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, Salzbourg, 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
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--)


**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

Ceauşescu was in power since 1965.

Ceauşescu re-elected at the 14th congress!
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

To expose and explain society functioning

Fact-checking (see next)
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
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 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”

Fact-checking (2012 – ongoing)

Did Democrats Once Support Border Wall?

By Robert Farley  Published on April 26, 2017

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 the construction of 706 miles of fencing and enhanced surveillance technology, such as unmanned drones, ground-based sensors, satellites, radar coverage and cameras. Sen. Chuck Schumer and then-Sen. 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 very general sense, the Democrats named by Mulvaney supported a bill to build more border wall in 2006. That’s it.

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.

“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 The Facts

Mick Mulvaney
Director, Office of Management and Budget

"We don't understand why the Democrats are so
wholeheartedly against [President Trump's border wall].
They voted for it in 2006.”
Fact-checking is a content management problem

- Media content
- Claim to be checked (text or data)
- Human actors (journalists, experts, crowd workers)
- Verification tool (query, match, source search...)
- Reference information source 1
- Reference information source 2
- Reference information source n
- Analysis result + proof « True / rather true / rather false / false »

See sources: http://dataref.com...
Fact-checking is a content management problem

Media content → Claim extraction

Claim to be checked (text or data) → Human actors (journalists, experts, crowd workers)

Media context → Social network analysis

Verification tool (query, match, source search...) → Analysis result + proof

« True / rather true / rather false / false

Source selection → Reference source construction, refinement, integration

Reference information source 1 → Reference information source 2 → ... → Reference information source n

Reconciliation, reputation

See sources: http://dataref.com... »
Fact-checking is a content management problem

- **Claim extraction**: Claim to be checked (text or data)
- **Human actors**: (journalists, experts, crowd workers)
- **Verification tool**: (query, match, source search...)
- **Analysis result + proof**: « True / rather true / rather false / false
- **Reference information sources**: n+1
- **Reference source construction, refinement, integration**

**Claim to be checked**

**Media content**

**Media context**

**Social network analysis**

**Source search / source selection**

**Reconciliation, reputation**

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)
Bobards en stock sur les plans de sauvetage européens

$2-3 bn

« 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

$10 bn

« Ça 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

$15 bn

« 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.
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
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
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.

The initial statements are mostly false.
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 $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
- The initial statements are mostly false
- The complete picture is complicated

The complete picture is complicated.
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:

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

The initial statements are mostly false.

The complete picture is complicated.

I would not have been able to give the answer.
The article goes on saying:
- The money is not given
- 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 ($16bn) is counted toward French public debt
- the remainder contribution ($12bn) is not

The initial statements are mostly false
The complete picture is complicated
I would not have been able to give the answer
Yet I vote!

IOANA MANOLESCU, INRIA AND ECOLE POLYTECHNIQUE / IIWAS, 2017
Saving Cyprus: how much does it cost?

The article goes on saying:

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

However, things are complicated:

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

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...)

Ioana Manolescu, INRIA and École Polytechnique / IIAW, 2017
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
Slowly getting there...
Back to the content management problem

Claim to be checked (text or data)

Human actors (journalists, experts, crowd workers)

Reconciliation, reputation

Verification tool (query, match, source search...)

Analysis result + proof

« True / rather true / rather false / false

See sources: http://dataref.com...

Source search / source selection

Reference information source 1

Reference information source 2

... Reference information source n

Reference source construction, refinement, integration

IOANA MANOLESCU, INRIA AND ECOLE POLYTECHNIQUE / IIWAS, 2017
Back to the content management problem

Media content → Claim extraction → Claim to be checked (text or data) → Human actors (journalists, experts, crowd workers) → Verification tool (query, match, source search...) → Reconciliation, reputation → Analysis result + proof « True / rather true / rather false / false See sources: http://dataref.com... »

Source search / source selection

Reference information source 1 → Reference information source 2 → ... → Reference information source n

Reference source construction, refinement, integration

IOANA MANOLESCU, INRIA AND ECOLE POLYTECHNIQUE / IIWAS, 2017
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)

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...


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), VLDB2016

So many data sources, so little time!

Data journalism; reference source enrichment

Data sources

- JSON sources (SOLR, Postgres...)
  - Full-text search
  - Indexing
- Relational sources (RDBMS)
  - SQL
  - Indexing
- RDF sources (Jena, Corese...)
  - SPARQL
  - Reasoning
- Streams (Twitter, RSS)
  - Subscribe
  - Query

Dataviz modules

Tatooine query engine

Custom processing
- Classification, ...

Demo

JOANA MANOLESCU, INRIA AND ECOLE POLYTECHNIQUE / IRIAS, 2017
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-4CSZD7bUl4gSPS
Merci / questions?