Data accessibility and exploration

Ioana Manolescu

Inria and Institut Polytechnique de Paris

htttps://pages.saclay.inria.fr/ioana.manolescu,
@ioanamanol

Who we are

Computer Science Researchers (research staff, faculty, engineer)

- Employed by **public organizations** (Inria, FR national research institute in CS, and Ecole Polytechnique, FR engineering school)
- Many students (M1, M2, PhD)

Research in **data management**

Since 2013, inspired by **data journalism** and **computational fact-checking**

ACM SIGMOD 2013

Fact Checking and Analyzing the Web

François Goasdoué[1] Konstantinos Karanasos[2]* Yannis Katsis[3]
Julien Leblay[1] Ioana Manolescu[1] Stamatis Zampetakis[1]

[1]OAK team, Inria Saclay & LRI Université Paris-Sud Orsay, France

[2]IBM Almaden Research Center San Jose, CA

[3]UCSD Database group & WebDam project, Inria Saclay San Diego, CA

firstname.lastname@inria.fr

ABSTRACT

Fact checking and *data journalism* are currently strong trends. The sheer amount of data at hand makes it difficult even for trained professionals to spot biased, outdated or simply incorrect information. We propose to demonstrate FactMinder, a fact checking and analysis assistance application. SIGMOD attendees will be able to analyze documents using FactMinder and experience how background knowledge and open data repositories help build insightful overviews of current topics.

of traffic on asthma cases) can comb the Web for bits of information, connect, interpret, annotate and re-share them. Such data gathering and fact checking have come at the core of "data journalism", pioneered, e.g., in Europe by The Guardian[2] and growing through efforts such as FactCheck[3], Politifact[4], and similar French sites[5].

Fact checking and analysis (*FCA*, for short), viewed as the process of analyzing a piece of information, crossing it with existing knowledge, verifying its accuracy and possibly enriching it with nuances, comments and connections to reputable sources, has an inherent part of human effort, thus

Our collaboration with Le Monde

Google Award in Computational Journalism (2014-2015), with U. Paris Sud, mostly data viz

ANR ContentCheck: Models, Algorithms and Tools for Data Journalism and Journalistic Fact-Checking (2015-2020), https://contentcheck.inria.fr
700 K€, w/ U. Paris Sud, U. Rennes, U. Lyon, and Les Décodeurs (fact-checking team from Le Monde):
Samuel Laurent, Maxime Ferrer, Adrien Sénécat

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

LES DÉCODEURS
VENONS-EN AUX FAITS

AI Chair SourcesSay: Intelligent Data Analysis and Interconnection in Digital Arenas (2020-2024), https://sourcessay.inria.fr
600 K€; Le Monde (Stéphane Horel) and WeDoData (Karine Bastien) as non-funded, supporting partners

Improving access to digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text

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

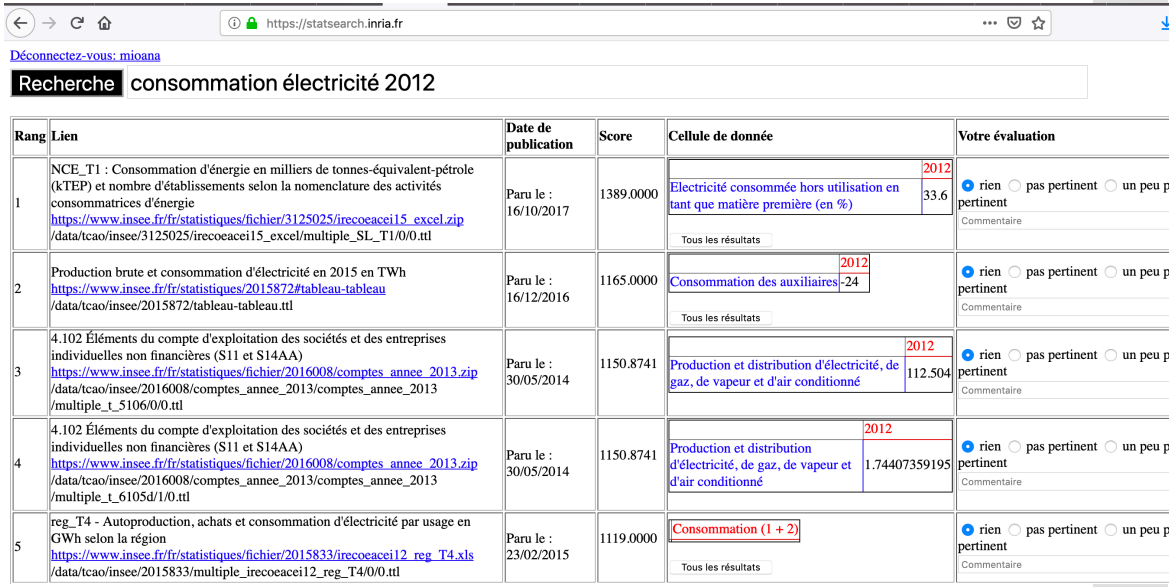
en %

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 digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text



https://statsearch.inria.fr

Déconnectez-vous: mioana

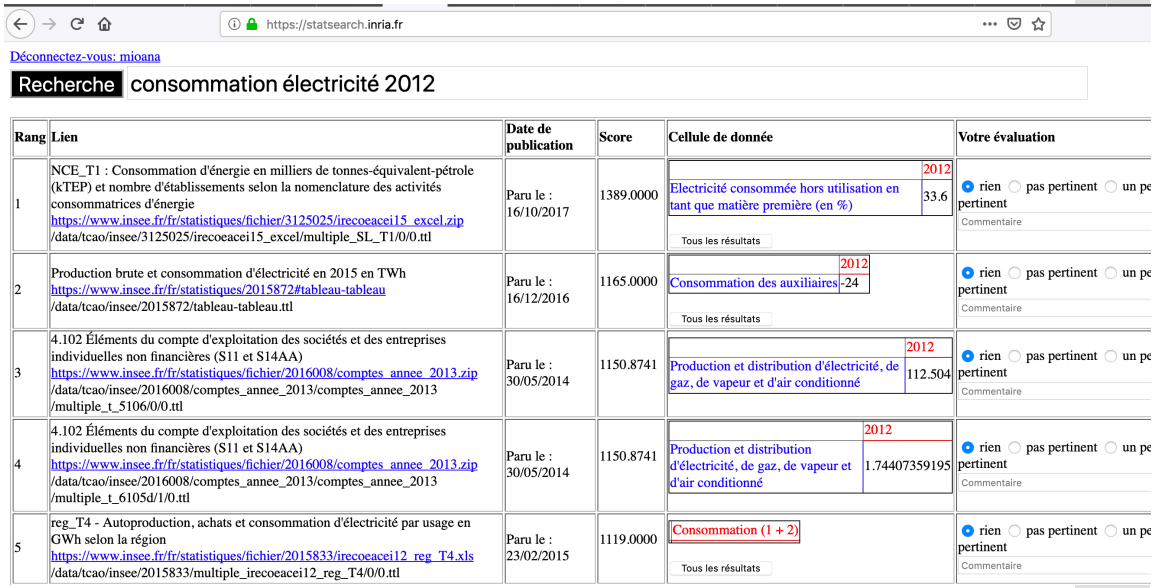
Recherche: consommation électricité 2012

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 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 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 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 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 Commentaire

Improving access to digital content

PhD of Tien-Duc Cao (2019):

1. **Crawl** all INSEE reports, turn into Linked Open Data
2. Tailored **search algorithm**, returning *cells* or *regions* + original page
3. **Statistic claim extraction** from text



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 pe 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 pe 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 pe 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 pe 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 pe pertinent — Commentaire

Follow-up: question answering on SDMX databases

Finding Interesting Aggregates in RDF Graphs

How to extract valuable insights from a Linked OpenData (RDF) graph?

RDF: Web standard for sharing open data

✔ Very flexible, no constraints on graph structure

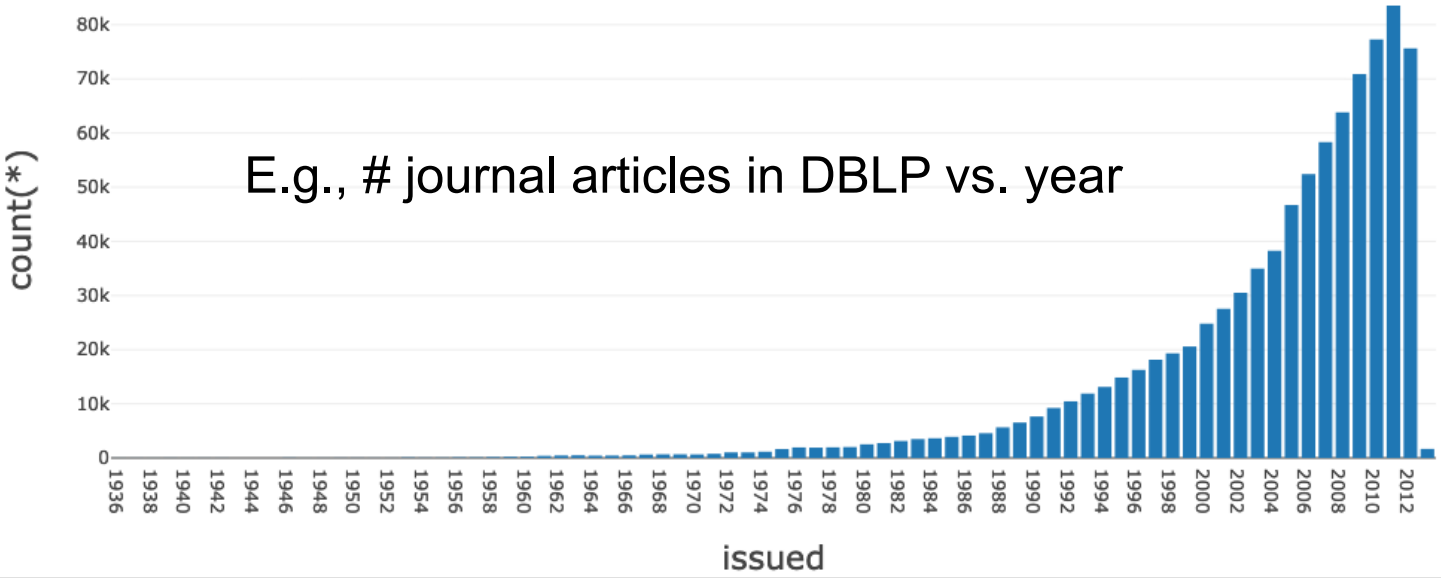
✗ Lack of schema: hard to get to know!

Graphs contain many values (e.g., numerical).
What story do they tell?

Finding Interesting Aggregates in RDF Graphs

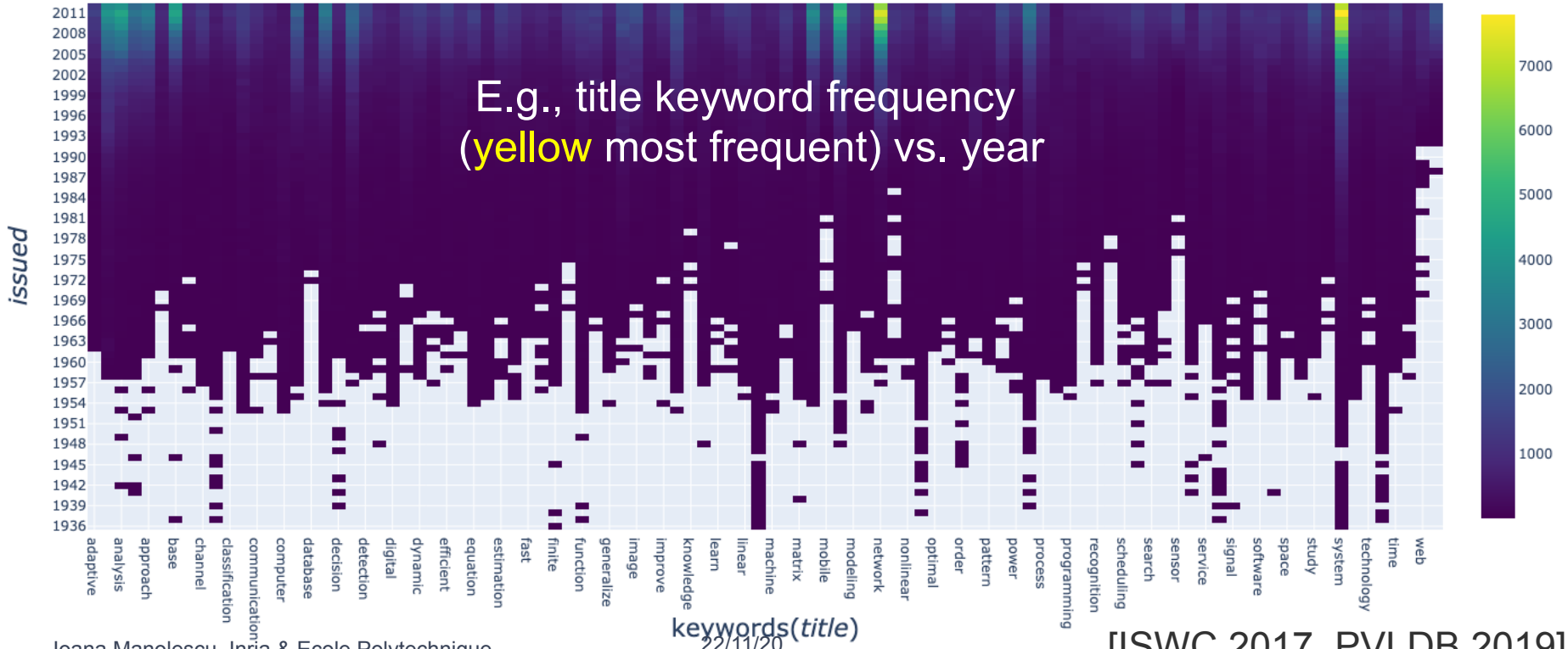
How to extract valuable insights from a Linked OpenData graph?

Automatically identify interesting aggregates to show



Finding Interesting Aggregates in RDF Graphs

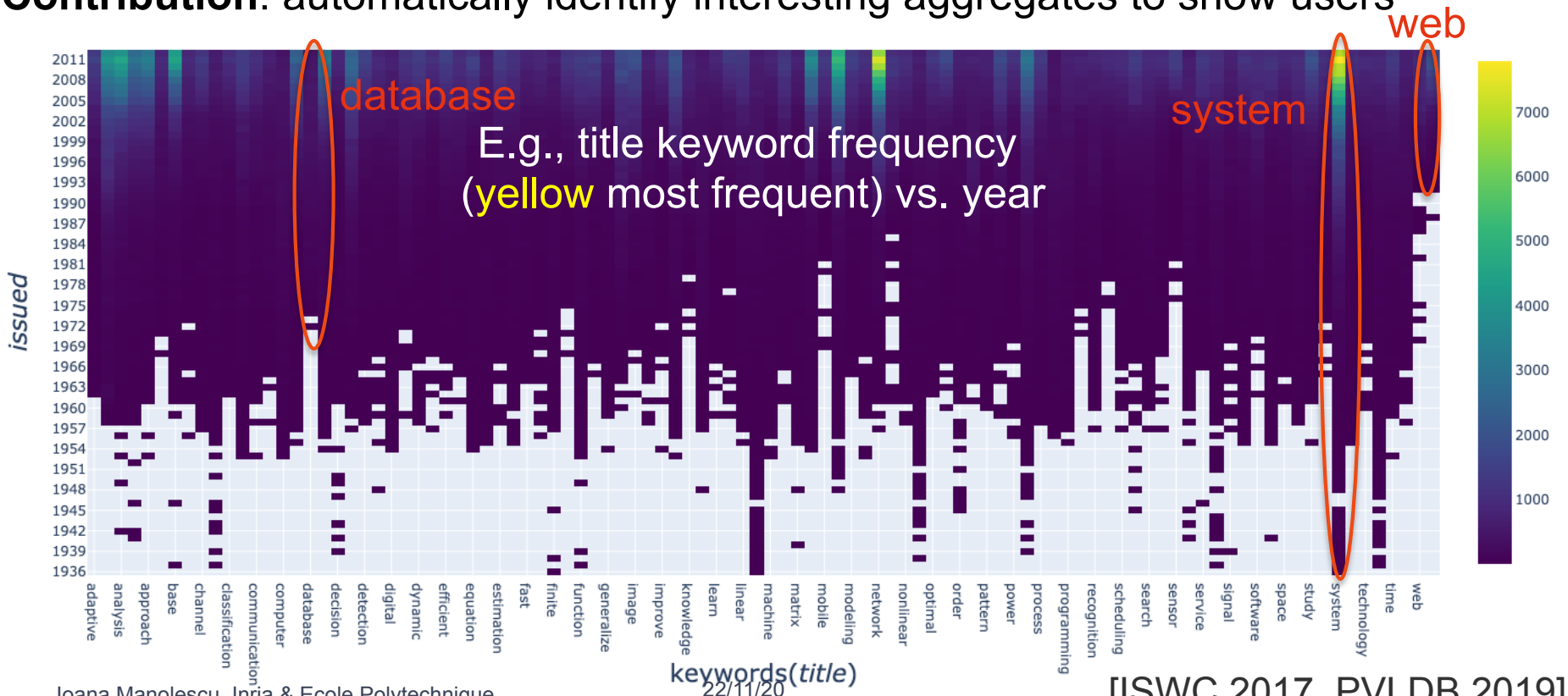
Contribution: automatically identify interesting aggregates to show users



[ISWC 2017, PVLDB 2019]

Finding Interesting Aggregates in RDF Graphs

Contribution: automatically identify interesting aggregates to show users



[ISWC 2017, PVLDB 2019]

Wrap-up

Research in data management and interest/usage of NLP

1. Making data accessible fast
2. Facilitating data usage:
   - ❖ To audience without CS skills: general audience, journalists, ...
   - ❖ To users not yet familiar with the data, regardless of their skills
   - ❖ Through friendly query interfaces (e.g., NL), exploration tools

Useful links

ContentCheck (ANR, 2015-2020)
https://contentcheck.inria.fr

SourcesSay (ANR + DGA, 2020-2024)
https://sourcessay.inria.fr

https://pages.saclay.inria.fr/ioana.manolescu/data-journalism-fact-checking.html