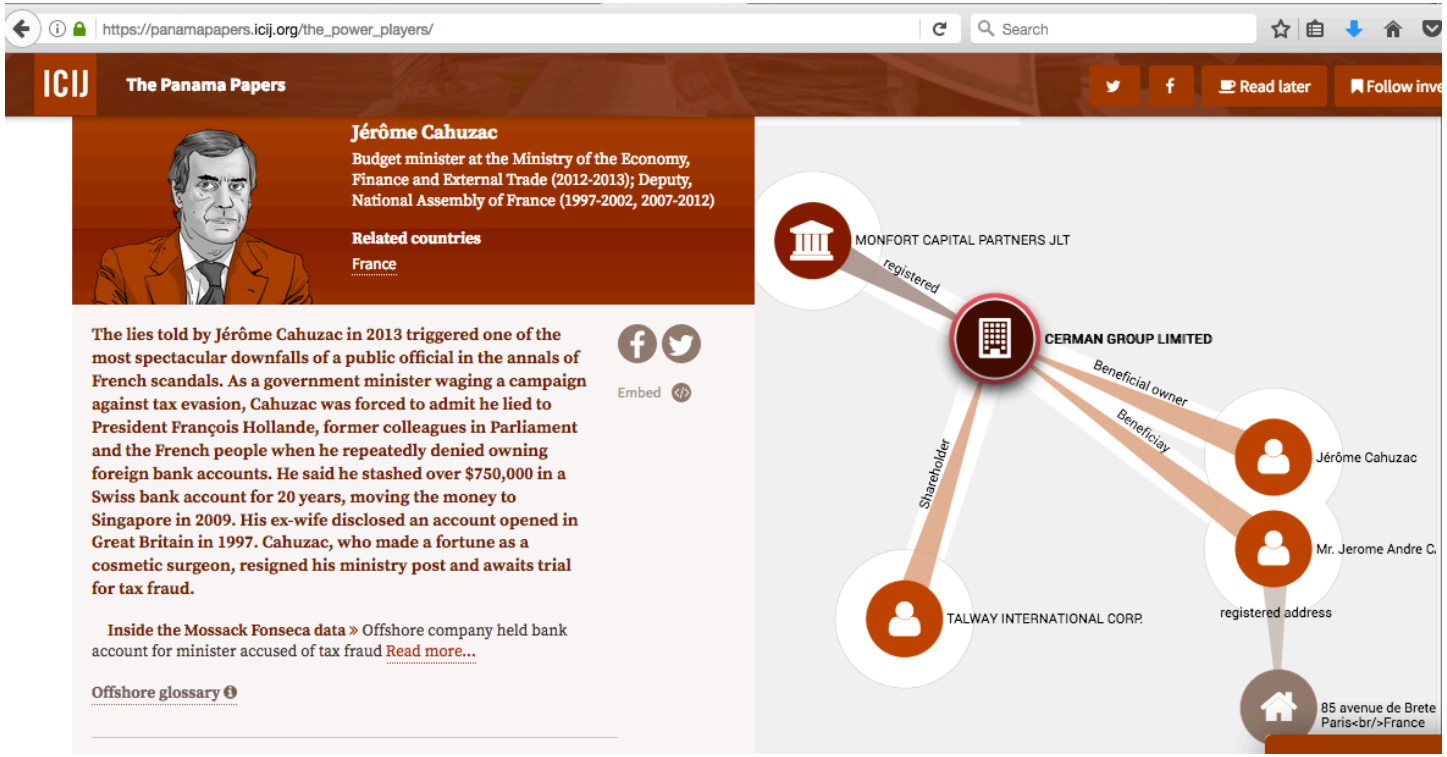
Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)



Data journalism

Panama Papers (International Consortium of Investigative Journalism, ICIJ)



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 existe 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

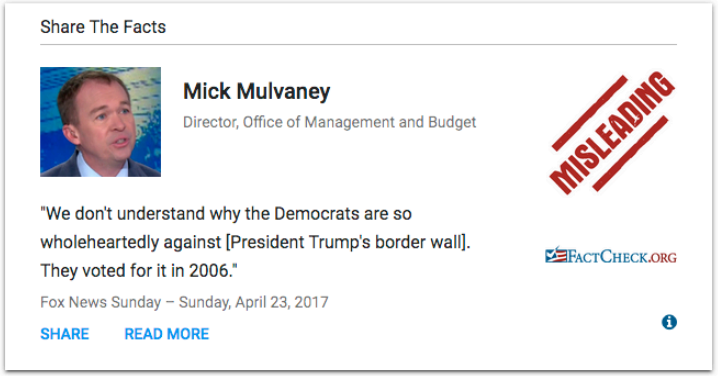
Fact-checking (2012 – ongoing)

Not everyone agrees, however, that Democrats are not flip-flopping on the issue.

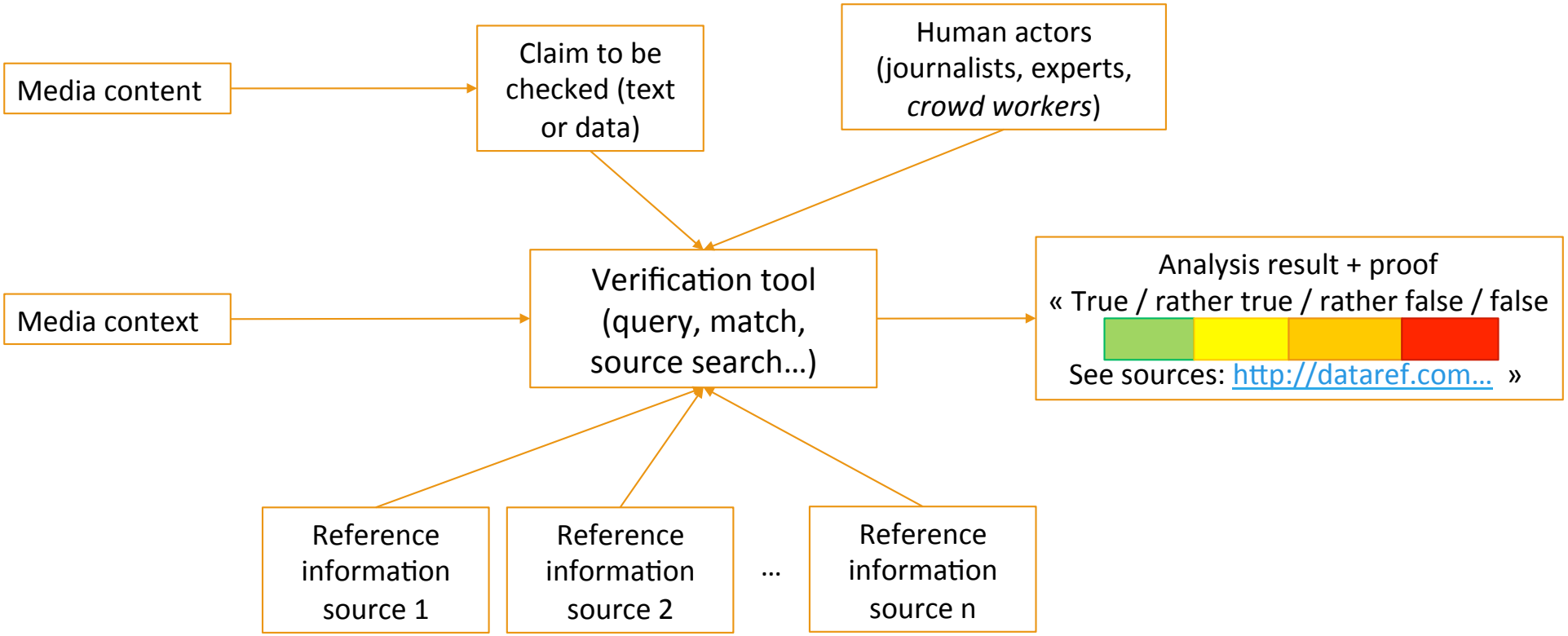
Mark Krikorian, executive director of the Center for Immigration Studies, a think tank that advocates for lower immigration, said that because the public doesn't know exactly what border barriers the Trump administration wants to build, Mulvaney's statement is not an "exact" comparison. But, he said, 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.

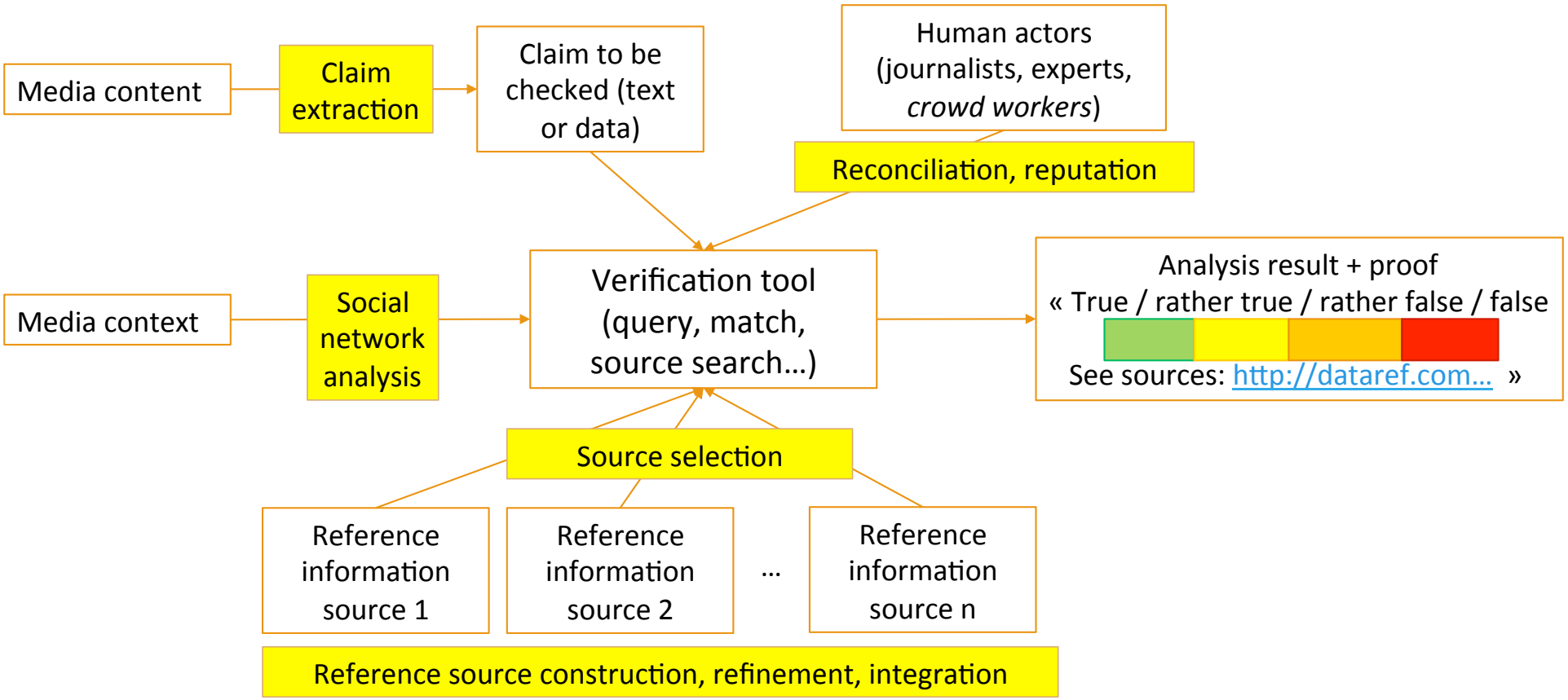
At the surface level, it is true in a broad sense that Democrats including Schumer, Obama and Clinton have in the past supported border fencing. All three voted for the Secure Fence Act of 2006, and all three supported the 2013 Senate immigration overhaul that passed the Senate, and which called for tougher border security including some additional fencing. But to claim that those measures are the same as what Trump is proposing is a stretch.



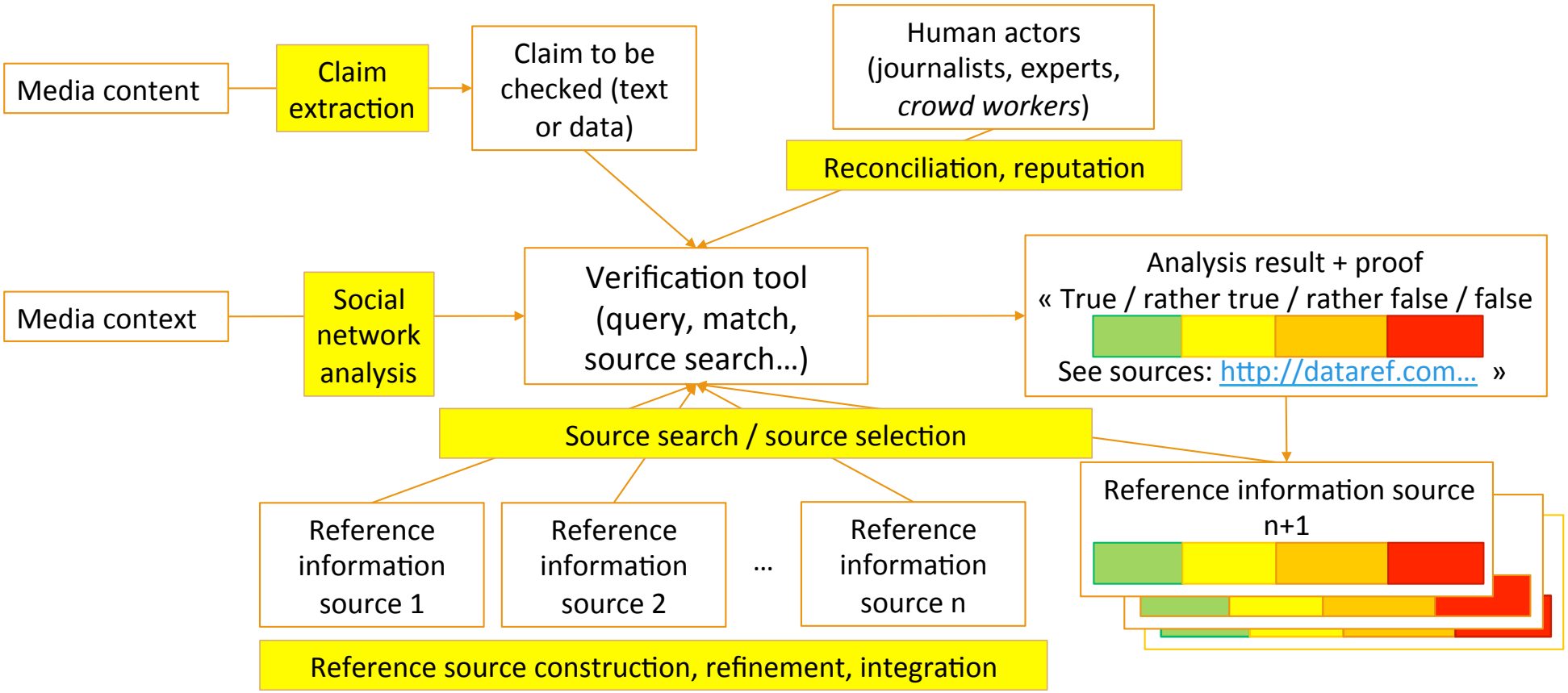
Fact-checking is a content management problem



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!

ContentCheck ANR project (2016-2019)

Libération *Désintox*: cost of saving Cyprus

$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 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 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 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

However, things are compli

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

The initial statements are mostly false

The complete picture is complicated

I would not have been able to give the answer

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 : *«C'est nous qui allons encore verser de* t engagé *à hauteur de 10 milliards d'euros,* France et des contribuables français». Un ment- à celui entendu quelque deux ans ine Le Pen, au sujet cette fois du plan d'aide ut-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~~ giv
- The European Mecha
- Out of which France

However, things are complicat

- the initial contribution of Franc toward French public debt
- the remainder contribution ($124) **is nc**

The initial statements are mostly false

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 UE , c'est-à-dire des Français, à des banques pourries à donne»*, dénonçait Nicolas Dupont-Aignan. En version : *«C'est nous qui allons encore verser de engagé à hauteur de 10 milliards d'euros, France et des contribuables français»*. Un ment- à celui entendu quelque deux ans ine Le Pen, au sujet cette fois du plan d'aide ut-on imposer aux Français, aux classes es, de nouvelles taxes, des 50 millions d'euros, et d'un supplémentaires en Grèce

DESINTOX ns 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~~ giv
- The European Mecha
- Out of which France

However, things are complicat

- the initial contribution of Franc toward French public debt
- the remainder contribution ($124) **is nc**

The initial statements are mostly false

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 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 donne»*, dénonçait Nicolas Dupont-Aignan. En version : *«C'est nous qui allons encore verser de engagé à hauteur de 10 milliards d'euros, France et des contribuables français»*. Un ment- à celui entendu quelque deux ans ine Le Pen, au sujet cette fois du plan d'aide ut-on imposer aux Français, aux classes es, de nouvelles taxes, des 0 millions d'euros, et d'un supplémentaires en Grèce

ns de sauvetage successifs tion de Chypre, ros?

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"

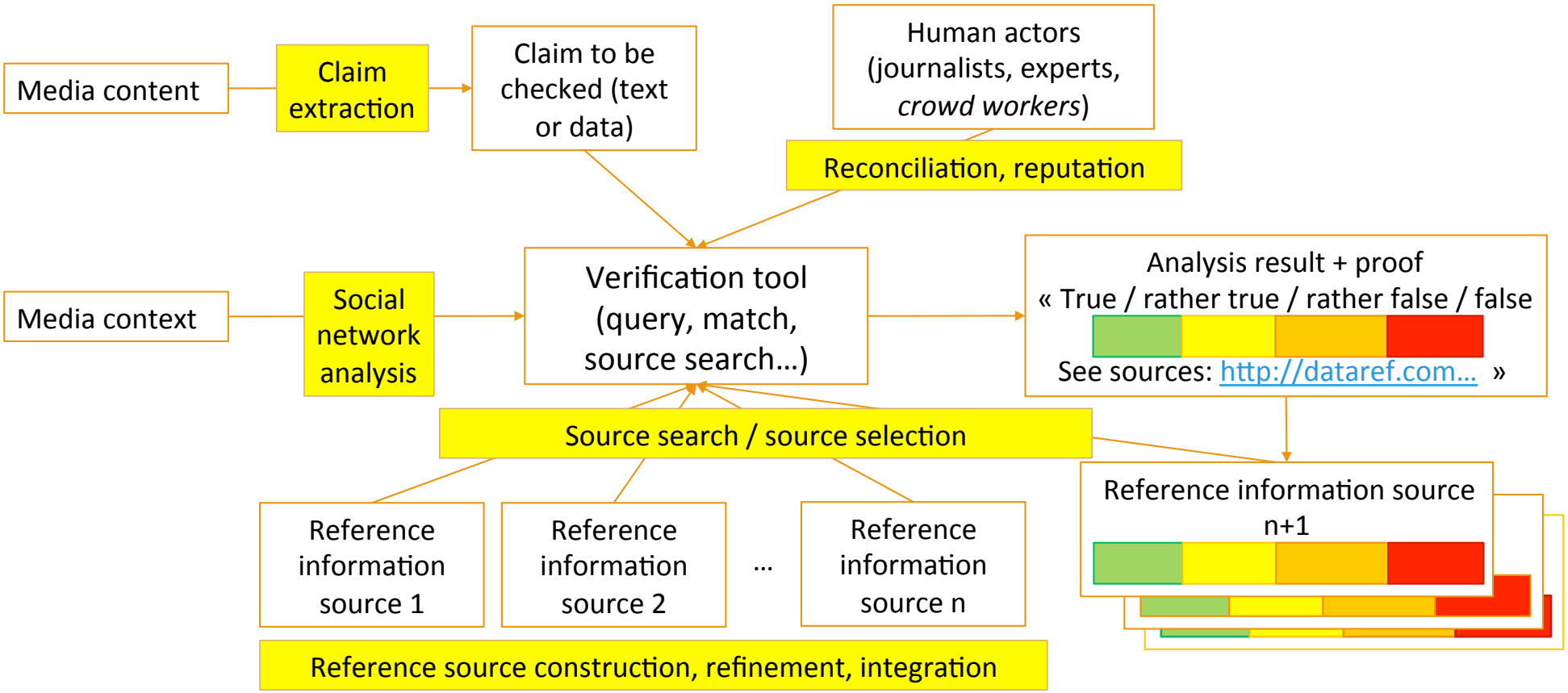
2016, Le Monde *old-school* journalist: "*the data is in my paper+pencil storage system*"

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

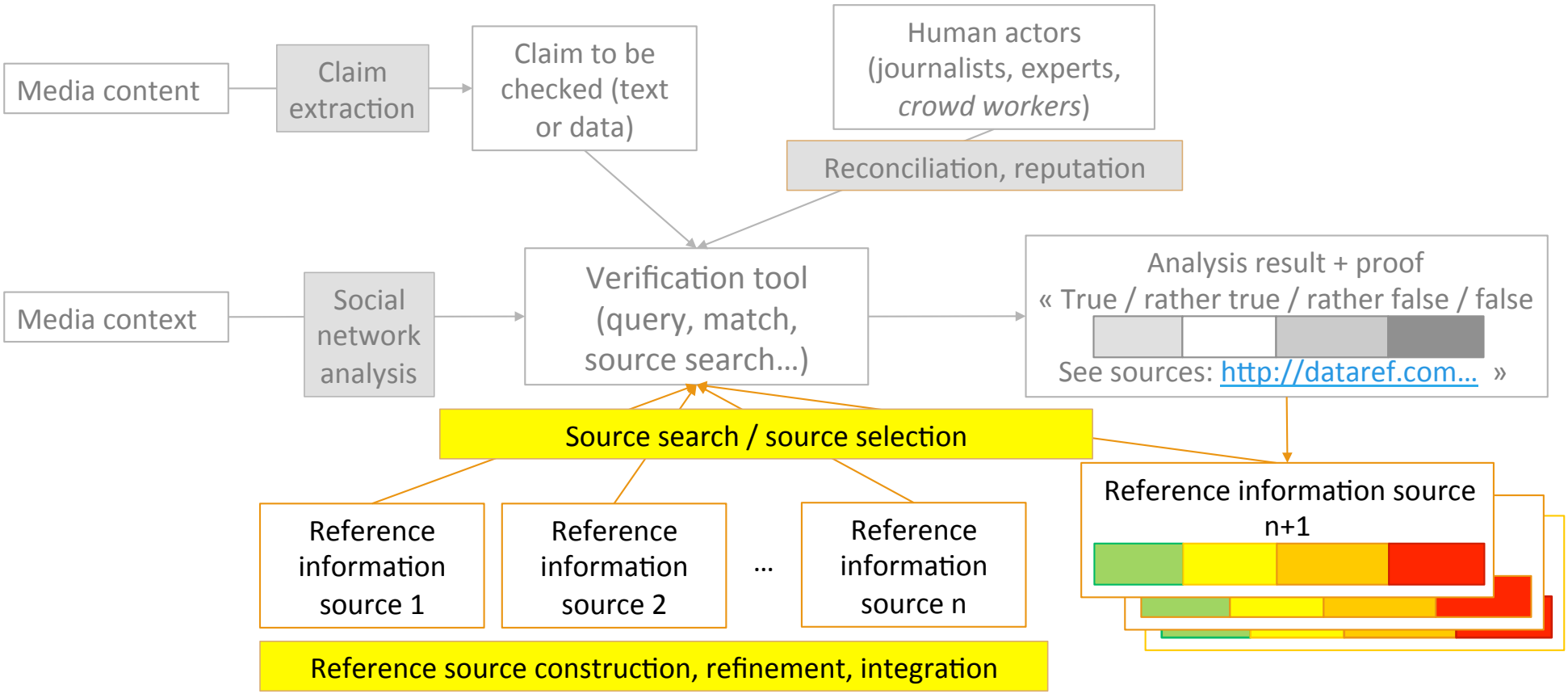
2017, Le Monde, Les Décodeurs: fact-checking brower plug-in, open source, open data

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:

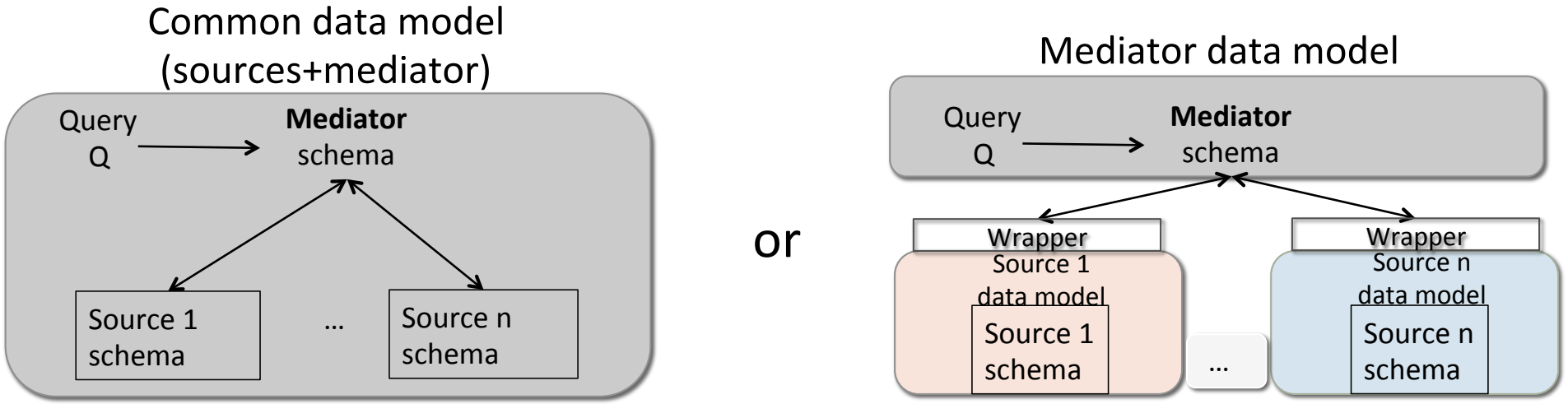
1. **Data warehouse**
   - Ingest all sources in a single system (Extract-Transform-Load)
   - Star product for relational database vendors
   - Schema design, maintenance, active rules...
   - Commerce, banking, telecom...
2. **Mediator systems** (revisited as: polystores)
3. **Dataspaces**
4. **Data 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

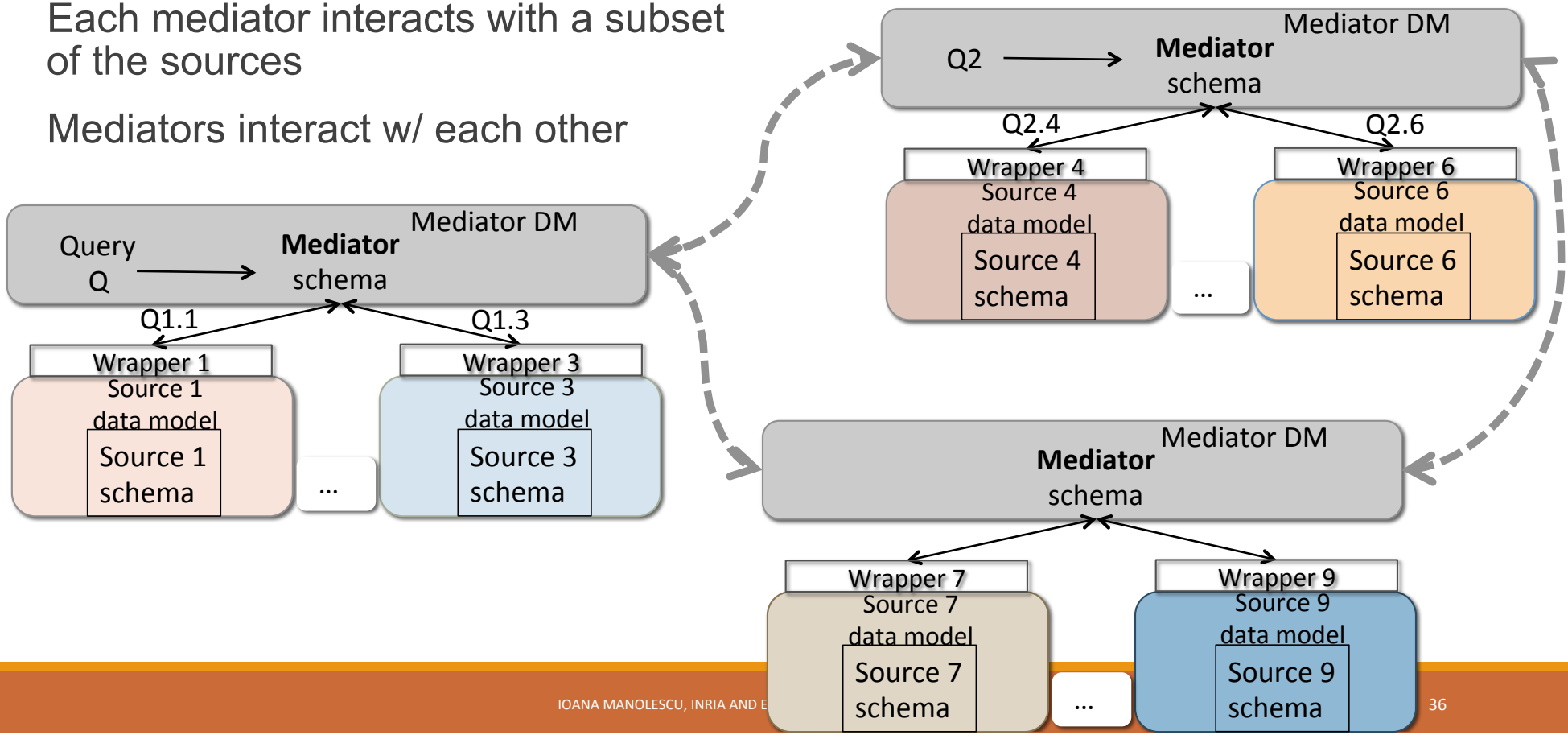
Independent sources; mediator-driven integration



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 ± 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 ≈ 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 »

Data lakes

"The closest thing to dataspaces 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 el. "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**? »

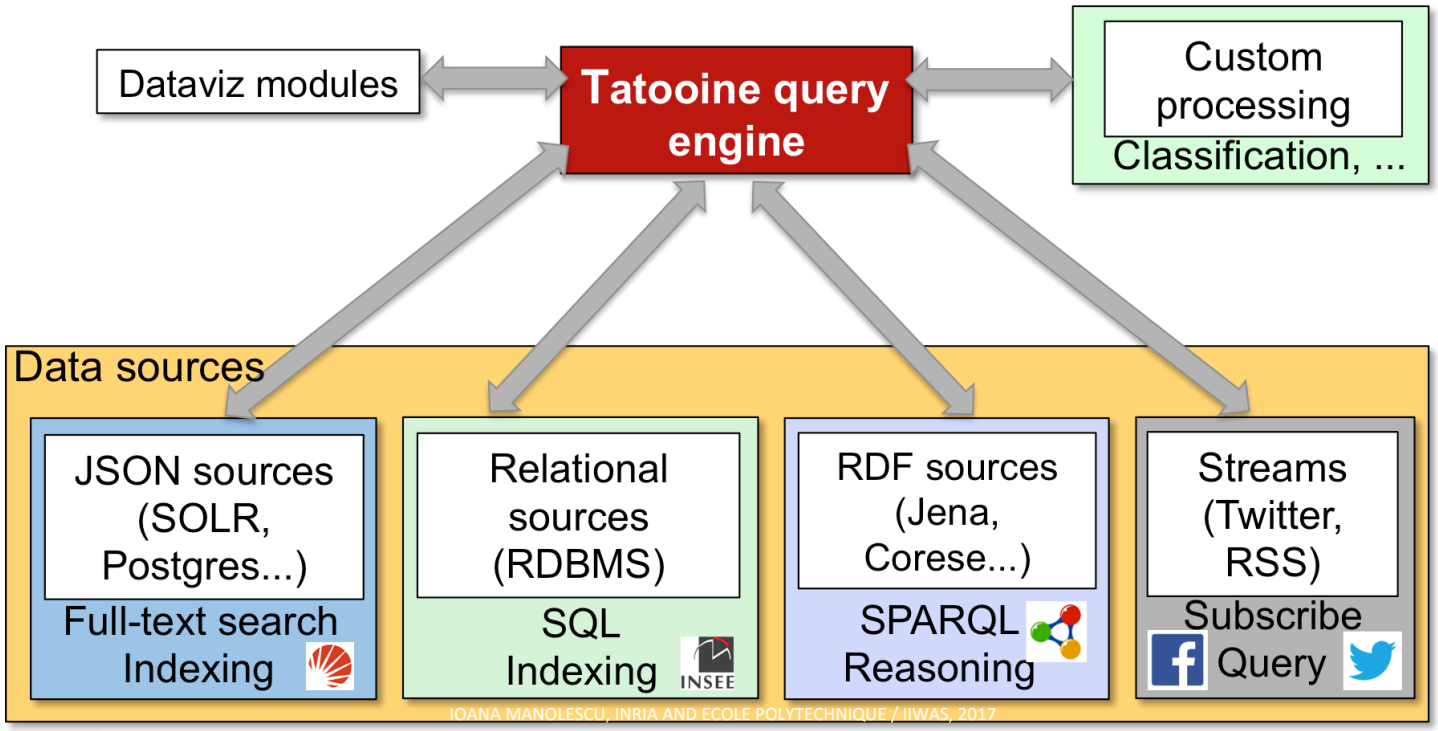
Can we **industrialize** answering such requests?

Content management for data journalism

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

Demo

So many data sources, so little time! Data journalism; reference source enrichment



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?