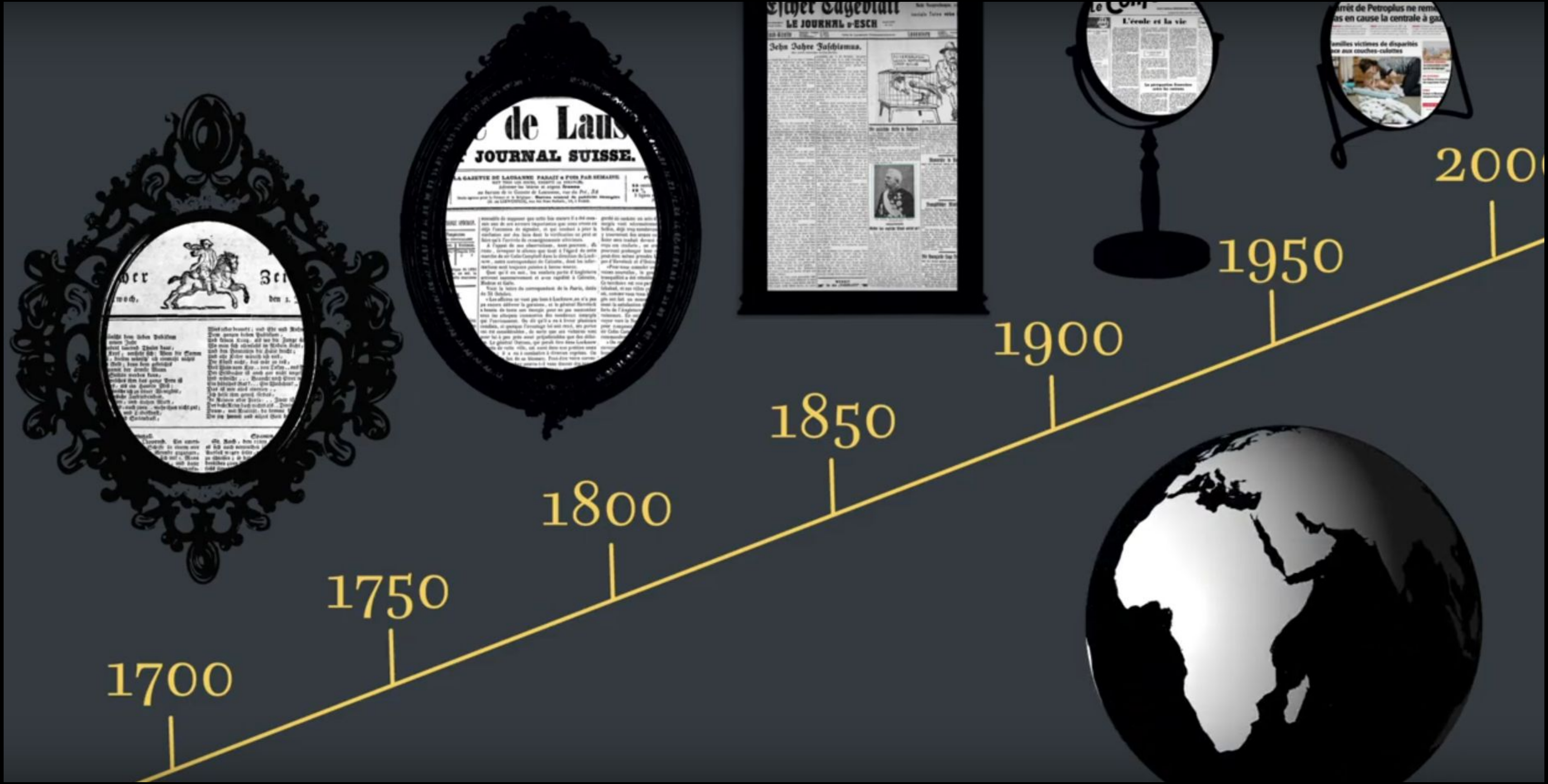
impresso - Media Monitoring of the Past I+II

Beyond Borders: Connecting Historical Newspapers and Radio

Marten Düring & impresso team

impresso Media Monitoring of the Past | EPFL | University of Zurich UZH | Unil UNIL | Université de Lausanne | C²DH | Swiss National Science Foundation | Luxembourg National Research Fund


1700
1750
1800
1850
1900
1950
200
de Laus
JOURNAL SUISSE.
Escher Tageblatt
LE JOURNAL d'ESCH
Zehn Jahre Faſchismus.
L'école et la vie
rrêt de Petroplus ne reme
as en cause la centrale à ga
amilles victimes de disparités
aux couches-culottes

für Wahrheit und Recht

Nr. 115 Mittwoch, 24. April 1912 65. Jahrgang.

Erſtes Blatt.

Serbiſche Wahlen.

Türkiſch-italieniſcher Krie

Erzbiſchof Dr. v. Abert †

Zum Untergang der Titanic.

Schloß Altenſtein.

Haute tension sur l'avenir électrique

ertains experts noncent une pénurie d'électricité pour la fin u millénaire. D'autres pécialistes affirment ue l'on néglige les conomies. En tendant, la Suisse ontinue d'investir ans le parc nucléaire ançais.

L'ère des incertitudes

ENERGIE DE L'OUEST SUISSE (EOS)

Christophe Babaiantz: le prix du renoncement à Kaiseraugst

Société anonyme „Le Piolet" Luxembourg.

Avis mortuaire.

Madame Alph. Diederich

La famille éplorée.

Wohnhaus-Verſteigerung

Ein Wohnhaus

Fett-, Halbfett- und Magerkohlen

Schmiedekohlen

Industriebrikets

Braunkohlenbrikets

Ruhr-Brechkoks

HENRI PUTZ, Ettelbrück.

Wir ſuchen zum ſofortigen Eintritt einen Laufburſchen

Warenhaus MONOPOL.

Herren = Konfektion

Volks-Kleider-Halle

Eſch an der Alzette

Berufs- und Arbeiter-Kleider

RESTAURANT Niedner-Schwer

Réouverture Café Belg

Waggon !Kinderwagen!

Bonneweg.

Kirchhofſtraße Nr. 31

Sargmagazin und Möbel-Lager

Félix Fiedler-Neys.

Persil

Für Krankenwäsche

Das selbsttätige Waschmittel.

Henkel's Bleich-Soda

J. SCHUMMER

Automobil

Chaussures Modernes

Peckels Steffen, 28 Pastorstr. 28.

MARA STIEFEL

Einheitspreis 10.50

Abteilung Möbel:

Möbelhaus Funck-Bauer

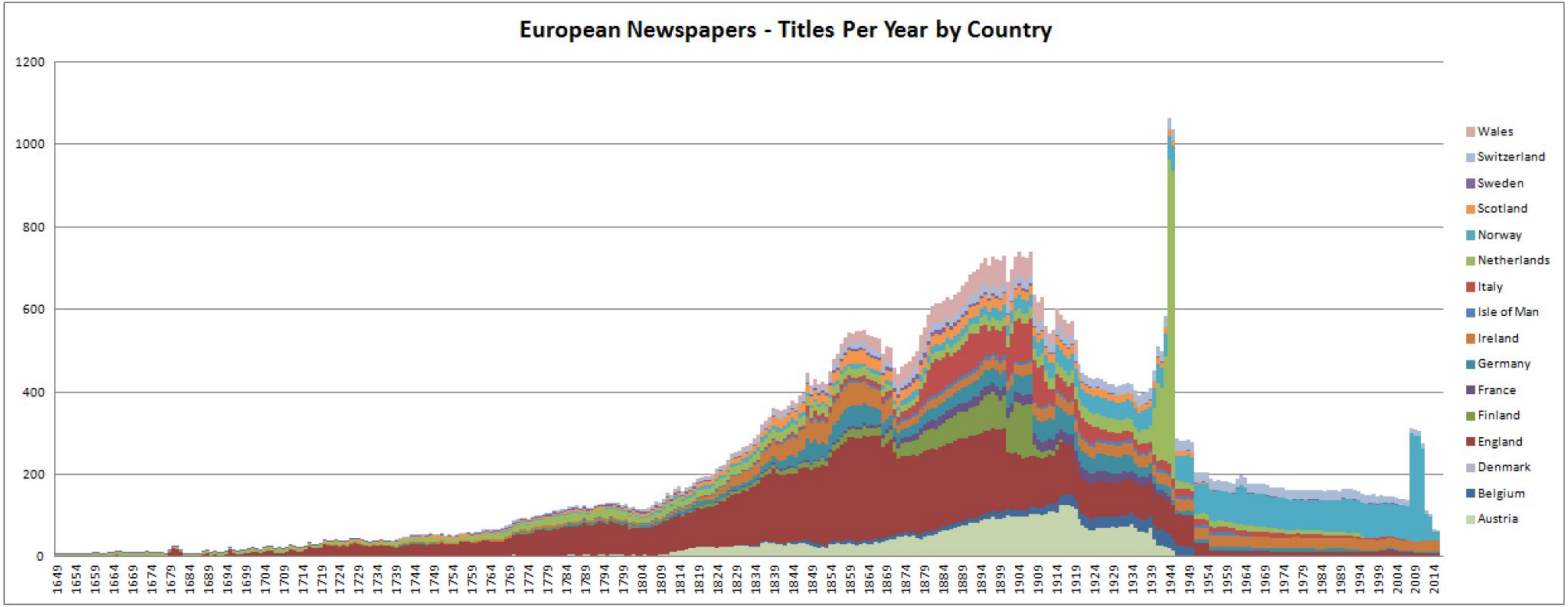
Automobil komplett für 2000 Fr.

Wasch- und Bleichmittel

Lessive Body

Spar-Einlagen

The landscape of historical newspapers



Cf. Center for Research Libraries. The State of the Art. A Comparative Analysis of Newspaper Digitization to Date, 2015

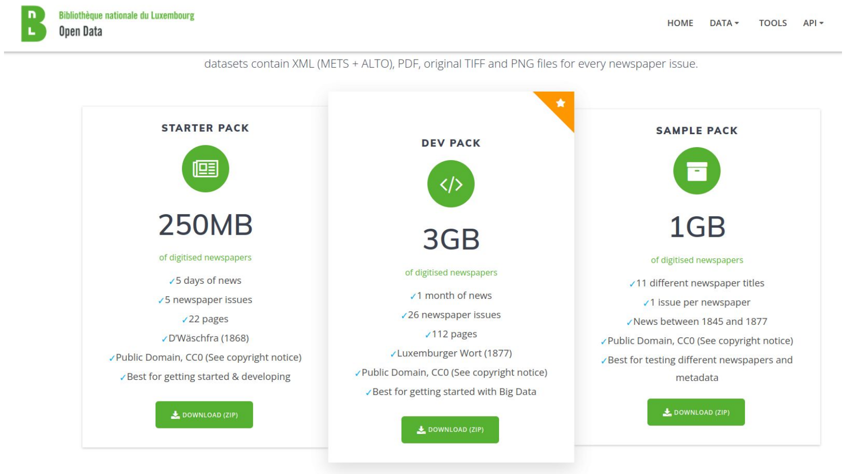
The challenging landscape of historical newspapers

1. Institutional silos
2. Big and messy data
3. Noisy historical text
4. Visualisation and exploration
5. Digital Scholarship

Newspapers as Data

‘*Collections as Data*’ as a “conceptual orientation to collections that renders them as ordered information, stored digitally, so that they are inherently amenable to computation”.

https://collectionsasdata.github.io/



Bibliothèque nationale du Luxembourg
Open Data

datasets contain XML (METS + ALTO), PDF, original TIFF and PNG files for every newspaper issue.

STARTER PACK

250MB

of digitised newspapers

✓5 days of news
✓5 newspaper issues
✓22 pages
✓D'Wäschfra (1868)
✓Public Domain, CC0 (See copyright notice)
✓Best for getting started & developing

DEV PACK

3GB

of digitised newspapers

✓1 month of news
✓26 newspaper issues
✓112 pages
✓Luxemburger Wort (1877)
✓Public Domain, CC0 (See copyright notice)
✓Best for getting started with Big Data

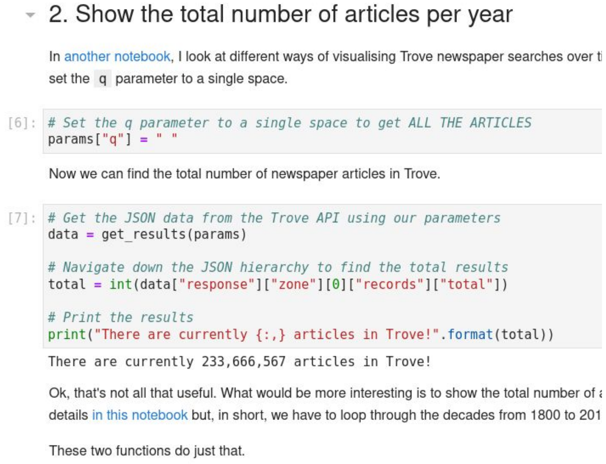
SAMPLE PACK

1GB

of digitised newspapers

✓11 different newspaper titles
✓1 issue per newspaper
✓News between 1845 and 1877
✓Public Domain, CC0 (See copyright notice)
✓Best for testing different newspapers and metadata

https://data.bnl.lu



2. Show the total number of articles per year

In another notebook, I look at different ways of visualising Trove newspaper searches over t
set the `q` parameter to a single space.

```
[6]: # Set the q parameter to a single space to get ALL THE ARTICLES
     params["q"] = " "
```

Now we can find the total number of newspaper articles in Trove.

```
[7]: # Get the JSON data from the Trove API using our parameters
     data = get_results(params)

     # Navigate down the JSON hierarchy to find the total results
     total = int(data["response"]["zone"][0]["records"]["total"])

     # Print the results
     print("There are currently {:,} articles in Trove!".format(total))
```

```
There are currently 233,666,567 articles in Trove!
```

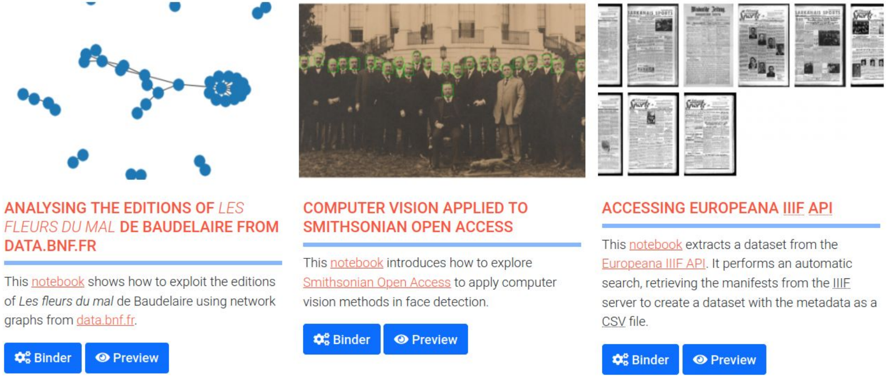
Ok, that's not all that useful. What would be more interesting is to show the total number of a
details in this notebook but, in short, we have to loop through the decades from 1800 to 201

These two functions do just that.

https://glam-workbench.net/trove-newspapers/

OPEN
A
GLAM
LAB

https://glamlabs.io/



ANALYSING THE EDITIONS OF *LES FLEURS DU MAL* DE BAUDELAIRE FROM DATA.BNF.FR

This notebook shows how to exploit the editions of *Les fleurs du mal* de Baudelaire using network graphs from data.bnf.fr.

COMPUTER VISION APPLIED TO SMITHSONIAN OPEN ACCESS

This notebook introduces how to explore Smithsonian Open Access to apply computer vision methods in face detection.

ACCESSING EUROPEANA IIIF API

This notebook extracts a dataset from the Europeana IIIF API. It performs an automatic search, retrieving the manifests from the IIIF server to create a dataset with the metadata as a CSV file.

https://data.cervantesvirtual.com/glam-jupyter-notebooks

In which areas of research is this method/approach currently being used and developed, and what are its merits?

What types of data and research interests fit well with this method/approach?

Interdisciplinary projects on newspapers

impresso
Media
Monitoring
of the Past

NEWS
EYE

LIVING WITH MACHINES

Oceanic Exchanges

NUMA
PRESSE

How can semantic indexing enhance the study and exploration of historical newspapers?

The team

+ *historical advisors*
+ *associated historians*

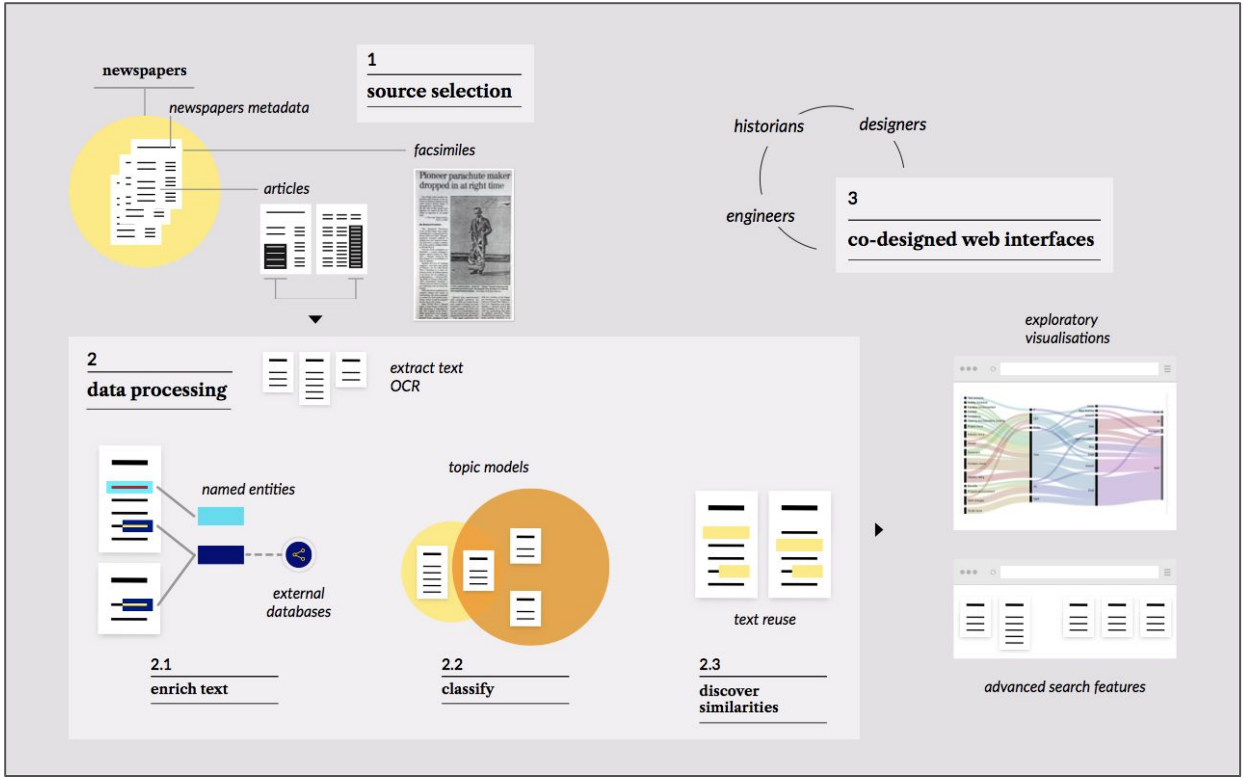
Estelle Bunout
Simon Clematide
Marten Duering
Maud Ehrmann
Andreas Fickers
Daniele Guido
Frédéric Kaplan
Peter Makarov
Matteo Romanello
Gerold Schneider
Paul Schroeder
Benoit Seguin
Phillip Stroëbel
Martin Volk
Thijs van Beek
Lars Wieneke

Objectives and research questions

1. How to adapt NLP tools to historical texts?

2. How to explore complex and vast amounts of data?

3. What is the impact of new tooling on digital scholarship?



newspapers

newspapers metadata

1
source selection

facsimiles

Pioneer parachute maker dropped in at right time

articles

historians

designers

engineers

3
co-designed web interfaces

2
data processing

extract text
OCR

named entities

external databases

topic models

text reuse

2.1
enrich text

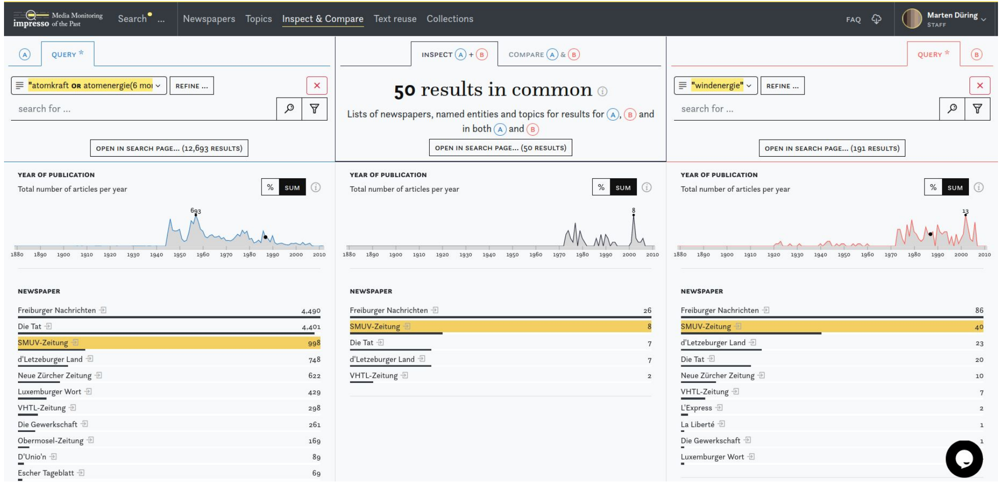
2.2
classify

2.3
discover similarities

exploratory visualisations

advanced search features

Target audience



50 results in common

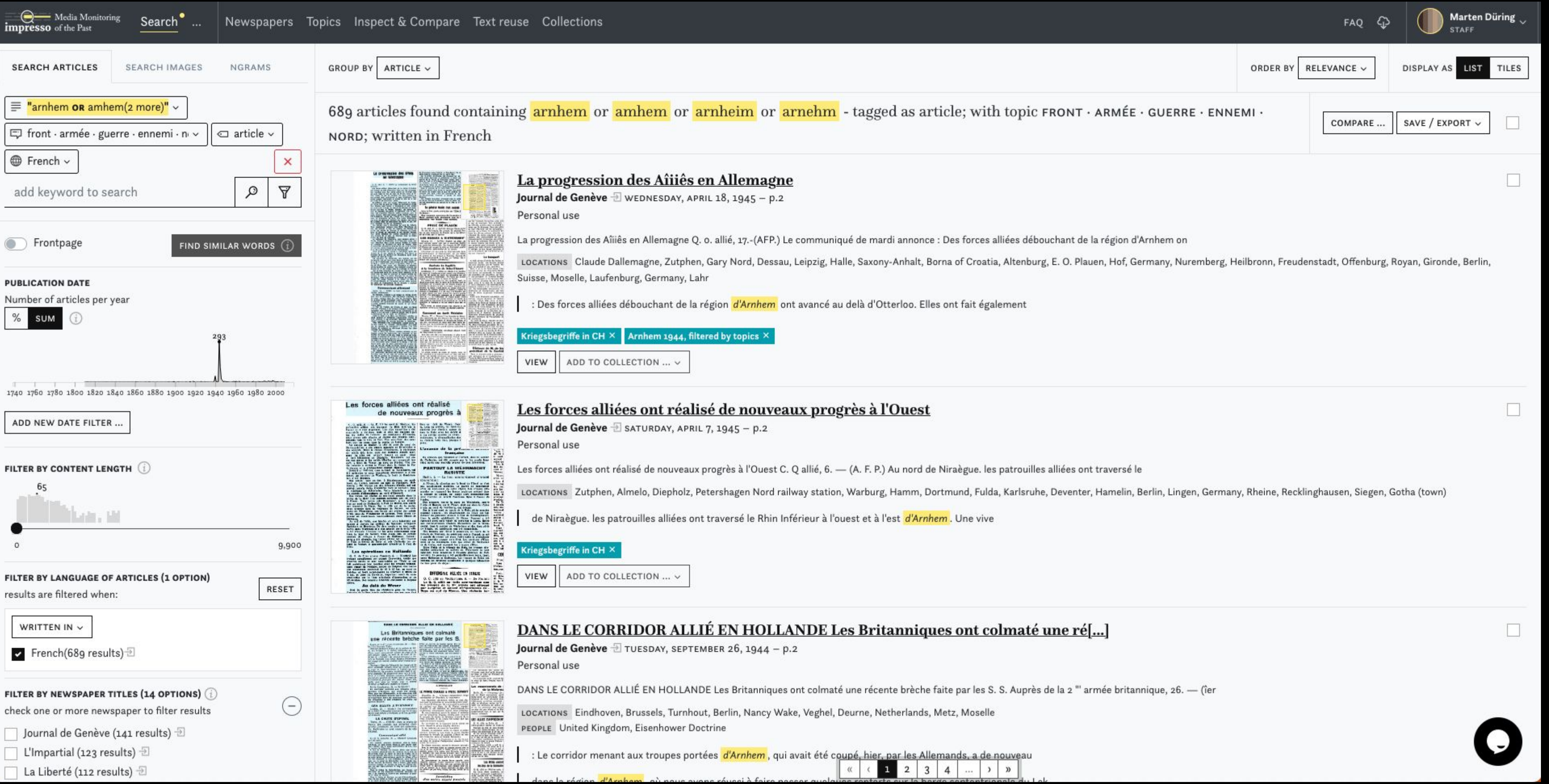
Lists of newspapers, named entities and topics for results for A, B and in both A and B

NEWSPAPER	
Freiburger Nachrichten	4,490
Die Tat	4,401
SMUV-Zeitung	998
d'Letzeburger Land	748
Neue Zürcher Zeitung	622
Luxemburger Wort	429
VHTL-Zeitung	298
Die Gewerkschaft	261
Obermosel-Zeitung	169
D'Unio'n	89
Escher Tageblatt	69

NEWSPAPER	
Freiburger Nachrichten	26
SMUV-Zeitung	8
Die Tat	7
d'Letzeburger Land	7
VHTL-Zeitung	2

NEWSPAPER	
Freiburger Nachrichten	86
SMUV-Zeitung	40
d'Letzeburger Land	23
Die Tat	20
Neue Zürcher Zeitung	10
VHTL-Zeitung	7
L'Express	2
La Liberté	1
Die Gewerkschaft	1
Luxemburger Wort	

Historians without specific digital literacy but curiosity to go beyond keyword search and the motivation to learn how to use a research tool.



"arnhem OR amhem(2 more)"

front · armée · guerre · ennemi · n

article

French

Frontpage

PUBLICATION DATE

Number of articles per year

% SUM

ADD NEW DATE FILTER ...

FILTER BY CONTENT LENGTH

FILTER BY LANGUAGE OF ARTICLES (1 OPTION)

results are filtered when:

RESET

WRITTEN IN

☑ French(689 results)

FILTER BY NEWSPAPER TITLES (14 OPTIONS)

check one or more newspaper to filter results

☐ Journal de Genève (141 results)

☐ L'Impartial (123 results)

☐ La Liberté (112 results)

689 articles found containing arnhem or amhem or arnheim or arnehm - tagged as article; with topic FRONT · ARMÉE · GUERRE · ENNEMI · NORD; written in French

La progression des Aîiiês en Allemagne

Journal de Genève WEDNESDAY, APRIL 18, 1945 – p.2

Personal use

La progression des Aîiiês en Allemagne Q. o. allié, 17.-(AFP.) Le communiqué de mardi annonce : Des forces alliées débouchant de la région d'Arnhem on

LOCATIONS Claude Dallemagne, Zutphen, Gary Nord, Dessau, Leipzig, Halle, Saxony-Anhalt, Borna of Croatia, Altenburg, E. O. Plauen, Hof, Germany, Nuremberg, Heilbronn, Freudenstadt, Offenburg, Royan, Gironde, Berlin, Suisse, Moselle, Laufenburg, Germany, Lahr

: Des forces alliées débouchant de la région d'Arnhem ont avancé au delà d'Otterloo. Elles ont fait également

Kriegsbegriffe in CH × Arnhem 1944, filtered by topics ×

Les forces alliées ont réalisé de nouveaux progrès à l'Ouest

Journal de Genève SATURDAY, APRIL 7, 1945 – p.2

Personal use

Les forces alliées ont réalisé de nouveaux progrès à l'Ouest C. Q allié, 6. — (A. F. P.) Au nord de Niraègue. les patrouilles alliées ont traversé le

LOCATIONS Zutphen, Almelo, Diepholz, Petershagen Nord railway station, Warburg, Hamm, Dortmund, Fulda, Karlsruhe, Deventer, Hamelin, Berlin, Lingen, Germany, Rheine, Recklinghausen, Siegen, Gotha (town)

de Niraègue. les patrouilles alliées ont traversé le Rhin Inférieur à l'ouest et à l'est d'Arnhem . Une vive

Kriegsbegriffe in CH ×

DANS LE CORRIDOR ALLIÉ EN HOLLANDE Les Britanniques ont colmaté une ré[...]

Journal de Genève TUESDAY, SEPTEMBER 26, 1944 – p.2

Personal use

DANS LE CORRIDOR ALLIÉ EN HOLLANDE Les Britanniques ont colmaté une récente brèche faite par les S. S. Auprès de la 2 '" armée britannique, 26. — (îer

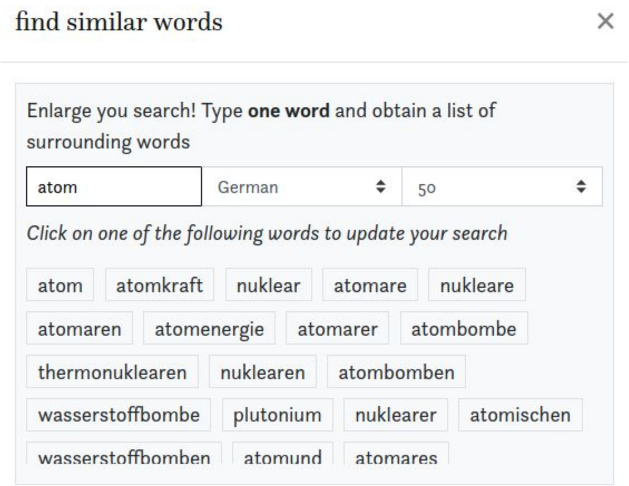
LOCATIONS Eindhoven, Brussels, Turnhout, Berlin, Nancy Wake, Veghel, Deurne, Netherlands, Metz, Moselle

PEOPLE United Kingdom, Eisenhower Doctrine

: Le corridor menant aux troupes portées d'Arnhem , qui avait été coupé, hier, par les Allemands, a de nouveau

dans la région d'Arnhem , où nous avons réussi à faire passer quelques renforts sur la berge septentrionale du Lek

Example codesign: Word embeddings



find similar words ×

Enlarge you search! Type **one word** and obtain a list of surrounding words

atom | German | 50

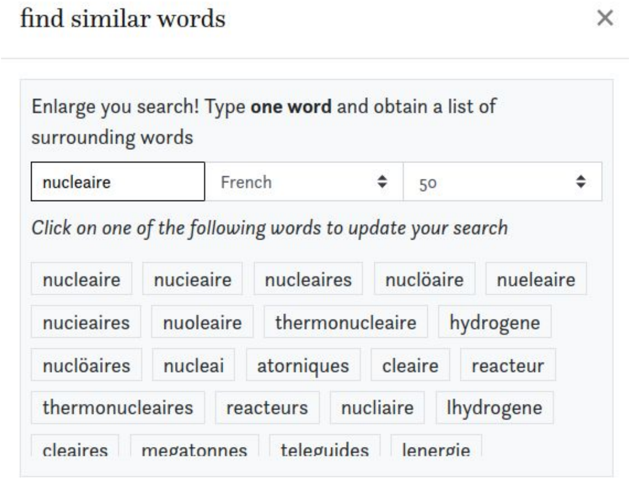
Click on one of the following words to update your search

atom atomkraft nuklear atomare nukleare
atomaren atomenergie atomarer atombombe
thermonuklearen nuklearen atombomben
wasserstoffbombe plutonium nuklearer atomischen
wasserstoffbomben atomund atomares

Word embeddings?

Keyword suggestions?

OCR mistakes!



find similar words ×

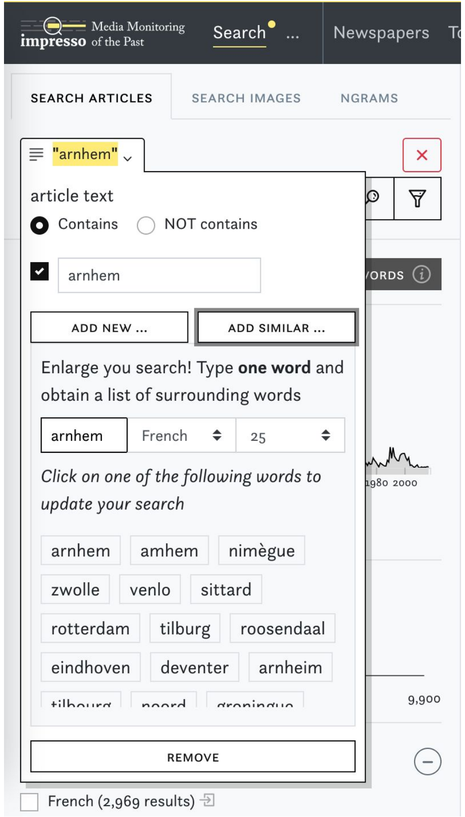
Enlarge you search! Type **one word** and obtain a list of surrounding words

nucleaire | French | 50

Click on one of the following words to update your search

nucleaire nucieaire nucleaires nuclöaire nueleaire
nucieaires nuoleaire thermonucleaire hydrogene
nuclöaires nucleai atorniques cleaire reacteur
thermonucleaires reacteurs nucliaire lhydrogene
cleaires megatonnes teleguides lenergie

Add similar



article text

● Contains ○ NOT contains

☑ arnhem

Enlarge you search! Type **one word** and obtain a list of surrounding words

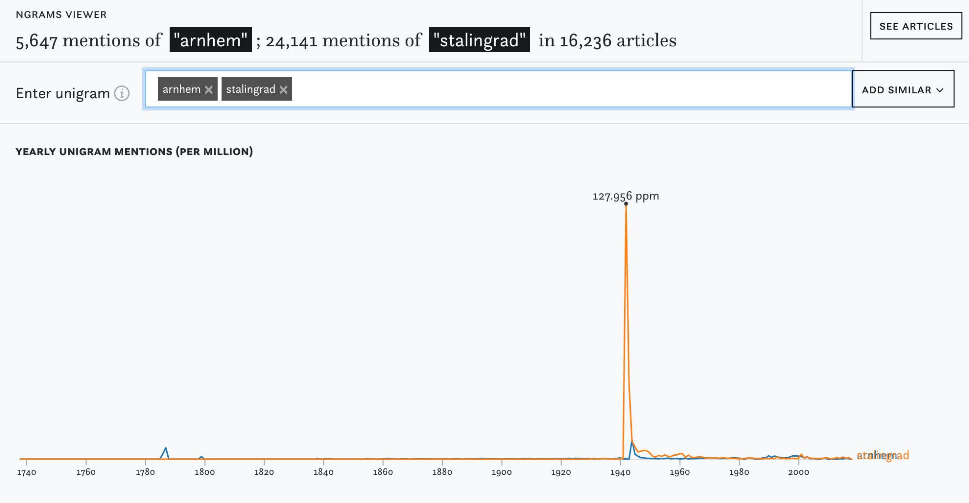
arnhem | French | 25

Click on one of the following words to update your search

arnhem amhem nimègue zwolle venlo sittard rotterdam tilburg roosendaal eindhoven deventer arnheim tilbourg noord groningue

□ French (2,969 results)

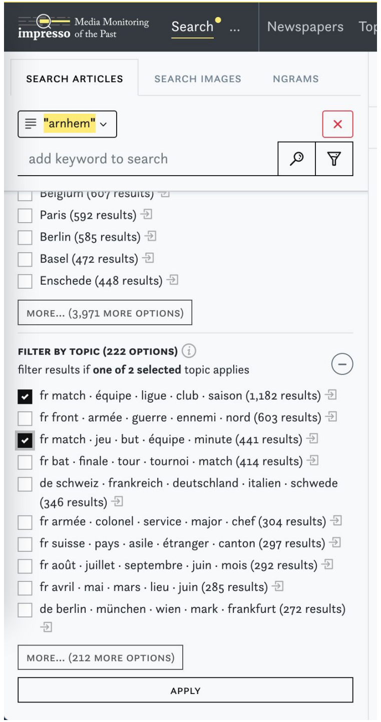
Ngrams



5,647 mentions of "arnhem" ; 24,141 mentions of "stalingrad" in 16,236 articles

Enter unigram: arnhem ✕ stalingrad ✕

Filter by topic



SEARCH ARTICLES | SEARCH IMAGES | NGRAMS

"arnhem"

add keyword to search

□ Belgium (607 results)
□ Paris (592 results)
□ Berlin (585 results)
□ Basel (472 results)
□ Enschede (448 results)

MORE... (3,971 MORE OPTIONS)

FILTER BY TOPIC (222 OPTIONS)

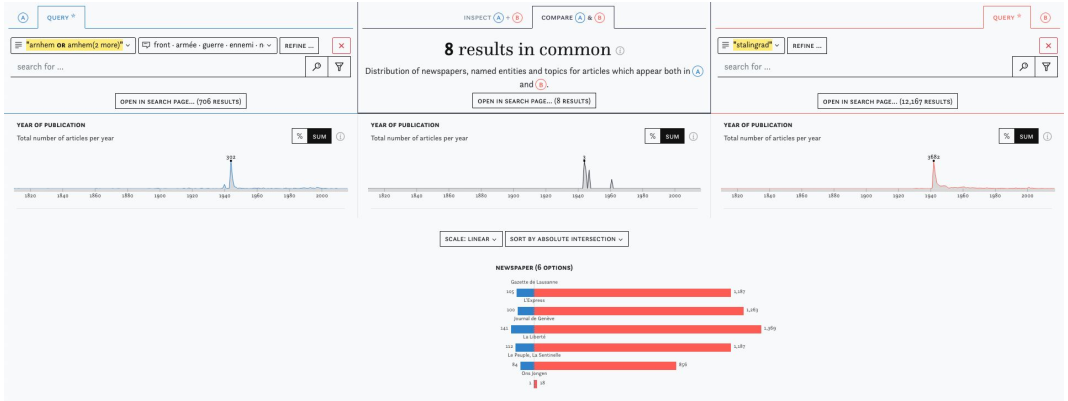
filter results if **one of 2 selected** topic applies

☑ fr match · équipe · ligue · club · saison (1,182 results)
□ fr front · armée · guerre · ennemi · nord (603 results)
☑ fr match · jeu · but · équipe · minute (441 results)
□ fr bat · finale · tour · tournoi · match (414 results)
□ de schweiz · frankreich · deutschland · italien · schwede (346 results)
□ fr armée · colonel · service · major · chef (304 results)
□ fr suisse · pays · asile · étranger · canton (297 results)
□ fr août · juillet · septembre · juin · mois (292 results)
□ fr avril · mai · mars · lieu · juin (285 results)
□ de berlin · münchen · wien · mark · frankfurt (272 results)

MORE... (212 MORE OPTIONS)

APPLY

Inspect & Compare



"arnhem OR amhem(2 more)" | front · armée · guerre · ennemi · n

search for ...

OPEN IN SEARCH PAGE... (706 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

8 results in common

Distribution of newspapers, named entities and topics for articles which appear both in A and B.

OPEN IN SEARCH PAGE... (8 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

"stalingrad"

search for ...

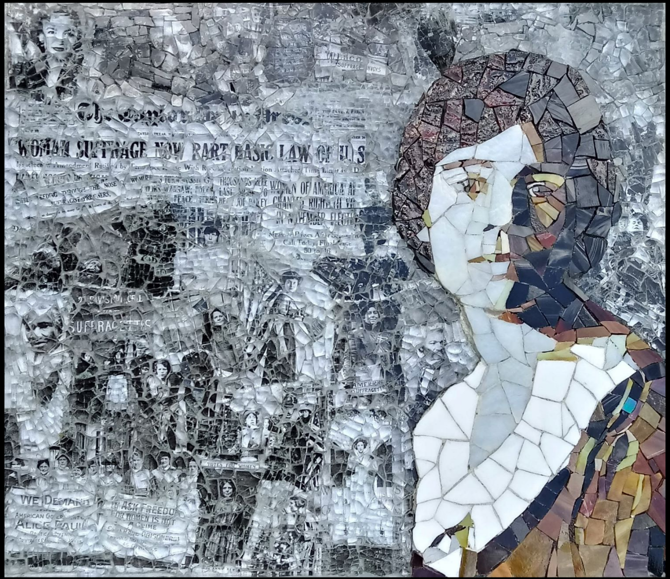
OPEN IN SEARCH PAGE... (12,167 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

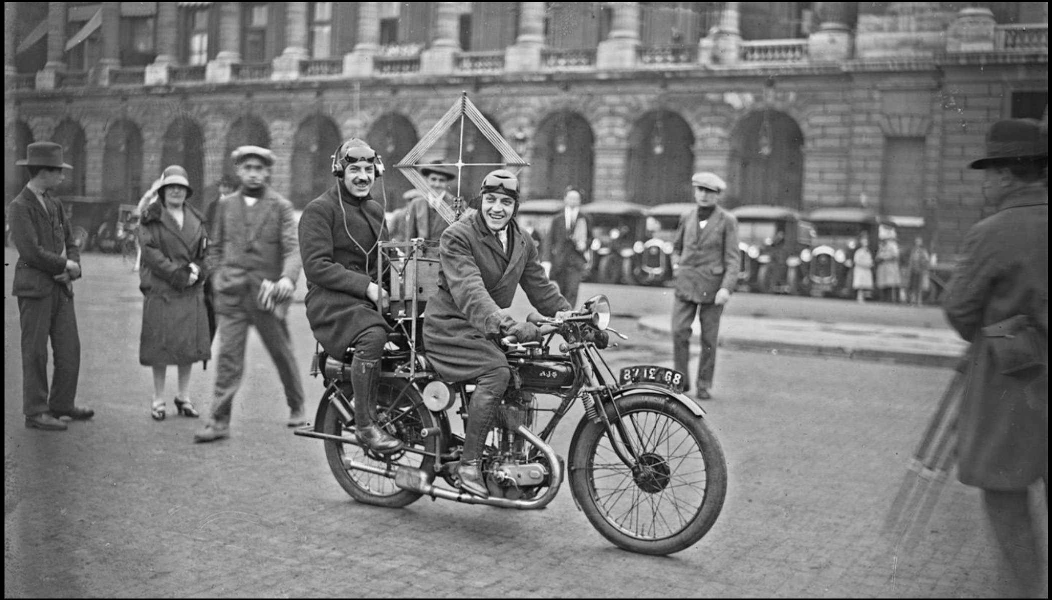
NEWSPAPER (6 OPTIONS)

Gianni Sarconi, The Master of Numbers (2006)



Jennifer Kuhns, Glass mosaic of Alice Paul, suffragist. 1885-1977 (2019)

How could this method/approach contribute to the study of historical processes of knowledge evolution?

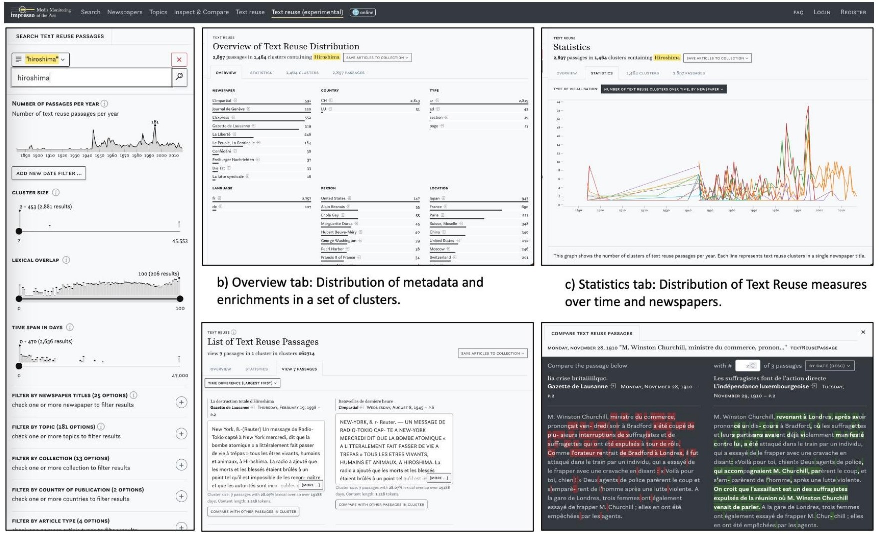


Outlook: What is next?

Text Reuse at Scale

Opportunity: integration of TR data with other semantic enrichments.

1. Expert workshop in Nov 2022 produced research objectives, tasks & mockups.
2. New prototype component.
3. User evaluation.
4. Research paper [preprint].
5. Revised prototype to be integrated in main app.


impresso Media Monitoring of the Past
Search
Newspapers
Topics
Inspect & Compare
Text reuse
Text reuse (experimental)
online
FAQ
LOGIN
REGISTER
SEARCH TEXT REUSE PASSAGES
hiroshima
NUMBER OF PASSAGES PER YEAR
CLUSTER SIZE
LEXICAL OVERLAP
TIME SPAN IN DAYS
FILTER BY NEWSPAPER TITLES (25 OPTIONS)
FILTER BY TOPIC (181 OPTIONS)
FILTER BY COLLECTION (13 OPTIONS)
FILTER BY COUNTRY OF PUBLICATION (2 OPTIONS)
FILTER BY ARTICLE TYPE (4 OPTIONS)
Overview of Text Reuse Distribution
2,897 passages in 1,464 clusters containing Hiroshima
NEWSPAPER
COUNTRY
TYPE
LANGUAGE
PERSON
LOCATION
Statistics
List of Text Reuse Passages
COMPARE TEXT REUSE PASSAGES

b) Overview tab: Distribution of metadata and enrichments in a set of clusters.

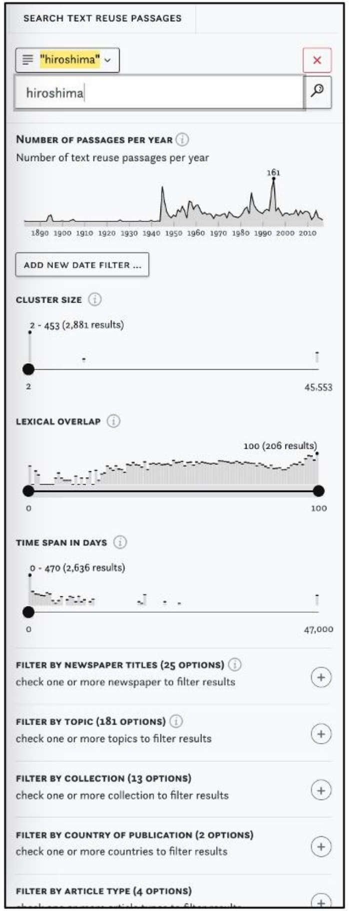
c) Statistics tab: Distribution of Text Reuse measures over time and newspapers.

a) Search and filters.

d) Passages tab: Passages within cluster c62714.

e) Passages tab: Comparative view.

impresso Media Monitoring of the Past | Search | Newspapers | Topics | Inspect & Compare | Text reuse | Text reuse (experimental) | online | FAQ | Login | Register



SEARCH TEXT REUSE PASSAGES

"hiroshima"

hiroshima

Number of passages per year

Number of text reuse passages per year

ADD NEW DATE FILTER ...

CLUSTER SIZE

2 - 453 (2,881 results)

LEXICAL OVERLAP

100 (206 results)

TIME SPAN IN DAYS

0 - 470 (2,636 results)

FILTER BY NEWSPAPER TITLES (25 OPTIONS)
check one or more newspaper to filter results

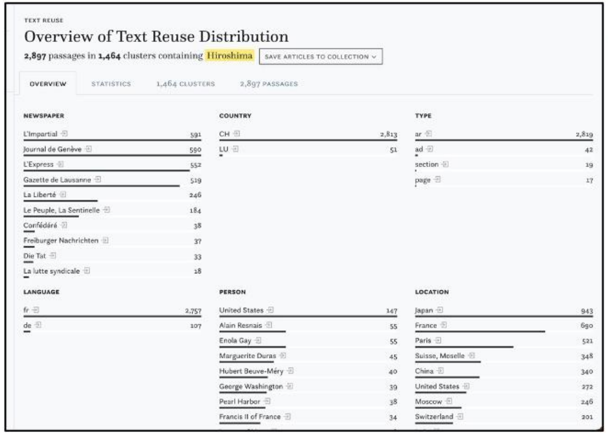
FILTER BY TOPIC (181 OPTIONS)
check one or more topics to filter results

FILTER BY COLLECTION (13 OPTIONS)
check one or more collection to filter results

FILTER BY COUNTRY OF PUBLICATION (2 OPTIONS)
check one or more countries to filter results

FILTER BY ARTICLE TYPE (4 OPTIONS)

a) Search and filters.



TEXT REUSE

Overview of Text Reuse Distribution

2,897 passages in **1,464** clusters containing Hiroshima

SAVE ARTICLES TO COLLECTION

OVERVIEW | STATISTICS | 1,464 CLUSTERS | 2,897 PASSAGES

NEWSPAPER	
L'Impartial	591
Journal de Genève	590
L'Express	552
Gazette de Lausanne	519
La Liberté	246
Le Peuple, La Sentinelle	184
Confédéré	38
Freiburger Nachrichten	37
Die Tat	33
La lutte syndicale	18

COUNTRY	
CH	2,813
LU	51

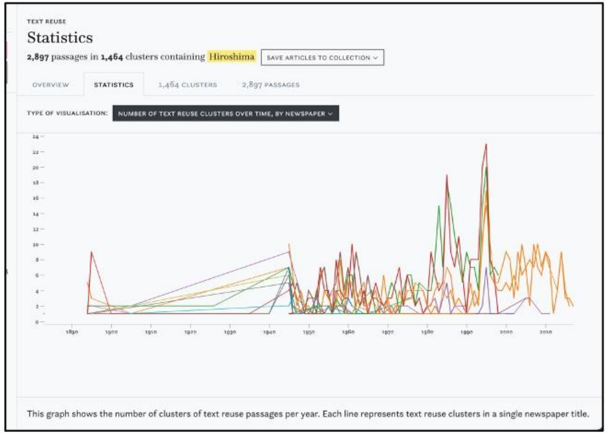
TYPE	
ar	2,819
ad	42
section	19
page	17

LANGUAGE	
fr	2,757
de	107

PERSON	
United States	147
Alain Resnais	55
Enola Gay	55
Marguerite Duras	45
Hubert Beuve-Méry	40
George Washington	39
Pearl Harbor	38
Francis II of France	34

LOCATION	
Japan	943
France	690
Paris	521
Suisse, Moselle	348
China	340
United States	272
Moscow	246
Switzerland	201

b) Overview tab: Distribution of metadata and enrichments in a set of clusters.



TEXT REUSE

Statistics

2,897 passages in **1,464** clusters containing Hiroshima

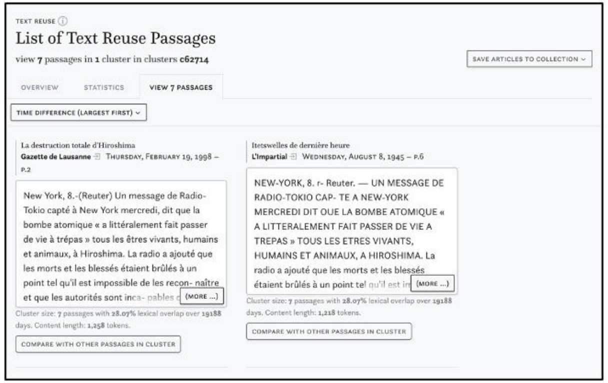
SAVE ARTICLES TO COLLECTION

OVERVIEW | STATISTICS | 1,464 CLUSTERS | 2,897 PASSAGES

TYPE OF VISUALISATION: NUMBER OF TEXT REUSE CLUSTERS OVER TIME, BY NEWSPAPER

This graph shows the number of clusters of text reuse passages per year. Each line represents text reuse clusters in a single newspaper title.

c) Statistics tab: Distribution of Text Reuse measures over time and newspapers.



TEXT REUSE

List of Text Reuse Passages

view **7** passages in **1** cluster in clusters **c62714**

SAVE ARTICLES TO COLLECTION

OVERVIEW | STATISTICS | VIEW 7 PASSAGES

TIME DIFFERENCE (LARGEST FIRST)

La destruction totale d'Hiroshima
Gazette de Lausanne Thursday, February 19, 1998 – P.2

New York, 8.-(Reuter) Un message de Radio-Tokio capté à New York mercredi, dit que la bombe atomique « a littéralement fait passer de vie à trépas » tous les êtres vivants, humains et animaux, à Hiroshima. La radio a ajouté que les morts et les blessés étaient brûlés à un point tel qu'il est impossible de les recon- naître et que les autorités sont inca- pables (MORE ...)

Cluster size: 7 passages with **28.07%** lexical overlap over **19188** days. Content length: **1,258** tokens.

COMPARE WITH OTHER PASSAGES IN CLUSTER

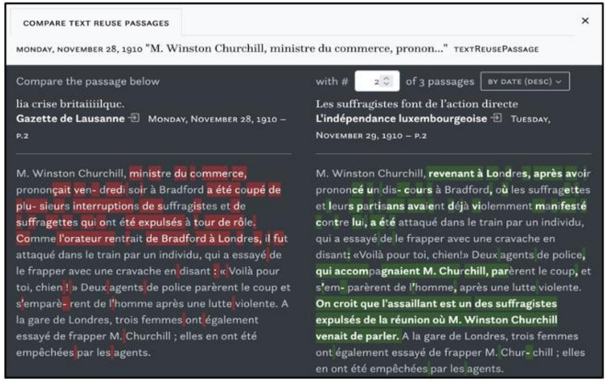
Ietswelles de dernière heure
L'Impartial Wednesday, August 8, 1945 – P.6

NEW-YORK, 8. r- Reuter. — UN MESSAGE DE RADIO-TOKIO CAP- TE A NEW-YORK MERCREDI DIT OUE LA BOMBE ATOMIQUE « A LITTERALEMENT FAIT PASSER DE VIE A TREPAS » TOUS LES ETRES VIVANTS, HUMAINS ET ANIMAUX, A HIROSHIMA. La radio a ajouté que les morts et les blessés étaient brûlés à un point tel qu'il est in (MORE ...)

Cluster size: 7 passages with **28.07%** lexical overlap over **19188** days. Content length: **1,218** tokens.

COMPARE WITH OTHER PASSAGES IN CLUSTER

d) Passages tab: Passages within cluster c62714.



COMPARE TEXT REUSE PASSAGES ×

MONDAY, NOVEMBER 28, 1910 "M. Winston Churchill, ministre du commerce, pronon..." TEXTREUSEPASSAGE

Compare the passage below

with # 2 of 3 passages BY DATE (DESC)

lia crise britaiiiilquc.
Gazette de Lausanne Monday, November 28, 1910 – P.2

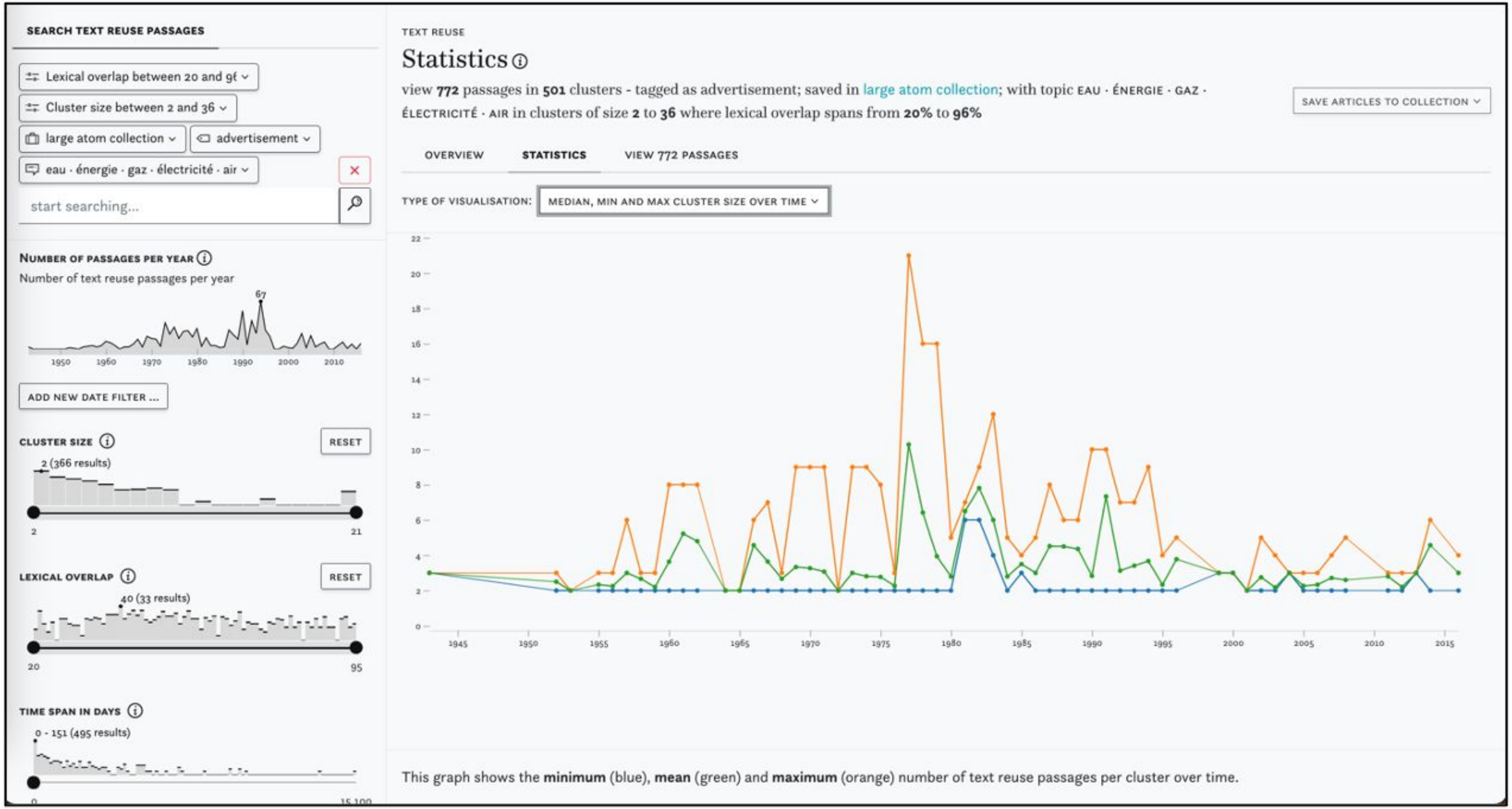
M. Winston Churchill, ministre du commerce, prononçait ven- dredi soir à Bradford a été coupé de plu- sieurs interruptions de suffragistes et de suffragettes qui ont été expulsés à tour de rôle. Comme l'orateur rentrait de Bradford à Londres, il fut attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant : « Voilà pour toi, chien ! » Deux agents de police parèrent le coup et s'emparè- rent de l'homme après une lutte violente. A la gare de Londres, trois femmes ont également essayé de frapper M. Churchill ; elles en ont été empêchées par les agents.

Les suffragistes font de l'action directe
L'indépendance luxembourgeoise Tuesday, November 29, 1910 – P.2

M. Winston Churchill, **revenant à Londres, après avoir** prononcé un dis- cours à Bradford, où les suffragettes et leurs partisans avaient déjà violemment manifesté contre lui, a été attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant: «Voilà pour toi, chien!» Deux agents de police, **qui accompagnaient M. Churchill, par**èrent le coup, et s'em- parèrent de l'homme, après une lutte violente. **On croit que l'assaillant est un des suffragistes expulsés de la réunion où M. Winston Churchill venait de parler.** A la gare de Londres, trois femmes ont également essayé de frapper M. Chur- chill ; elles en ont été empêchées par les agents.

e) Passages tab: Comparative view.

Example: When did nuclear power ads circulate?



SEARCH TEXT REUSE PASSAGES

Lexical overlap between 20 and 9€

Cluster size between 2 and 36

large atom collection

advertisement

eau · énergie · gaz · électricité · air

start searching...

NUMBER OF PASSAGES PER YEAR

Number of text reuse passages per year

ADD NEW DATE FILTER ...

CLUSTER SIZE

RESET

LEXICAL OVERLAP

RESET

TIME SPAN IN DAYS

TEXT REUSE

Statistics

view **772** passages in **501** clusters - tagged as advertisement; saved in large atom collection; with topic EAU · ÉNERGIE · GAZ · ÉLECTRICITÉ · AIR in clusters of size **2** to **36** where lexical overlap spans from **20%** to **96%**

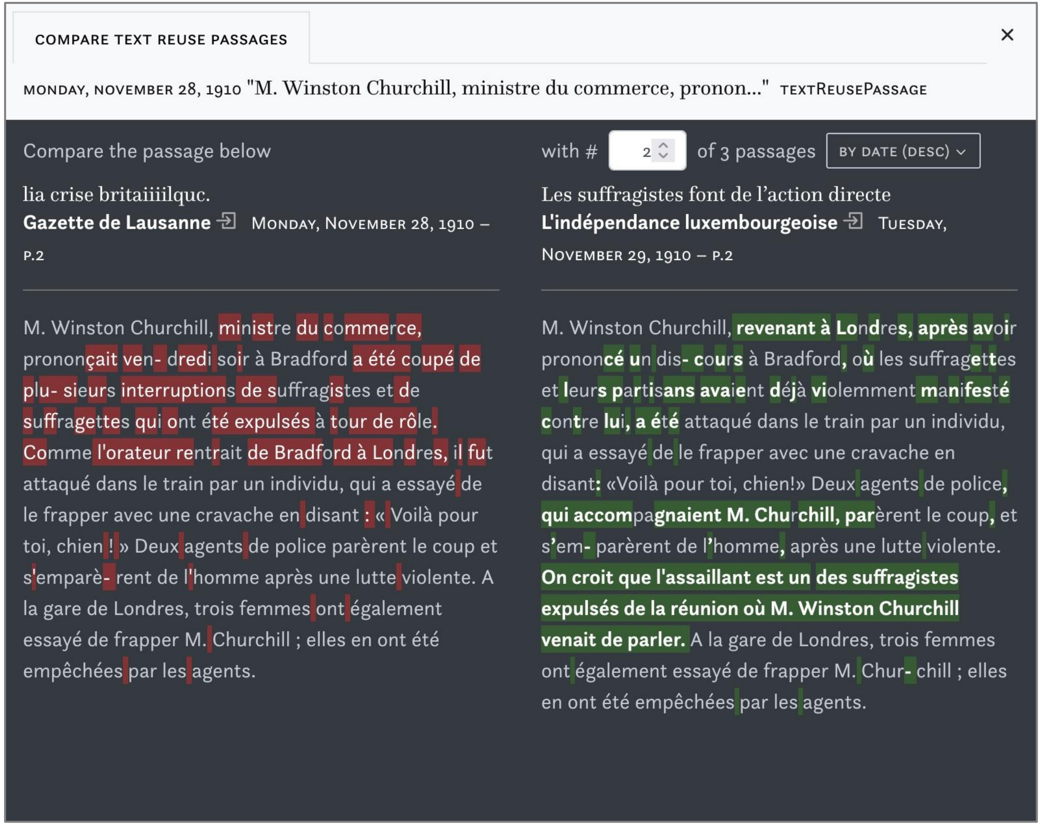
SAVE ARTICLES TO COLLECTION

OVERVIEW | **STATISTICS** | VIEW 772 PASSAGES

TYPE OF VISUALISATION: MEDIAN, MIN AND MAX CLUSTER SIZE OVER TIME

This graph shows the **minimum** (blue), **mean** (green) and **maximum** (orange) number of text reuse passages per cluster over time.

Editing press agency content



MONDAY, NOVEMBER 28, 1910 "M. Winston Churchill, ministre du commerce, pronon..." TEXTREUSEPASSAGE

Compare the passage below

lia crise britaiiiilquc.
Gazette de Lausanne MONDAY, NOVEMBER 28, 1910 – P.2

M. Winston Churchill, ministre du commerce, prononçait ven- dredi soir à Bradford a été coupé de plu- sieurs interruptions de suffragistes et de suffragettes qui ont été expulsés à tour de rôle. Comme l'orateur rentrait de Bradford à Londres, il fut attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant : « Voilà pour toi, chien ! » Deux agents de police parèrent le coup et s'emparè- rent de l'homme après une lutte violente. A la gare de Londres, trois femmes ont également essayé de frapper M. Churchill ; elles en ont été empêchées par les agents.

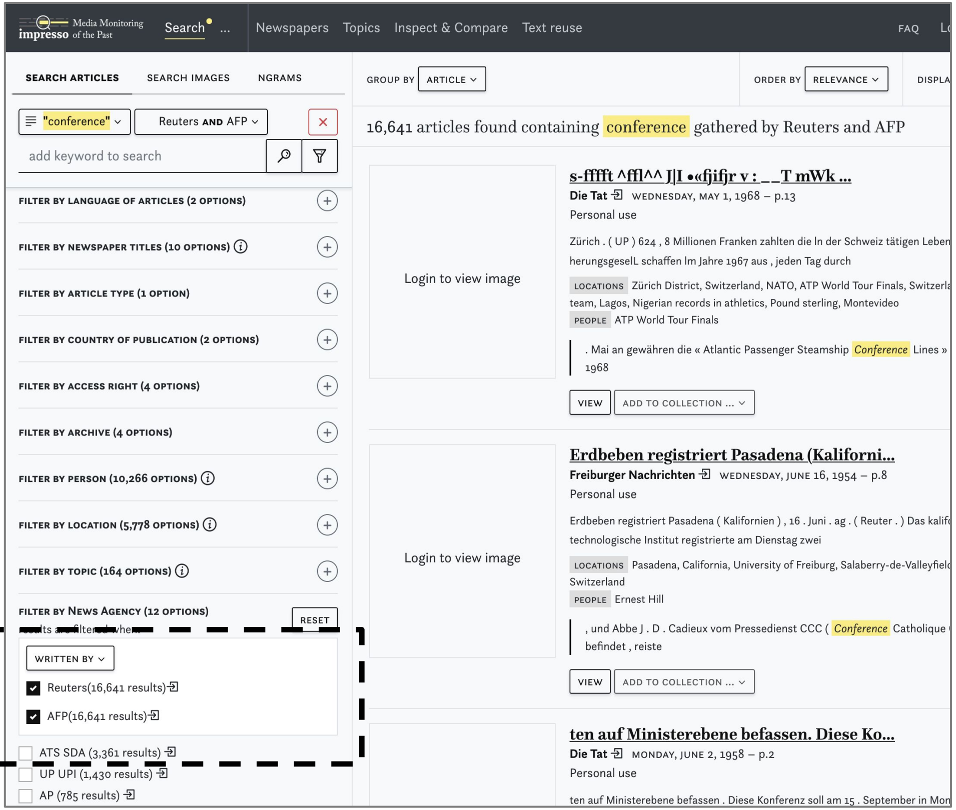
Les suffragistes font de l'action directe
L'indépendance luxembourgeoise TUESDAY, NOVEMBER 29, 1910 – P.2

M. Winston Churchill, revenant à Londres, après avoir prononcé un dis- cours à Bradford, où les suffragettes et leurs partisans avaient déjà violemment manifesté contre lui, a été attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant: «Voilà pour toi, chien!» Deux agents de police, qui accompagnaient M. Churchill, parèrent le coup, et s'em- parèrent de l'homme, après une lutte violente. On croit que l'assaillant est un des suffragistes expulsés de la réunion où M. Winston Churchill venait de parler. A la gare de Londres, trois femmes ont également essayé de frapper M. Chur- chill ; elles en ont été empêchées par les agents.

Where did the news come from?

News Agency Recognition

- Construction of an annotated dataset (27 agencies, ca 2000 articles, fr and de);
- Training and evaluation of models to recognise news agencies;
- Application on the whole *impresso* corpus;
- First analyses.



"conference" | Reuters AND AFP

add keyword to search

FILTER BY LANGUAGE OF ARTICLES (2 OPTIONS)

FILTER BY NEWSPAPER TITLES (10 OPTIONS)

FILTER BY ARTICLE TYPE (1 OPTION)

FILTER BY COUNTRY OF PUBLICATION (2 OPTIONS)

FILTER BY ACCESS RIGHT (4 OPTIONS)

FILTER BY ARCHIVE (4 OPTIONS)

FILTER BY PERSON (10,266 OPTIONS)

FILTER BY LOCATION (5,778 OPTIONS)

FILTER BY TOPIC (164 OPTIONS)

FILTER BY News Agency (12 OPTIONS)

results are filtered when...

WRITTEN BY

☑ Reuters(16,641 results)

☑ AFP(16,641 results)

□ ATS SDA (3,361 results)

□ UP UPI (1,430 results)

□ AP (785 results)

16,641 articles found containing conference gathered by Reuters and AFP

s-fffft ^ffl^^ J|I •«fjifjr v : _ _T mWk ...

Die Tat WEDNESDAY, MAY 1, 1968 – p.13

Personal use

Login to view image

Zürich . (UP) 624 , 8 Millionen Franken zahlten die ln der Schweiz tätigen Leben herungsgesellL schaffen lm Jahre 1967 aus , jeden Tag durch

LOCATIONS Zürich District, Switzerland, NATO, ATP World Tour Finals, Switzerla team, Lagos, Nigerian records in athletics, Pound sterling, Montevideo

PEOPLE ATP World Tour Finals

> . Mai an gewähren die « Atlantic Passenger Steamship *Conference* Lines » 1968

Erdbeben registriert Pasadena (Kaliforni...

Freiburger Nachrichten WEDNESDAY, JUNE 16, 1954 – p.8

Personal use

Login to view image

Erdbeben registriert Pasadena (Kalifornien) , 16 . Juni . ag . (Reuter .) Das kalifo technologische Institut registrierte am Dienstag zwei

LOCATIONS Pasadena, California, University of Freiburg, Salaberry-de-Valleyfiel Switzerland

PEOPLE Ernest Hill

> , und Abbe J . D . Cadieux vom Pressedienst CCC (*Conference* Catholique befindet , reiste

ten auf Ministerebene befassen. Diese Ko...

Die Tat MONDAY, JUNE 2, 1958 – p.2

Personal use

ten auf Ministerebene befassen . Diese Konferenz soll am 15 . September in Mon

News agency mentions

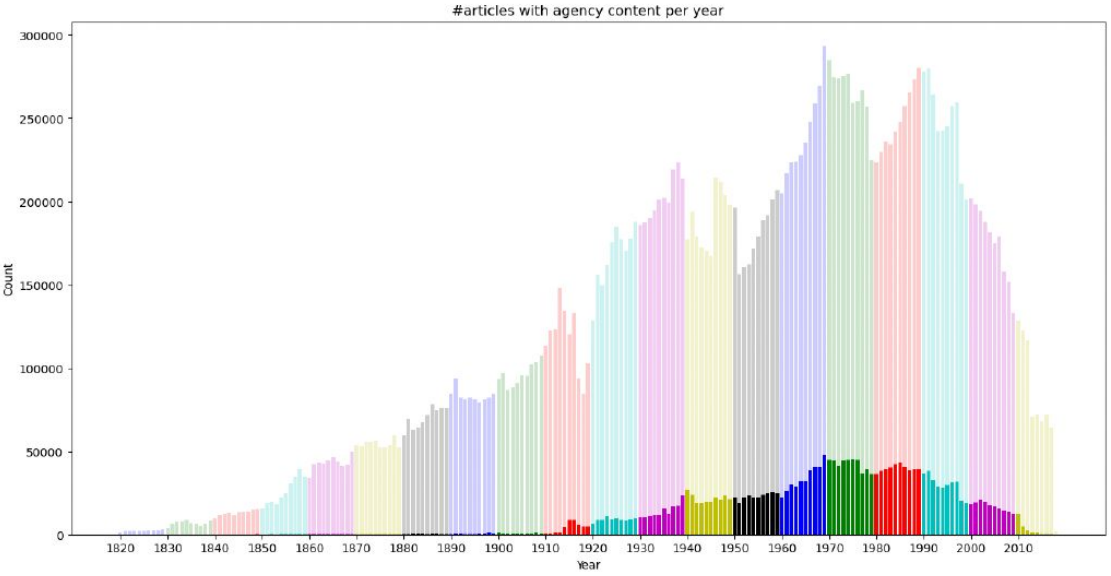


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

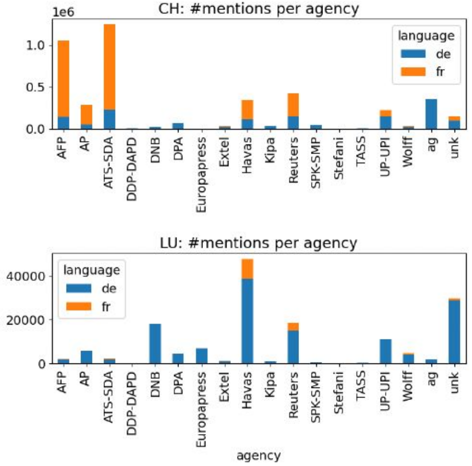


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

In Swiss newspapers

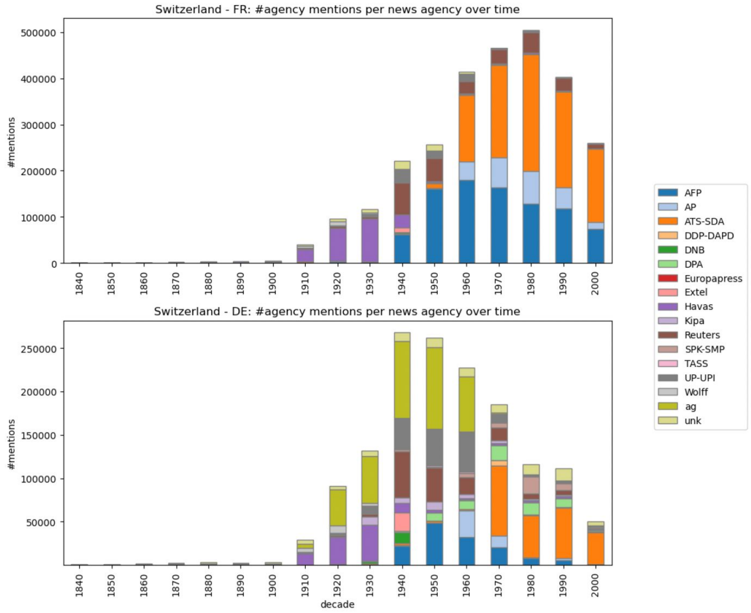


FIGURE 5.8
News agency mentions in Swiss newspapers over time, split by language (French above, German below).

In Luxembourgish newspapers (1940-1944)

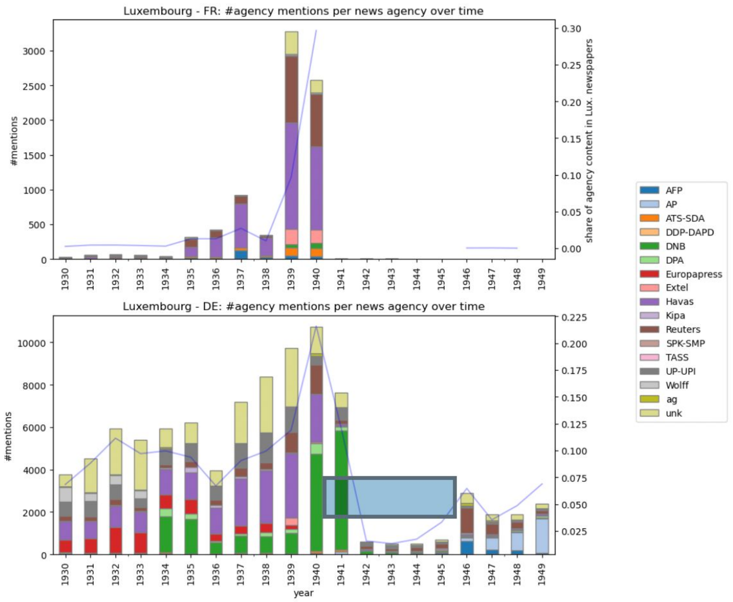
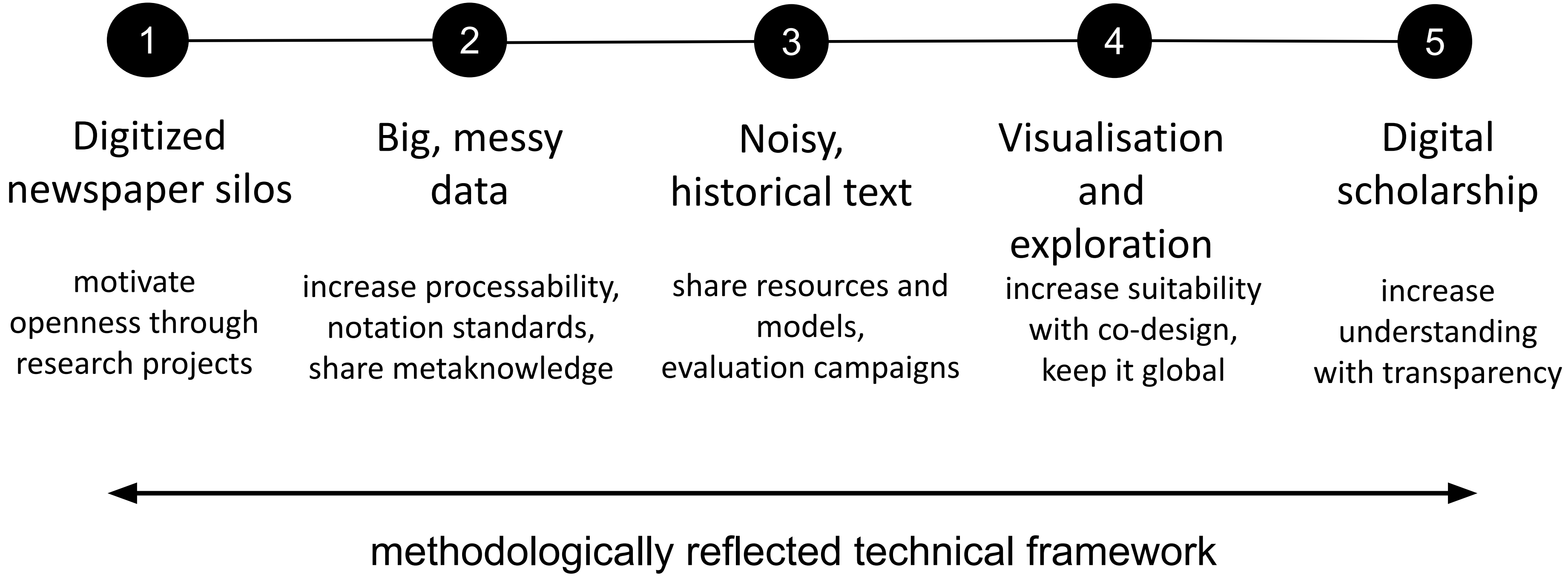


FIGURE 5.15

The development of agency mentions in Luxembourg in the years 1930-1949, split by language (French above, German below). The bars show the distribution of the different agency mentions (left y-axis), while the blue line indicates the general share of articles with agency mentions in the Luxembourgish corpus (right y-axis).

Reliable Semantic Indexing of Historical Newspapers at Scale: Are We There Yet?



1

Digitized newspaper silos

motivate openness through research projects

2

Big, messy data

increase processability, notation standards, share metaknowledge

3

Noisy, historical text

share resources and models, evaluation campaigns

4

Visualisation and exploration

increase suitability with co-design, keep it global

5

Digital scholarship

increase understanding with transparency

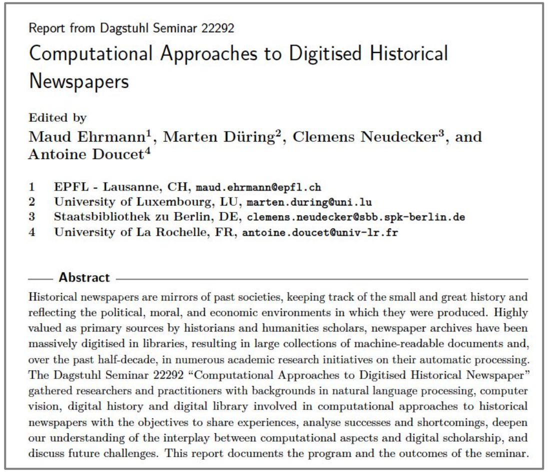
methodologically reflected technical framework

What’s next ? (from the community)

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

What are the main challenges we need to address in relation with historical newspapers?

- Document processing
- Text and image processing
- **Evaluation** (digitisation and content mining)
- Exploration of enriched, **global** collections
- **Working with data**
- **Workflows**
- **Criticism, inclusivity**
- Legal matters



Report from Dagstuhl Seminar 22292

Computational Approaches to Digitised Historical Newspapers

Edited by
Maud Ehrmann[1], Marten Düring[2], Clemens Neudecker[3], and Antoine Doucet[4]

1 EPFL - Lausanne, CH, maud.ehrmann@epfl.ch
2 University of Luxembourg, LU, marten.during@uni.lu
3 Staatsbibliothek zu Berlin, DE, clemens.neudecker@sbb.spk-berlin.de
4 University of La Rochelle, FR, antoine.doucet@univ-lr.fr

Abstract

Historical newspapers are mirrors of past societies, keeping track of the small and great history and reflecting the political, moral, and economic environments in which they were produced. Highly valued as primary sources by historians and humanities scholars, newspaper archives have been massively digitised in libraries, resulting in large collections of machine-readable documents and, over the past half-decade, in numerous academic research initiatives on their automatic processing. The Dagstuhl Seminar 22292 “Computational Approaches to Digitised Historical Newspaper” gathered researchers and practitioners with backgrounds in natural language processing, computer vision, digital history and digital library involved in computational approaches to historical newspapers with the objectives to share experiences, analyse successes and shortcomings, deepen our understanding of the interplay between computational aspects and digital scholarship, and discuss future challenges. This report documents the program and the outcomes of the seminar.

What’s next ? (from impresso)

impresso

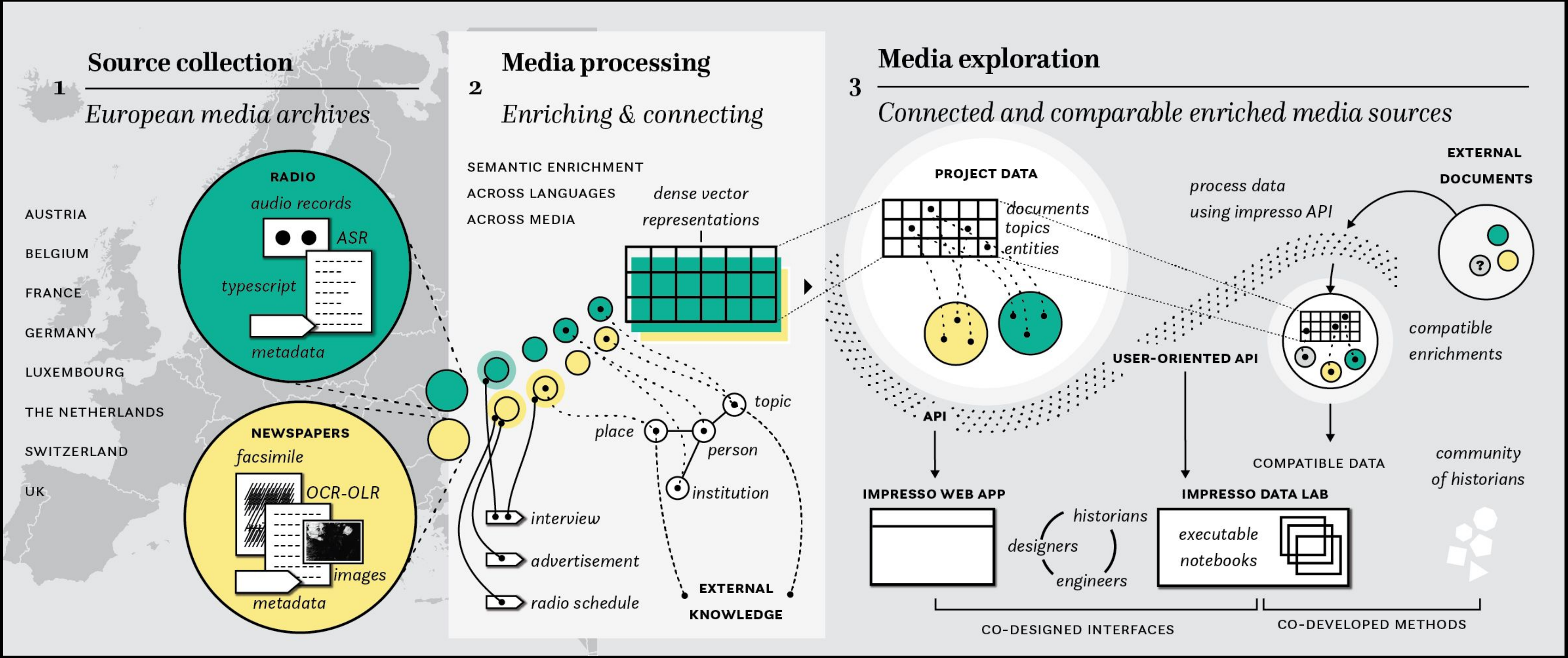
Media Monitoring of the Past

Beyond Borders: Connecting Historical Newspapers and Radio

Main objectives

- Enrichment and integration of **newspaper and radio sources in a single semantic space;**
- Expand the corpus to Western Europe;
- Develop interfaces for exploratory and computational research;
- Conduct case studies in (media) history, theme "influences."

09/2023 - 02/2027



1 Source collection

European media archives

AUSTRIA
BELGIUM
FRANCE
GERMANY
LUXEMBOURG
THE NETHERLANDS
SWITZERLAND
UK

RADIO
audio records
ASR
typescript
metadata

NEWSPAPERS
facsimile
OCR-OLR
images
metadata

2 Media processing

Enriching & connecting

SEMANTIC ENRICHMENT
ACROSS LANGUAGES
ACROSS MEDIA

dense vector representations

place
person
topic
institution

interview
advertisement
radio schedule

EXTERNAL KNOWLEDGE

3 Media exploration

Connected and comparable enriched media sources

PROJECT DATA
documents
topics
entities

API

USER-ORIENTED API

process data using impresso API

EXTERNAL DOCUMENTS

compatible enrichments

COMPATIBLE DATA

community of historians

IMPRESSO WEB APP

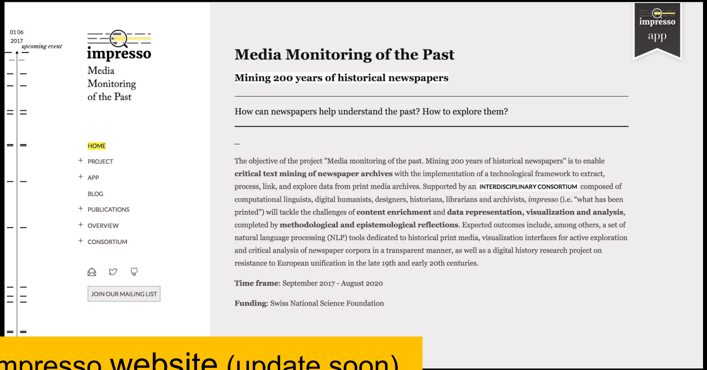
historians
designers
engineers

IMPRESSO DATA LAB
executable notebooks

CO-DESIGNED INTERFACES

CO-DEVELOPED METHODS

For more information



01 06
2017
upcoming event

impresso
Media
Monitoring
of the Past

HOME
+ PROJECT
+ APP
BLOG
+ PUBLICATIONS
+ OVERVIEW
+ CONSORTIUM

JOIN OUR MAILING LIST

impresso app

Media Monitoring of the Past

Mining 200 years of historical newspapers

How can newspapers help understand the past? How to explore them?

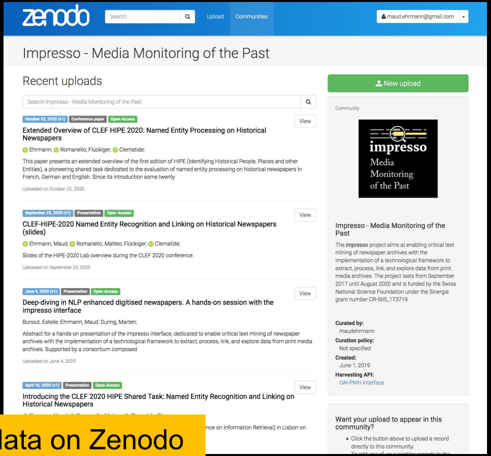
—

The objective of the project "Media monitoring of the past. Mining 200 years of historical newspapers" is to enable **critical text mining of newspaper archives** with the implementation of a technological framework to extract, process, link, and explore data from print media archives. Supported by an INTERDISCIPLINARY CONSORTIUM composed of computational linguists, digital humanists, designers, historians, librarians and archivists, *impresso* (i.e. "what has been printed") will tackle the challenges of **content enrichment** and **data representation, visualization and analysis**, completed by **methodological and epistemological reflections**. Expected outcomes include, among others, a set of natural language processing (NLP) tools dedicated to historical print media, visualization interfaces for active exploration and critical analysis of newspaper corpora in a transparent manner, as well as a digital history research project on resistance to European unification in the late 19th and early 20th centuries.

Time frame: September 2017 - August 2020

Funding: Swiss National Science Foundation

impresso website (update soon)



zenodo Search Upload Communities maud.ehrmann@gmail.com

Impresso - Media Monitoring of the Past

Recent uploads

Search Impresso - Media Monitoring of the Past

October 22, 2020 (v1) Conference paper Open Access View

Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers

Ehrmann; Romanello; Flückiger; Clematide;

This paper presents an extended overview of the first edition of HIPE (Identifying Historical People, Places and other Entities), a pioneering shared task dedicated to the evaluation of named entity processing on historical newspapers in French, German and English. Since its introduction some twenty

Uploaded on October 22, 2020

September 23, 2020 (v1) Presentation Open Access View

CLEF-HIPE-2020 Named Entity Recognition and Linking on Historical Newspapers (slides)

Ehrmann, Maud; Romanello, Matteo; Flückiger; Clematide;

Slides of the HIPE-2020 Lab overview during the CLEF 2020 conference.

Uploaded on September 23, 2020

June 4, 2020 (v1) Presentation Open Access View

Deep-diving in NLP enhanced digitised newspapers. A hands-on session with the impresso interface

Bunout, Estelle; Ehrmann, Maud; During, Marten;

Abstract for a hands-on presentation of the impresso interface, dedicated to enable critical text mining of newspaper archives with the implementation of a technological framework to extract, process, link, and explore data from print media archives. Supported by a consortium composed

Uploaded on June 4, 2020

April 16, 2020 (v1) Presentation Open Access View

Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers

ence on Information Retrieval) in Lisbon on

New upload

Community

impresso Media Monitoring of the Past

Impresso - Media Monitoring of the Past

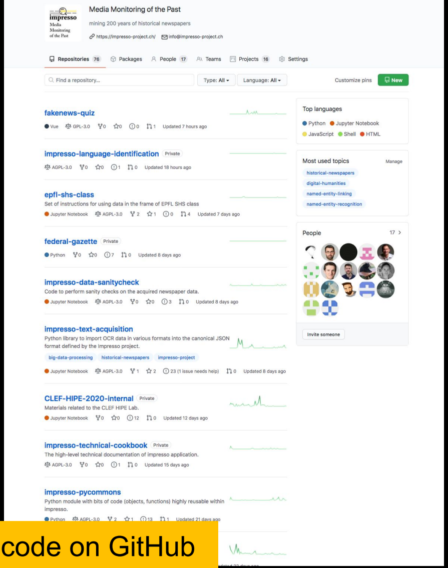
The *impresso* project aims at enabling critical text mining of newspaper archives with the implementation of a technological framework to extract, process, link, and explore data from print media archives. The project lasts from September 2017 until August 2020 and is funded by the Swiss National Science Foundation under the Sinergia grant number CR-SII5_173719.

Curated by: maudehrmann
Curation policy: Not specified
Created: June 1, 2019
Harvesting API: OAI-PMH Interface

Want your upload to appear in this community?

- Click the button above to upload a record directly to this community.

data on Zenodo



impresso Media Monitoring of the Past

Media Monitoring of the Past
mining 200 years of historical newspapers
https://impresso-project.ch/ info@impresso-project.ch

Repositories 76 Packages People 17 Teams Projects 16 Settings

Find a repository... Type: All Language: All Customize pins New

fakenews-quiz
Vue GPL-3.0 0 0 0 1 Updated 7 hours ago

impresso-language-identification Private
AGPL-3.0 0 0 1 0 Updated 18 hours ago

epfl-shs-class
Set of instructions for using data in the frame of EPFL SHS class
Jupyter Notebook AGPL-3.0 2 1 0 4 Updated 7 days ago

federal-gazette Private
Python 0 0 7 0 Updated 8 days ago

impresso-data-sanitycheck
Code to perform sanity checks on the acquired newspaper data.
Jupyter Notebook AGPL-3.0 0 0 3 0 Updated 8 days ago

impresso-text-acquisition
Python library to import OCR data in various formats into the canonical JSON format defined by the Impresso project.
big-data-processing historical-newspapers impresso-project
Jupyter Notebook AGPL-3.0 1 2 23 (1 issue needs help) 0 Updated 8 days ago

CLEF-HIPE-2020-internal Private
Materials related to the CLEF HIPE Lab.
Jupyter Notebook 0 0 12 0 Updated 12 days ago

impresso-technical-cookbook Private
The high-level technical documentation of impresso application.
AGPL-3.0 0 0 1 0 Updated 15 days ago

impresso-pycommons
Python module with bits of code (objects, functions) highly reusable within impresso.
Python AGPL-3.0 2 1 13 1 Updated 21 days ago

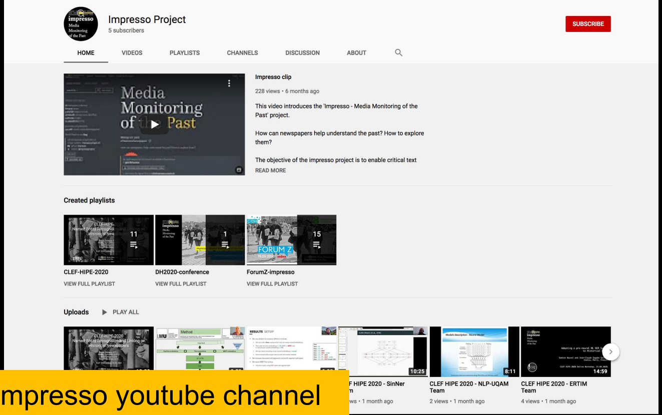
Top languages
● Python ● Jupyter Notebook ● JavaScript ● Shell ● HTML

Most used topics Manage
historical-newspapers
digital-humanities
named-entity-linking
named-entity-recognition

People 17

Invite someone

code on GitHub



impresso Media Monitoring of the Past

Impresso Project
5 subscribers

SUBSCRIBE

HOME VIDEOS PLAYLISTS CHANNELS DISCUSSION ABOUT

Impresso clip
228 views • 6 months ago

This video introduces the 'Impresso - Media Monitoring of the Past' project.

How can newspapers help understand the past? How to explore them?

The objective of the impresso project is to enable critical text

READ MORE

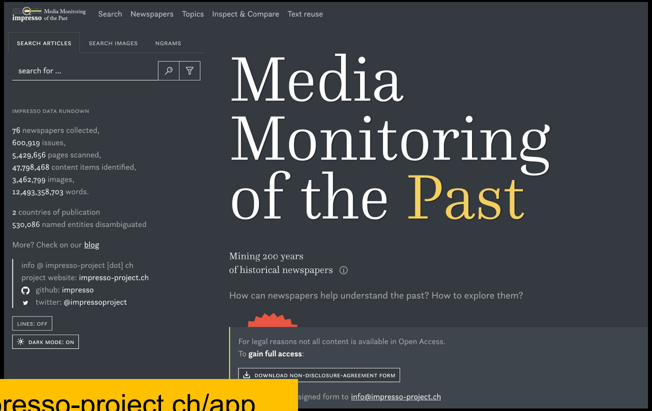
Created playlists

CLEF-HIPE-2020 (11) VIEW FULL PLAYLIST
DH2020-conference (1) VIEW FULL PLAYLIST
ForumZ-impresso (15) VIEW FULL PLAYLIST

Uploads PLAY ALL

F HIPE 2020 - SinNer m — 10:25 — ws • 1 month ago
CLEF HIPE 2020 - NLP-UQAM Team — 8:11 — 2 views • 1 month ago
CLEF HIPE 2020 - ERTIM Team — 14:59 — 4 views • 1 month ago

impresso youtube channel



impresso Media Monitoring of the Past

Search Newspapers Topics Inspect & Compare Text reuse

SEARCH ARTICLES SEARCH IMAGES NGRAMS

search for ...

IMPRESSO DATA RUNDOWN

76 newspapers collected,
600,919 issues,
5,429,656 pages scanned,
47,798,468 content items identified,
3,462,799 images,
12,493,358,703 words.

2 countries of publication
530,086 named entities disambiguated

More? Check on our blog

info @ impresso-project [dot] ch
project website: impresso-project.ch
github: impresso
twitter: @impressoproject

LINES: OFF
DARK MODE: ON

Media Monitoring of the Past

Mining 200 years
of historical newspapers

How can newspapers help understand the past? How to explore them?

For legal reasons not all content is available in Open Access.
To **gain full access**:

DOWNLOAD NON-DISCLOSURE-AGREEMENT FORM

signed form to info@impresso-project.ch

impresso-project.ch/app