From Data to the Press: Data Management for Journalism and Fact-Checking

Ioana Manolescu

▶ **To cite this version:**

Ioana Manolescu. From Data to the Press: Data Management for Journalism and Fact-Checking. DATA 2020 - 9th International Conference on Data Science, Technology and Applications, Jul 2020, Paris / Virtuel, France. hal-02895316

HAL Id: hal-02895316
https://inria.hal.science/hal-02895316

Submitted on 9 Jul 2020

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.

From Data to the Press: Data Management for Journalism and Fact-Checking

Ioana Manolescu

CEDAR team, Inria Saclay and Ecole Polytechnique
http://pages.saclay.inria.fr/ioana.manolescu, @ioanamanol

DATA Conference, July 2020

MOTIVATION

Bad memories: Romania, 1989

Bad memories: Romania, 1989

Ceauşescu re-elected at the 14th congress!

Bad memories: Romania, 1989

Ceauşescu re-elected at the 14th congress!

He had held power since 1965.

Bad memories: Romania, 1989

Bad memories: Romania, 1985

1990: Things get better

... kind of

FIZICA
ASE
JOS
COMUNISMUL
1000 dead (approx.)
No one convicted.

Democratic societies crucially need the press

- To debate and express dissent

- To analyze, confim or refute public statements

Fact-checking

(Data) journalism

- 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

(Data) journalism

- To expose and explain society functioning

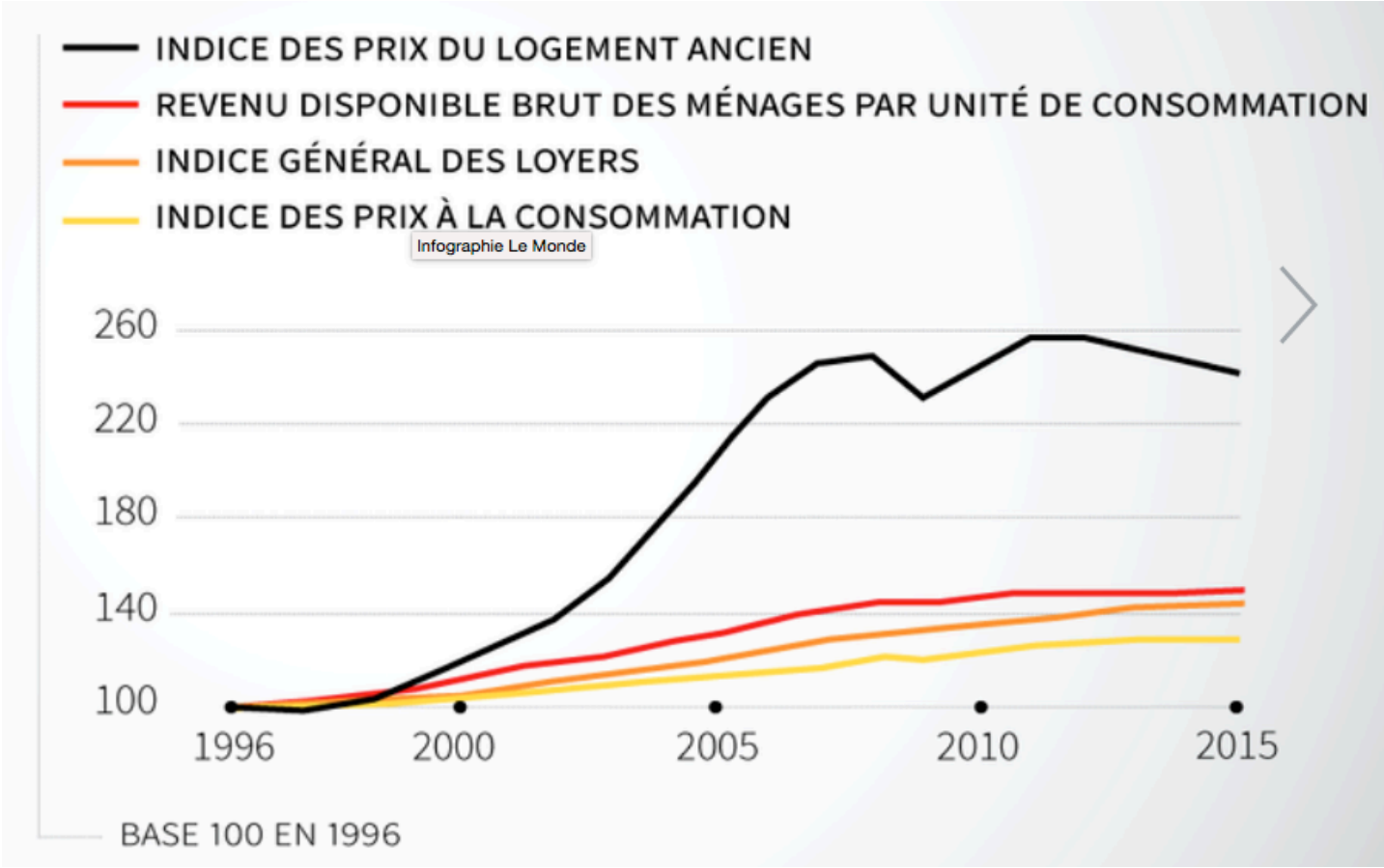
DATA JOURNALISM,

JOURNALISTIC FACT-CHECKING,

FAKE NEWS DETECTION

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

Fact-checking (since 1930 approx.)

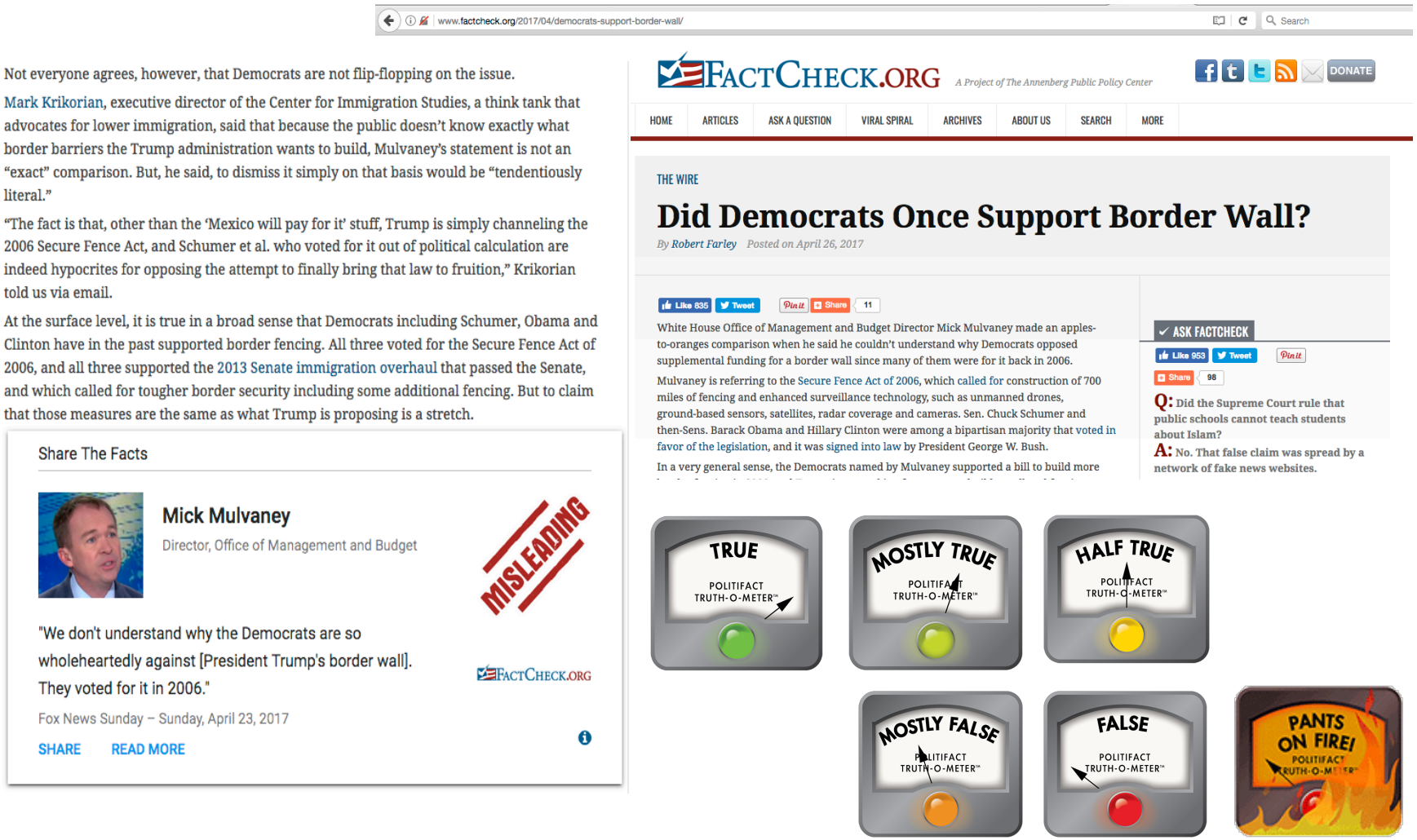
Fact-checking: verification of facts mentioned in media content

- To protect media reputation and avoid legal action

"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"

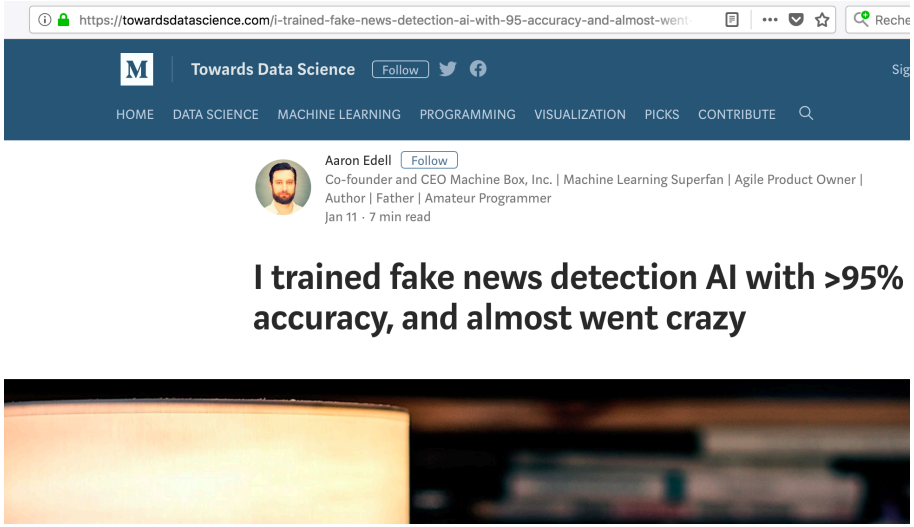
http://www.nytimes.com/2010/08/22/magazine/22FOB-medium-t.html

Fact-checking (2012 – ongoing)



Fact checking vs. fake news detection

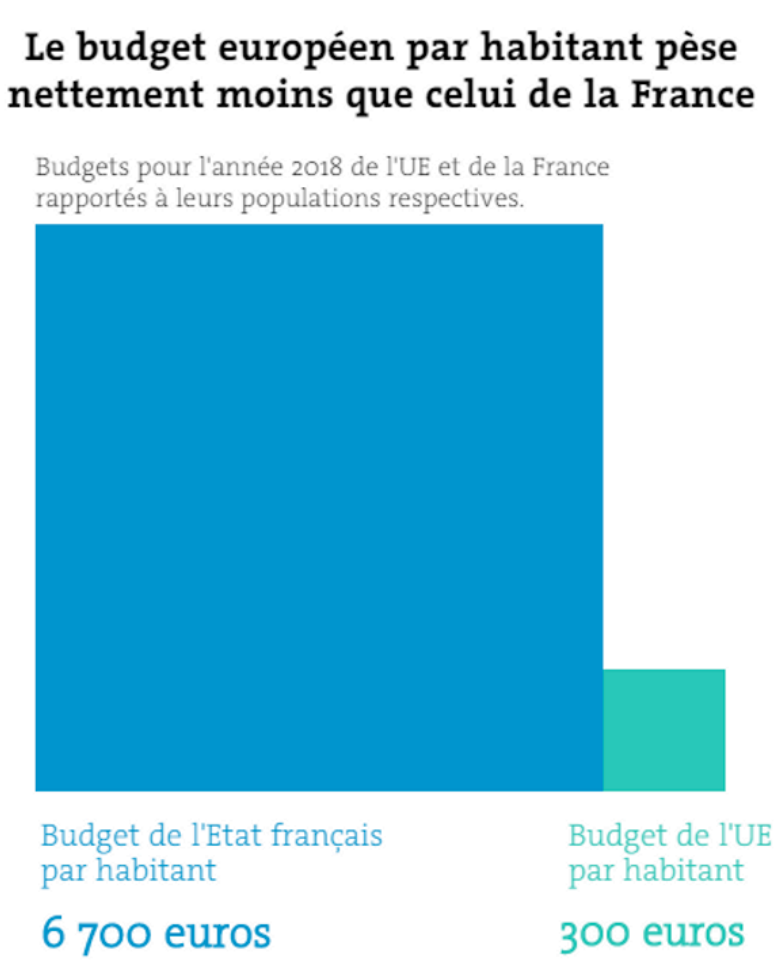
- Fact checking is based on some **background information source**
  - **Truth commonly agreed upon**
- Fake news detection may or may not use a source
  - E.g., *text classifier* (true, fake) trained on *text style*

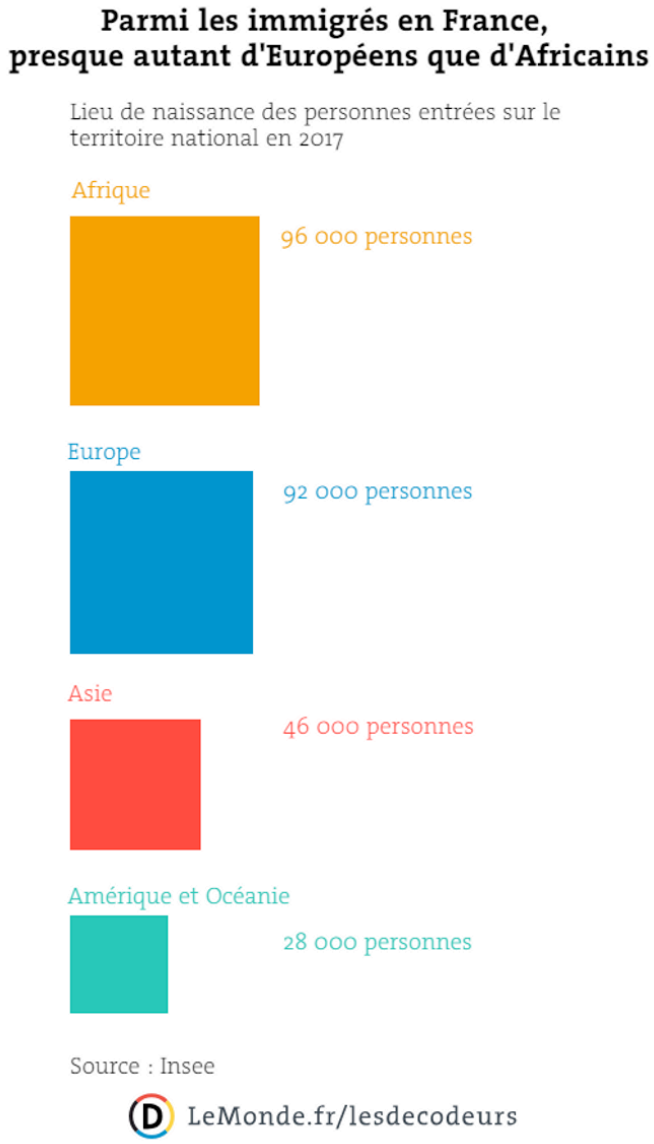


Common fact-checking scenarios

- "What is the **value** of **metric X** in **space Y** at **time T**"?
  - X=youth unemployment, Y=Germany, T=2018
  - X=illegal immigrants, Y=Italy, T=[2015-2018]
  - X=budget for research, Y=France, T=2019
- Comparison patterns
  - X1 against X2; Y1 against Y2; T1 against T2; temporal trend etc.

Common fact-checking scenarios





Common fact-checking scenarios

- "What did **X** say about **Y [at time T]**"?

- "Is **X** related **[in sense S]** to **Y**?'

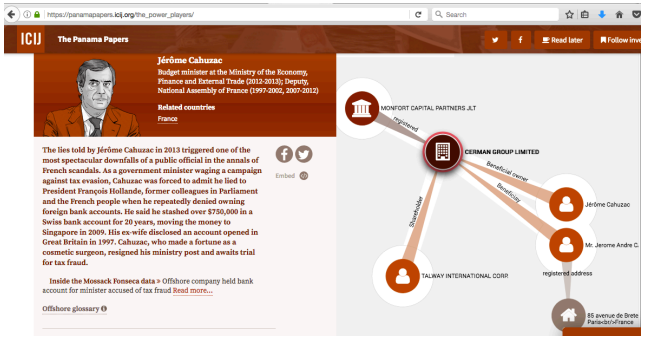
FRANCE — ENQUÊTE

Nicolas Sarkozy a bien servi les intérêts de Kadhafi. Voici les preuves

4 AVR. 2018 | PAR FABRICE ARFI ET KARL LASKE

Nicolas Sarkozy et Mouammar Kadhafi devant la maison du second bombarbée par les Américains. © Reuters

Contrairement à ce qu'il a affirmé devant les juges puis dans les médias, Nicolas Sarkozy, actuellement mis en examen pour corruption dans l'affaire des financements libyens, a objectivement servi les intérêts du régime de Kadhafi entre 2005 et 2011. La preuve en cinq actes.



Common fact-checking scenarios

- The French catholic priest beaten by thugs in 2017 that the media won't tell you about

LES DÉCODEURS

Favoris Partage

Des militants dénoncent l'omerta des médias sur l'agression d'un prêtre à Avignon… qui date de 2013

« Agression d'un prêtre lundi dernier, aucune info des médias », s'insurge un message d'un militant pro-FN, partagé des dizaines de milliers de fois sur Facebook.

Par Adrien Sénécat · Publié le 27 février 2017 à 13h01 - Mis à jour le 13 mars 2017 à 12h36

Lecture 2 min.

« Si les médias n'en parlent pas, on va s'en charger… » Corentin Clamens, un internaute qui se présente comme président du collectif France Jeunesses (proche du FN) et soutien de Marine Le Pen, a partagé, mercredi 15 février, sur Facebook, un message dénonçant l'agression d'un prêtre *« lundi dernier »* à Avignon (Vaucluse). Fait dont aucun média n'aurait parlé, selon lui. Mais, contrairement à ce que ce message laisse entendre, l'histoire, qui date de 2013, a été largement évoquée par la presse à l'époque. Explications.

CE QUE DIT LA RUMEUR

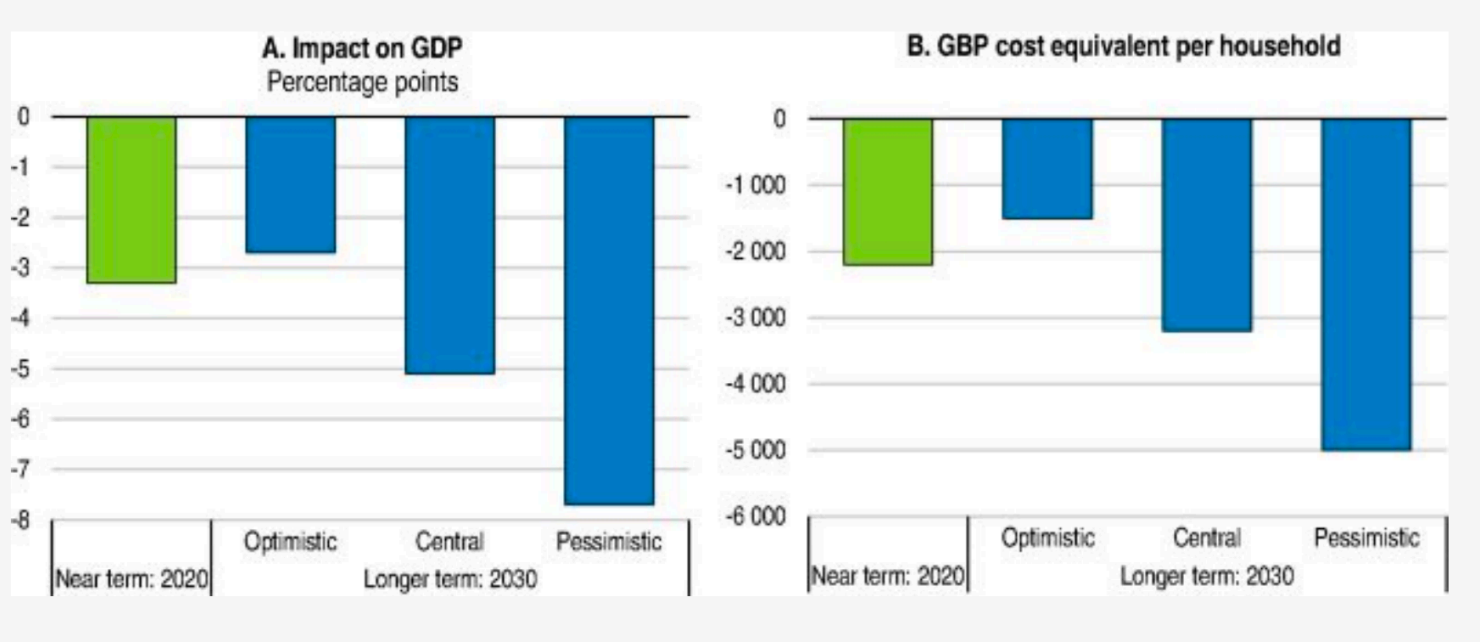
Fact-checking and/or fake news detection apply here

- The pope and the tablecloth

https://www.youtube.com/watch?v=1KFj6b1Xfe8

Where fact checking meets data journalism

- Most aspects of modern reality are **complex**
- **Explaining** can be as important and useful as checking
  - Also to analyze the future, e.g., Brexit impact:



http://www.oecd.org/economy/the-economic-consequences-of-brexit-a-taxing-decision.htm

Old-style fact-checking: Radio Yerevan jokes

Famous in the Eastern block during communism

Q1: *Is it true* that in Moscow, Mercedes cars are being given to citizens?"

A1: "*Yes, but* it is not Moscow but Leningrad, not Mercedes but Ladas, and not given to but stolen from."

Q2: *Is it true* that in Romania…

A2: Yes it is.

A CONTENT MANAGEMENT PERSPECTIVE

Lines of past and current research

1. **Model fact-checking** through a data and information management perspective

2. Identify **applicable tools and techniques**
   - In a special journalistic context (see next)

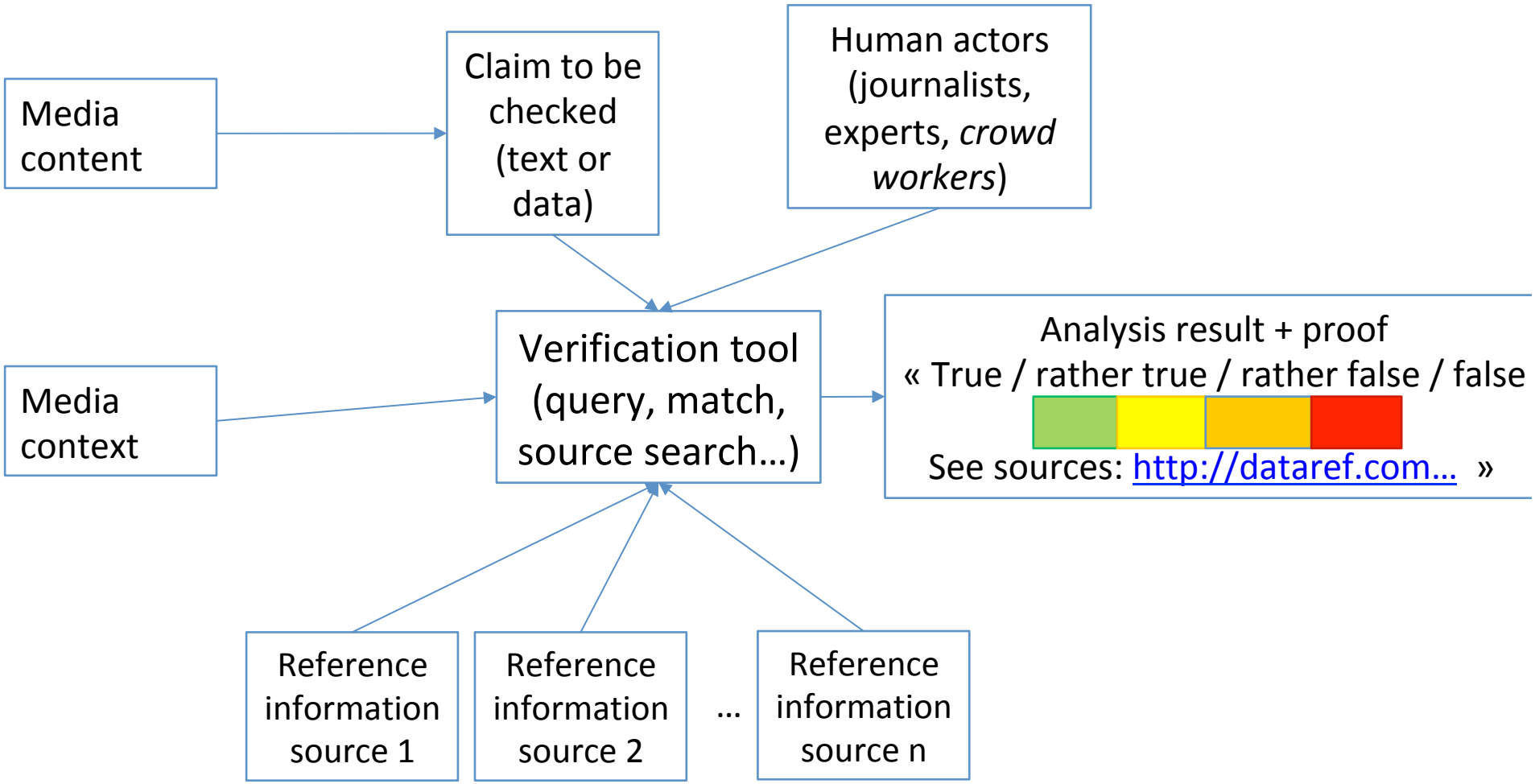
3. Devise **new models, tools and techniques** for fact-checking and data journalism problems

Past projects and collaborations

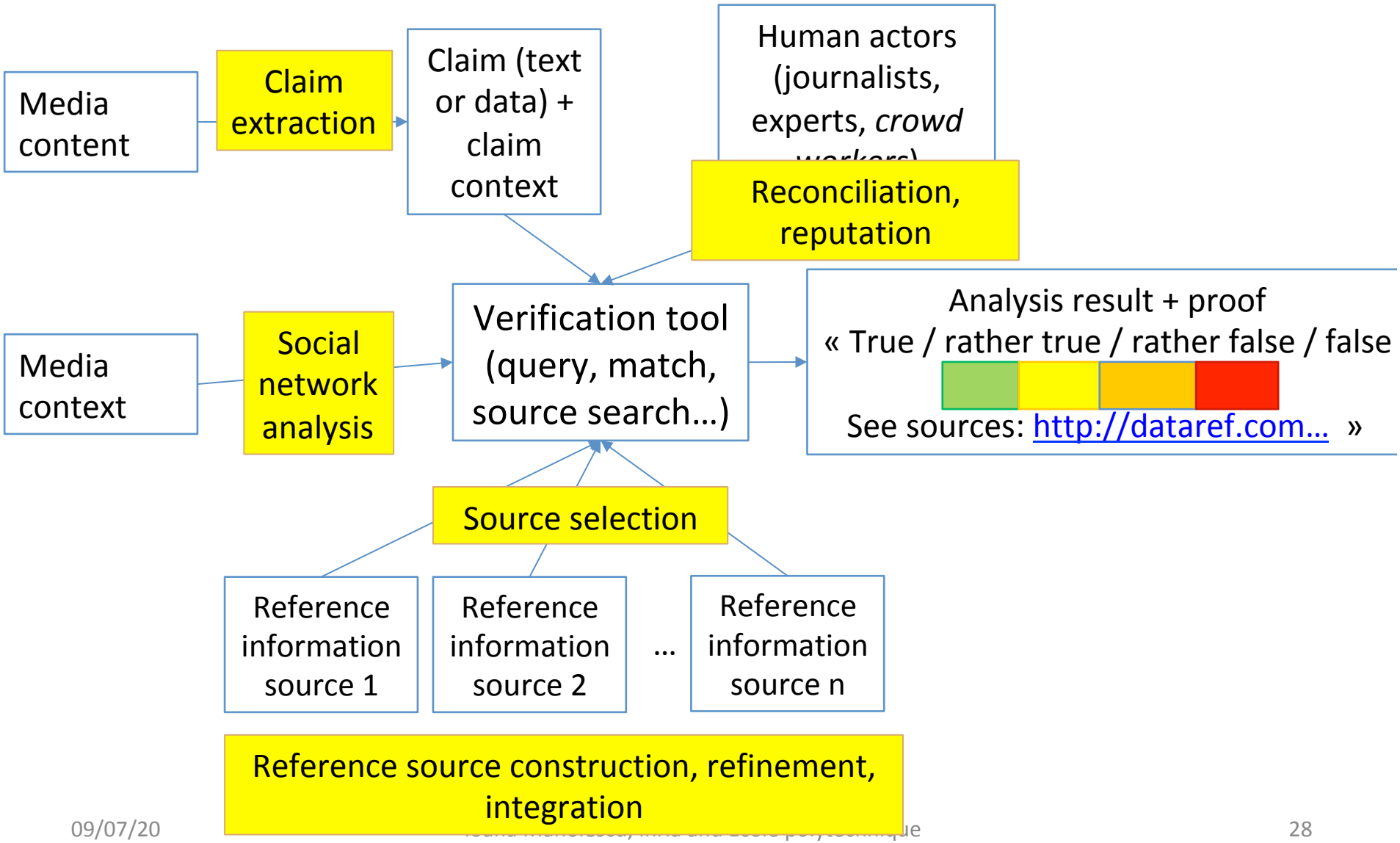
- **Google Award** (2015) with X. Tannier (LIMSI)
- **ANR ContentCheck** (2016-2019) with X. Tannier (Sorbonne Université), S. Cazalens, P. Lamarre, J.-M. Petit, M. Plantevit (U. Lyon), F. Goasdoué (U. Rennes 1), Les Décodeurs (Le Monde)

  http://contentcheck.inria.fr
- **Inria Associated Team WebClaimExplain** with AIST Japan (Julien Leblay)
- **Collaborations** with H. Galhardas (University of Lisbon), former PhD S. Zampetakis and others

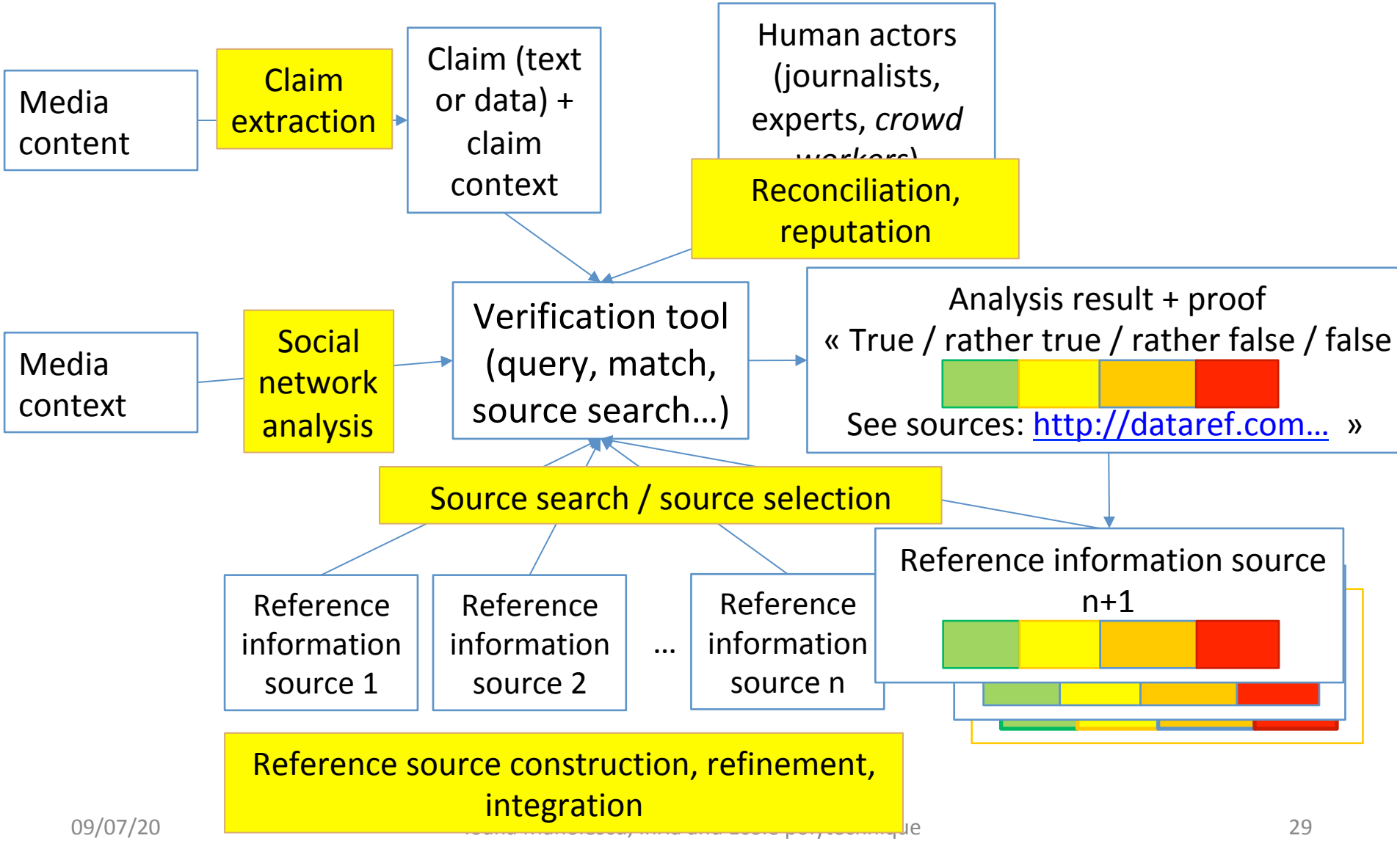
Fact-checking as a content management problem



Fact-checking as a content management problem

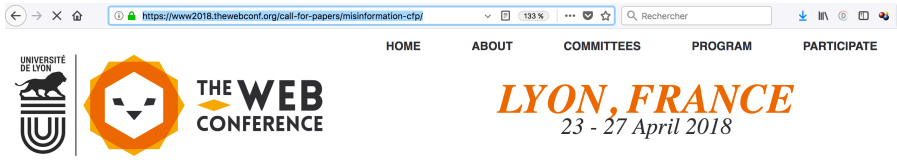


Fact-checking as a content management problem



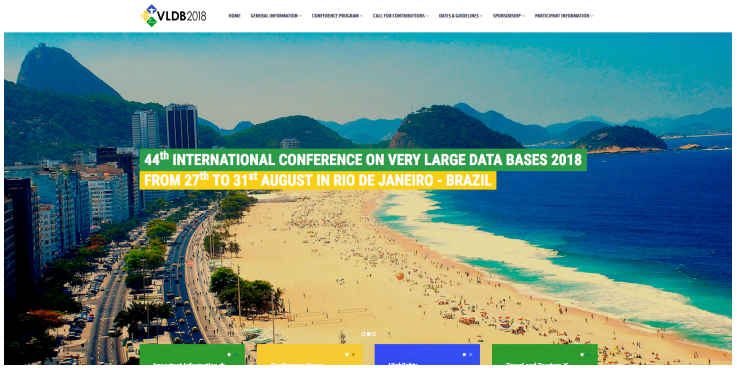
Fact-checking as a content management problem

[WWW2018] "A Content Management Perspective on Fact-Checking", S. Cazalens, J. Leblay, I. Manolescu, X. Tannier (fact-checking track)



[WWW2018 tutorial] "Computational fact-checking: problems, state of the art, and perspectives", J. Leblay, I. Manolescu, X. Tannier

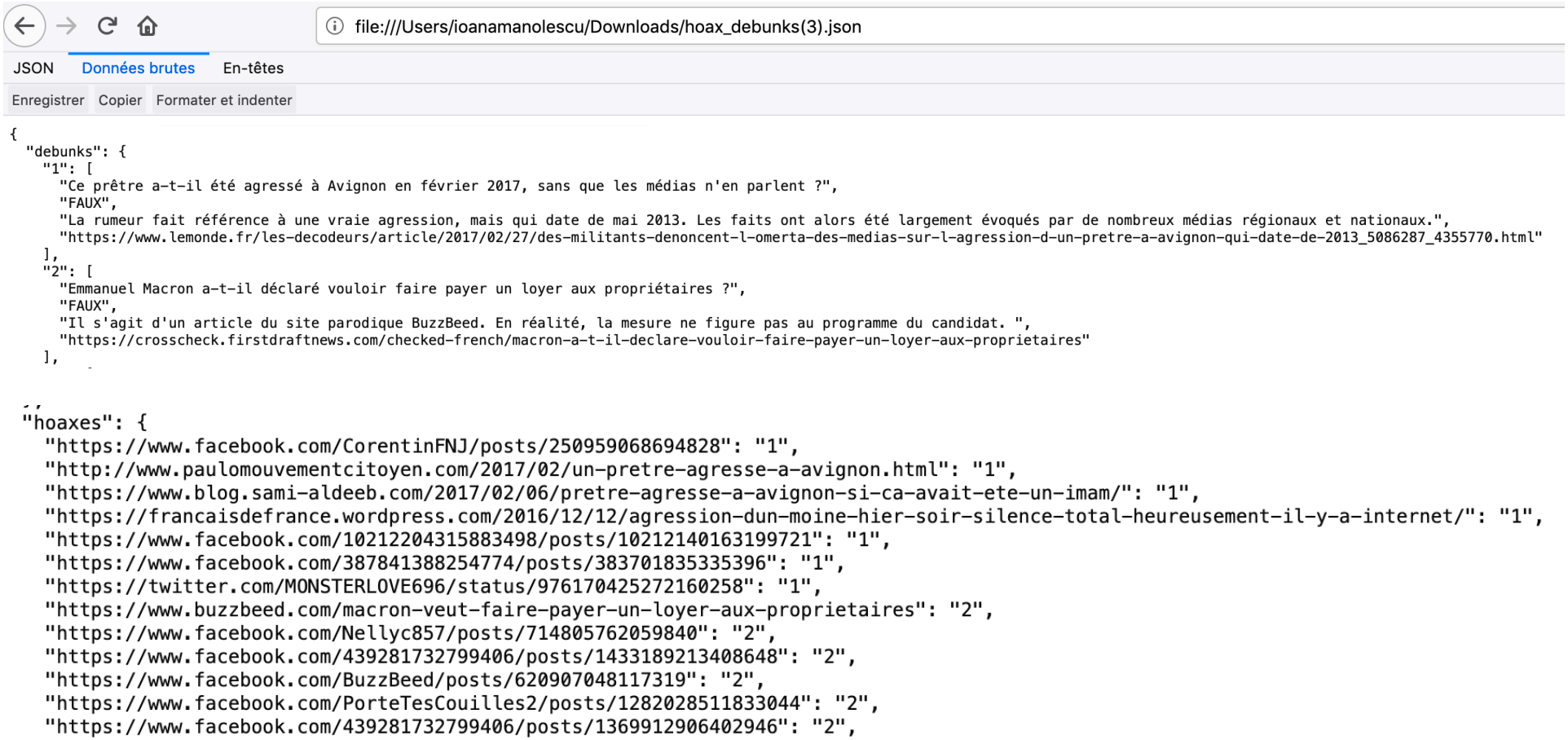
[VLDB2018 tutorial] "Computational fact-checking: a content management perspective", S. Cazalens, J. Leblay, P. Lamarre, I. Manolescu, X. Tannier



http://contentcheck.inria.fr/

Fact-checks shared as open data!

http://s1.lemde.fr/mmpub/data/decodex/hoax/hoax_debunks.json

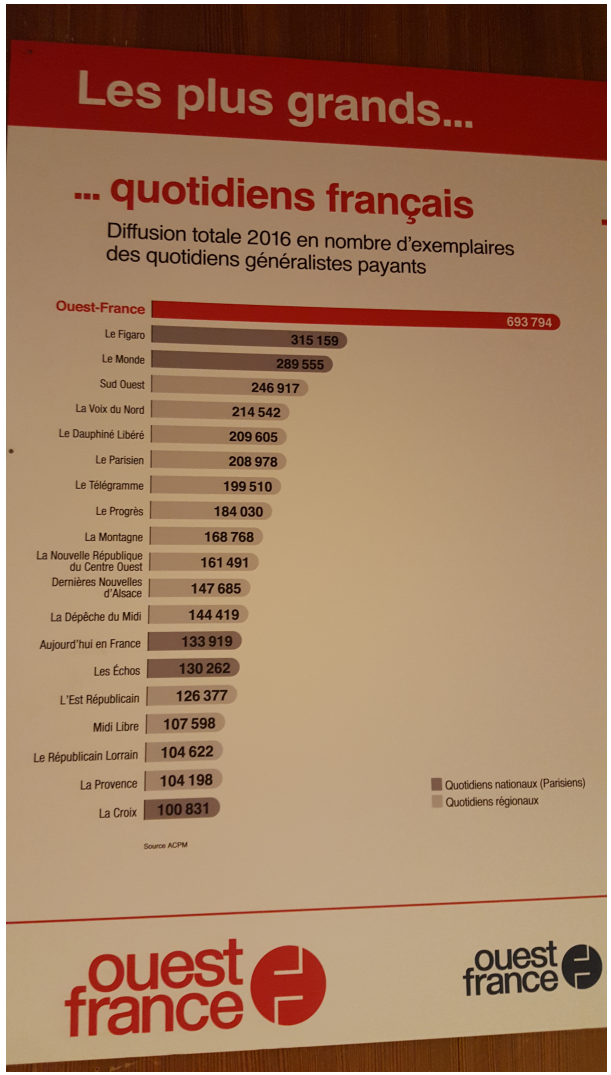



```
{
  "debunks": {
    "1": [
      "Ce prêtre a-t-il été agressé à Avignon en février 2017, sans que les médias n'en parlent ?",
      "FAUX",
      "La rumeur fait référence à une vraie agression, mais qui date de mai 2013. Les faits ont alors été largement évoqués par de nombreux médias régionaux et nationaux.",
      "https://www.lemonde.fr/les-decodeurs/article/2017/02/27/des-militants-denoncent-l-omerta-des-medias-sur-l-agression-d-un-pretre-a-avignon-qui-date-de-2013_5086287_4355770.html"
    ],
    "2": [
      "Emmanuel Macron a-t-il déclaré vouloir faire payer un loyer aux propriétaires ?",
      "FAUX",
      "Il s'agit d'un article du site parodique BuzzBeed. En réalité, la mesure ne figure pas au programme du candidat. ",
      "https://crosscheck.firstdraftnews.com/checked-french/macron-a-t-il-declare-vouloir-faire-payer-un-loyer-aux-proprietaires"
    ],

  },
  "hoaxes": {
    "https://www.facebook.com/CorentinFNJ/posts/250959068694828": "1",
    "http://www.paulomouvementcitoyen.com/2017/02/un-pretre-agresse-a-avignon.html": "1",
    "https://www.blog.sami-aldeeb.com/2017/02/06/pretre-agresse-a-avignon-si-ca-avait-ete-un-imam/": "1",
    "https://francaisdefrance.wordpress.com/2016/12/12/agression-dun-moine-hier-soir-silence-total-heureusement-il-y-a-internet/": "1",
    "https://www.facebook.com/10212204315883498/posts/10212140163199721": "1",
    "https://www.facebook.com/387841388254774/posts/383701835335396": "1",
    "https://twitter.com/MONSTERLOVE696/status/976170425272160258": "1",
    "https://www.buzzbeed.com/macron-veut-faire-payer-un-loyer-aux-proprietaires": "2",
    "https://www.facebook.com/Nellyc857/posts/714805762059840": "2",
    "https://www.facebook.com/439281732799406/posts/1433189213408648": "2",
    "https://www.facebook.com/BuzzBeed/posts/620907048117319": "2",
    "https://www.facebook.com/PorteTesCouilles2/posts/1282028511833044": "2",
    "https://www.facebook.com/439281732799406/posts/1369912906402946": "2",
```

Which data sources for fact-checking and data journalism ?

- Journalists do not, historically, build databases.
  - "Writers, not techies"
  - "Not part of our job"
  - Persisting data is novel to some
  - Journal information systems not always helpful
- However, some journals (e.g., OuestFrance) have valuable (reference) databases



Which data sources for fact-checking and data journalism ?

- **Curse of the coverage**: they need to cover (almost) any topic
- **Curse of noteworthiness**: write about hot topics of today (or tomorrow)
- They work under **strong time pressure**
- Good journalists are very **picky** with their data sources

Which data sources for fact-checking and data journalism ?

- Good journalists are very **picky** with their data sources...
- Which are **heterogeneous**: HTML, JSON, Excel, XML, CSV etc.
- Journalists won't write queries.
  They may not know what they're looking for
- The fact-checking process and result must be **explainable**
  - Not (only) "an ML algorithm said so"

Improving access to reference data sources

[WebDB2018, SBD2017] T. D. Cao, I. Manolescu, X. Tannier. "Extracting Linked Data from statistic spreadsheets"

"Créations d'entreprises en France en 2015" → return:

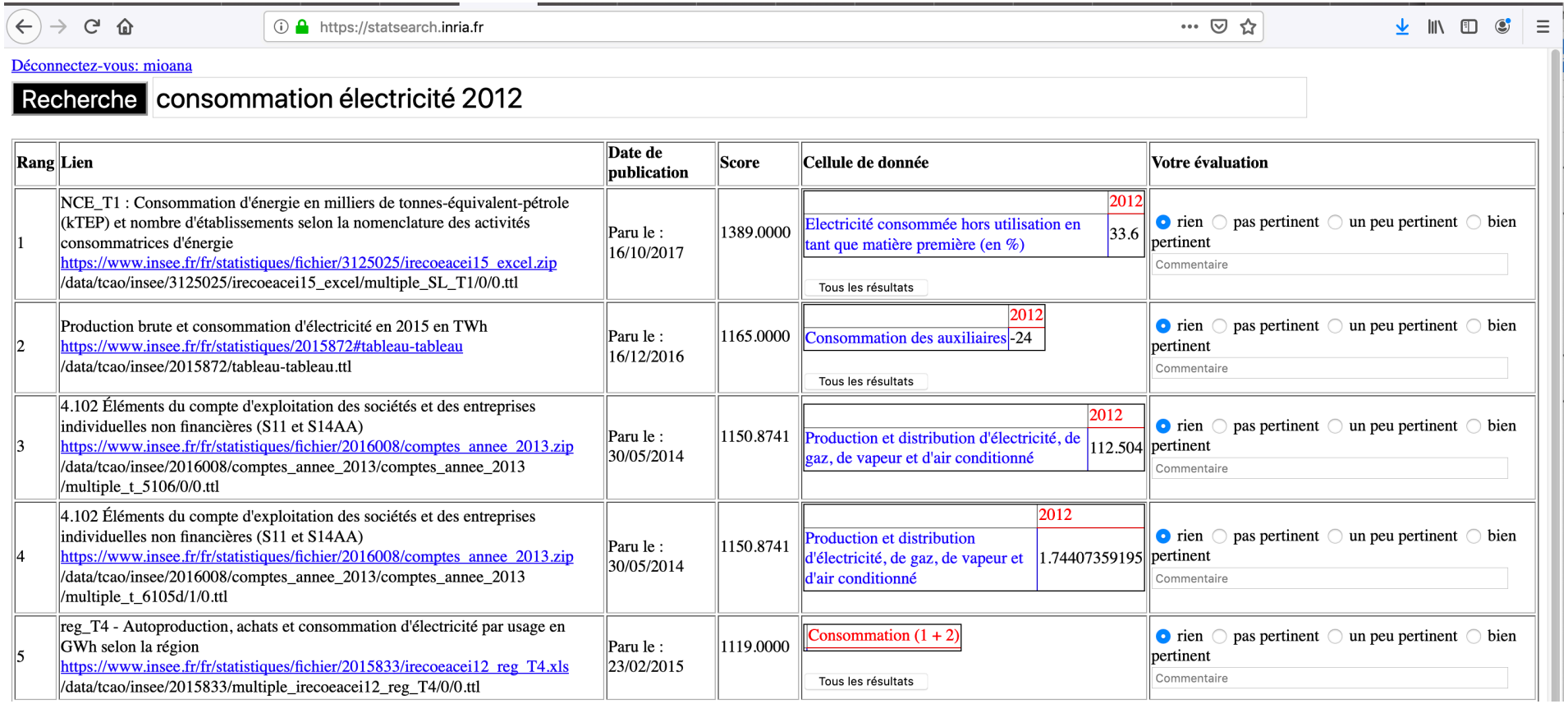
Créations d'entreprises dans quelques pays de l'Union européenne en 2015

en %

Pays	Taux de création
Allemagne	7,1
Belgique	6,2
Espagne	9,5
France (1)	**9,5**
Italie	7,5
Pays-Bas	10,1
Portugal	15,7
République tchèque	8,2
Royaume-Uni	14,3

Improving access to reference data sources

[WebDB2018, SBD2017] T. D. Cao, I. Manolescu, X. Tannier. "Extracting Linked Data from statistic spreadsheets"



https://statsearch.inria.fr

Déconnectez-vous: mioana

Recherche consommation électricité 2012

Rang	Lien	Date de publication	Score	Cellule de donnée	Votre évaluation
1	NCE_T1 : Consommation d'énergie en milliers de tonnes-équivalent-pétrole (kTEP) et nombre d'établissements selon la nomenclature des activités consommatrices d'énergie https://www.insee.fr/fr/statistiques/fichier/3125025/irecoeacei15_excel.zip /data/tcao/insee/3125025/irecoeacei15_excel/multiple_SL_T1/0/0.ttl	Paru le : 16/10/2017	1389.0000	2012 Electricité consommée hors utilisation en tant que matière première (en %) 33.6 Tous les résultats	rien pas pertinent un peu pertinent bien pertinent Commentaire
2	Production brute et consommation d'électricité en 2015 en TWh https://www.insee.fr/fr/statistiques/2015872#tableau-tableau /data/tcao/insee/2015872/tableau-tableau.ttl	Paru le : 16/12/2016	1165.0000	2012 Consommation des auxiliaires -24 Tous les résultats	rien pas pertinent un peu pertinent bien pertinent Commentaire
3	4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_5106/0/0.ttl	Paru le : 30/05/2014	1150.8741	2012 Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 112.504	rien pas pertinent un peu pertinent bien pertinent Commentaire
4	4.102 Éléments du compte d'exploitation des sociétés et des entreprises individuelles non financières (S11 et S14AA) https://www.insee.fr/fr/statistiques/fichier/2016008/comptes_annee_2013.zip /data/tcao/insee/2016008/comptes_annee_2013/comptes_annee_2013 /multiple_t_6105d/1/0.ttl	Paru le : 30/05/2014	1150.8741	2012 Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné 1.74407359195	rien pas pertinent un peu pertinent bien pertinent Commentaire
5	reg_T4 - Autoproduction, achats et consommation d'électricité par usage en GWh selon la région https://www.insee.fr/fr/statistiques/fichier/2015833/irecoeacei12_reg_T4.xls /data/tcao/insee/2015833/multiple_irecoeacei12_reg_T4/0/0.ttl	Paru le : 23/02/2015	1119.0000	Consommation (1 + 2) Tous les résultats	rien pas pertinent un peu pertinent bien pertinent Commentaire

Recognizing statistic mentions in tweets [NLDB2019]

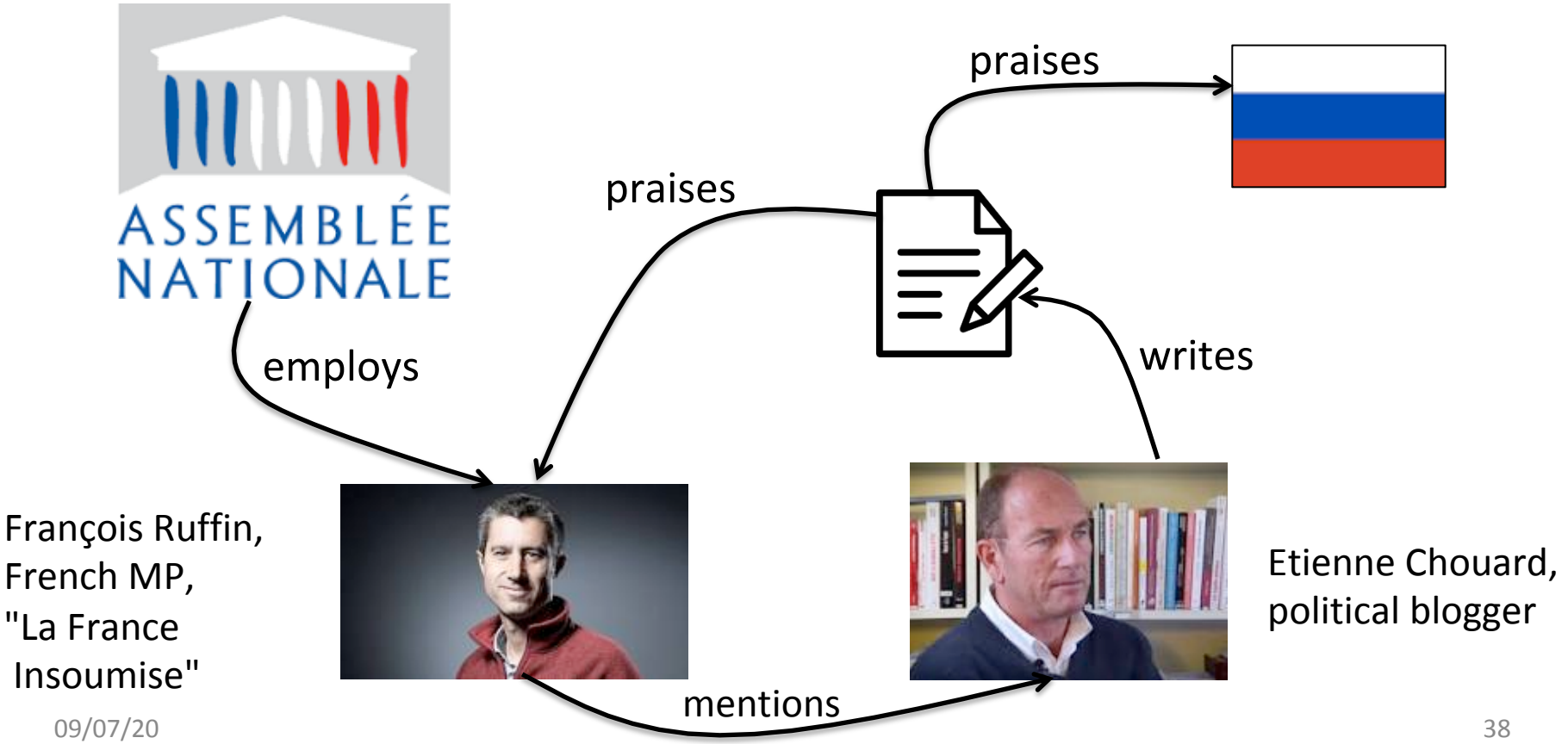
2020-03-05
En Marche
https://twitter.com/enmarchefr/status/1235501467907260416
L'index de l'égalité professionnelle change concrètement la vie des 4,5 millions de femmes en entreprise. En un an, la part des entreprises qui accordent une augmentation de salaire à la sortie du congé maternité est passée de 67% à 90%. Muriel Pénicaud Egalite F H
En un an , la part de les entreprises qui accordent une augmentation de salaire à la sortie du congé maternité est passée de 67 % à 90 % .
3/11
part de les entreprises qui accordent une augmentation de salaire à la sortie du congé materni | Recherche

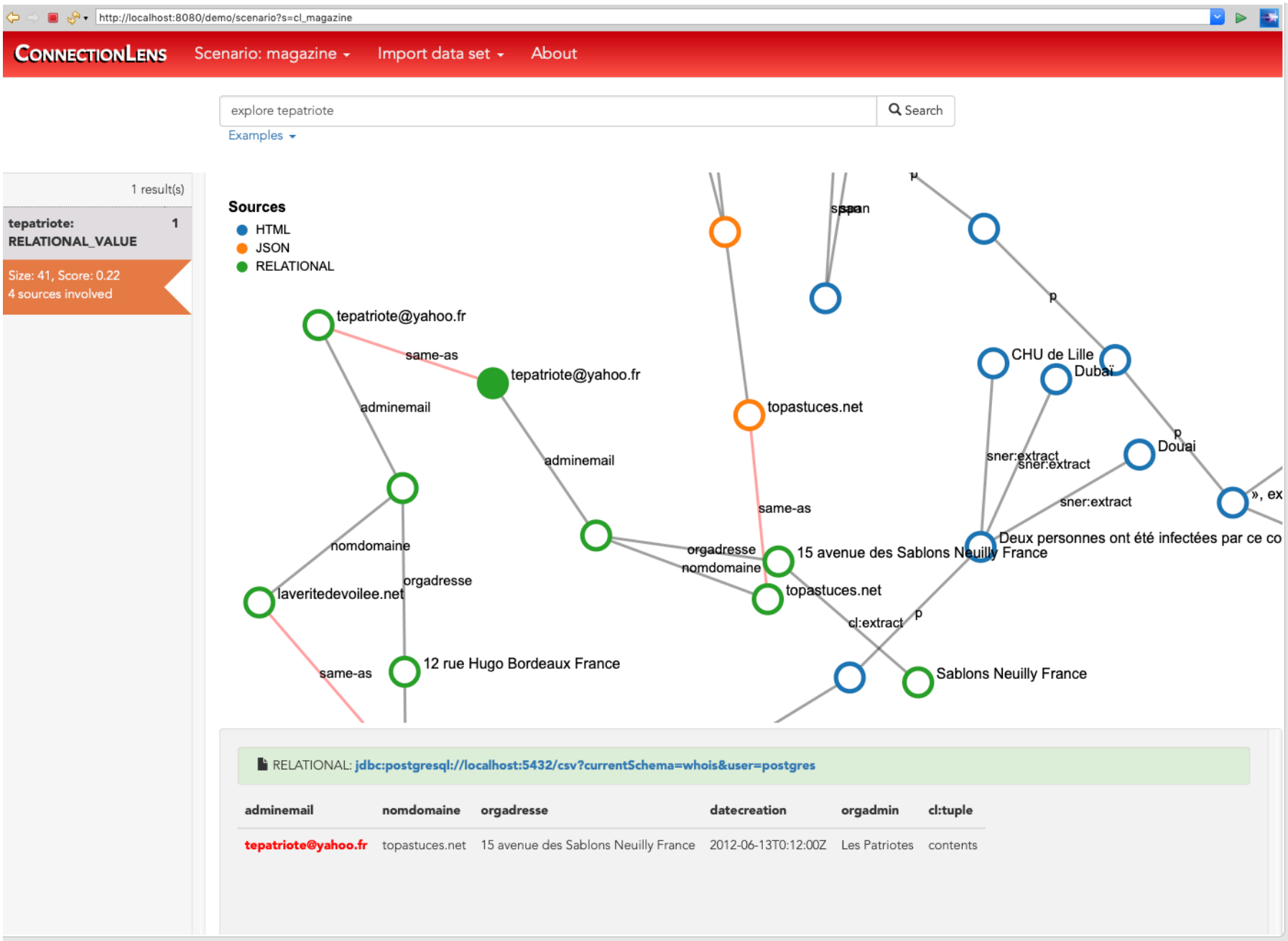
2020-03-05
Bruno Bonnell
https://twitter.com/BrunoBonnellOff/status/1235500886669045761
Moins de 5% des voies françaises portent un nom de femmes! Une inégalité facile à rattraper. En baptisant les rues de noms de modèles féminins, nous encouragerons plus de jeunes filles à oser. Mvt Radical En Marche …
Moins de 5 % de les voies françaises portent un nom de femmes !
4/13
femmes 2020 | Recherche

Automatically recognizes numerical mentions involving entities known from the INSEE corpus
Proposes keyword query to send to the previous statistics search algorithm.

ConnectionLens: interconnecting heterogeneous data sources **[VLDB2018]**

"What are the connections between elected members of the French National Assembly, and Russia?"




http://localhost:8080/demo/scenario?s=cl_magazine
ConnectionLens
Scenario: magazine
Import data set
About
explore tepatriote
Search
Examples
1 result(s)
tepatriote: 1
RELATIONAL_VALUE
Size: 41, Score: 0.22
4 sources involved
Sources
HTML
JSON
RELATIONAL
tepatriote@yahoo.fr
same-as
tepatriote@yahoo.fr
adminemail
adminemail
topastuces.net
same-as
nomdomaine
orgadresse
laveritedevoilee.net
same-as
12 rue Hugo Bordeaux France
orgadresse
nomdomaine
15 avenue des Sablons Neuilly France
topastuces.net
cl:extract
Sablons Neuilly France
CHU de Lille
Dubaï
Douai
sner:extract
sner:extract
sner:extract
Deux personnes ont été infectées par ce co
RELATIONAL: jdbc:postgresql://localhost:5432/csv?currentSchema=whois&user=postgres
adminemail
nomdomaine
orgadresse
datecreation
orgadmin
cl:tuple
tepatriote@yahoo.fr
topastuces.net
15 avenue des Sablons Neuilly France
2012-06-13T0:12:00Z
Les Patriotes
contents

A data model for facts, statement and beliefs

Modeling who said what when

POLITICS January 19, 2018 7:19 pm Updated: January 19, 2018 7:24 pm

10 times that Trump has contradicted himself in his first year in office

By Darlene Superville The Associated Press

Comments 8 | Facebook 398 | Twitter | Email | Print | ...

https://globalnews.ca/news/3976740/trump-contradicted-first-year/

[MisInfo@WWW2019, CIKM2019]
L. Duroyon, F. Goasdoué, I. Manolescu
"A Linked Data Model for Facts, Statements and Beliefs"

- Fact **F** (holds according to the database)
- Statement of **X** about **F**
  - It's only according to **X**
- Statement of **Y** about statement of **X** about **F**
  - Only according to **Y**

Facts: application-dependent

Statements: writes, says, declares, ...

A data model for facts, statement and beliefs

Modeling who said what when

POLITICS January 19, 2018 7:19 pm Updated: January 19, 2018 7:24 pm

10 times that Trump has contradicted himself in his first year in office

By Darlene Superville The Associated Press

Comments 8 | Facebook 398 | Twitter | Email | Print | ...

https://globalnews.ca/news/3976740/trump-contradicted-first-year/

[MisInfo@WWW2019, CIKM2019]
L. Duroyon, F. Goasdoué, I. Manolescu
"A Linked Data Model for Facts, Statements and Beliefs"

- ❑ Fact **F**
- ❑ Statement of **X** about **F**
- ❑ Statement of **Y** about statement of **X** about **F**

Also:

- ❑ **Time** (for facts and statements)
- ❑ **Propagation of information** through communication →
- ❑ Who **has heard of** what when

Lessons learned

- Work with **the right data**
  - The trusted data
- Work with the data **as it is**
  - Heterogeneous
  - Evolving usage and requirements prevent schema design, consolidation
  - **Extract**, **structure**, **connect**
- **Simplify** use
  - Keyword queries, canned queries

PERSPECTIVES

Google FactCheck tools

Google Fact Check Tools

Explorer

Markup Tool

APIs

Fact Check Explorer

list:recent

French

Claim by Sources multiples:

La photo d'un homme qui urine dans le métro parisien

AFP Factuel rating: Faux

Ce cliché d'un homme qui urine a été pris dans le métro de New York

1 day ago

AFP Factuel

Claim by الحملة الوطنية لتأميم ثروات البلاد:

L'ambassadeur de France en Tunisie évoque la possibilité d'amener des militaires français dans le sud de la Tunisie

Factuel AFP rating: Faux

Non, l'ambassadeur de France en Tunisie n'a pas évoqué la possibilité de recourir à l'armée française dans le sud du pays

1 day ago

Tunisie

Ambassade de France / سفارة فرنسا

France

Factuel AFP

Twitter approach to limiting false information spreading

Yoel Roth
@yoyoel
Head of Site Integrity

Nick Pickles
@nickpickles
Director, Global Public Policy Strategy & Development

Product

Updating our Approach to Misleading Information

By Yoel Roth and Nick Pickles
Monday, 11 May 2020

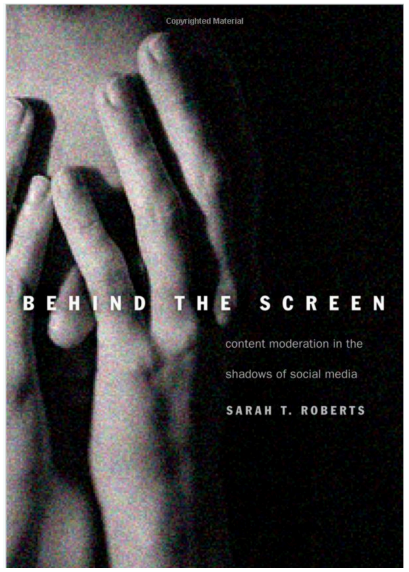
In serving the public conversation, our goal is to make it easy to find credible information on Twitter and to limit the spread of potentially harmful and misleading content. Starting today, we're introducing new labels and warning messages that will provide additional context and information on some Tweets containing disputed or misleading information related to COVID-19.

In March, we broadened our policy guidance to address content that goes directly against guidance on COVID-19 from authoritative sources of global and local public health information. Moving forward, we may use these labels and warning messages to provide additional explanations or clarifications in situations where the risks of harm associated with a Tweet are less severe but where people may still be confused or misled by the content. This will make it easier to find facts and make informed decisions about what people see on Twitter.

New labels and warnings

ML-based fake news detection

- Good use case for learning based on labeled examples
- Problems:
  - Manual effort required by labeling
  - Scale
  - Fake news detection speed
- All these problems are solved (or easy to solve) by Web giants who already employ tens of thousands of *commercial content moderators*



CS research for fact-checking

- DB, KR, IR, NLP
- In its most general statement, fact-checking supposes perfect NLP → study **sub-problems**!
- In fact-checking journalism, **human writers** chose topics, angle, style...
  - "A story wrapped around a query"
- Vision: build "**perfect data machines**" and give them to talented writers

A vision of journalistic dataspaces

- "Dataspaces": Franklin, Halevy and Maier, SIGMOD Record 2005
- Ingest **data of any nature**: structured (relational), semistructured (JSON, XML, (social) graphs), unstructured (text), KB...
- Storage, indexing
- Search across the data sources

Requirements for journalistic dataspaces

- Time
  - Of data acquisition
  - Of events described in the data
- Provenance
  - Authorship metadata
  - Annotation by users
  - Access control based on provenance and annotations
- Ability to "derive" content (à la views)
- Semantic annotation and classification
- Social connections analysis
- Friendly interfaces
- Scalability

A vision of journalistic dataspaces

ANR AI Chair "SourcesSay": Intelligent Analysis and Interconnexion of Heterogeneous Data in Digital Arenas (2020-2024) (co-financed by DGA)

- With support from **Le Monde** and **WeDoData**
- https://project.inria.fr/sourcessay/

PhD funding available!

- 1 or 2 PhDs to start in 2020 on SourcesSay
- 1 PhD funded by Région Île-de-France in collaboration with WeDoData