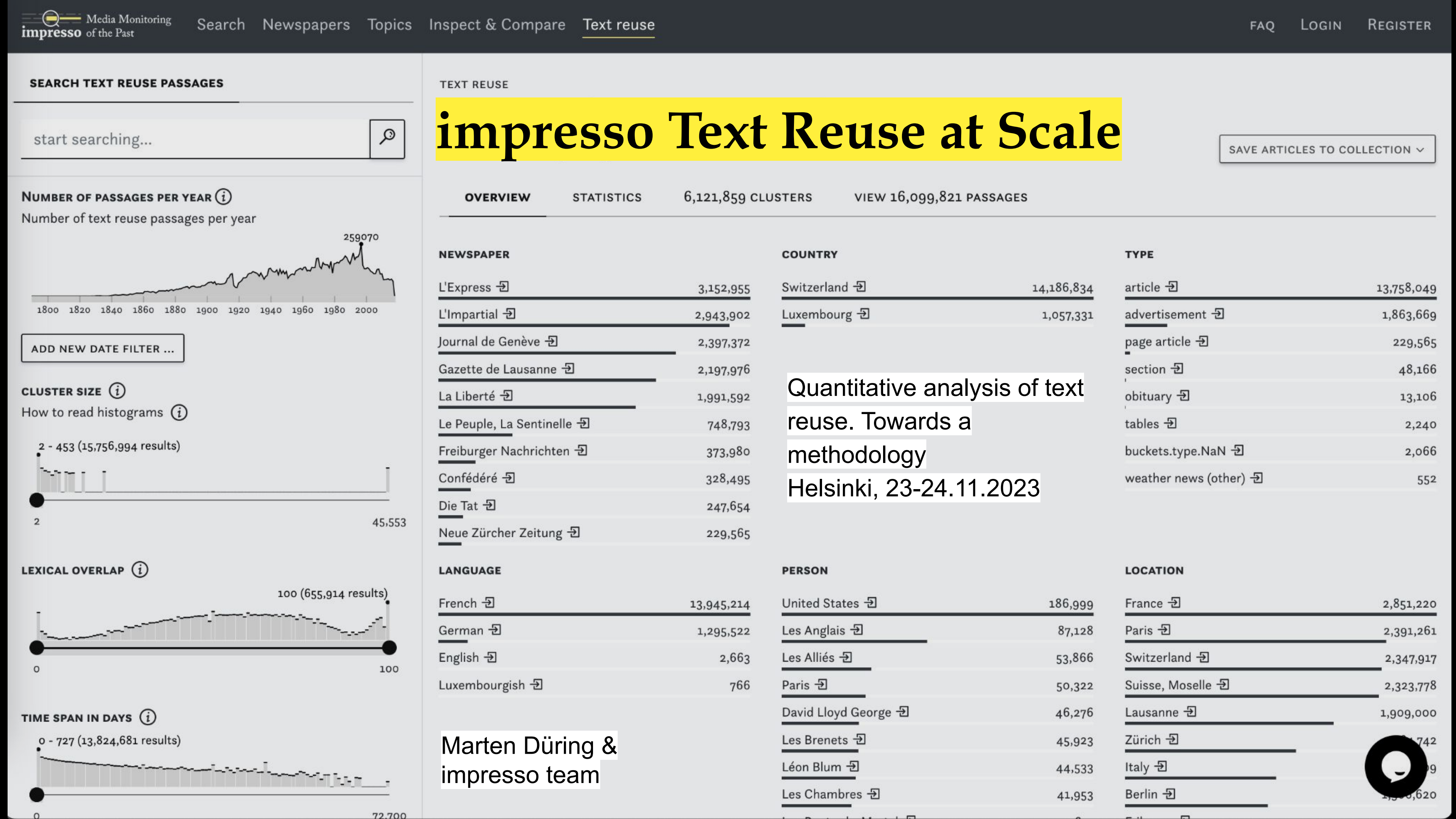


TEXT REUSE

impresso Text Reuse at Scale

Quantitative analysis of text reuse. Towards a methodology
Helsinki, 23-24.11.2023

Marten Düring &
impresso team

OVERVIEW | STATISTICS | 6,121,859 CLUSTERS | VIEW 16,099,821 PASSAGES

Number of passages per year

Number of text reuse passages per year

Cluster size

How to read histograms

2 - 453 (15,756,994 results)

Lexical overlap

100 (655,914 results)

Time span in days

0 - 727 (13,824,681 results)

NEWSPAPER	
L'Express	3,152,955
L'Impartial	2,943,902
Journal de Genève	2,397,372
Gazette de Lausanne	2,197,976
La Liberté	1,991,592
Le Peuple, La Sentinelle	748,793
Freiburger Nachrichten	373,980
Confédéré	328,495
Die Tat	247,654
Neue Zürcher Zeitung	229,565

COUNTRY	
Switzerland	14,186,834
Luxembourg	1,057,331

TYPE	
article	13,758,049
advertisement	1,863,669
page article	229,565
section	48,166
obituary	13,106
tables	2,240
buckets.type.NaN	2,066
weather news (other)	552

LANGUAGE	
French	13,945,214
German	1,295,522
English	2,663
Luxembourgish	766

PERSON	
United States	186,999
Les Anglais	87,128
Les Alliés	53,866
Paris	50,322
David Lloyd George	46,276
Les Brenets	45,923
Léon Blum	44,533
Les Chambres	41,953

LOCATION	
France	2,851,220
Paris	2,391,261
Switzerland	2,347,917
Suisse, Moselle	2,323,778
Lausanne	1,909,000
Zürich	[illegible],742
Italy	[illegible]9
Berlin	1,[illegible],620

impresso - Media Monitoring of the Past II

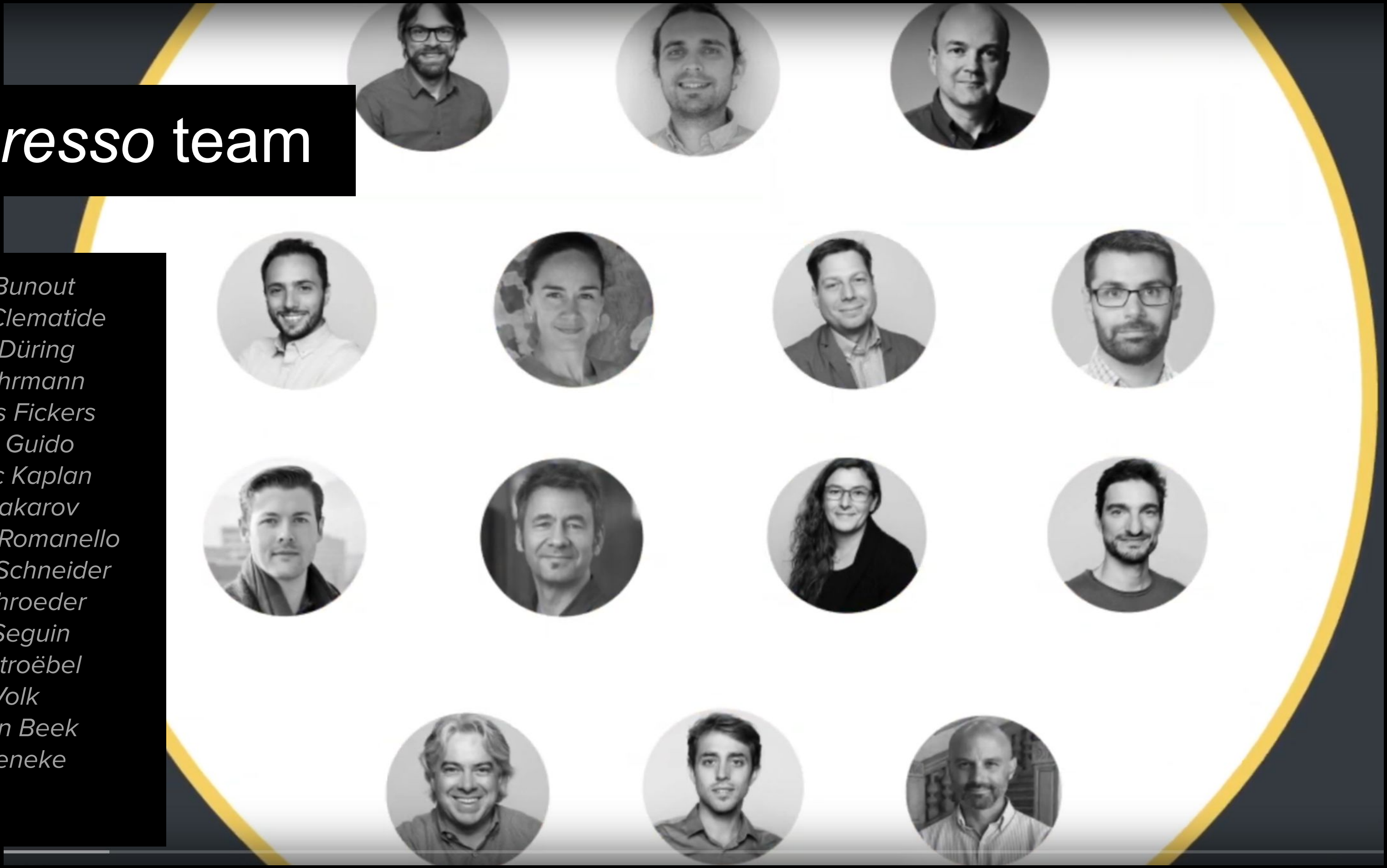
Beyond Borders: Connecting Historical Newspapers and Radio

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

How can semantic enrichments enhance the study of digitised historical newspapers?

impresso team

Estelle Bunout
Simon Clematide
Marten Düring
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



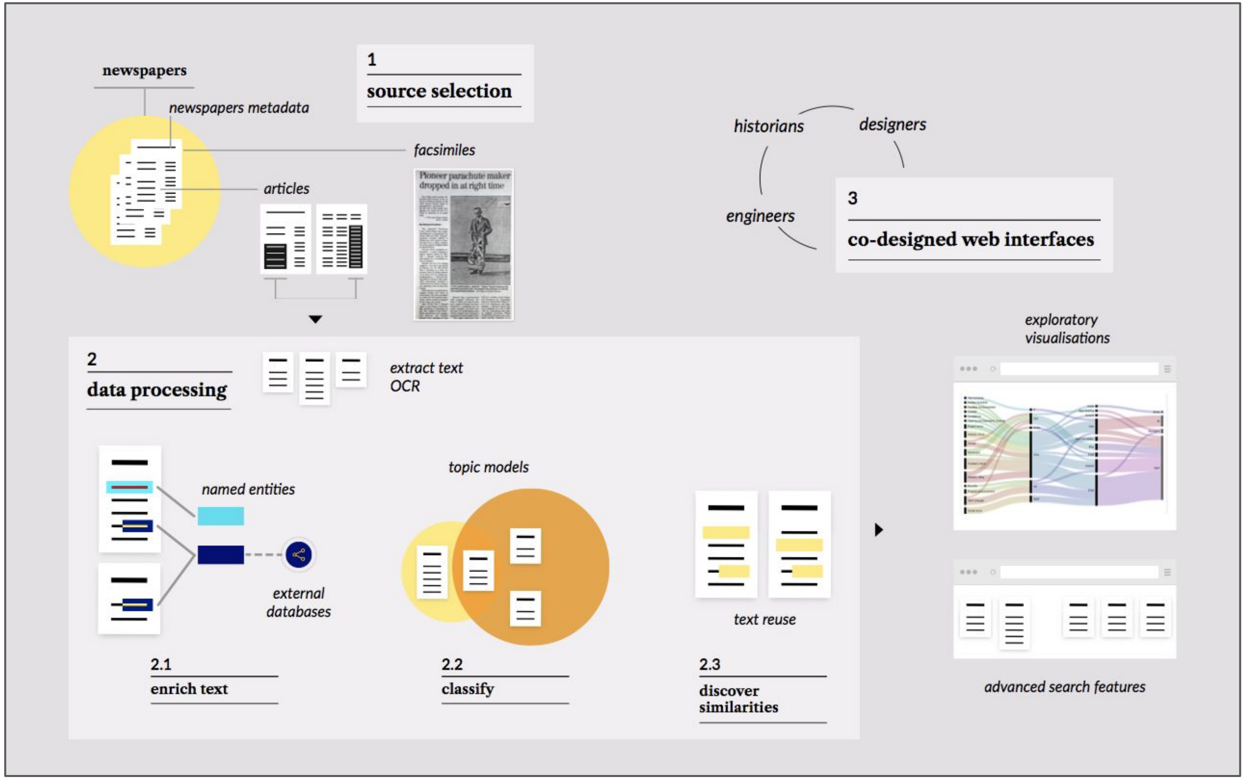
+ a team of historical advisors and associated researchers in (media) history

Objectives and research questions

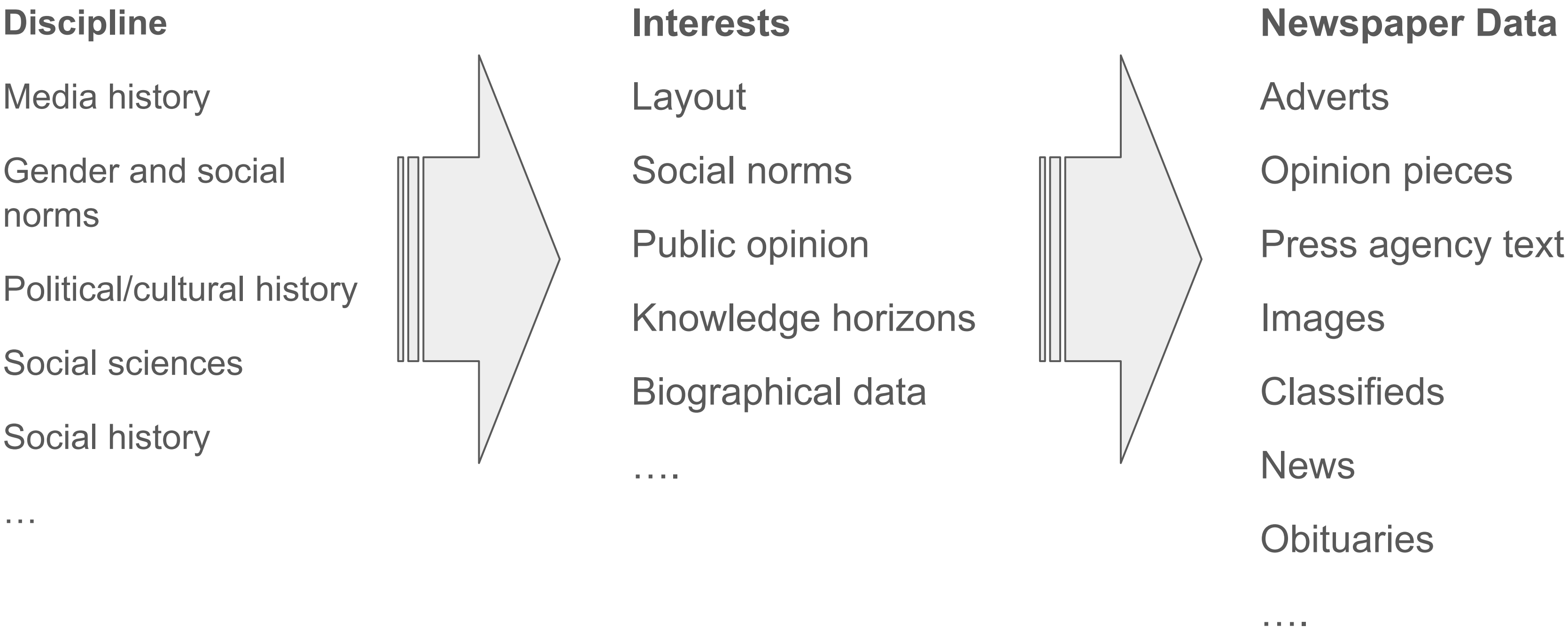
1. How to adapt NLP tools to historical texts?

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

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



Research interests



Discipline

Media history

Gender and social norms

Political/cultural history

Social sciences

Social history

…

Interests

Layout

Social norms

Public opinion

Knowledge horizons

Biographical data

….

Newspaper Data

Adverts

Opinion pieces

Press agency text

Images

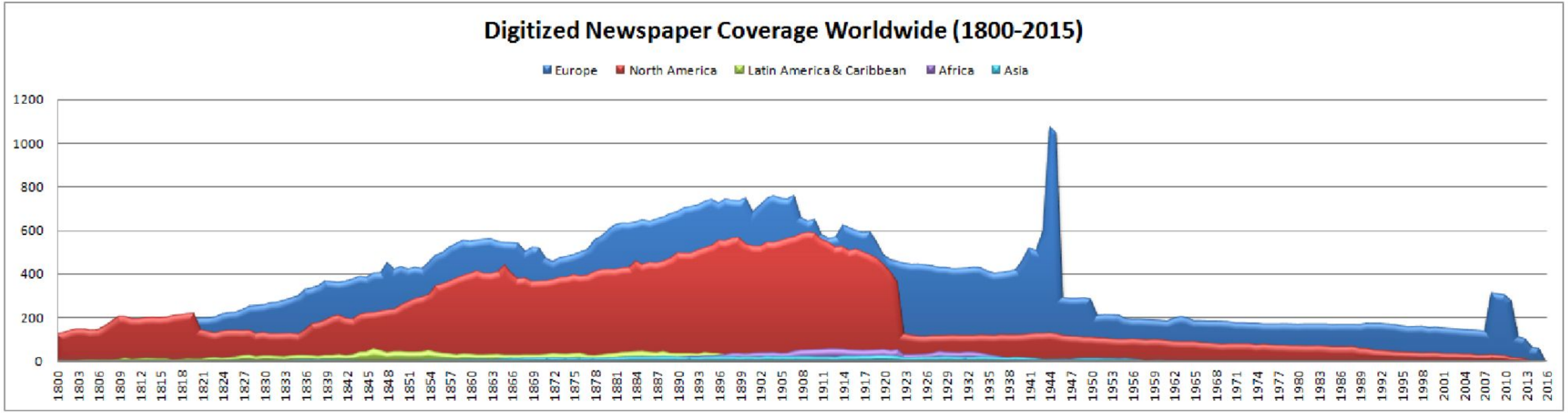
Classifieds

News

Obituaries

….

The big tip of a hidden iceberg



Centre for Research Libraries. 2015. *The "State of the Art". A Comparative Analysis of Newspaper Digitization to Date*.

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

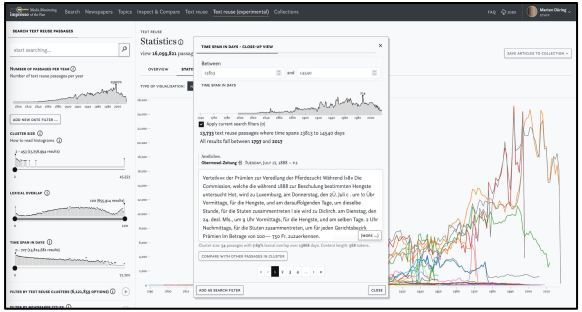
The impresso app

Based on digitised newspaper collections provided by partners.

Target audience are historians and other humanities scholars.

Iterative exploration from multiple perspectives, not an analytical workbench.

Interfaces for overview, exploration, limits & opportunities, flaws in the data.



Number of text reuse passages per year

2 - 453 (15,756,994 results)

100 (655,914 results)

0 - 727 (13,824,681 results)

view **16,099,821** passag

Between 13813 and 14540

☑ Apply current search filters (0)

13,733 text reuse passages where time spans 13813 to 14540 days

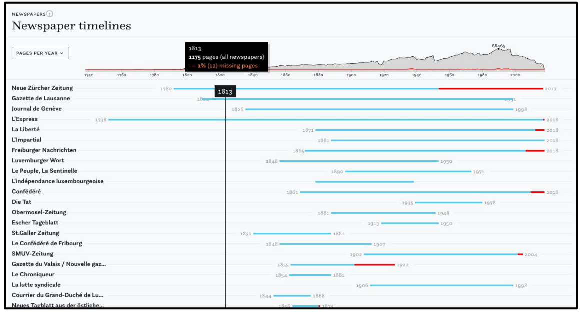
All results fall between **1797** and **2017**

Amtliches.

Obermosel-Zeitung TUESDAY, JULY 17, 1888 – P.1

Verteil«« der Prümlen zur Veredlung der Pferdezucht Während l»8» Die Commission, welche die während 1888 zur Beschulung bestimmten Hengste untersucht Hot, wird zu Luxemburg, am Donnerstag, den 2Ü. Juli c . um !o Übr Vormittags, für die Hengste, und am darauffolgenden Tage, um dieselbe Stunde, für die Stuten zusammentreten ! sie wird zu Diclirch. am Dienstag, den 24. desl. Mls., um 9 Uhr Vormittags, für die Hengste, und am selben Tage. 2 Uhr Nachmittags, für die Stuten zusammentreten, um für jeden Gerichtsbezirk Prämien Im Betrage von 100— 750 Fr. zuzuerkennen.

Cluster size: **54** passages with **7.69%** lexical overlap over **13868** days. Content length: **568** tokens.

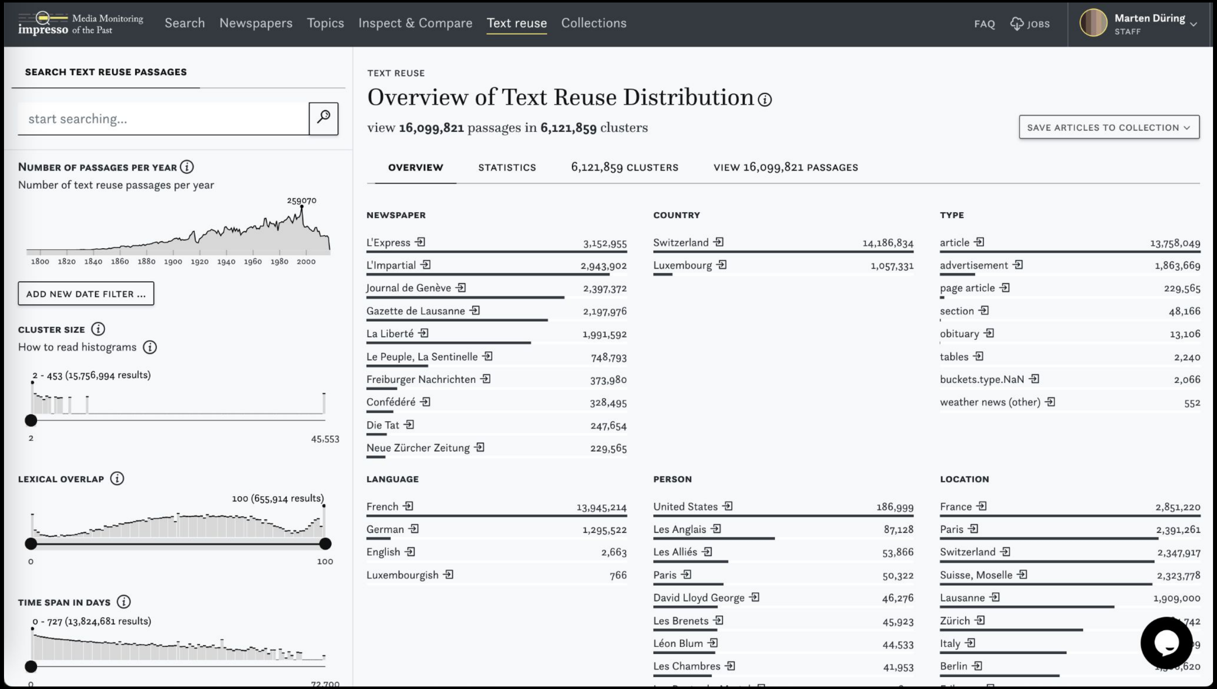


Newspaper timelines

1813
1175 pages (all newspapers)
— **1%** (12) missing pages

Neue Zürcher Zeitung	1780	2017
Gazette de Lausanne	1804	1991
Journal de Genève	1826	1998
L'Express	1738	2018
La Liberté	1871	2018
L'Impartial	1881	2018
Freiburger Nachrichten	1865	2018
Luxemburger Wort	1848	1950
Le Peuple, La Sentinelle	1890	1971
L'indépendance luxembourgeoise		
Confédéré	1861	2018
Die Tat	1935	1978
Obermosel-Zeitung	1881	1948
Escher Tageblatt	1913	1950
St.Galler Zeitung	1831	1881
Le Confédéré de Fribourg	1848	1907
SMUV-Zeitung	1902	2004
Gazette du Valais / Nouvelle gaz...	1855	1922
Le Chroniqueur	1854	1881
La lutte syndicale	1906	1998
Courrier du Grand-Duché de Lu...	1844	1868
Neues Tagblatt aus der östliche...	1856	1874

About Text reuse at Scale



SEARCH TEXT REUSE PASSAGES

start searching...

NUMBER OF PASSAGES PER YEAR

Number of text reuse passages per year

ADD NEW DATE FILTER ...

CLUSTER SIZE

How to read histograms

LEXICAL OVERLAP

TIME SPAN IN DAYS

TEXT REUSE

Overview of Text Reuse Distribution

view **16,099,821** passages in **6,121,859** clusters

SAVE ARTICLES TO COLLECTION

OVERVIEW | STATISTICS | 6,121,859 CLUSTERS | VIEW 16,099,821 PASSAGES

NEWSPAPER	
L'Express	3,152,955
L'Impartial	2,943,902
Journal de Genève	2,397,372
Gazette de Lausanne	2,197,976
La Liberté	1,991,592
Le Peuple, La Sentinelle	748,793
Freiburger Nachrichten	373,980
Confédéré	328,495
Die Tat	247,654
Neue Zürcher Zeitung	229,565

COUNTRY	
Switzerland	14,186,834
Luxembourg	1,057,331

TYPE	
article	13,758,049
advertisement	1,863,669
page article	229,565
section	48,166
obituary	13,106
tables	2,240
buckets.type.NaN	2,066
weather news (other)	552

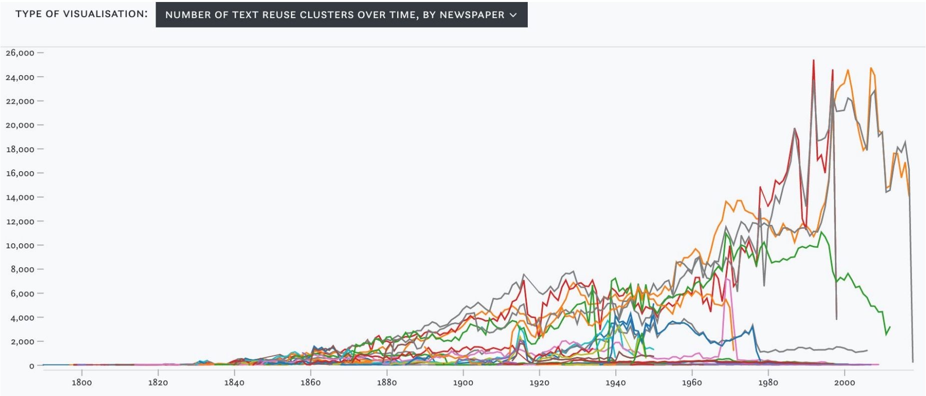
LANGUAGE	
French	13,945,214
German	1,295,522
English	2,663
Luxembourgish	766

PERSON	
United States	186,999
Les Anglais	87,128
Les Alliés	53,866
Paris	50,322
David Lloyd George	46,276
Les Brenets	45,923
Léon Blum	44,533
Les Chambres	41,953

LOCATION	
France	2,851,220
Paris	2,391,261
Switzerland	2,347,917
Suisse, Moselle	2,323,778
Lausanne	1,909,000
Zürich	[illegible]742
Italy	[illegible]9
Berlin	1,[illegible],620

impresso text reuse data

Using passim and applied to 76 newspapers from CH and LU; yielded 16 mio passages in 6 mio clusters



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

Romanello, Matteo. *Text re-use detection in a nutshell*, Blog post, impresso, 2018 <http://impresso-project.ch/news/2018/06/12/tradingzone-tr.html>.

Matteo Romanello and Simon Hengchen, "Detecting Text Reuse with Passim," *Programming Historian* 10 (2021), https://doi.org/10.46430/phen0092.

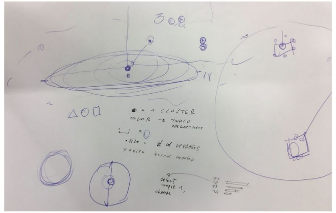
How can we enable scalable reading of text reuse data within the impresso web app?

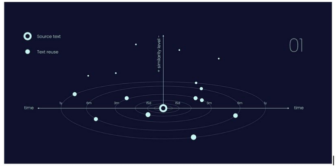
Workshop

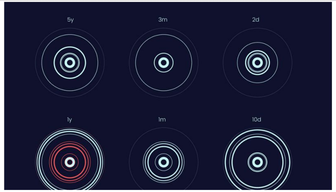
Funded by a mini grant by C2DH to conduct experimental research.

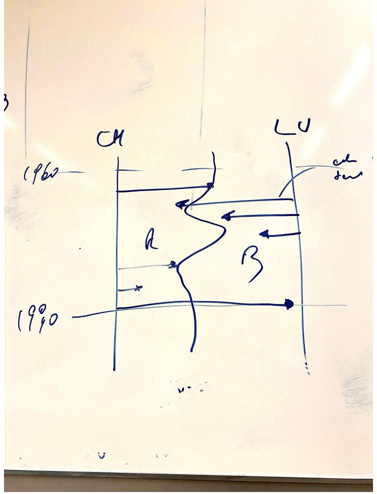
Workshop with 10 NLP, history, design experts in Nov 2022 which produced:

- A set of historical research objectives based on case studies
- Focus on the integration of TR with other semantic enrichments
- Three mockups to support envisioned tasks





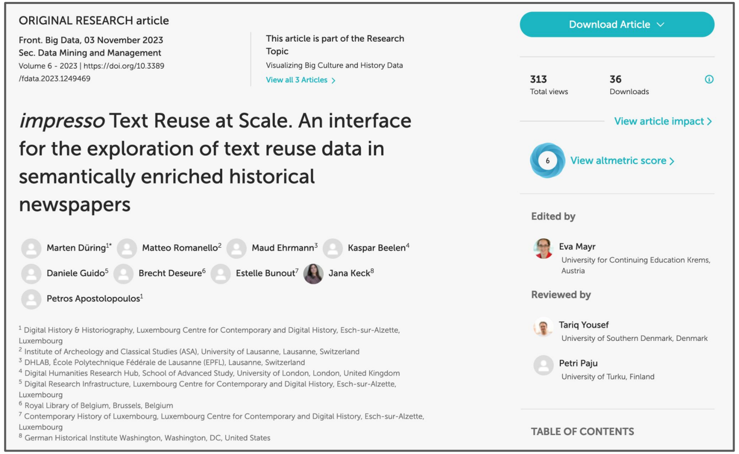




Outputs

Mockups, objectives and tasks inspired:

- Text reuse at Scale interface for impresso Web app
- Press agency detection project (impresso + Lea Marxen)
- User evaluation (Zenodo)
- Paper: *impresso* Text Reuse at Scale (Frontiers in Big Data, 6/2023)



ORIGINAL RESEARCH article

Front. Big Data, 03 November 2023
Sec. Data Mining and Management
Volume 6 - 2023 | https://doi.org/10.3389/fdata.2023.1249469

This article is part of the Research Topic
Visualizing Big Culture and History Data
View all 3 Articles

313 Total views
36 Downloads

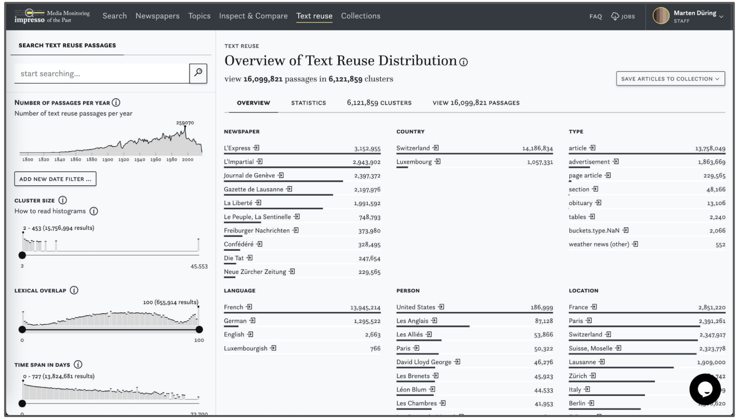
impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers

Marten Düring[1*] Matteo Romanello[2] Maud Ehrmann[3] Kaspar Beelen[4] Daniele Guido[5] Brecht Deseure[6] Estelle Bunout[7] Jana Keck[8] Petros Apostolopoulos[1]

[1] Digital History & Historiography, Luxembourg Centre for Contemporary and Digital History, Esch-sur-Alzette, Luxembourg
[2] Institute of Archeology and Classical Studies (ASA), University of Lausanne, Lausanne, Switzerland
[3] DHLAB, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[4] Digital Humanities Research Hub, School of Advanced Study, University of London, London, United Kingdom
[5] Digital Research Infrastructure, Luxembourg Centre for Contemporary and Digital History, Esch-sur-Alzette, Luxembourg
[6] Royal Library of Belgium, Brussels, Belgium
[7] Contemporary History of Luxembourg, Luxembourg Centre for Contemporary and Digital History, Esch-sur-Alzette, Luxembourg
[8] German Historical Institute Washington, Washington, DC, United States

Edited by
Eva Mayr
University for Continuing Education Krems, Austria

Reviewed by
Tariq Yousef
University of Southern Denmark, Denmark
Petri Paju
University of Turku, Finland



SEARCH TEXT REUSE PASSAGES

NUMBER OF PASSAGES PER YEAR
Number of text reuse passages per year

CLUSTER SIZE
How to read histograms
2 - 453 (15,756,994 results)

LEXICAL OVERLAP
100 (655,914 results)

TIME SPAN IN DAYS
0 - 727 (13,824,681 results)

TEXT REUSE

Overview of Text Reuse Distribution

view **16,099,821** passages in **6,121,859** clusters

NEWSPAPER		COUNTRY		TYPE	
L'Express	3,152,955	Switzerland	14,186,834	article	13,758,049
L'Impartial	2,943,902	Luxembourg	1,057,331	advertisement	1,863,669
Journal de Genève	2,397,372			page article	229,565
Gazette de Lausanne	2,197,976			section	48,166
La Liberté	1,991,592			obituary	13,106
Le Peuple, La Sentinelle	748,793			tables	2,240
Freiburger Nachrichten	373,980			buckets.type.NaN	2,066
Confédéré	328,495			weather news (other)	552
Die Tat	247,654				
Neue Zürcher Zeitung	229,565				

LANGUAGE		PERSON		LOCATION	
French	13,945,214	United States	186,999	France	2,851,220
German	1,295,522	Les Anglais	87,128	Paris	2,391,261
English	2,663	Les Alliés	53,866	Switzerland	2,347,917
Luxembourgish	766	Paris	50,322	Suisse, Moselle	2,323,778
		David Lloyd George	46,276	Lausanne	1,909,000
		Les Brenets	45,923	Zürich	[illegible]
		Léon Blum	44,533	Italy	[illegible]
		Les Chambres	41,953	Berlin	[illegible]

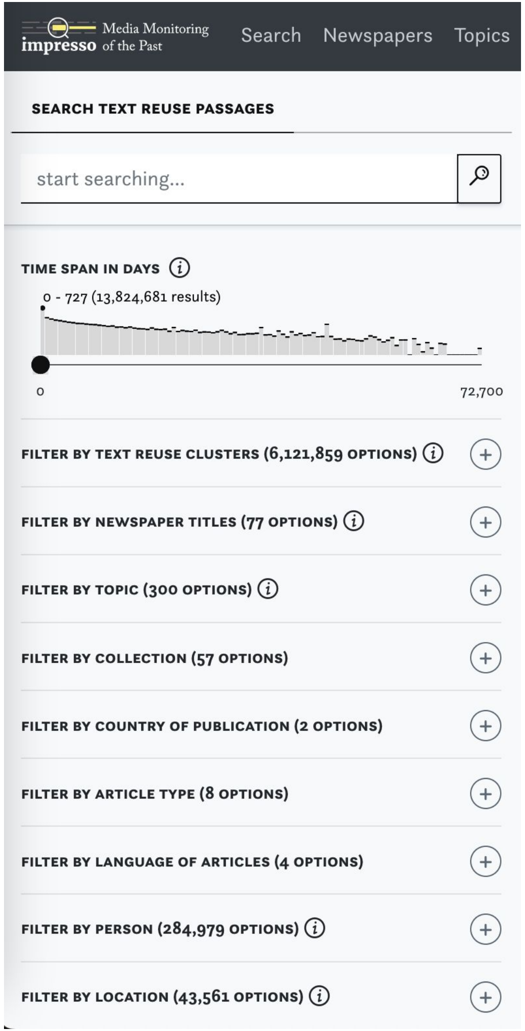
Prios and desiderata

Main Priorities

1. Scalable reading of text reuse data
2. Integration in impresso Web app and its component to enable general exploration
3. Integration with other semantic enrichments for more precision
~~4. Keep Frontiers deadline~~

Main Desiderata

1. Passage-based search
2. Passage-based comparison
~~3. Cluster-level view~~
4. TR data export



SEARCH TEXT REUSE PASSAGES

TIME SPAN IN DAYS

0 - 727 (13,824,681 results)

FILTER BY TEXT REUSE CLUSTERS (6,121,859 OPTIONS)

FILTER BY NEWSPAPER TITLES (77 OPTIONS)

FILTER BY TOPIC (300 OPTIONS)

FILTER BY COLLECTION (57 OPTIONS)

FILTER BY COUNTRY OF PUBLICATION (2 OPTIONS)

FILTER BY ARTICLE TYPE (8 OPTIONS)

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

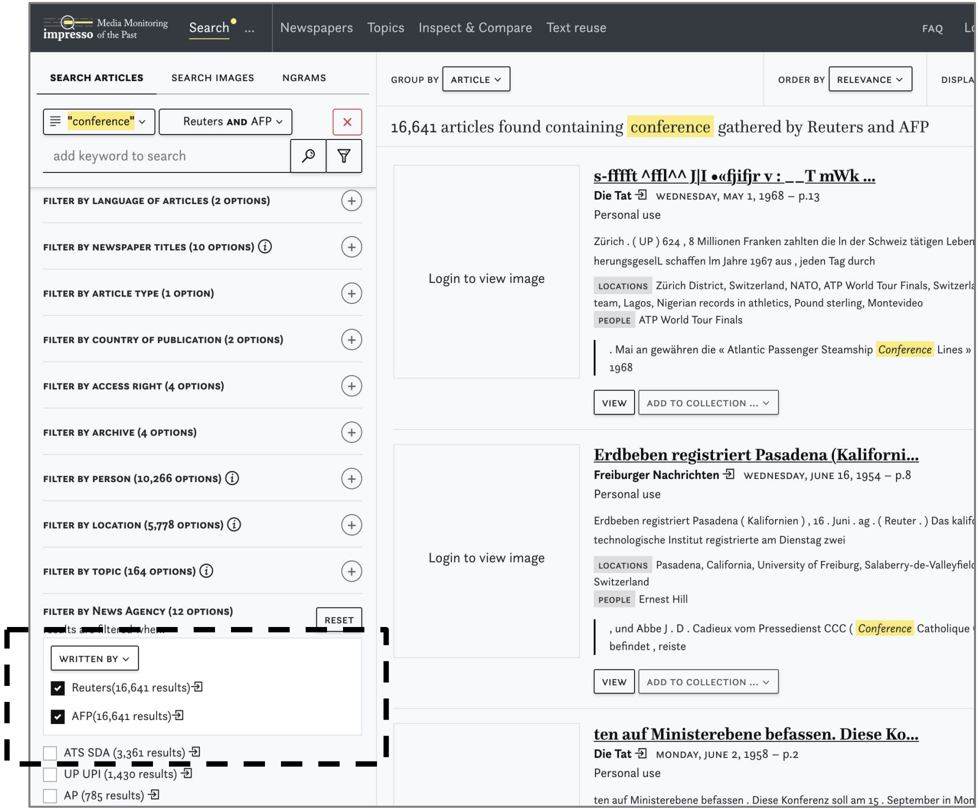
FILTER BY PERSON (284,979 OPTIONS)

FILTER BY LOCATION (43,561 OPTIONS)

Side project: 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.



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

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

☑ Reuters(16,641 results)
☑ AFP(16,641 results)
□ ATS SDA (3,361 results)
□ UP UPI (1,430 results)
□ AP (785 results)

s-fffft ^ffl^^ J|I •«fjifjr v : _ _T mWk ...
Die Tat WEDNESDAY, MAY 1, 1968 – p.13
Personal use

Zürich . (UP) 624 , 8 Millionen Franken zahlten die ln der Schweiz tätigen Leben herungsgeselL 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

Erdbeben registriert Pasadena (Kalifornien) , 16 . Juni . ag . (Reuter .) Das kalif 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

Side project: News agency mentions

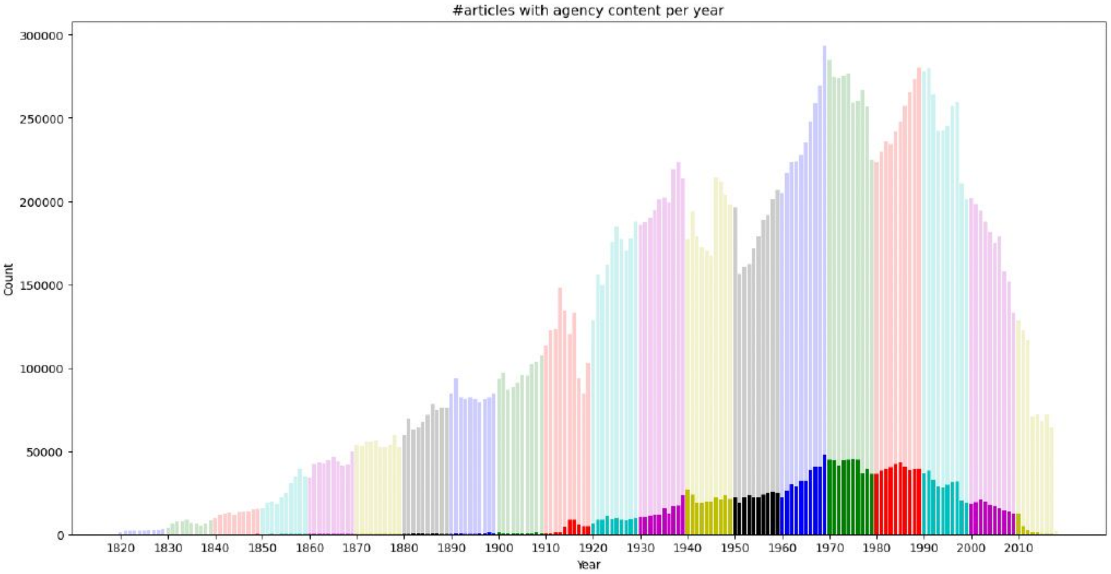


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

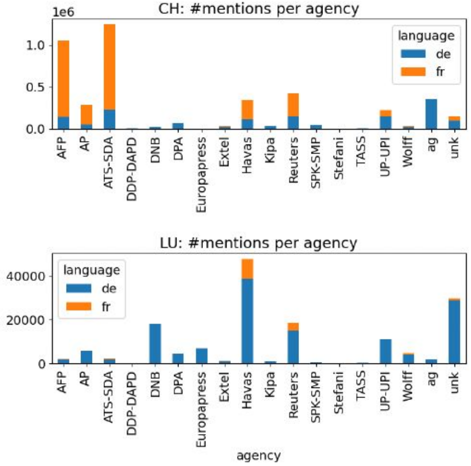


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

Side project: In Swiss newspapers

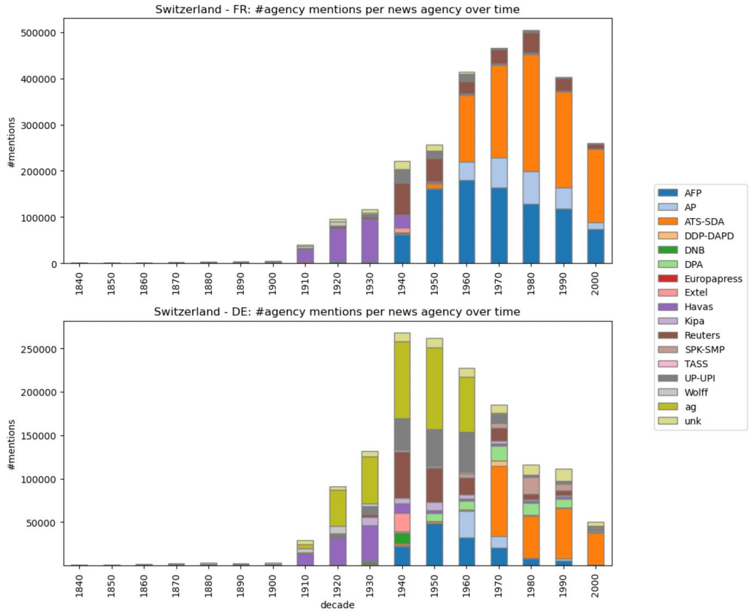
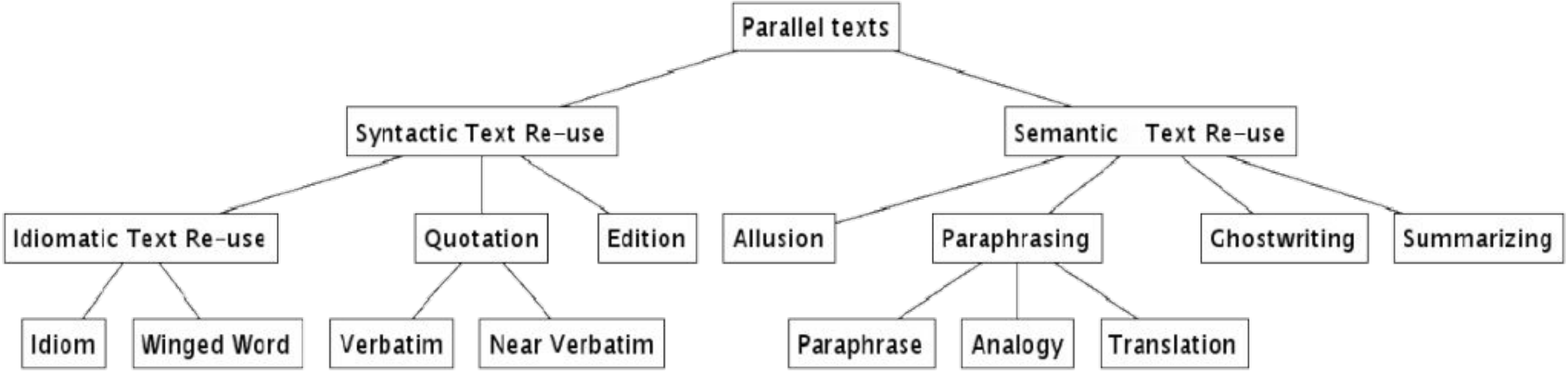


FIGURE 5.8

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

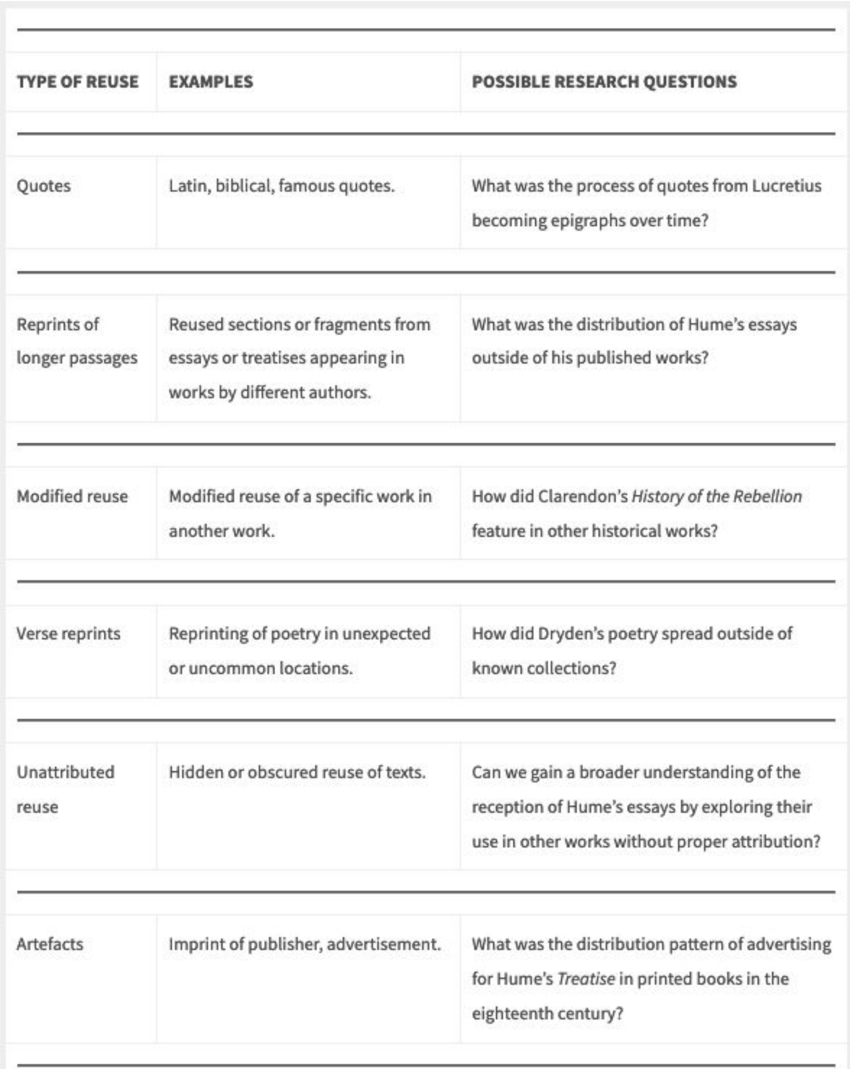
Why do we want to work TR in newspapers?

Types of Text Reuse



Marco Büchler, Historical text reuse: what is it?, https://www.etrap.eu/historical-text-re-use/

Types of Text Reuse



TYPE OF REUSE	EXAMPLES	POSSIBLE RESEARCH QUESTIONS
Quotes	Latin, biblical, famous quotes.	What was the process of quotes from Lucretius becoming epigraphs over time?
Reprints of longer passages	Reused sections or fragments from essays or treatises appearing in works by different authors.	What was the distribution of Hume's essays outside of his published works?
Modified reuse	Modified reuse of a specific work in another work.	How did Clarendon's *History of the Rebellion* feature in other historical works?
Verse reprints	Reprinting of poetry in unexpected or uncommon locations.	How did Dryden's poetry spread outside of known collections?
Unattributed reuse	Hidden or obscured reuse of texts.	Can we gain a broader understanding of the reception of Hume's essays by exploring their use in other works without proper attribution?
Artefacts	Imprint of publisher, advertisement.	What was the distribution pattern of advertising for Hume's *Treatise* in printed books in the eighteenth century?

Rosson et al. Reception Reader: Exploring Text Reuse in Early Modern British Publications, 2023, https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.101

	Research questions	Purpose of TR	Examples	Match with Rosson et al.
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.	
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries	Reprints, verse reprints
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022	
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.	Modified reuse
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts	quotes
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism	Unattributed reuse
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication	Artefacts

	Research questions	Purpose of TR	Examples
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

	Research questions	**Purpose of TR**	**Examples**
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

	Research questions	**Purpose of TR**	**Examples**
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

	Research questions	Purpose of TR	Examples
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

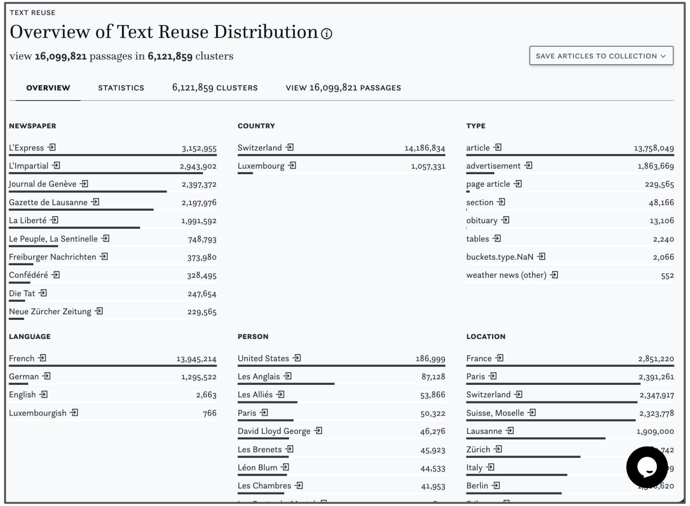
	Research questions	**Purpose of TR**	**Examples**
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

	Research questions	**Purpose of TR**	**Examples**
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

	Research questions	Purpose of TR	Examples
Media ecosystems	How did historical media function? How did transnational information flow shape historical media?	TR as indicator of content flow across titles and borders.	Train and telegraph lines predict the travel of news items.
Bricolage	Where does content in historical newspapers come from?	TR to identify the patterns and fragments which constitute content and their evolution over time.	Death notices, weather reports, event anniversaries
Virality	Which content spreads?	TR as an indicator of the distribution of content within media across time, space, titles.	Paju et al. 2022
Events	How did journalists shape the content they published?	TR to reconstruct the changes in texts spreading across media along editorial lines.	Adjustments of press agency content: title, adjectives, omissions, additions.
Zeitgeist	How did ideas co-evolve?	TR to trace co-evolving ideas.	“A shared way” to write text, design adverts
Unattributed reuse*	Which content circulated without attribution?	TR as evidence for undeclared reuse.	Plagiarism
Data cleaning*	When does TR reduce data quality?	TR as indicators of unwanted duplicates	Mastheads, co-publication

What do we want to do with text reuse data?

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



TEXT REUSE

Overview of Text Reuse Distribution

view **16,099,821** passages in **6,121,859** clusters

NEWSPAPER	
L'Express	3,152,955
L'Impartial	2,943,902
Journal de Genève	2,397,372
Gazette de Lausanne	2,197,976
La Liberté	1,991,592
Le Peuple, La Sentinelle	748,793
Freiburger Nachrichten	373,980
Confédéré	328,495
Die Tat	247,654
Neue Zürcher Zeitung	229,565

COUNTRY	
Switzerland	14,186,834
Luxembourg	1,057,331

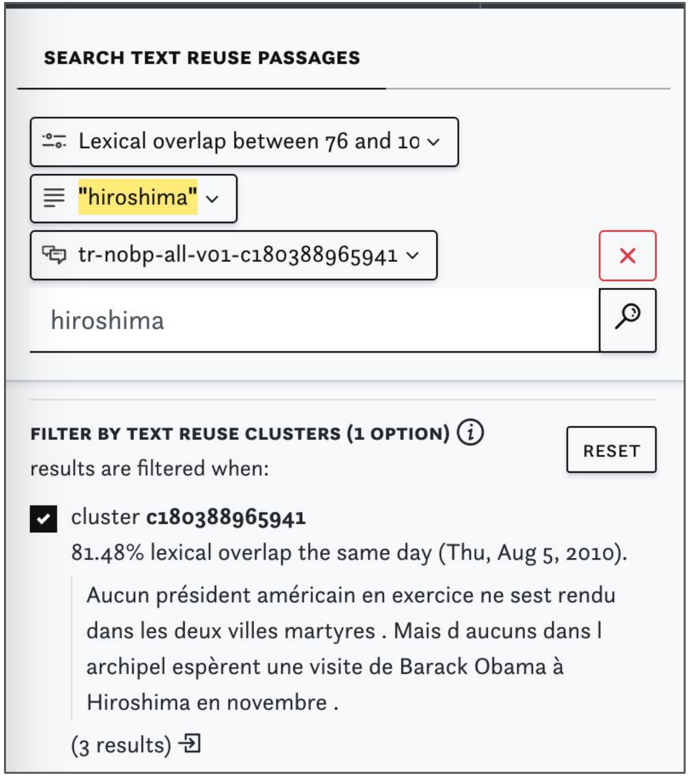
TYPE	
article	13,758,049
advertisement	1,863,669
page article	229,565
section	48,166
obituary	13,106
tables	2,240
buckets.type.NaN	2,066
weather news (other)	552

LANGUAGE	
French	13,945,214
German	1,295,522
English	2,663
Luxembourgish	766

PERSON	
United States	186,999
Les Anglais	87,128
Les Alliés	53,866
Paris	50,322
David Lloyd George	46,276
Les Brenets	45,923
Léon Blum	44,533
Les Chambres	41,953

LOCATION	
France	2,851,220
Paris	2,391,261
Switzerland	2,347,917
Suisse, Moselle	2,323,778
Lausanne	1,909,000
Zürich	[illegible],742
Italy	[illegible]9
Berlin	[illegible],620

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



SEARCH TEXT REUSE PASSAGES

Lexical overlap between 76 and 10

"hiroshima"

tr-nobp-all-v01-c180388965941

hiroshima

FILTER BY TEXT REUSE CLUSTERS (1 OPTION)

RESET

results are filtered when:

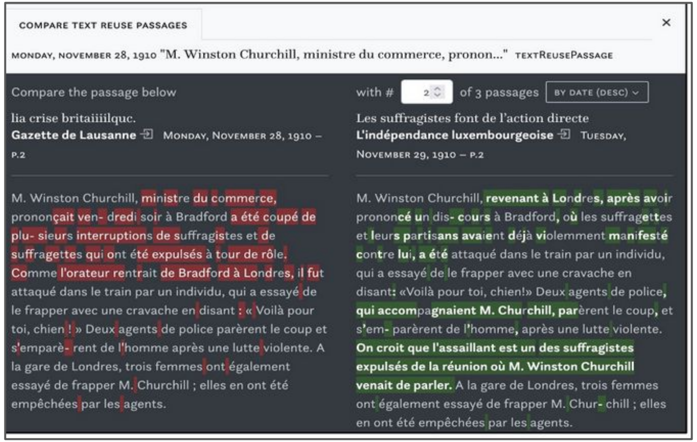
☑ cluster **c180388965941**

81.48% lexical overlap the same day (Thu, Aug 5, 2010).

> Aucun président américain en exercice ne sest rendu dans les deux villes martyres . Mais d aucuns dans l archipel espèrent une visite de Barack Obama à Hiroshima en novembre .

(3 results)

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



COMPARE TEXT REUSE PASSAGES ×

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

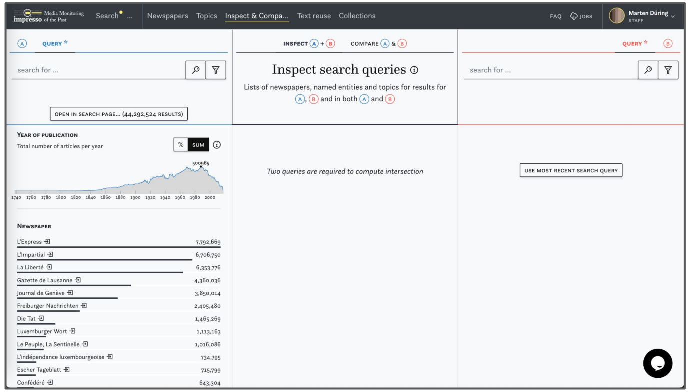
M. Winston Churchill, ministre du commerce, prononçait ven- dredi soir à Bradford a été coupé de plu- sieurs interruptions de suffragistes et de suffragettes qui ont été expulsés à tour de rôle. Comme l'orateur rentrait de Bradford à Londres, il fut attaqué dans le train par un individu, qui a essayé 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.

with # 2 of 3 passages BY DATE (DESC)

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.

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



Inspect search queries

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

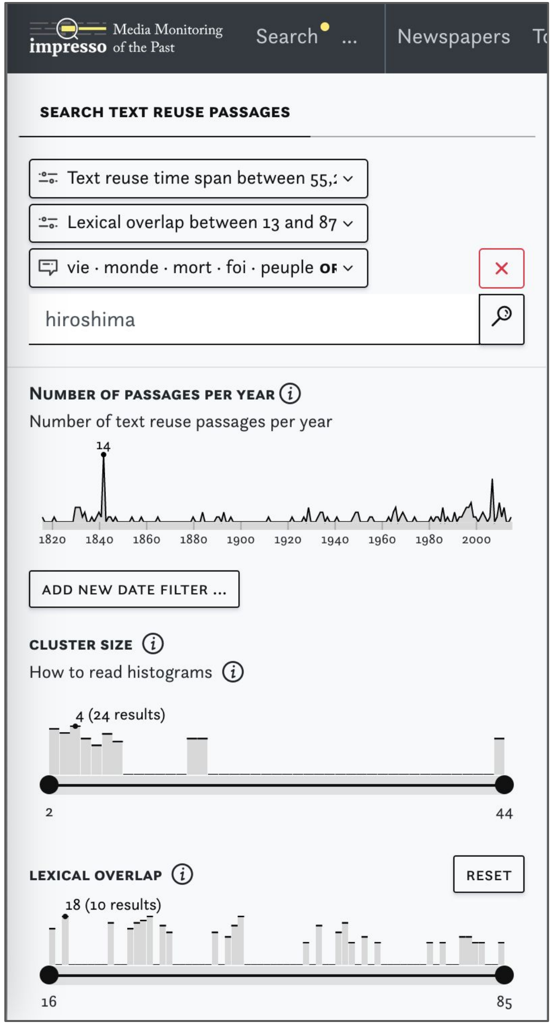
Year of publication

Total number of articles per year

Two queries are required to compute intersection

Newspaper	
L'Express	7,792,669
L'Impartial	6,706,750
La Liberté	6,353,776
Gazette de Lausanne	4,360,036
Journal de Genève	3,850,014
Freiburger Nachrichten	2,405,480
Die Tat	1,465,269
Luxemburger Wort	1,113,163
Le Peuple, La Sentinelle	1,016,086
L'indépendance luxembourgeoise	734,795
Escher Tageblatt	715,799
Confédéré	643,304

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned


impresso Media Monitoring of the Past
Search
...
Newspapers
To
SEARCH TEXT REUSE PASSAGES

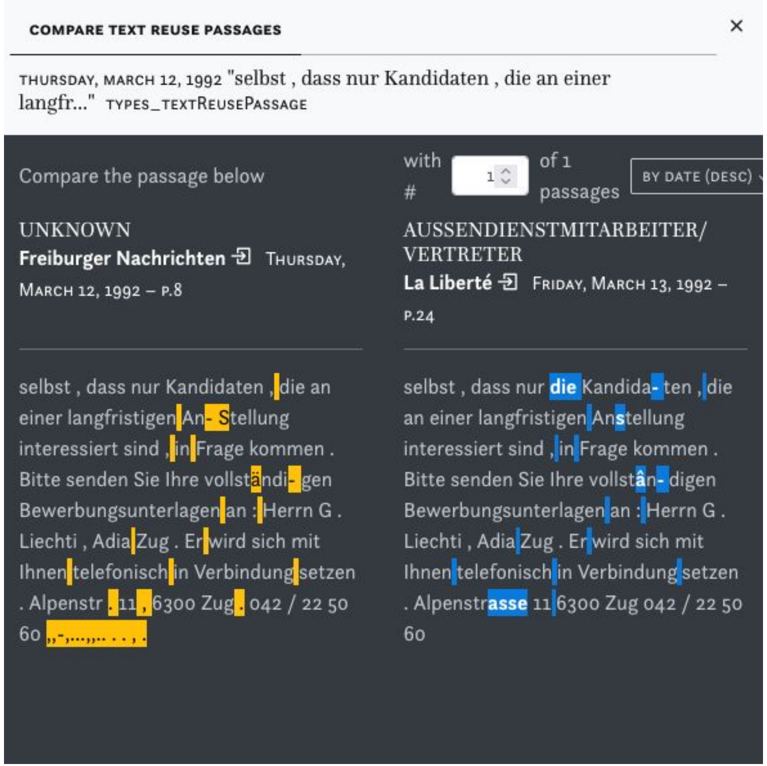
Text reuse time span between 55,:

Lexical overlap between 13 and 87

vie · monde · mort · foi · peuple OR

hiroshima

Job adverts - intentional distribution



COMPARE TEXT REUSE PASSAGES ×

THURSDAY, MARCH 12, 1992 "selbst , dass nur Kandidaten , die an einer langfr..." TYPES_TEXTREUSEPASSAGE

Compare the passage below

with # 1 of 1 passages

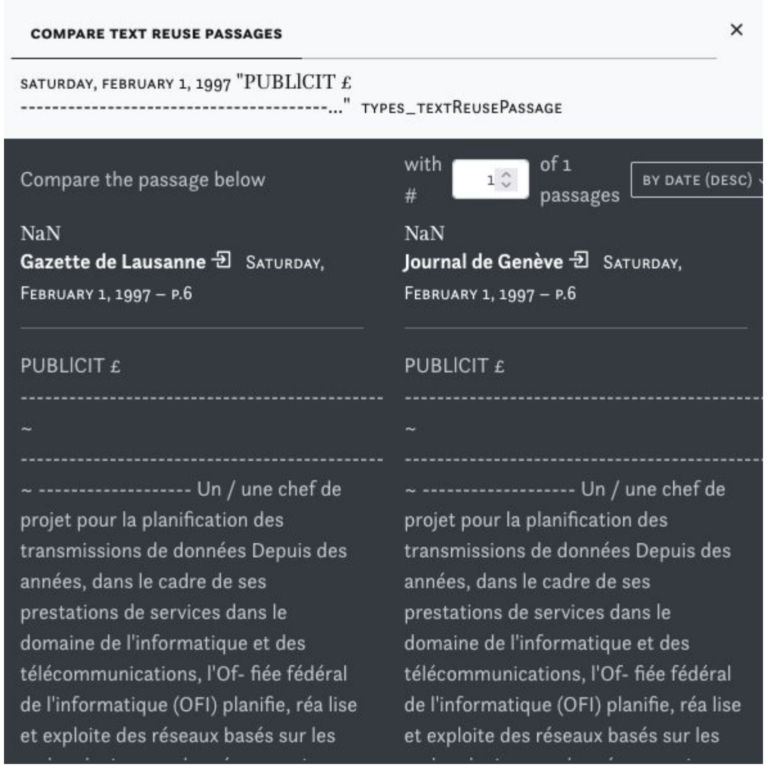
BY DATE (DESC)

UNKNOWN
Freiburger Nachrichten THURSDAY, MARCH 12, 1992 – P.8

selbst , dass nur Kandidaten , die an einer langfristigen An- Stellung interessiert sind , in Frage kommen . Bitte senden Sie Ihre vollständi- gen Bewerbungsunterlagen an : Herrn G . Liechti , Adia Zug . Er wird sich mit Ihnen telefonisch in Verbindung setzen . Alpenstr . 11 , 6300 Zug . 042 / 22 50 60 ,,-,...,,.. . . , .

AUSSENDIENSTMITARBEITER/ VERTRETER
La Liberté FRIDAY, MARCH 13, 1992 – P.24

selbst , dass nur **die** Kandida- ten , die an einer langfristigen Anstellung interessiert sind , in Frage kommen . Bitte senden Sie Ihre vollstân- digen Bewerbungsunterlagen an : Herrn G . Liechti , Adia Zug . Er wird sich mit Ihnen telefonisch in Verbindung setzen . Alpenstr**asse** 11 6300 Zug 042 / 22 50 60



COMPARE TEXT REUSE PASSAGES ×

SATURDAY, FEBRUARY 1, 1997 "PUBLlCIT £ --..." TYPES_TEXTREUSEPASSAGE

Compare the passage below

with # 1 of 1 passages

BY DATE (DESC)

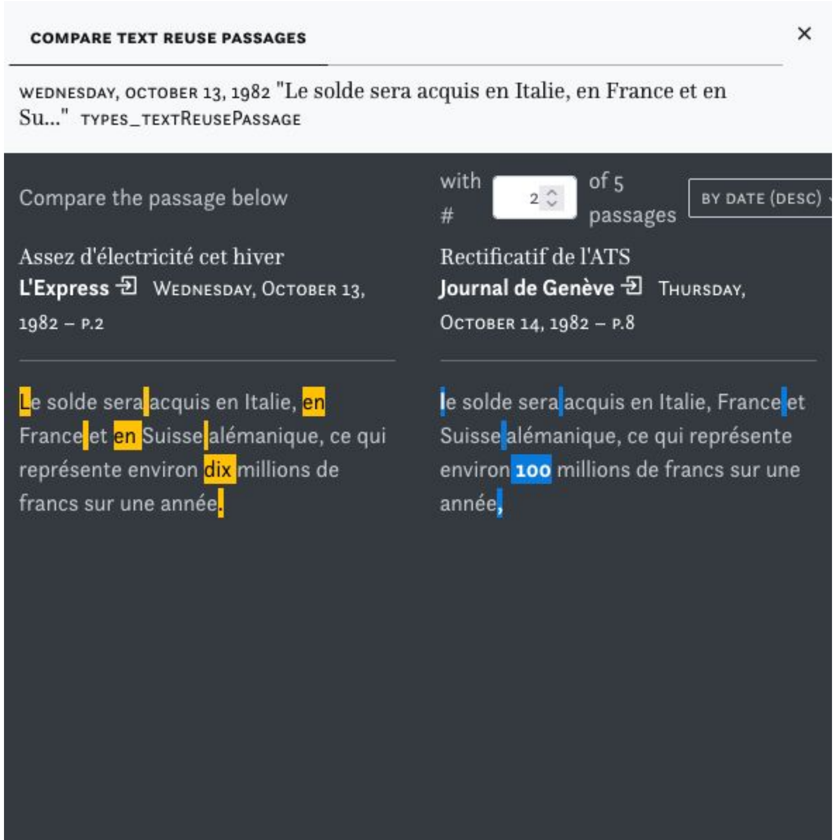
NaN
Gazette de Lausanne SATURDAY, FEBRUARY 1, 1997 – P.6

PUBLICIT £ -- ~ -- ~ ------------------ Un / une chef de projet pour la planification des transmissions de données Depuis des années, dans le cadre de ses prestations de services dans le domaine de l'informatique et des télécommunications, l'Of- fiée fédéral de l'informatique (OFI) planifie, réa lise et exploite des réseaux basés sur les

NaN
Journal de Genève SATURDAY, FEBRUARY 1, 1997 – P.6

PUBLICIT £ -- ~ -- ~ ------------------ Un / une chef de projet pour la planification des transmissions de données Depuis des années, dans le cadre de ses prestations de services dans le domaine de l'informatique et des télécommunications, l'Of- fiée fédéral de l'informatique (OFI) planifie, réa lise et exploite des réseaux basés sur les

Rectification - modified reuse



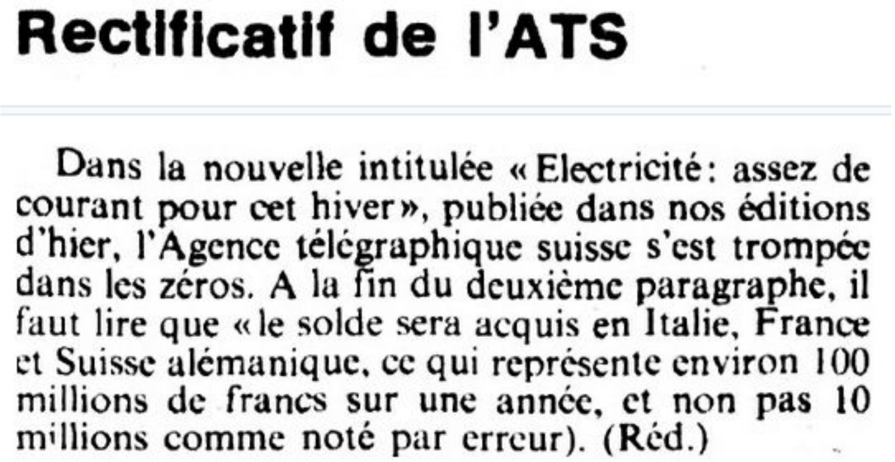
WEDNESDAY, OCTOBER 13, 1982 "Le solde sera acquis en Italie, en France et en Su..." TYPES_TEXTREUSEPASSAGE

Assez d'électricité cet hiver
L'Express WEDNESDAY, OCTOBER 13, 1982 – P.2

Le solde sera acquis en Italie, en France et en Suisse alémanique, ce qui représente environ dix millions de francs sur une année.

Rectificatif de l'ATS
Journal de Genève THURSDAY, OCTOBER 14, 1982 – P.8

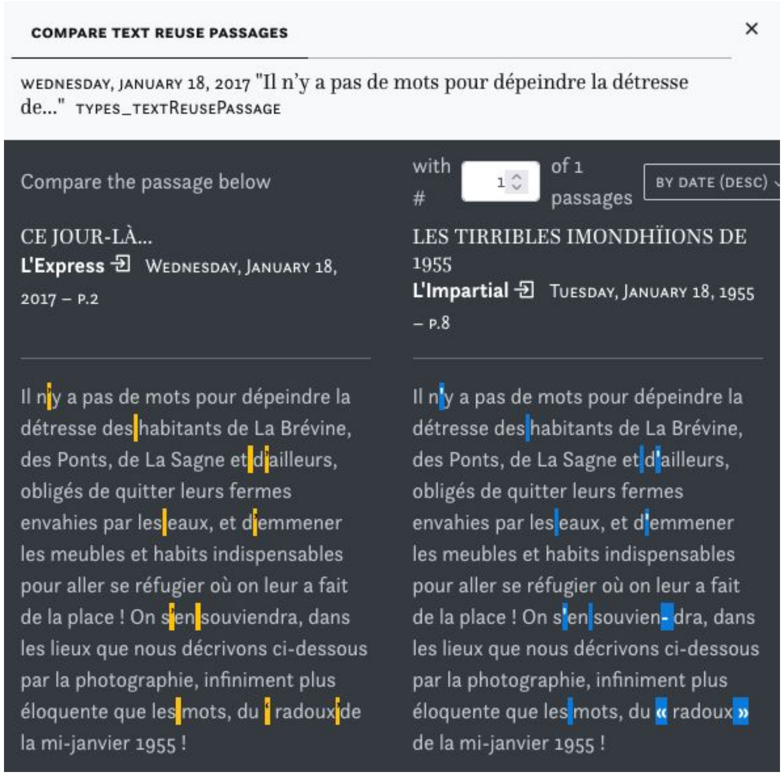
le solde sera acquis en Italie, France et Suisse alémanique, ce qui représente environ 100 millions de francs sur une année,



Rectificatif de l'ATS

Dans la nouvelle intitulée « Electricité : assez de courant pour cet hiver », publiée dans nos éditions d'hier, l'Agence télégraphique suisse s'est trompée dans les zéros. A la fin du deuxième paragraphe, il faut lire que « le solde sera acquis en Italie, France et Suisse alémanique, ce qui représente environ 100 millions de francs sur une année, et non pas 10 millions comme noté par erreur). (Réd.)

Event anniversaries - reprints



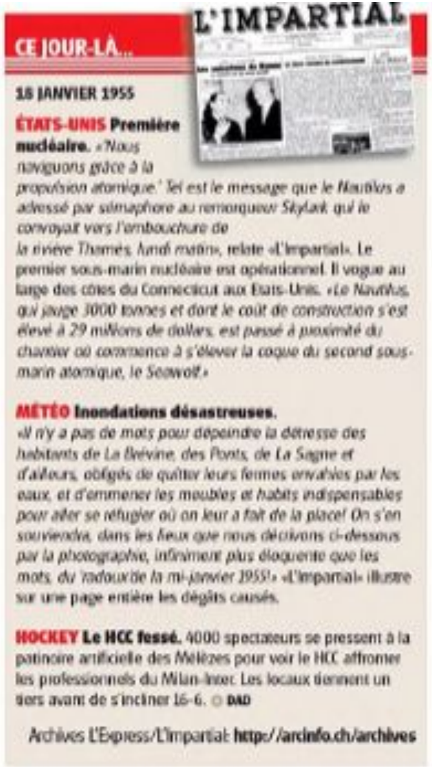
WEDNESDAY, JANUARY 18, 2017 "Il n'y a pas de mots pour dépeindre la détresse de..." TYPES_TEXTREUSEPASSAGE

CE JOUR-LÀ...
L'Express WEDNESDAY, JANUARY 18, 2017 – P.2

Il n'y a pas de mots pour dépeindre la détresse des habitants de La Brévine, des Ponts, de La Sagne et d'ailleurs, obligés de quitter leurs fermes envahies par les eaux, et d'emmener les meubles et habits indispensables pour aller se réfugier où on leur a fait de la place ! On s'en souviendra, dans les lieux que nous décrivons ci-dessous par la photographie, infiniment plus éloquente que les mots, du ‘radoux de la mi-janvier 1955 !

LES TIRRIBLES IMONDHÏIONS DE 1955
L'Impartial TUESDAY, JANUARY 18, 1955 – P.8

Il n'y a pas de mots pour dépeindre la détresse des habitants de La Brévine, des Ponts, de La Sagne et d'ailleurs, obligés de quitter leurs fermes envahies par les eaux, et d'emmener les meubles et habits indispensables pour aller se réfugier où on leur a fait de la place ! On s'en souvien- dra, dans les lieux que nous décrivons ci-dessous par la photographie, infiniment plus éloquente que les mots, du « radoux » de la mi-janvier 1955 !



L'IMPARTIAL

CE JOUR-LÀ...

18 JANVIER 1955

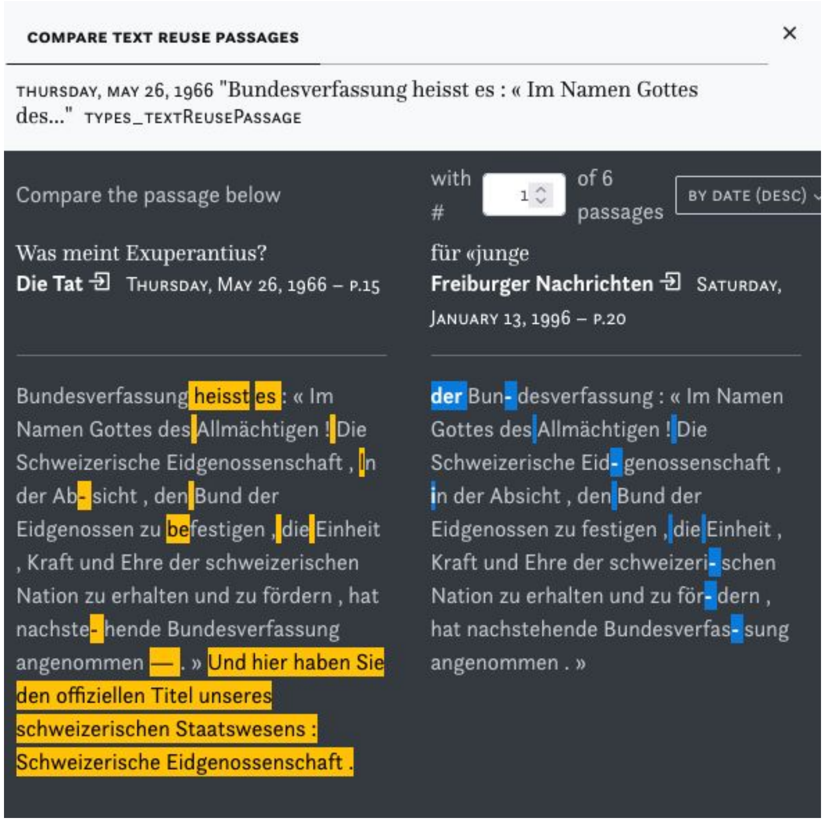
ÉTATS-UNIS Première nucléaire. *«?Nous naviguons grâce à la propulsion atomique.' Tel est le message que le Nautilus a adressé par sémaphore au remorqueur Skylark qui le convoyait vers l'embouchure de la rivière Thamès, lundi matin»*, relate «L'Impartial». Le premier sous-marin nucléaire est opérationnel. Il vogue au large des côtes du Connecticut aux Etats-Unis. *«Le Nautilus, qui jauge 3000 tonnes et dont le coût de construction s'est élevé à 29 millions de dollars, est passé à proximité du chantier où commence à s'élever la coque du second sous-marin atomique, le Seawolf.»*

MÉTÉO Inondations désastreuses. *«Il n'y a pas de mots pour dépeindre la détresse des habitants de La Brévine, des Ponts, de La Sagne et d'ailleurs, obligés de quitter leurs fermes envahies par les eaux, et d'emmener les meubles et habits indispensables pour aller se réfugier où on leur a fait de la place! On s'en souviendra, dans les lieux que nous décrivons ci-dessous par la photographie, infiniment plus éloquente que les mots, du 'radoux'de la mi-janvier 1955!»* «L'Impartial» illustre sur une page entière les dégâts causés.

HOCKEY Le HCC fessé. 4000 spectateurs se pressent à la patinoire artificielle des *Mélèzes* pour voir le HCC affronter les professionnels du *Milan-Inter*. Les locaux tiennent un tiers avant de s'incliner 16-6. © DAD

Archives L'Express/L'Impartial: **http://arcinfo.ch/archives**

Quotes



COMPARE TEXT REUSE PASSAGES ×

THURSDAY, MAY 26, 1966 "Bundesverfassung heisst es : « Im Namen Gottes des..." TYPES_TEXTREUSEPASSAGE

Compare the passage below

with # 1 of 6 passages

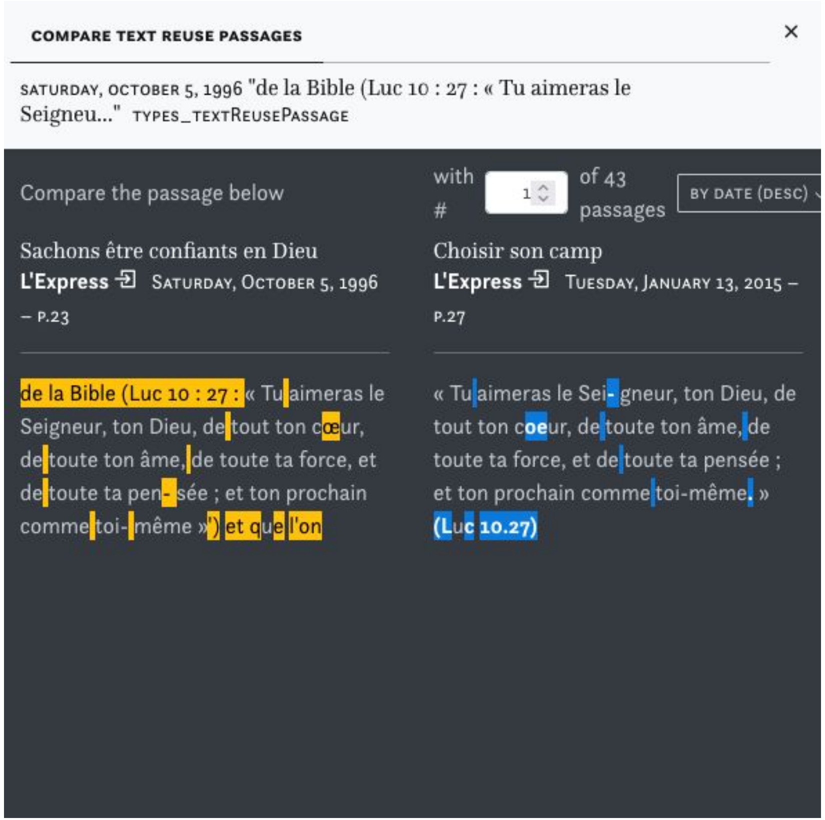
BY DATE (DESC)

Was meint Exuperantius?
Die Tat THURSDAY, MAY 26, 1966 – P.15

für «junge
Freiburger Nachrichten SATURDAY, JANUARY 13, 1996 – P.20

Bundesverfassung heisst es : « Im Namen Gottes des Allmächtigen ! Die Schweizerische Eidgenossenschaft , In der Ab- sicht , den Bund der Eidgenossen zu befestigen , die Einheit , Kraft und Ehre der schweizerischen Nation zu erhalten und zu fördern , hat nachste- hende Bundesverfassung angenommen — . » Und hier haben Sie den offiziellen Titel unseres schweizerischen Staatswesens : Schweizerische Eidgenossenschaft .

der Bun- desverfassung : « Im Namen Gottes des Allmächtigen ! Die Schweizerische Eid- genossenschaft , in der Absicht , den Bund der Eidgenossen zu festigen , die Einheit , Kraft und Ehre der schweizeri- schen Nation zu erhalten und zu för- dern , hat nachstehende Bundesverfas- sung angenommen . »



COMPARE TEXT REUSE PASSAGES ×

SATURDAY, OCTOBER 5, 1996 "de la Bible (Luc 10 : 27 : « Tu aimeras le Seigneu..." TYPES_TEXTREUSEPASSAGE

Compare the passage below

with # 1 of 43 passages

BY DATE (DESC)

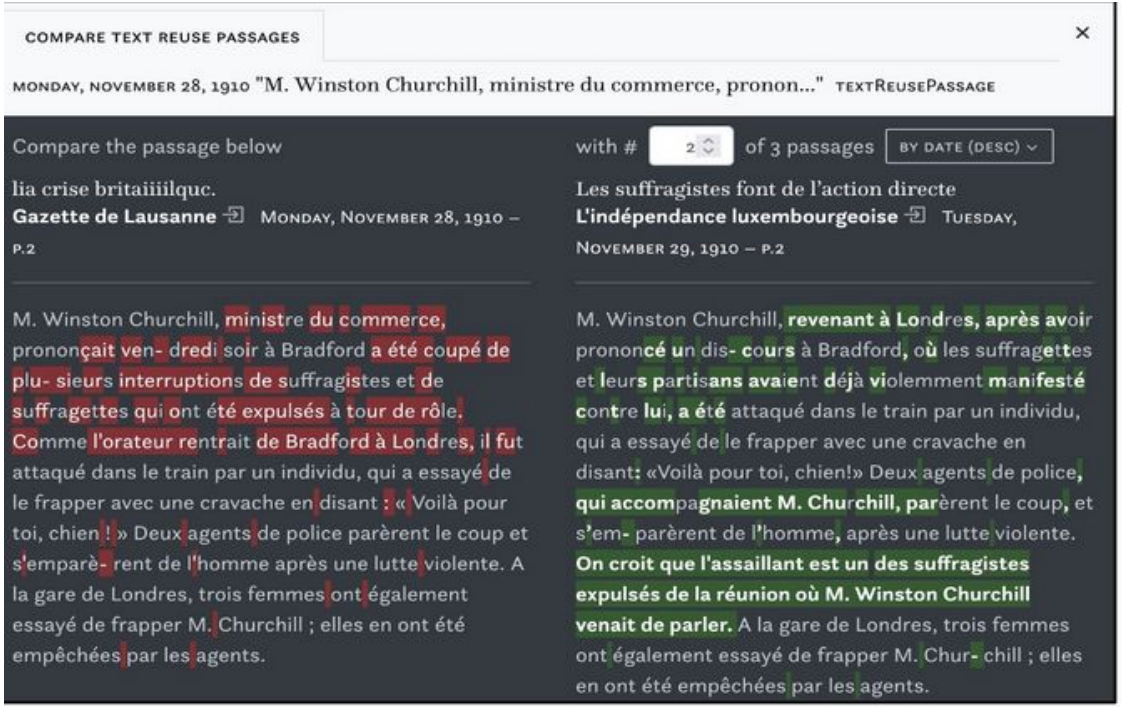
Sachons être confiants en Dieu
L'Express SATURDAY, OCTOBER 5, 1996 – P.23

Choisir son camp
L'Express TUESDAY, JANUARY 13, 2015 – P.27

de la Bible (Luc 10 : 27 : « Tu aimeras le Seigneur, ton Dieu, de tout ton cœur, de toute ton âme, de toute ta force, et de toute ta pen- sée ; et ton prochain comme toi- même »") et que l'on

« Tu aimeras le Sei- gneur, ton Dieu, de tout ton coeur, de toute ton âme, de toute ta force, et de toute ta pensée ; et ton prochain comme toi-même. » (Luc 10.27)

Edits of press agency content



Monday, November 28, 1910 "M. Winston Churchill, ministre du commerce, pronon..." TextReusePassage

Compare the passage below

with # 2 of 3 passages

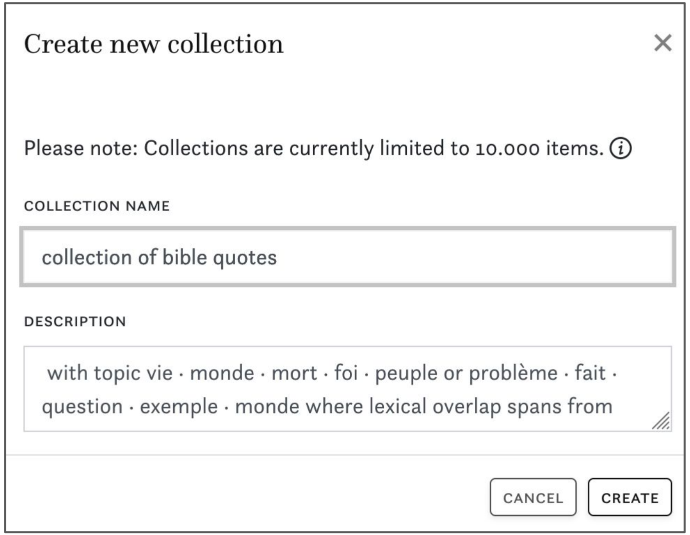
lia crise britaiiiilquc.
Gazette de Lausanne Monday, November 28, 1910 – p.2

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.

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



Create new collection

Please note: Collections are currently limited to 10.000 items.

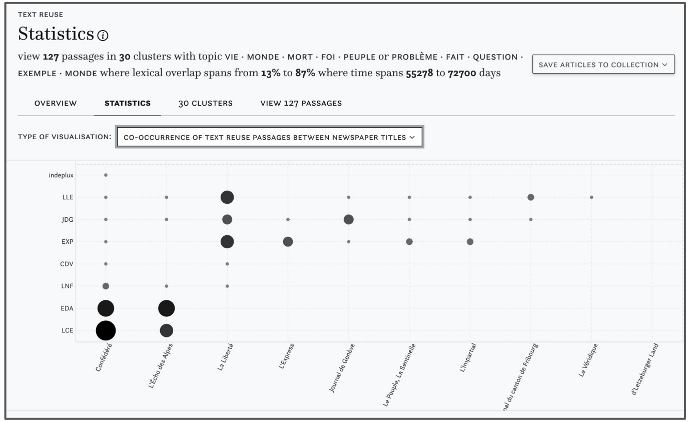
COLLECTION NAME

collection of bible quotes

DESCRIPTION

with topic vie · monde · mort · foi · peuple or problème · fait · question · exemple · monde where lexical overlap spans from

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned

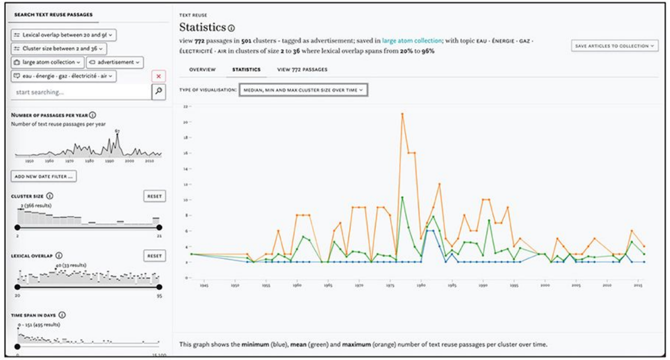

TEXT REUSE

Statistics ⓘ

view **127** passages in **30** clusters with topic VIE · MONDE · MORT · FOI · PEUPLE or PROBLÈME · FAIT · QUESTION · EXEMPLE · MONDE where lexical overlap spans from **13%** to **87%** where time spans **55278** to **72700** days

TYPE OF VISUALISATION: CO-OCCURRENCE OF TEXT REUSE PASSAGES BETWEEN NEWSPAPER TITLES

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



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

TEXT REUSE

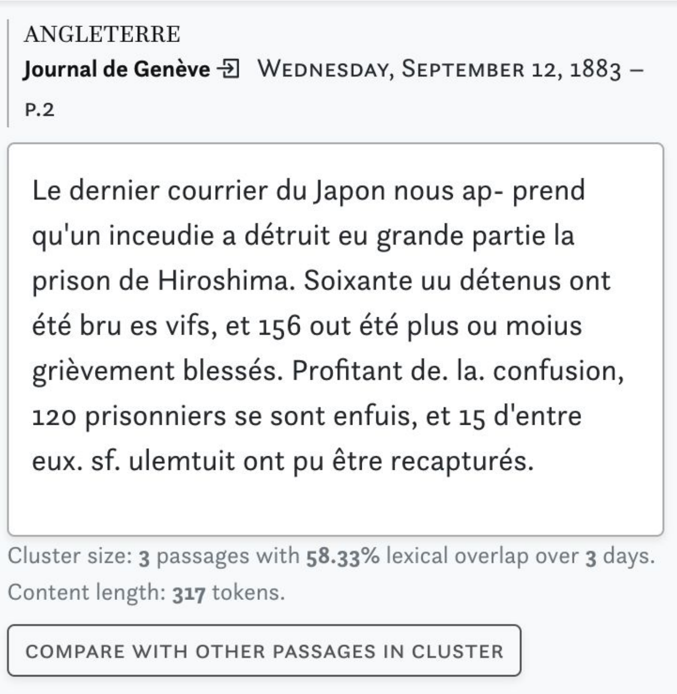
Statistics

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%**

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.

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



ANGLETERRE

Journal de Genève WEDNESDAY, SEPTEMBER 12, 1883 – P.2

Le dernier courrier du Japon nous ap- prend qu'un inceudie a détruit eu grande partie la prison de Hiroshima. Soixante uu détenus ont été bru es vifs, et 156 out été plus ou moius grièvement blessés. Profitant de. la. confusion, 120 prisonniers se sont enfuis, et 15 d'entre eux. sf. ulemtuit ont pu être recapturés.

Cluster size: **3** passages with **58.33%** lexical overlap over **3** days.

Content length: **317** tokens.

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned

JOURNAL DE GENÈVE

NATIONAL, POLITIQUE ET LITTÉRAIRE

LIBERTÉ ET PATRIE

N° 135 — Jeudi 10 juin 1954

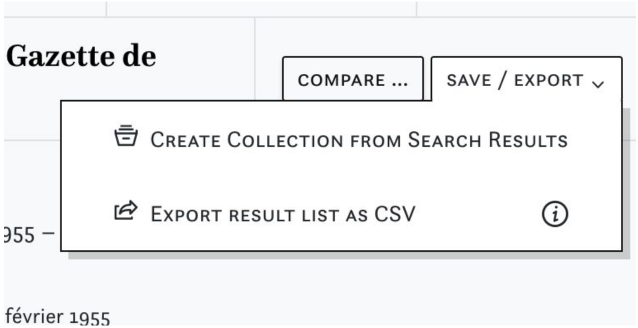
Gazette de Lausanne

ET JOURNAL SUISSE

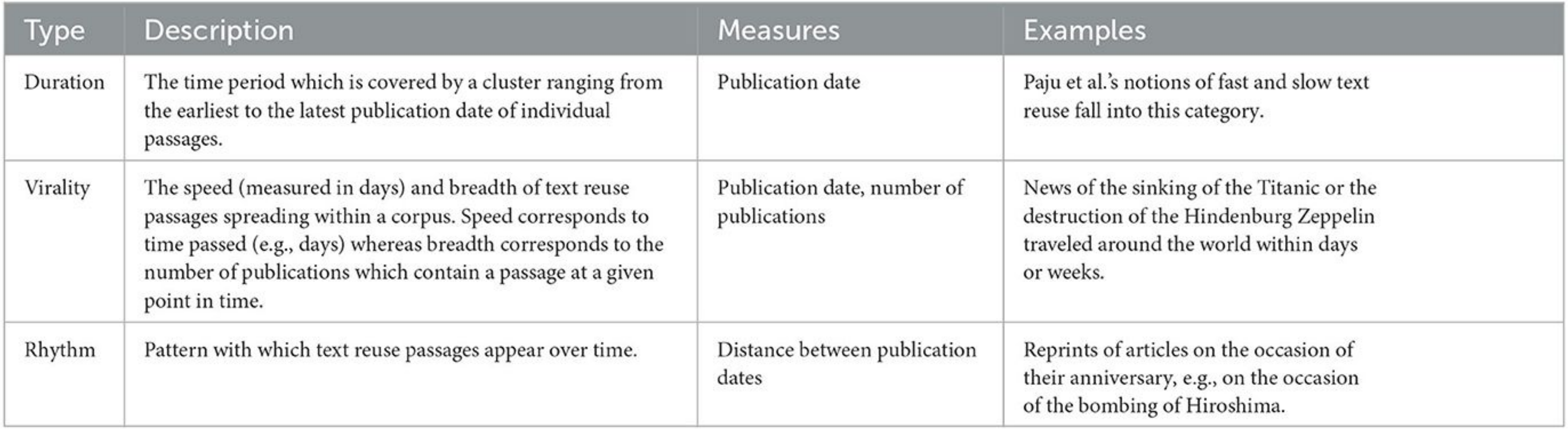
QUOTIDIEN FONDÉ EN 1798

Kohler
Confection pour dames
Rue Centrale - Lausanne
CRÉATIONS EXCLUSIVES
PROJETS ETUDES DEVIS

Task	Title	Level	Support
1	Obtain an overview of text reuse in a corpus, collection or query	Corpus	Yes
2	Obtain an overview of a single cluster Cluster	Cluster	Yes
3	Compare passages	Passages	Yes
4	Compare clusters	Cluster	Yes
5	Identify different types of text reuse	Corpus	Yes
6	Generate research corpora based on text reuse clusters	Corpus	Yes
7	Identify connections	Corpus	Partial
8	Detect and trace virality	Corpus	No
9	Search for passages	Passages	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned

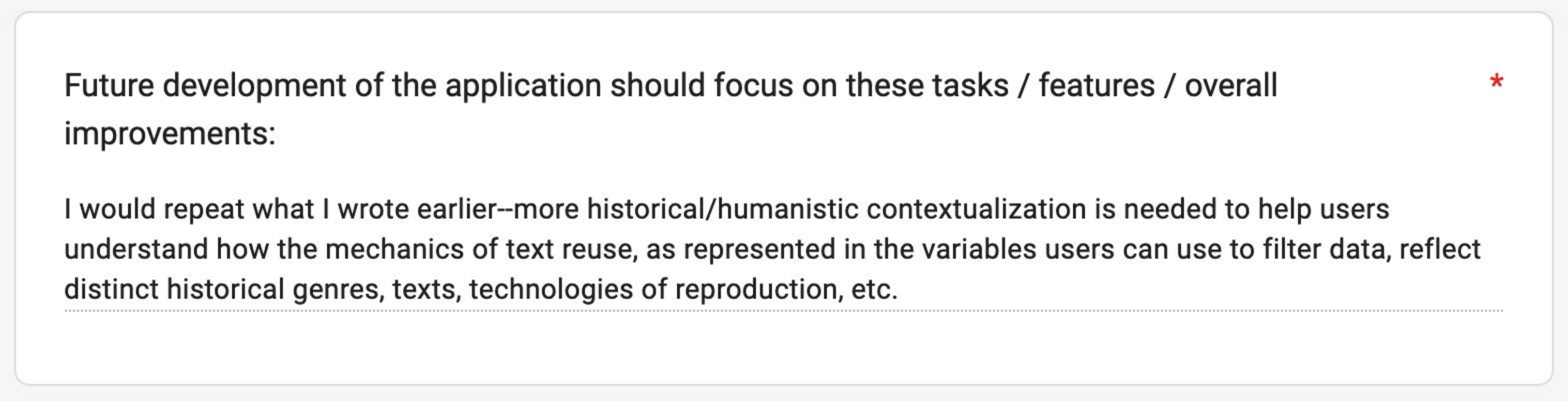

Gazette de
COMPARE ...
SAVE / EXPORT
Create Collection from Search Results
Export result list as CSV
955 –
février 1955

Temporalities in text reuse data



Type	Description	Measures	Examples
Duration	The time period which is covered by a cluster ranging from the earliest to the latest publication date of individual passages.	Publication date	Paju et al.'s notions of fast and slow text reuse fall into this category.
Virality	The speed (measured in days) and breadth of text reuse passages spreading within a corpus. Speed corresponds to time passed (e.g., days) whereas breadth corresponds to the number of publications which contain a passage at a given point in time.	Publication date, number of publications	News of the sinking of the Titanic or the destruction of the Hindenburg Zeppelin traveled around the world within days or weeks.
Rhythm	Pattern with which text reuse passages appear over time.	Distance between publication dates	Reprints of articles on the occasion of their anniversary, e.g., on the occasion of the bombing of Hiroshima.

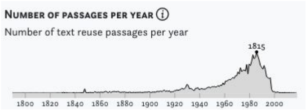
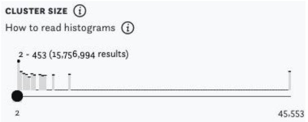
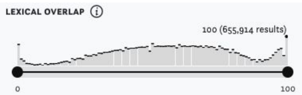
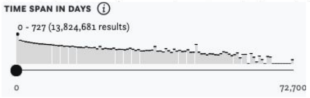
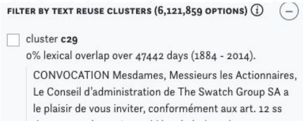
Next steps?



Future development of the application should focus on these tasks / features / overall improvements: *

I would repeat what I wrote earlier--more historical/humanistic contextualization is needed to help users understand how the mechanics of text reuse, as represented in the variables users can use to filter data, reflect distinct historical genres, texts, technologies of reproduction, etc.

Capture the characteristics of text reuse through filters

Measure	Description	Implementation in interface prototype	
Passages per year	Number of passages counted in a given year.	Line chart which displays the count of passages per year for a given query or filter operation. This gives a first indication, during which years text reuse occurred more commonly. Time sliders and precise date entry allow users to filter for exact date ranges to inspect.	
Cluster size	The number of passages contained in a cluster.	Histogram which shows the distribution of text reuse cluster sizes and indicates the highest score. The histogram groups clusters of size n and displays their sum. This gives a first indication of averages as well as outliers. Sliders can be used to specify a cluster size range of interest. Filtering by cluster size allows to exclude or explicitly focus on outliers but different cluster sizes may also correspond to different types of content.	
Lexical overlap	The percentage of unique tokens that all passages in a cluster have in common. All text was lowercased and punctuation was stripped.	Histogram which shows the distribution of lexical overlap in percent and indicates the largest number of clusters for a given score. Extremely low lexical overlap decreases the chance to discover meaningful text reuse whilst extremely high overlap will only reveal near-copies of content and may be too restrictive for some purposes.	
Time span	The time window covered by documents in the cluster, measured in number of days.	Histogram which shows the gap between the earliest publication date of an article in a text reuse cluster and the latest measured in days and indicates the largest number of passages for a given score. This is an efficient approach to discover or filter for instances of slow, mid-range and rapid text reuse. The histogram groups clusters by the number of days in between publication dates and displays their sum.	
Text reuse clusters	Clusters store text segments (or passages) that are reused in different units of a corpus.	List of text reuse clusters which match a given query, sorted by number of passages. Each cluster is characterized with basic information (passages count, lexical overlap, time periods and years covered) as well as a snippet preview of the passage. Clusters are sorted by the number of matching passages. Clusters can be selected manually for further inspection in the Text Reuse app or in other *impresso* components such as Search.	 □ cluster **c29** 0% lexical overlap over 47442 days (1884 - 2014). CONVOCATION Mesdames, Messieurs les Actionnaires, Le Conseil d'administration de The Swatch Group SA a le plaisir de vous inviter, conformément aux art. 12 ss

Data-driven text reuse classification?

	Passages per year	**Cluster size**	**Lexical overlap**	**Time span (=duration)**
Reprints of historical materials	Few	Small	Very high	Very large
Co-publication	Many	Rather small	Very high	Very short
Advertising campaign	Many	Large	High	Varies

Finally..

Do we need a(nother) universal typology of historical TR types?

Do TR types have distinct properties (lex overlap, temporality, cluster/passage size, links to other sem enrichments) across source types?

Can we classify them analogue to the way we can predict content types?
E.g. quotes, unattributed reuse, modified reuse, rhythmic

Impresso use case: “Which types of TR can I find in my collection of articles?”

Can we link texts between between spoken and written text and languages?

How do we do TRD in the LLM age?

impresso - Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio

09/2023 - 02/2027

Main objectives

1. **Enrichment and integration** of newspaper and radio sources
2. **Expand** the corpus to Western Europe
3. **Develop** interfaces for exploratory and computational research
4. **Conduct** case studies in (media) history, theme “influences”

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

Partners

National or state libraries (holding digitised newspaper collections)

Bibliothèque Nationale Suisse, BN
Bibliothèque Nationale du Luxembourg, BNL
Österreichische Nationalbibliothek, ONB
Staatsbibliothek zu Berlin, SBB
The British Library (BL)
Bibliothèque nationale de France, BnF
Staats- und Universitätsbibliothek Hamburg, HUB
Bibliothèque royale de Belgique/Koninklijke Bibliotheek van België, KBR
Koninklijke Bibliotheek, KB

Newspapers

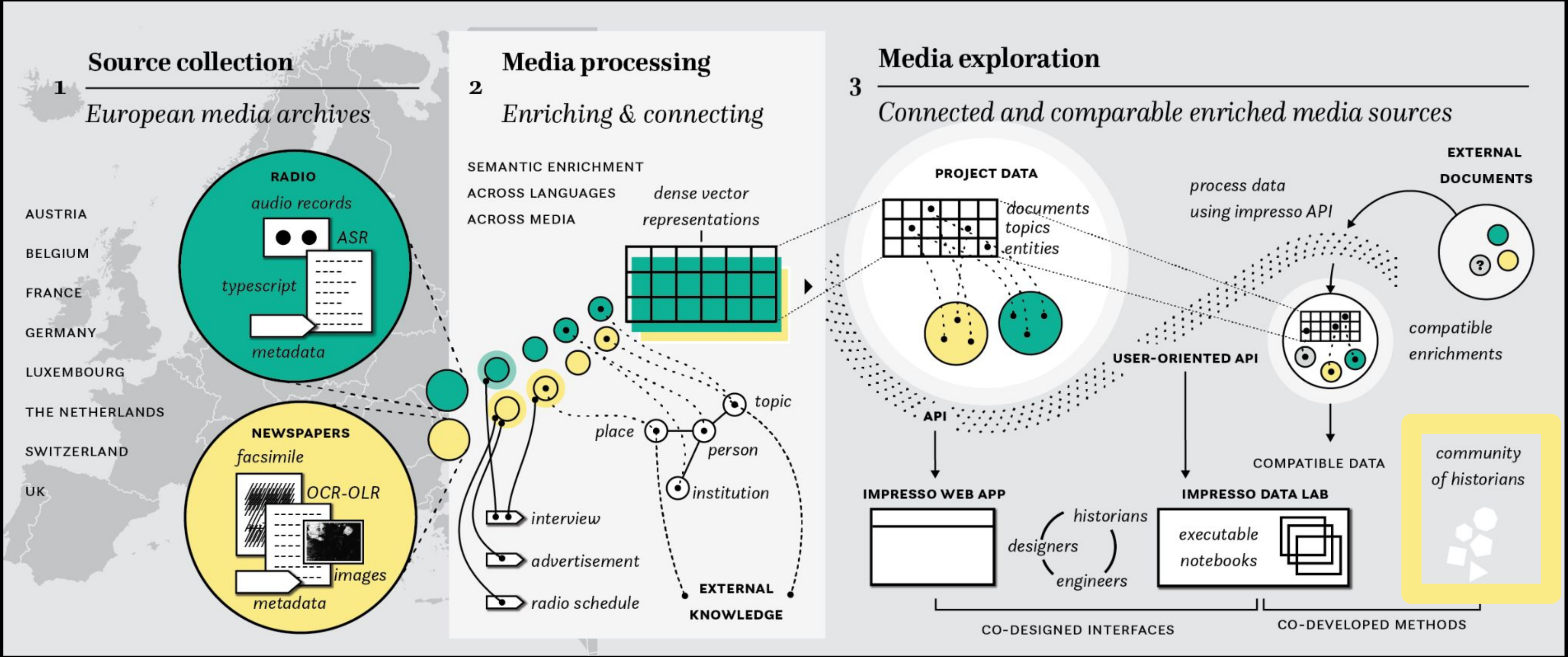
Le Temps
Neue Zürcher Zeitung

Audiovisual heritage institutions and archives (holding digitised radio collections)

Radio Television Suisse (RTS)
Österreichischer Rundfunk, ORF
British Broadcasting Corporation (BBC)
DeutschlandRadio
Institut National de l'Audiovisuel, INA
Nederlands Instituut voor Beeld en Geluid, NISV

Research Networks

Entangled Media Histories Research Network for European media historians (EMHIS)
Memoriav, the Swiss network for audiovisual cultural heritage preservation
infoclio.ch



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

topic

place

person

institution

interview

advertisement

radio schedule

EXTERNAL KNOWLEDGE

3 Media exploration

Connected and comparable enriched media sources

PROJECT DATA

documents

topics

entities

API

process data using impresso API

EXTERNAL DOCUMENTS

compatible enrichments

USER-ORIENTED API

COMPATIBLE DATA

IMPRESSO WEB APP

historians

designers

engineers

IMPRESSO DATA LAB

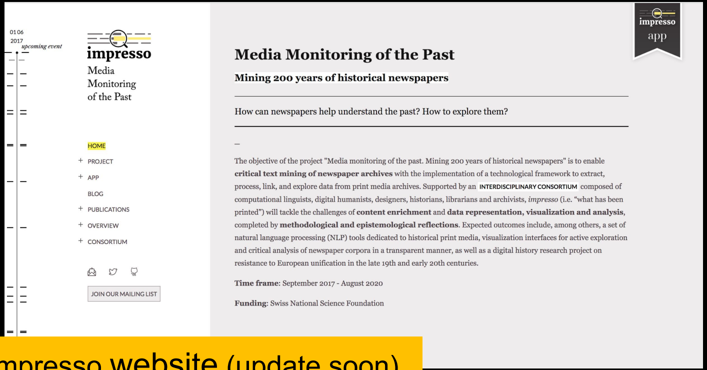
executable notebooks

community of historians

CO-DESIGNED INTERFACES

CO-DEVELOPED METHODS

Thank you



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?

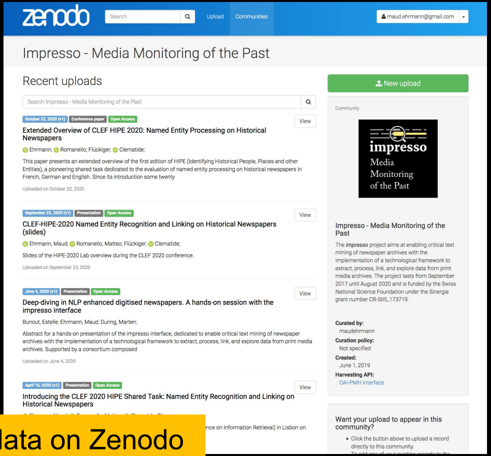
—

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

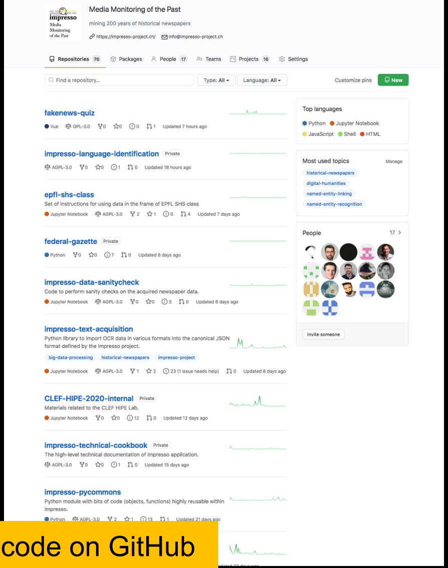
The *impresso* project aims at enabling critical text mining of newspaper archives with the implementation of a technological framework to extract, process, link, and explore data from print media archives. The project lasts from September 2017 until August 2020 and is funded by the Swiss National Science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



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

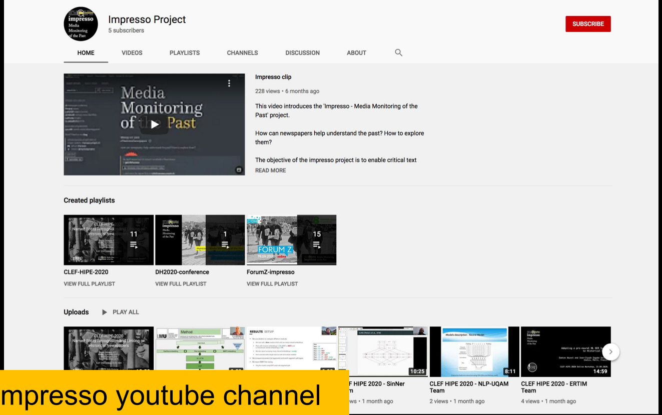
Top languages
● Python ● Jupyter Notebook ● JavaScript ● Shell ● HTML

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

People 17

Invite someone

code on GitHub



Impresso 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	DH2020-conference	ForumZ-impresso
11	1	15
VIEW FULL PLAYLIST	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST

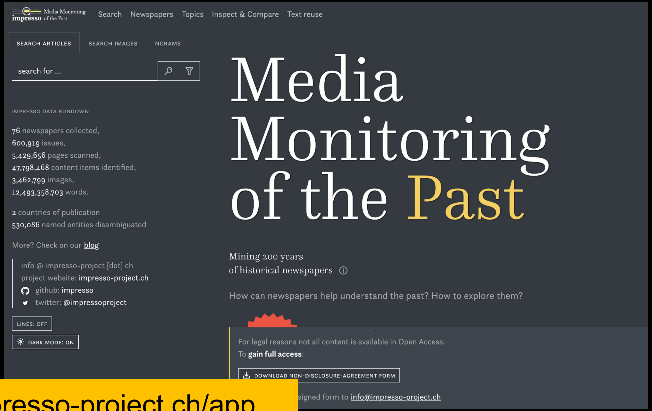
Uploads ▶ PLAY ALL

F HIPE 2020 - SinNer m
ws • 1 month ago

CLEF HIPE 2020 - NLP-UQAM Team
2 views • 1 month ago

CLEF HIPE 2020 - ERTIM Team
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