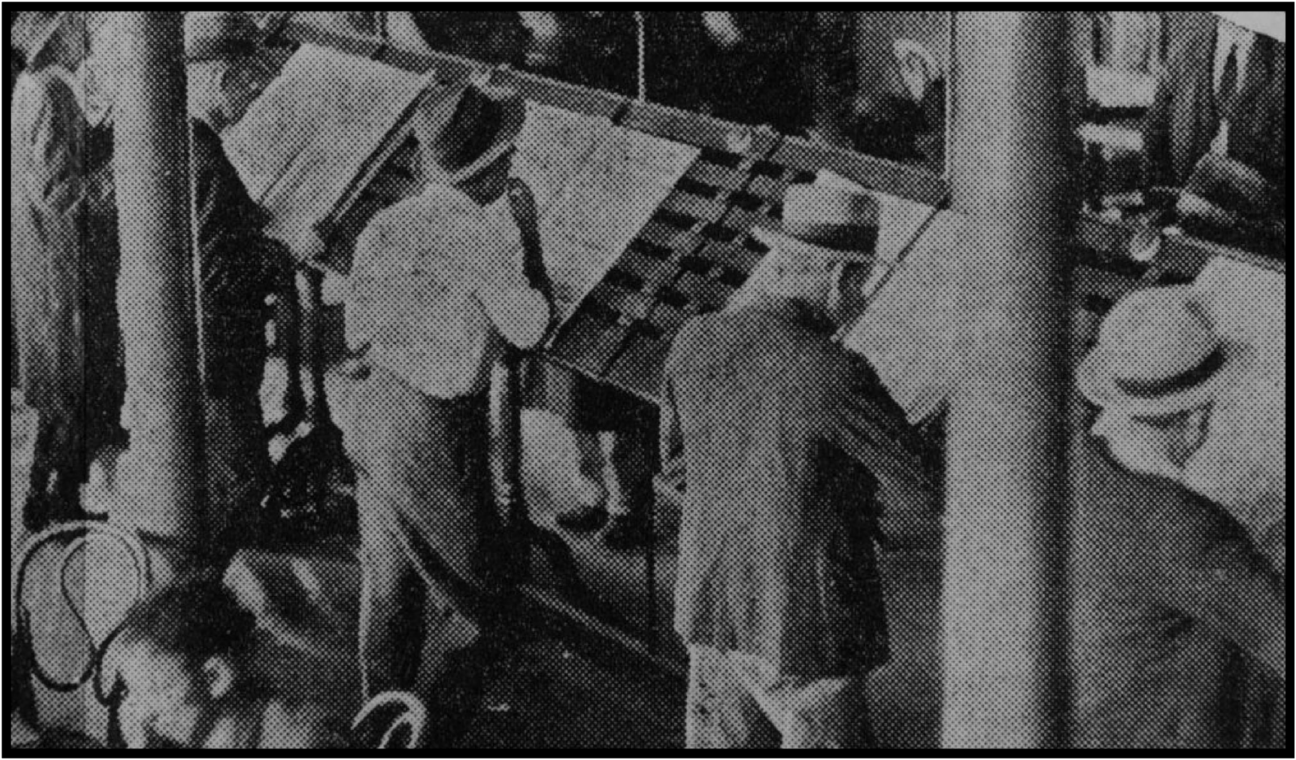
impresso Text Reuse at Scale

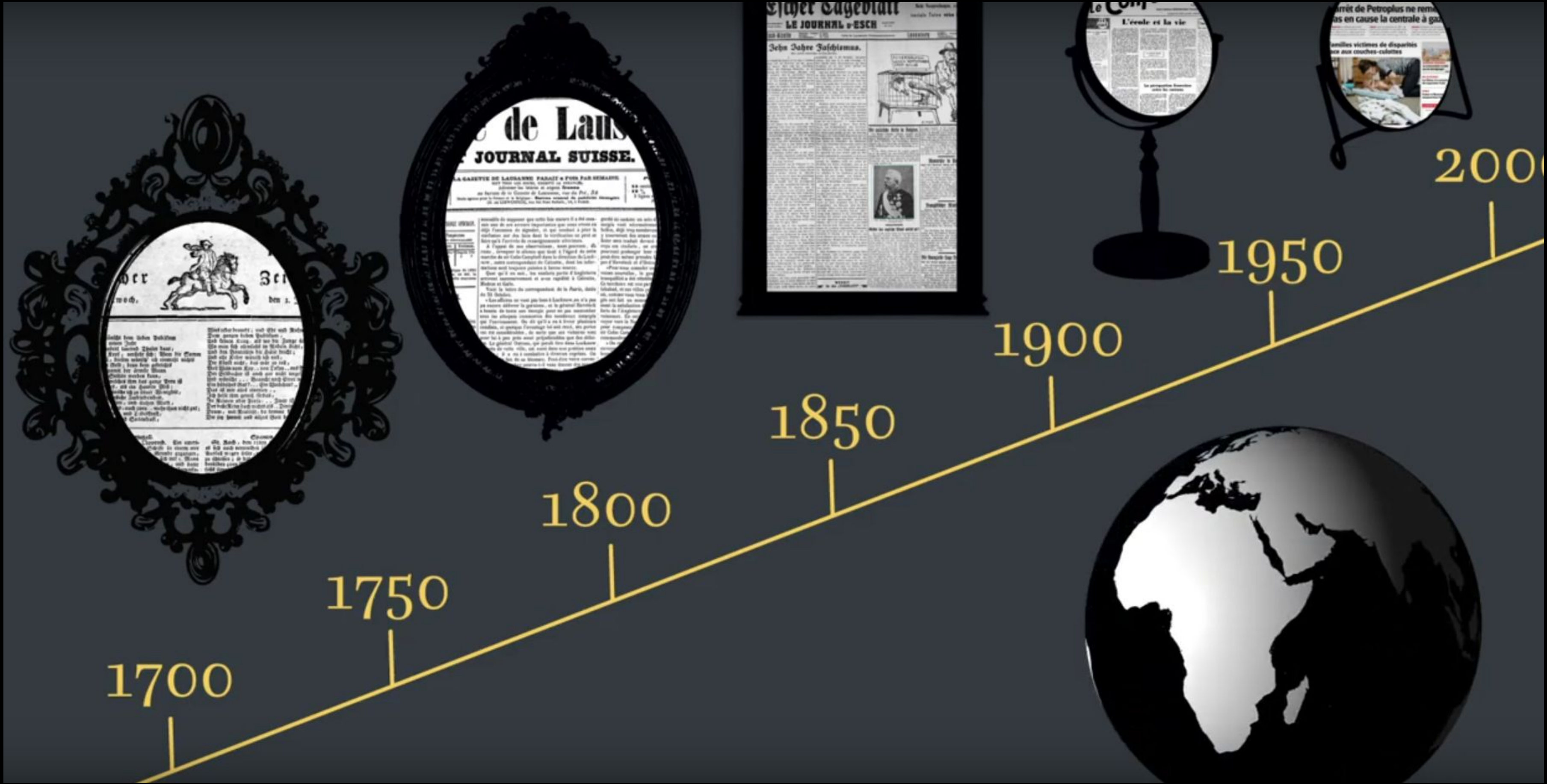
Kampagnen, Wiederverwertung und Plagiate in historischen Zeitungsammlungen

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



1. Ein paar Worte über historische Zeitungen


de Laus
JOURNAL SUISSE.
LE JOURNAL d'ESCH
Zehn Jahre Faschismus.
L'école et la vie
1700
1750
1800
1850
1900
1950
200

für Wahrheit und Recht

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

Erſtes Blatt.

Serbiſche Wahlen.

Türkiſch-italieniſcher Krieg

Erzbiſchof Dr. v. Abert †

Zum Untergang der Titanic.

Schloß Altenſtein.

Haute tension sur l'avenir électrique

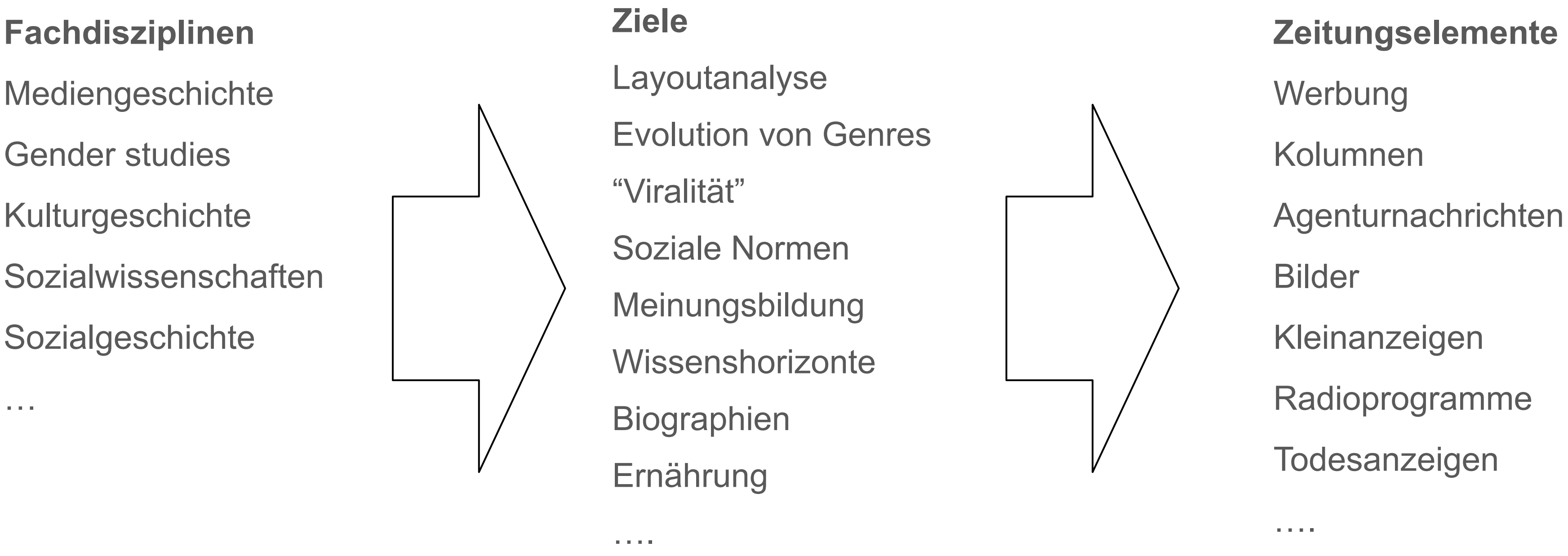
Certains experts annoncent une pénurie d'électricité pour la fin du millénaire. D'autres spécialistes affirment que l'on néglige les économies. En attendant, la Suisse continue d'investir dans le parc nucléaire français.

ENERGIE DE L'OUEST SUISSE (EOS)

Christophe Babaiantz: le prix du renoncement à Kaiseraugst

L'ère des incertitudes

Forschungsinteressen: Geschichte der Medien und Medien als Quellen



Fachdisziplinen

Mediengeschichte

Gender studies

Kulturgeschichte

Sozialwissenschaften

Sozialgeschichte

…

Ziele

Layoutanalyse

Evolution von Genres

“Viralität”

Soziale Normen

Meinungsbildung

Wissenshorizonte

Biographien

Ernährung

….

Zeitungselemente

Werbung

Kolumnen

Agenturnachrichten

Bilder

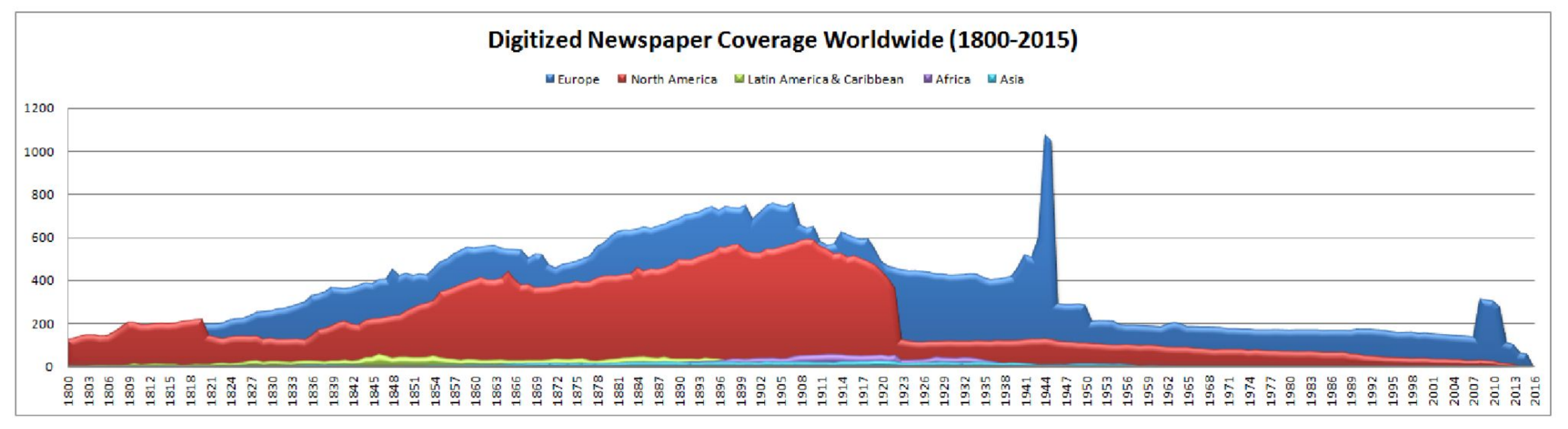
Kleinanzeigen

Radioprogramme

Todesanzeigen

….

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*.

Historische Zeitungen als Herausforderung

1. Institutionelle Silos
2. Big and messy data
3. Noisy historical text
4. Visualisierung und Exploration
5. Digitale Forschungskultur

Critical content mining of 200 years of historical newspapers

Impresso

Inwieweit dienen semantische Anreicherungen der Analyse und Exploration historischer Zeitungen?

Das Team

+ beratende HistorikerInnen
+ assoziierte WissenschaftlerInnen

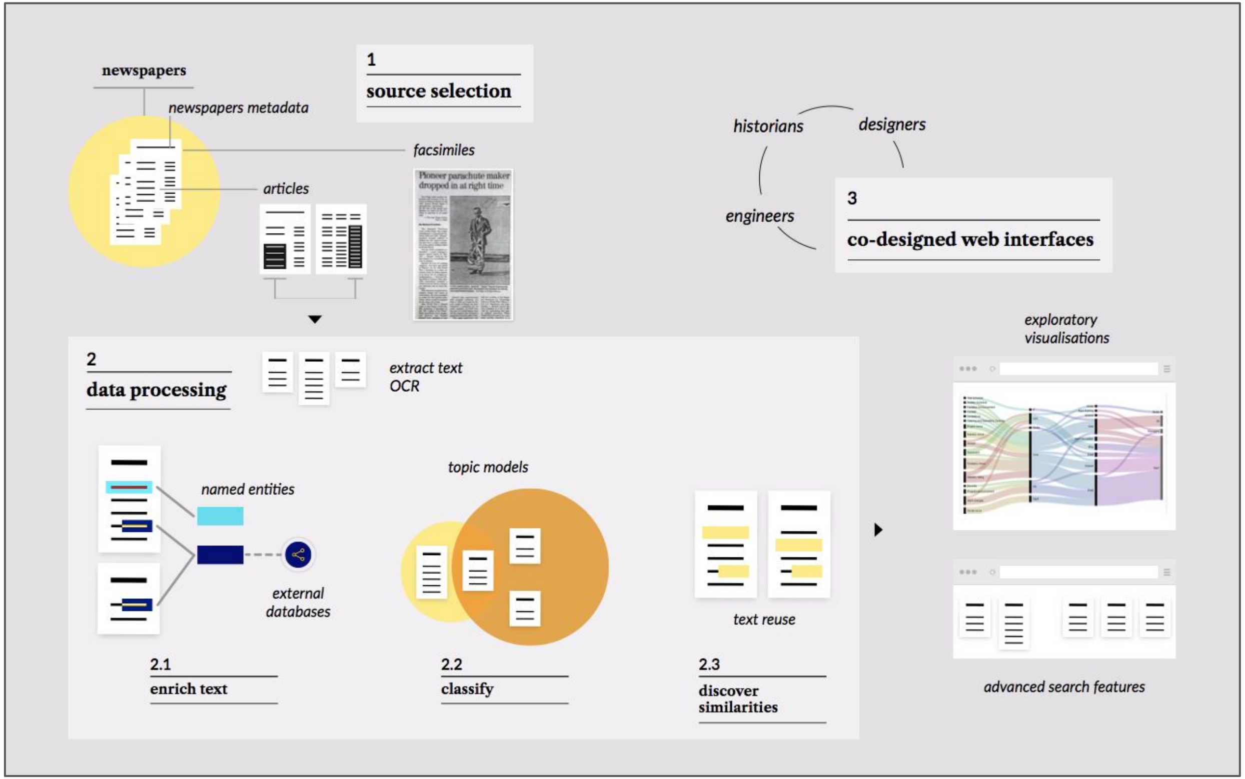
Estelle Bunout
Simon Clematide
Marten Duering
Maud Ehrmann
Andreas Fickers
Daniele Guido
Frédéric Kaplan
Peter Makarov
Matteo Romanello
Gerold Schneider
Paul Schroeder
Benoit Seguin
Phillip Stroëbel
Martin Volk
Thijs van Beek
Lars Wieneke

Ziele und Forschungsfragen

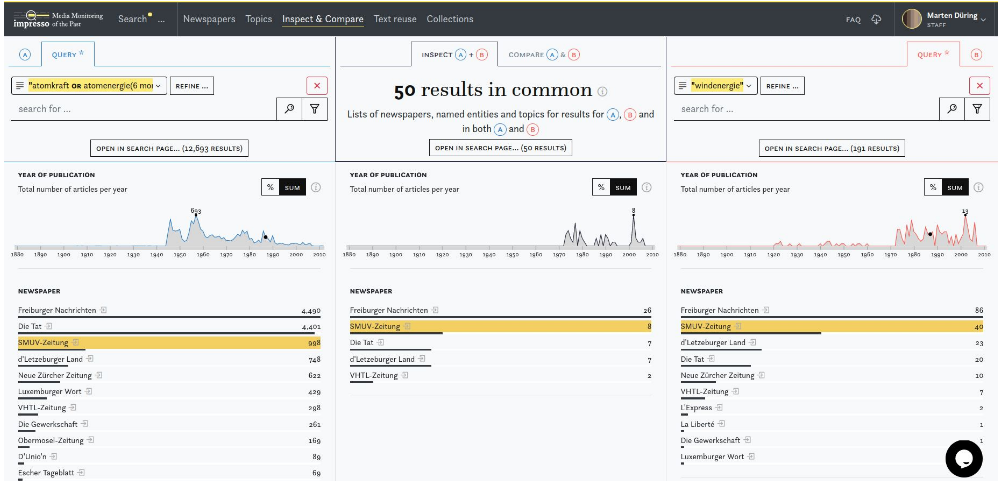
1. Wie passt man NLP-Werkzeuge an historischen Text an?

2. Wie erforscht man große und komplex Datenbestände?

3. Welche Rückwirkungen hat dies auf historische Forschung?



Impressos Zielgruppe



50 results in common

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

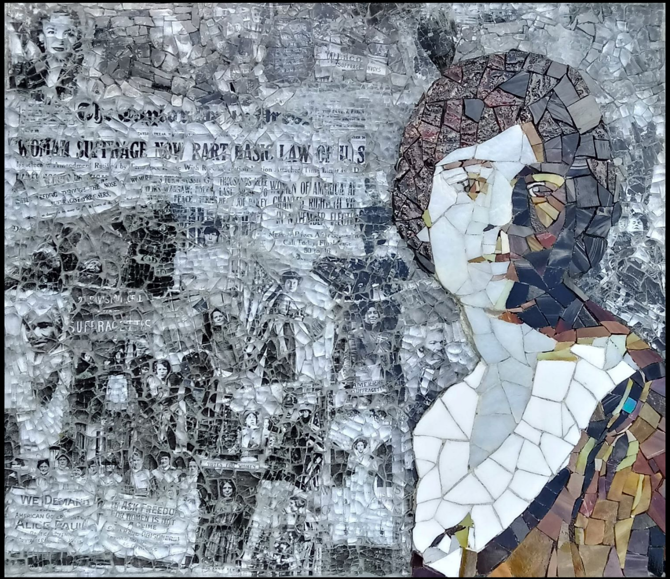
NEWSPAPER	
Freiburger Nachrichten	4,490
Die Tat	4,401
SMUV-Zeitung	998
d'Letzeburger Land	748
Neue Zürcher Zeitung	622
Luxemburger Wort	429
VHTL-Zeitung	298
Die Gewerkschaft	261
Obermosel-Zeitung	169
D'Unio'n	89
Escher Tageblatt	69

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	

HistorikerInnen ohne besondere digital Vorbildung aber mit der Neugier, Stichwortsuchen zu überwinden und der Motivation, ein neues Tool kennenzulernen.

[Gianni Sarconi, The Master of Numbers (2006)](#)



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

Impresso II

09/2023 - 02/2027

impresso
Media
Monitoring
of the Past

Beyond Borders: Connecting Historical Newspapers and Radio

Ziele

- Anreicherung und Integration von **Zeitungs- und Radioquellen in einem semantischen Raum;**
- Korpuserweiterung auf Westeuropa zusammen mit 20 Projektpartnern;
- Interfaces für die Exploration und datengetriebene Analyse;
- Fallstudien in (Medien) Geschichte mit dem Leitthema "influences."

Partner

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

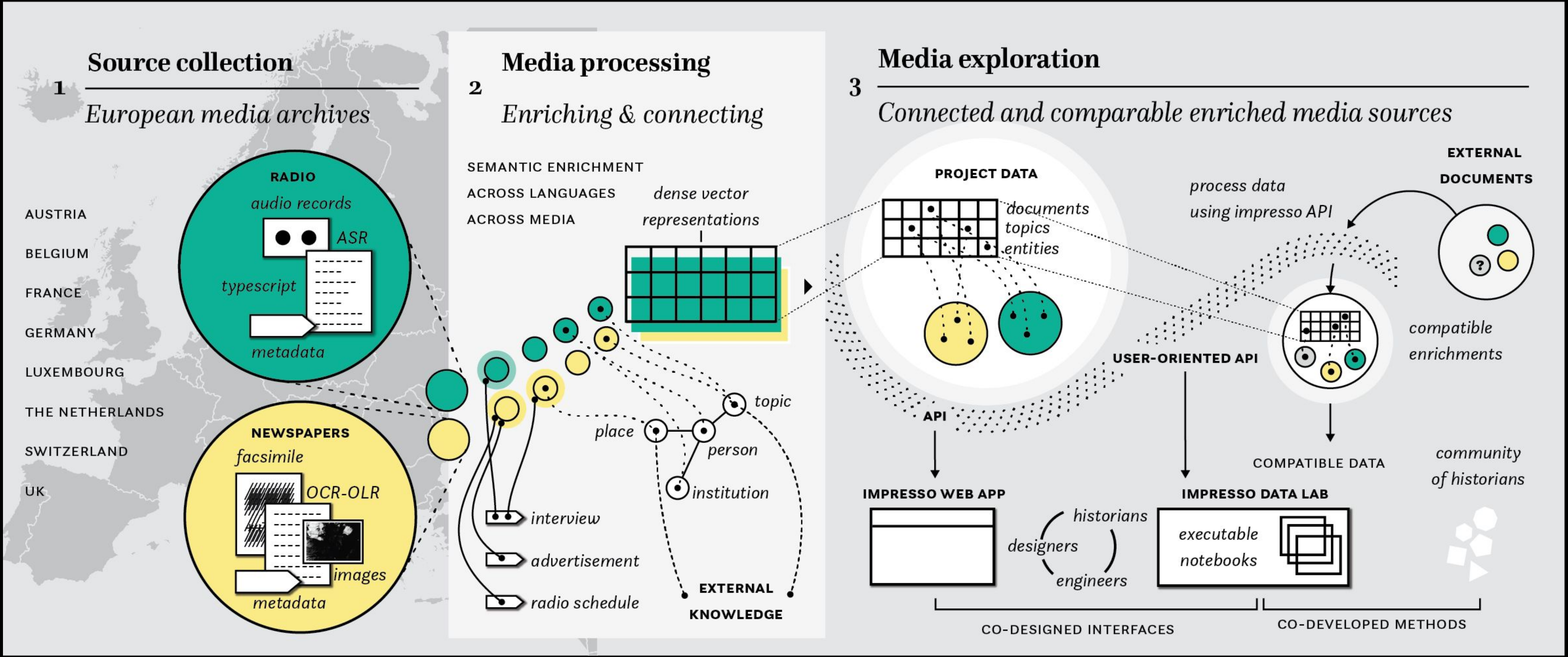
Le Temps
Neue Zürcher Zeitung

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

USER-ORIENTED API

process data using impresso API

EXTERNAL DOCUMENTS

compatible enrichments

COMPATIBLE DATA

community of historians

IMPRESSO WEB APP

historians
designers
engineers

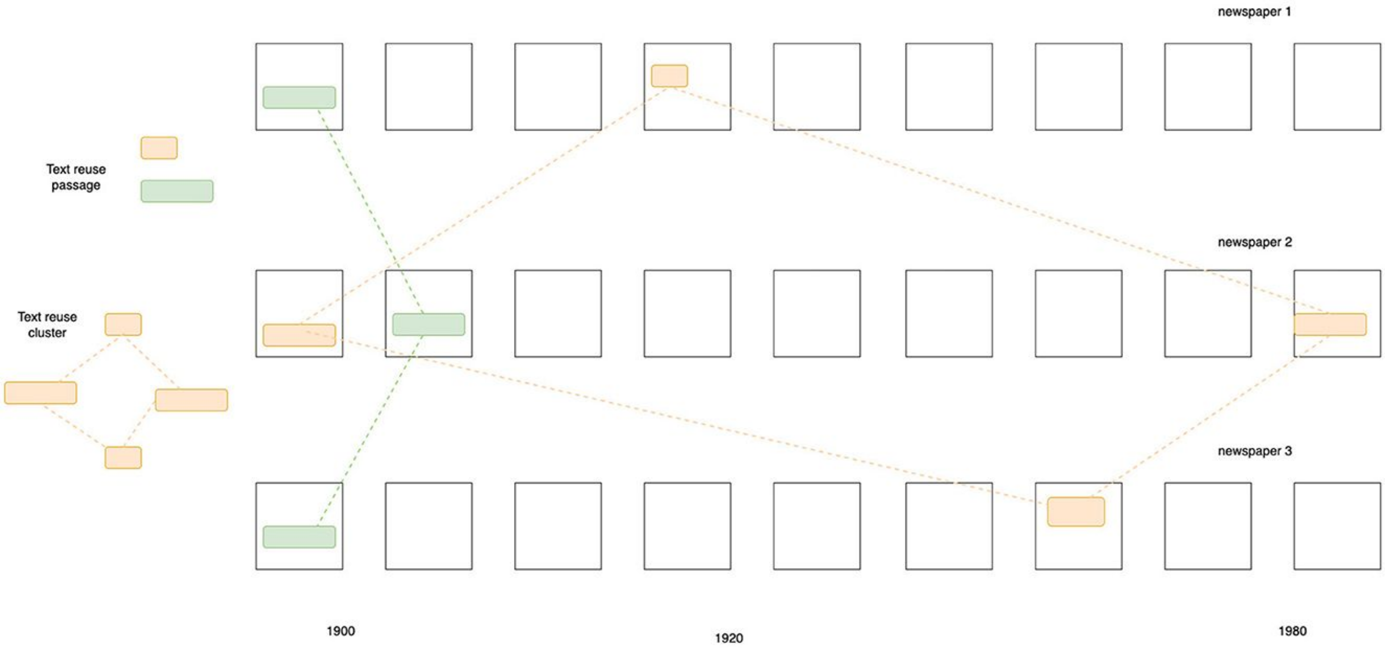
IMPRESSO DATA LAB

executable notebooks

CO-DESIGNED INTERFACES

CO-DEVELOPED METHODS

3. Text reuse


newspaper 1
Text reuse passage
newspaper 2
Text reuse cluster
newspaper 3
1900
1920
1980

Über Text Reuse

Text reuse ist die “the meaningful reiteration of text, usually beyond the simple repetition of common language” (Romanello et al. 2014).

Warum text reuse?

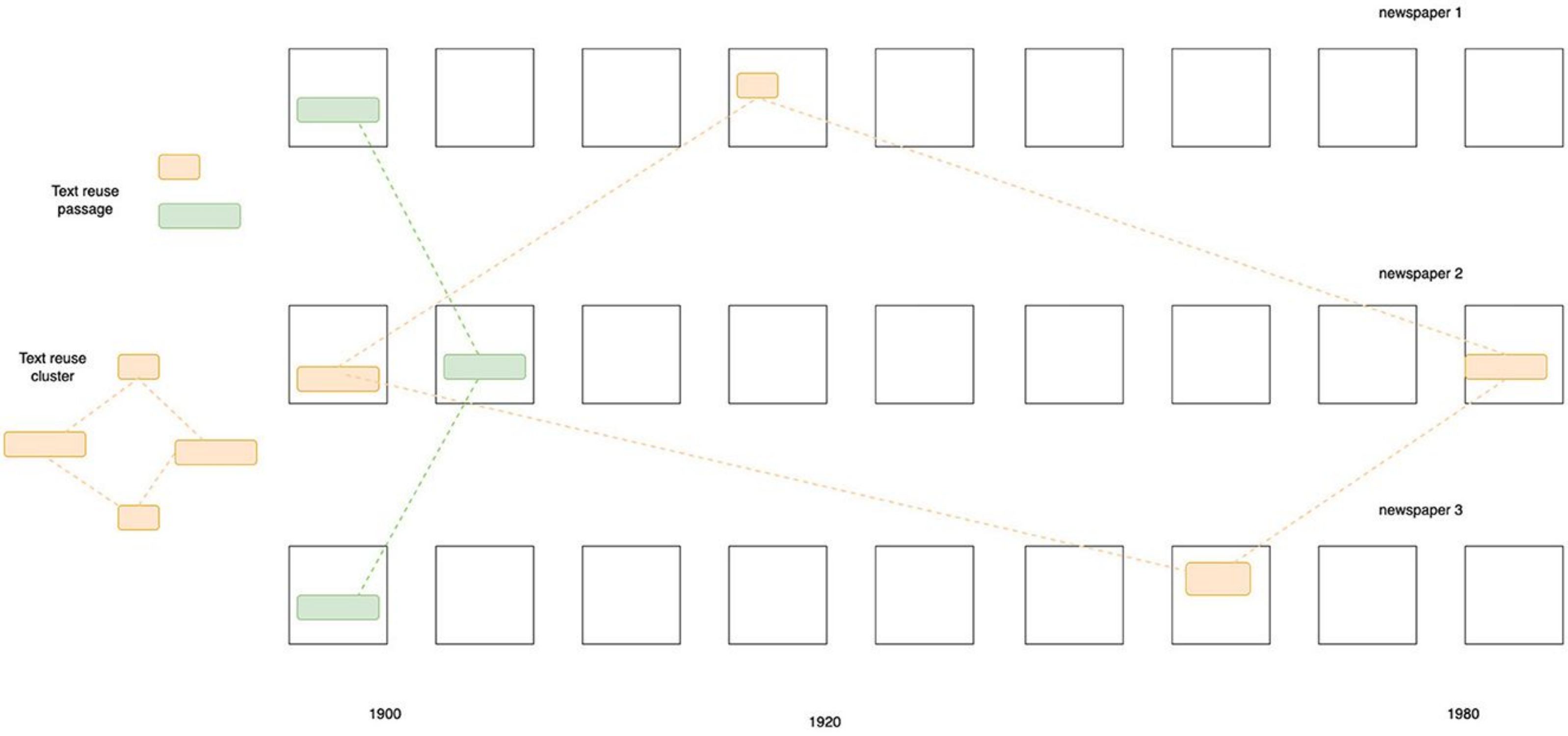
- Um festzustellen, ob eine digitale Bibliothek mehrere Ausgaben desselben Werks bzw. derselben Werke enthält.

- Um Zitate in einem Text zu finden, vorausgesetzt, die Zielwerke sind bekannt (z. B. Zitate aus der Bibel in der englischen Literatur des 17. Jahrhunderts finden) Untersuchung der Viralität und Verbreitung von Texten (z. B. Viral Texts von Cordell und Smith für historische Zeitungen)

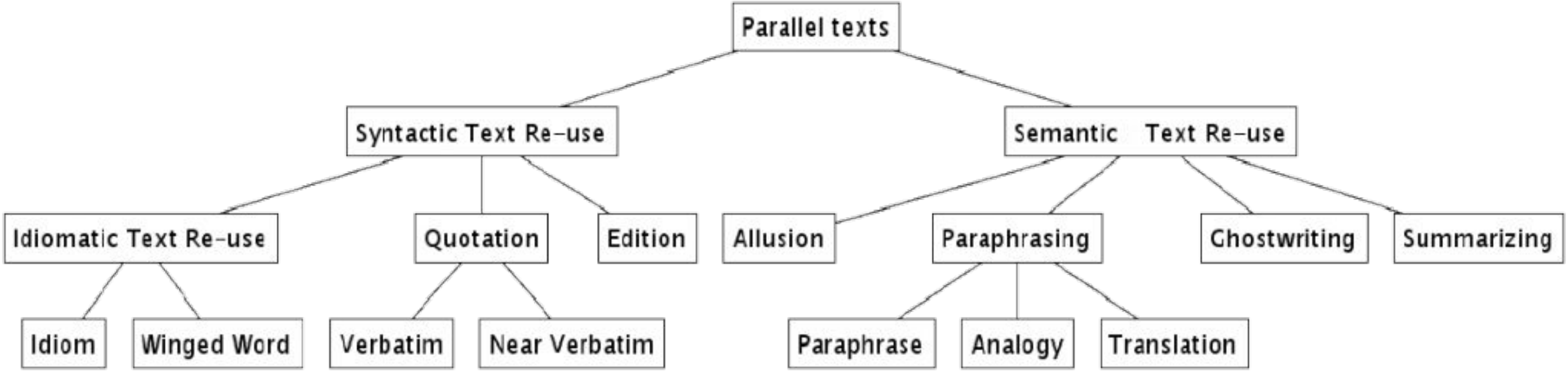
- Um Duplikate innerhalb einer Textsammlung zu identifizieren (und möglicherweise herauszufiltern), bevor weitere Verarbeitungsschritte durchgeführt werden (z. B. topic modeling, wie von Schofield et al. (2017) dargestellt)

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

Text reuse Passagen und Cluster

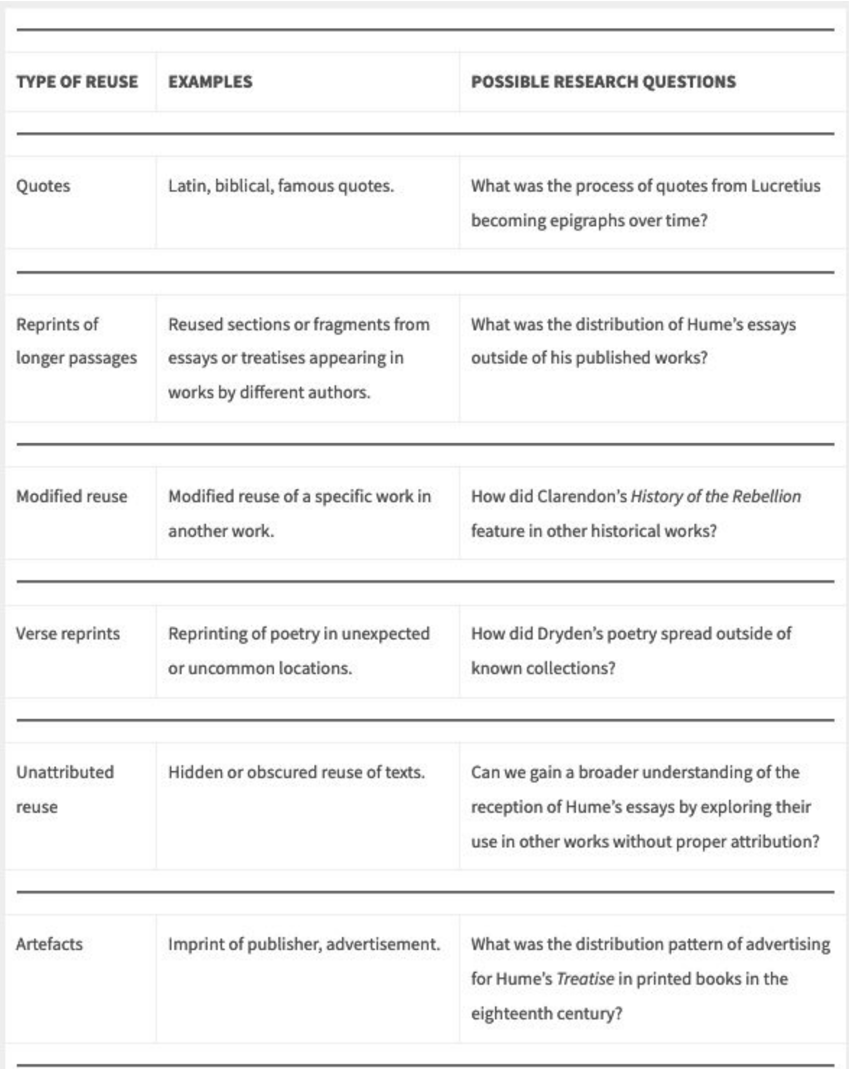


Typen von Text Reuse



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

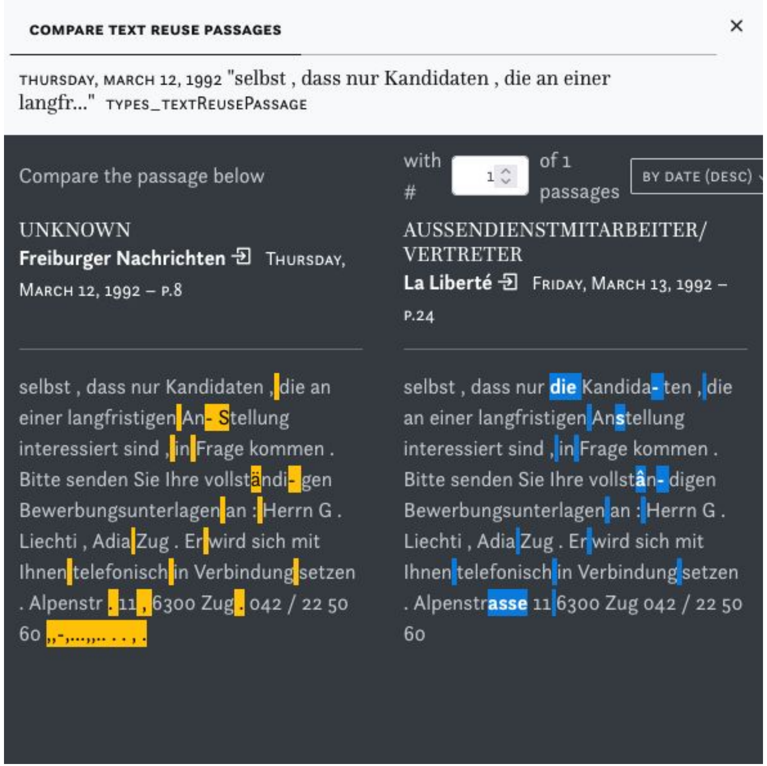
Typen von Text Reuse



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

Beispiele: Job-Anzeigen - intendierte Zirkulation



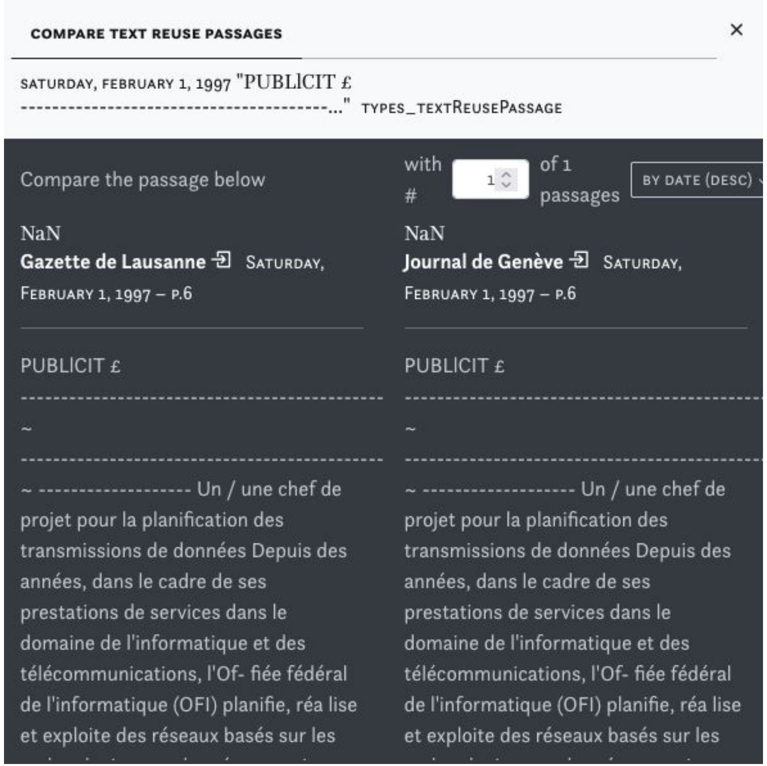
THURSDAY, MARCH 12, 1992 "selbst , dass nur Kandidaten , die an einer langfr..." TYPES_TEXTREUSEPASSAGE

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 . Alpenstrasse 11 6300 Zug 042 / 22 50 60



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

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

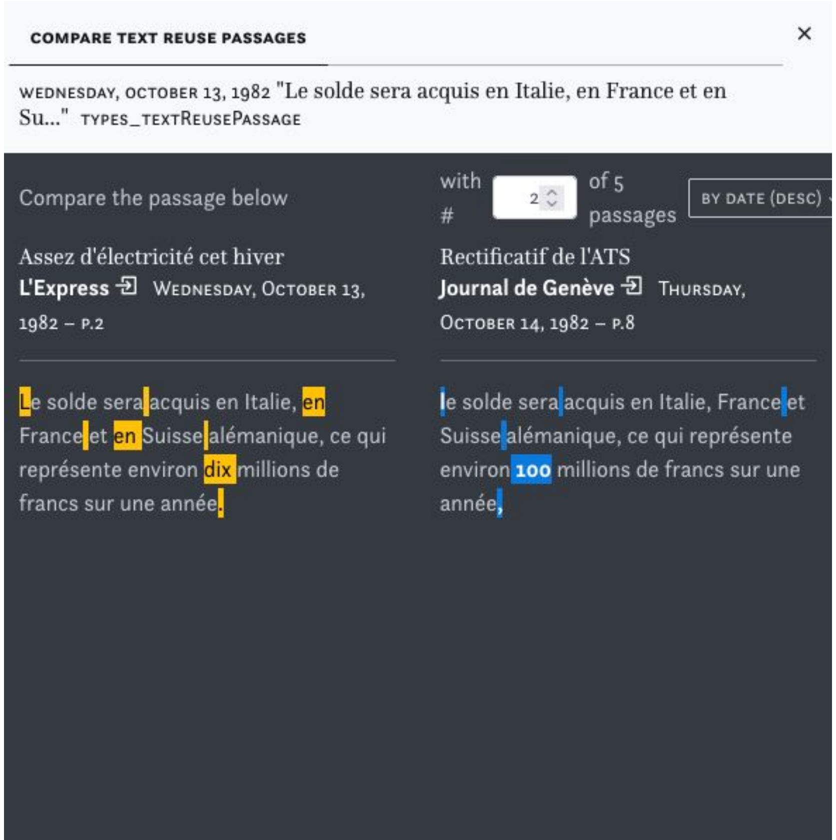
NaN
Journal de Genève SATURDAY, FEBRUARY 1, 1997 – P.6

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

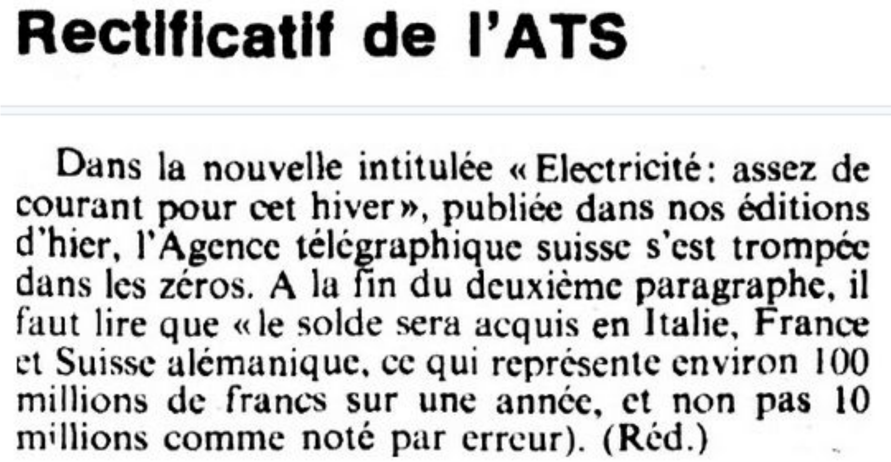
Beispiele: Publikation und Korrektur (modified reuse)



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

Compare the passage below with # 2 of 5 passages BY DATE (DESC)

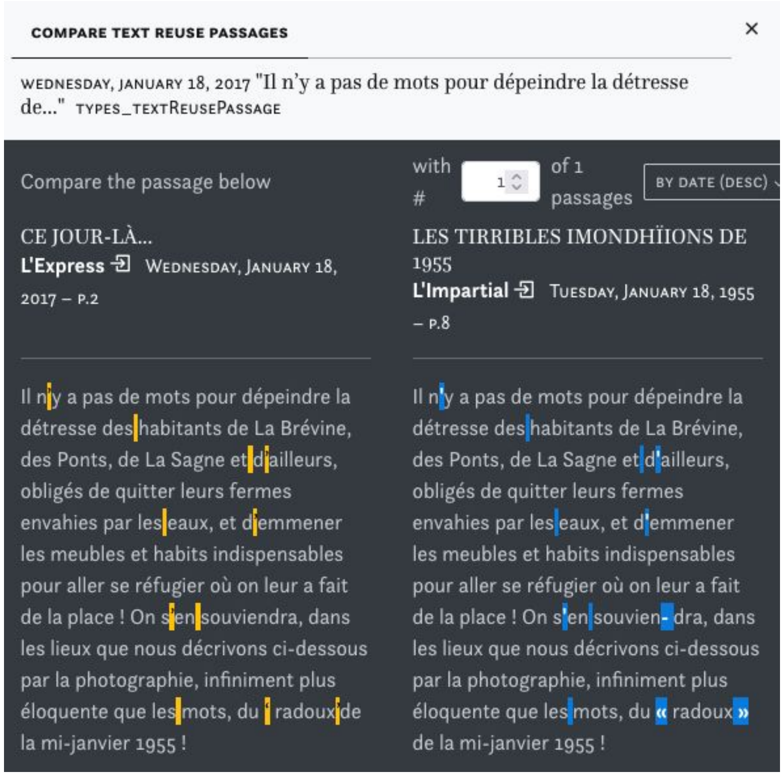
Assez d'électricité cet hiver L'Express WEDNESDAY, OCTOBER 13, 1982 – P.2	Rectificatif de l'ATS Journal de Genève THURSDAY, OCTOBER 14, 1982 – P.8
Le solde sera acquis en Italie, en France et en Suisse alémanique, ce qui représente environ dix millions de francs sur une année.	le solde sera acquis en 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.)

Beispiele: Nachdrucke zu Anlässen (reprints)



COMPARE TEXT REUSE PASSAGES ×

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

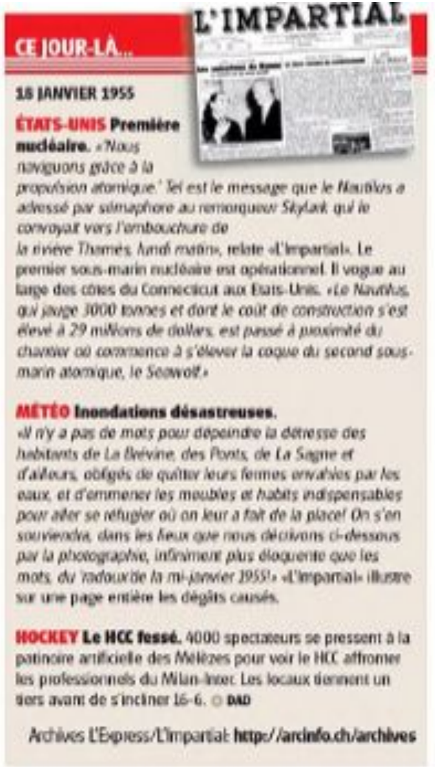
Compare the passage below with # 1 of 1 passages BY DATE (DESC)

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

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 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 !

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 !



CE JOUR-LÀ...

L'IMPARTIAL

18 JANVIER 1955

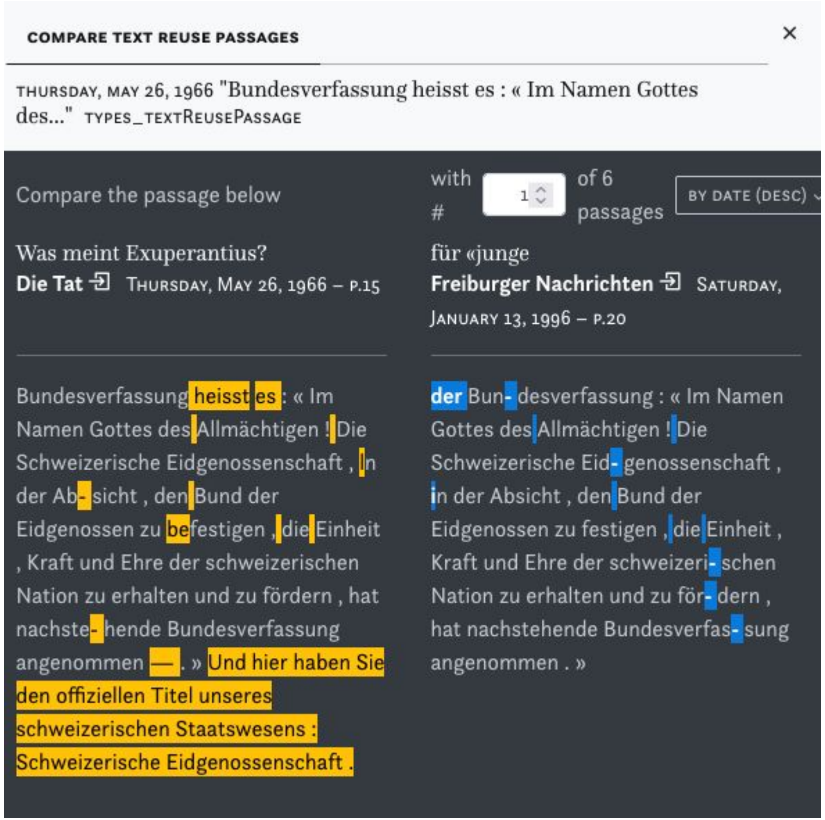
ÉTATS-UNIS Première nucléaire. *«'Nous naviguons grâce à la propulsion atomique.' Tel est le message que le Nautilus a adressé par sémaphore au remorqueur Skylark qui le convoyait vers l'embouchure de la rivière Thames, lundi matin»*, relate «L'Impartial». Le premier sous-marin nucléaire est opérationnel. 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**

Beispiele: Zitate (quotes)



COMPARE TEXT REUSE PASSAGES ×

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

Compare the passage below

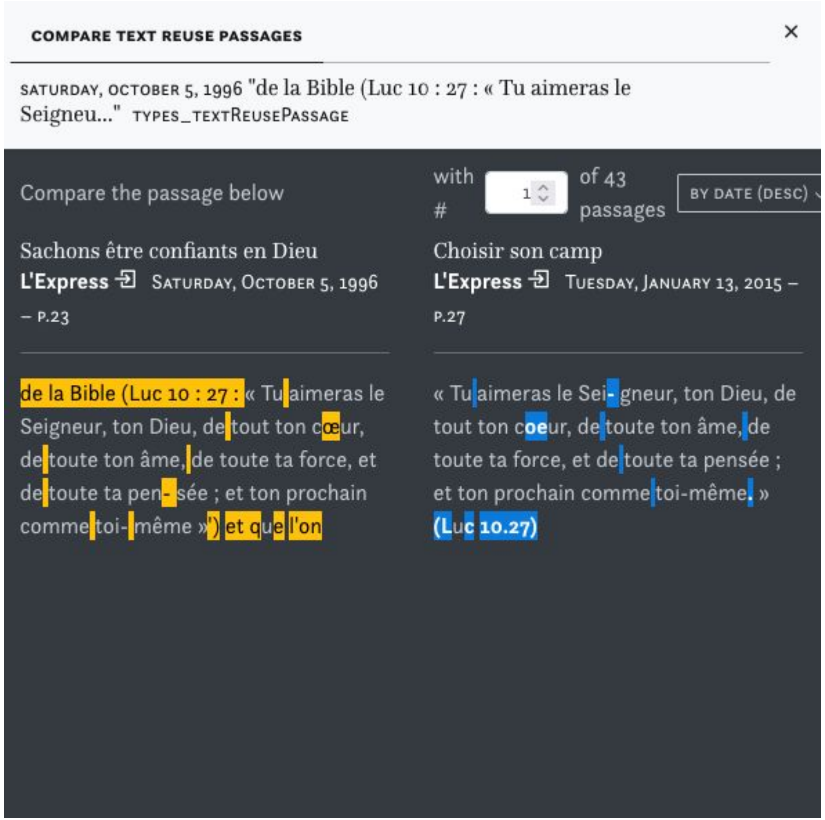
with # 1 of 6 passages BY DATE (DESC)

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)

Text reuse and historical research objectives

1. (Trans-) national media ecosystems
2. Newspaper content as *bricolage*
3. Historicising virality
4. Tracing historical events
5. Capturing historical Zeitgeist

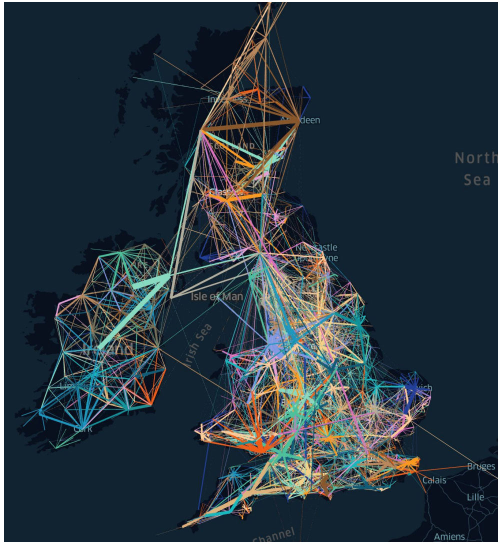
(Trans-) nationale Medien-Ökosysteme

Goal: Highlight the connectivity of historical media across borders and languages. Which ideological, commercial, and financial structures made this work and how did they shape media?

TR: (How) does content flow through the international network of newspapers?

Focus e.g. on cut-paste practices, the relevance of information infrastructure, geography, cross-border cultural affinities.

- Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational methods for uncovering reprinted texts in antebellum newspapers. *Am. Liter. Hist*. 27, E1–E15. doi: 10.1093/alh/ajv029
- Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news: a computational and multilingual approach to digitized newspaper collections. *Media History* 29, 287–304. doi: 10.1080/13688804.2022.2146905
- Salmi, H., Rantala, H., Vesanto, A., and Ginter, F. (2019). "The long-term reuse of text in the finnish press, 1771–1920," in *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, eds. C. Navarretta, M. Agirrezabal, and B. Maegaard (Copenhagen, Denmark: CEUR Workshop Proceedings), 253–273.
- Paju, P., Salmi, H., Rantala, H., Lundell, P., and Marjanen, Vesanto, A. (2022). "Textual migration across the baltic sea: Creating a database of text reuse between Finland and Sweden," in *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR Workshop Proceedings*, eds. K. Berglund, M. La Mela, and I. Zwart (Aachen: CEUR-WS.org), 361–369.
- Beelen, K. Digitising newspapers press directories to understand the landscape of historical newspapers, 2023, https://livingwithmachines.ac.uk/digitising-newspapers-press-directories-to-understand-the-landscape-of-historical-newspapers/.



Beelen, K. Digitising newspapers press directories to understand the landscape of historical newspapers

Newspaper content as bricolage

From which textual resources were newspapers compiled?

TR: How did formulaic ways of representing information emerge and spread (e.g. weather reports)? Can we determine the origins of content? Which content was influential?

Focuses on parallel developments across titles and the emergence of common practices.

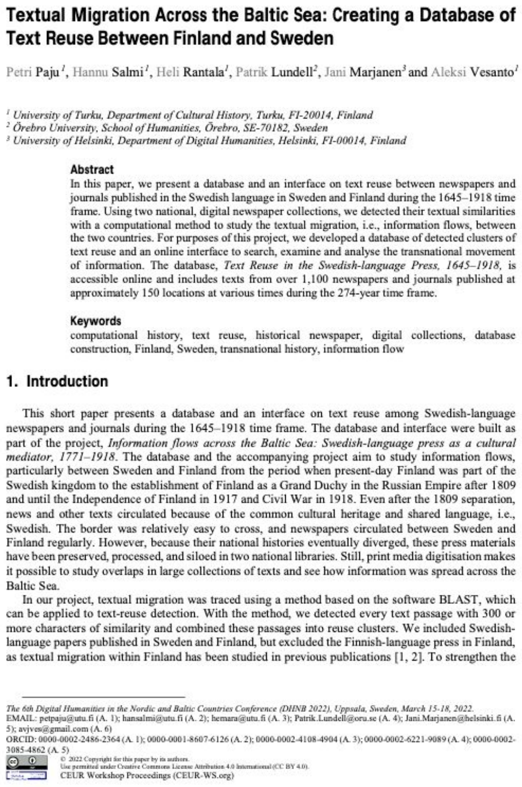
- Walma, L. W. B. (2015). Filtering the “news:” Uncovering morphine's multiple meanings on delpher's dutch newspapers and the need to distinguish more article types. *Tijdschrift voor Tijdschriftstudies*. 38, 61–78. doi: 10.18352/ts.345
- Thèrenty, M.-E., and Venayre, S. (2021). *Le monde à la une. Une histoire de la presse par ses rubriques*. Anamosa, illustrated èdition edition. doi: 10.3917/anamo.there.2021.02

Historicising virality

Which conditions (geography, type of information, infrastructure) enable rapid dissemination?

Focus is on the breadth and speed with which content spreads. Pioneering work of Paju et al. who define a virality score based on the number of titles within a cluster, the number of unique printing locations, and the distance in days between the first and last passage publication date.

Paju, P., Salmi, H., Rantala, H., Lundell, P., and Marjanen, Vesanto, A. (2022). “Textual migration across the baltic sea: Creating a database of text reuse between Finland and Sweden,” in *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR Workshop Proceedings*, eds. K. Berglund, M. La Mela, and I. Zwart (Aachen: CEUR-WS.org), 361–369.



Textual Migration Across the Baltic Sea: Creating a Database of Text Reuse Between Finland and Sweden

Petri Paju[1], Hannu Salmi[1], Heli Rantala[1], Patrik Lundell[2], Jani Marjanen[3] and Aleksi Vesanto[1]

[1] *University of Turku, Department of Cultural History, Turku, FI-20014, Finland*
[2] *Örebro University, School of Humanities, Örebro, SE-70182, Sweden*
[3] *University of Helsinki, Department of Digital Humanities, Helsinki, FI-00014, Finland*

Abstract

In this paper, we present a database and an interface on text reuse between newspapers and journals published in the Swedish language in Sweden and Finland during the 1645–1918 time frame. Using two national, digital newspaper collections, we detected their textual similarities with a computational method to study the textual migration, i.e., information flows, between the two countries. For purposes of this project, we developed a database of detected clusters of text reuse and an online interface to search, examine and analyse the transnational movement of information. The database, *Text Reuse in the Swedish-language Press, 1645–1918*, is accessible online and includes texts from over 1,100 newspapers and journals published at approximately 150 locations at various times during the 274-year time frame.

Keywords

computational history, text reuse, historical newspaper, digital collections, database construction, Finland, Sweden, transnational history, information flow

1. Introduction

This short paper presents a database and an interface on text reuse among Swedish-language newspapers and journals during the 1645–1918 time frame. The database and interface were built as part of the project, *Information flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–1918*. The database and the accompanying project aim to study information flows, particularly between Sweden and Finland from the period when present-day Finland was part of the Swedish kingdom to the establishment of Finland as a Grand Duchy in the Russian Empire after 1809 and until the Independence of Finland in 1917 and Civil War in 1918. Even after the 1809 separation, news and other texts circulated because of the common cultural heritage and shared language, i.e., Swedish. The border was relatively easy to cross, and newspapers circulated between Sweden and Finland regularly. However, because their national histories eventually diverged, these press materials have been preserved, processed, and siloed in two national libraries. Still, print media digitisation makes it possible to study overlaps in large collections of texts and see how information was spread across the Baltic Sea.

In our project, textual migration was traced using a method based on the software BLAST, which can be applied to text-reuse detection. With the method, we detected every text passage with 300 or more characters of similarity and combined these passages into reuse clusters. We included Swedish-language papers published in Sweden and Finland, but excluded the Finnish-language press in Finland, as textual migration within Finland has been studied in previous publications [1, 2]. To strengthen the

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: petpaju@utu.fi (A. 1); hansalmi@utu.fi (A. 2); hemara@utu.fi (A. 3); Patrik.Lundell@oru.se (A. 4); Jani.Marjanen@helsinki.fi (A. 5); avjves@gmail.com (A. 6)
ORCID: 0000-0002-2486-2364 (A. 1); 0000-0001-8607-6126 (A. 2); 0000-0002-4108-4904 (A. 3); 0000-0002-6221-9089 (A. 4); 0000-0002-3085-4862 (A. 5)

Tracing historical events

How do individual media titles situate events in the political, economic, social, and cultural context relevant to them? How does this affect the perception of these events by their audiences?

Focus is on the appropriation of often global news on the local scale.

- Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., et al. (2020). Spreading news in 1904. *Media History* 26, 391–407. doi: 10.1080/13688804.2019.1652090
- Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news: a computational and multilingual approach to digitized newspaper collections. *Media History* 29, 287–304. doi: 10.1080/13688804.2022.2146905

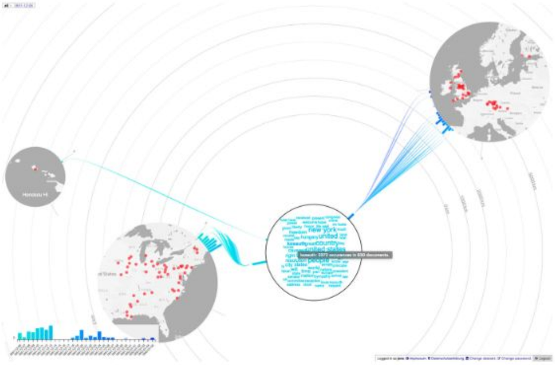
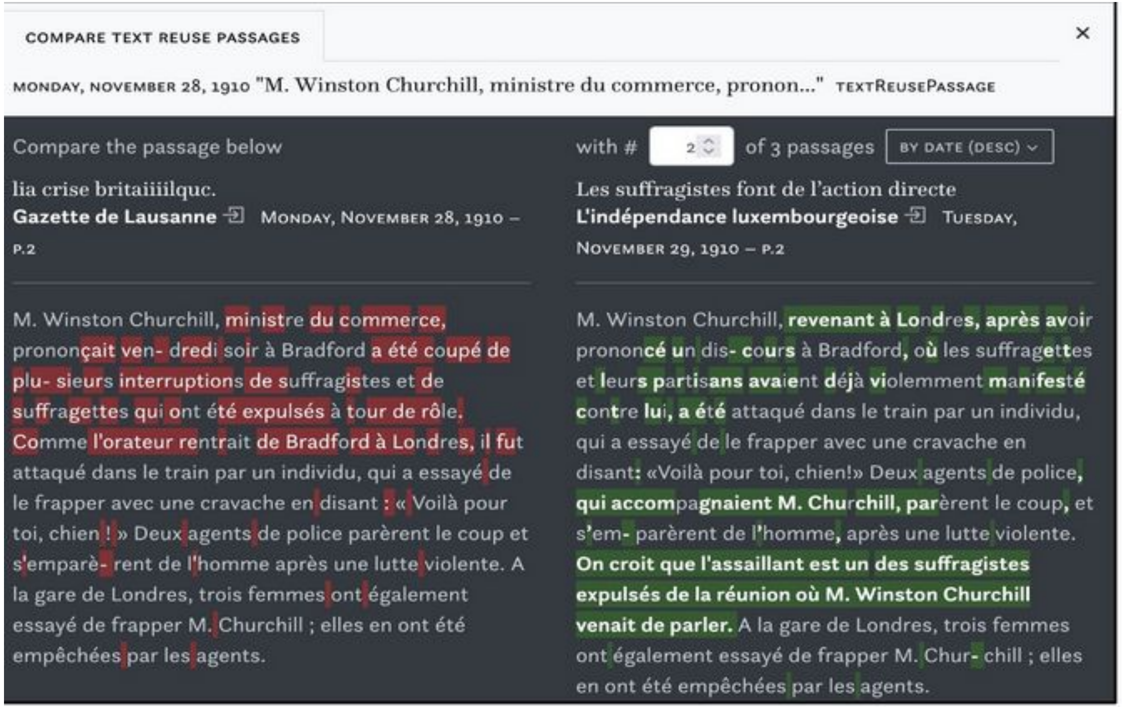


FIGURE 2.
The bar plot of the number of reprinted articles mentioning Kossuth in OcEx corpus shows that the American tour established Kossuth as an international celebrity.

Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news

Tracing historical events



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.

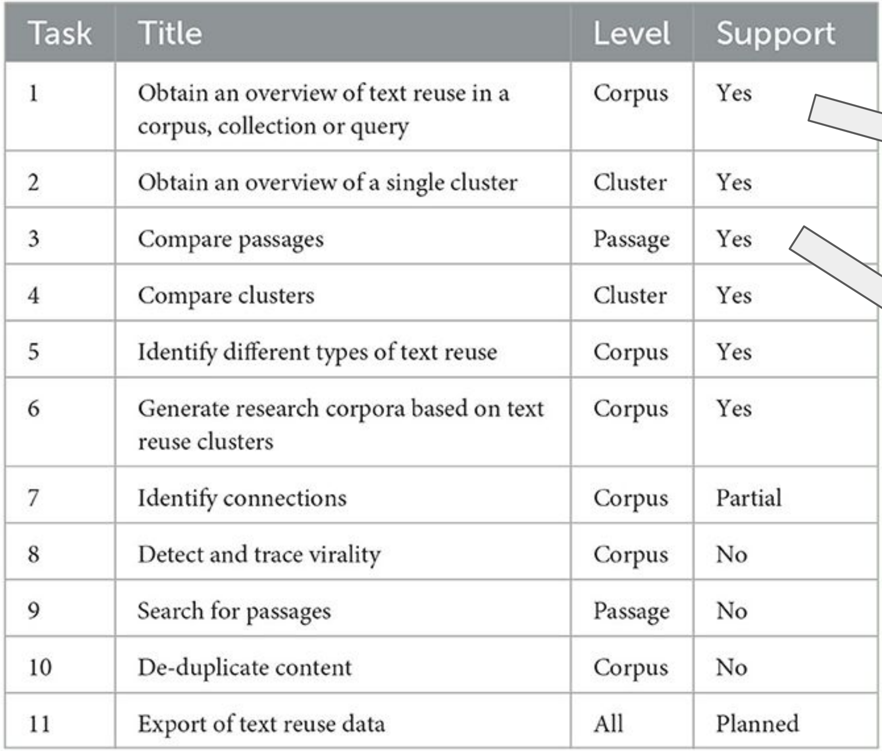
Capturing historical Zeitgeist

Historical media partially capture the attitudes, norms, beliefs, moods and feelings of past generations, or Zeitgeist.

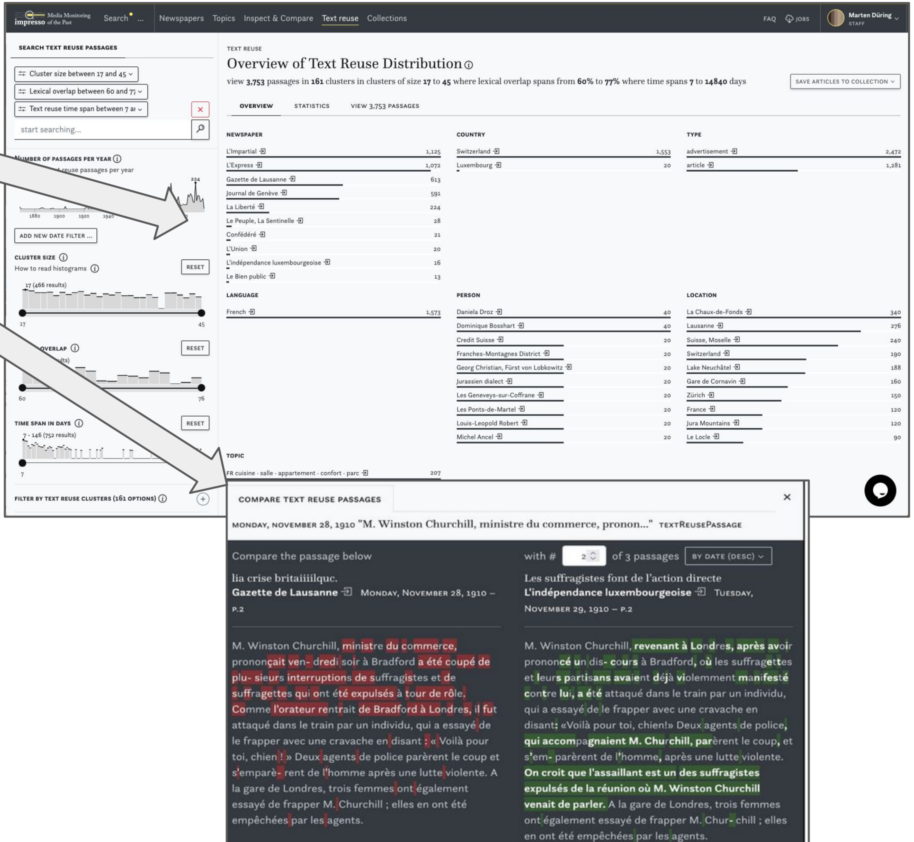
Focus is on texts which were produced independently but still share certain characteristics.

- Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., et al. (2022). Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914. *Dig. Human. Quart*. 16, 7445. Available online at: https://www.digitalhumanities.org/dhq/vol/16/2/000550/000550.html
- Paasikivi, S., Salmi, H., Vesanto, A., and Ginter, F. (2022). Infectious media: Cholera and the circulation of texts in the finnish press, 1860–1920. *Media Hist*. 29, 17–38. doi: 10.1080/13688804.2022.2054408

Tasks to support research objectives



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	Yes
3	Compare passages	Passage	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	Passage	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



TEXT REUSE

Overview of Text Reuse Distribution

view **3,753** passages in **161** clusters in clusters of size **17** to **45** where lexical overlap spans from **60%** to **77%** where time spans **7** to **14840** days

NEWSPAPER	
L'Impartial	1,125
L'Express	1,072
Gazette de Lausanne	613
Journal de Genève	591
La Liberté	224
Le Peuple, La Sentinelle	28
Confédéré	21
L'Union	20
L'indépendance luxembourgeoise	16
Le Bien public	13

COUNTRY	
Switzerland	1,553
Luxembourg	20

TYPE	
advertisement	2,472
article	1,281

LANGUAGE	
French	1,573

PERSON	
Daniela Droz	40
Dominique Bosshart	40
Credit Suisse	20
Franches-Montagnes District	20
Georg Christian, Fürst von Lobkowitz	20
Jurassien dialect	20
Les Geneveys-sur-Coffrane	20
Les Ponts-de-Martel	20
Louis-Leopold Robert	20
Michel Ancel	20

LOCATION	
La Chaux-de-Fonds	340
Lausanne	276
Suisse, Moselle	240
Switzerland	190
Lake Neuchâtel	188
Gare de Cornavin	160
Zürich	150
France	120
Jura Mountains	120
Le Locle	90

TOPIC	
FR cuisine · salle · appartement · confort · parc	207

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

Compare the passage below with # 2 of 3 passages

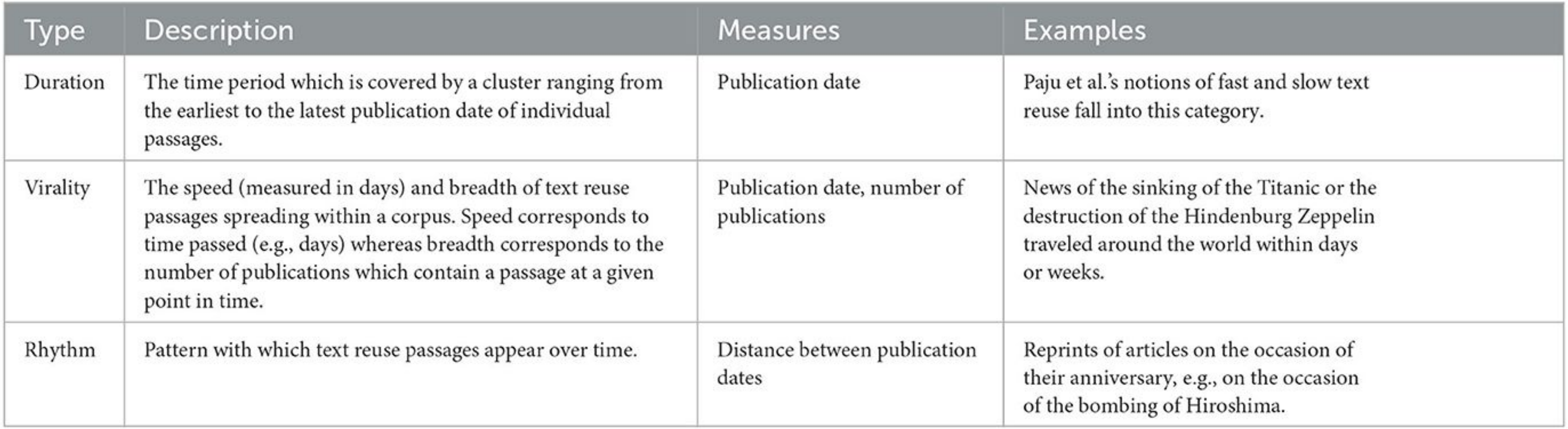
lia crise britaiiiilquc.
Gazette de Lausanne MONDAY, NOVEMBER 28, 1910 – P.2

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

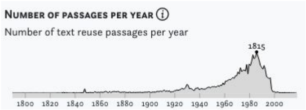
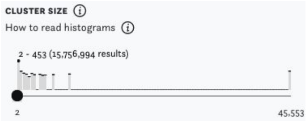
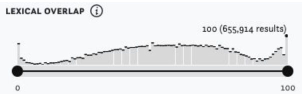
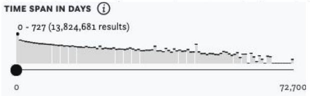
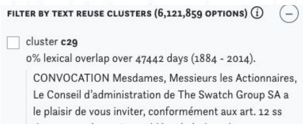
M. Winston Churchill, revenant à Londres, après avoir prononcé un dis- cours à Bradford, où les suffragettes et leurs partisans avaient déjà violemment manifesté contre lui, a été attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant: «Voilà pour toi, chien!» Deux 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.

Temporalities in text reuse data

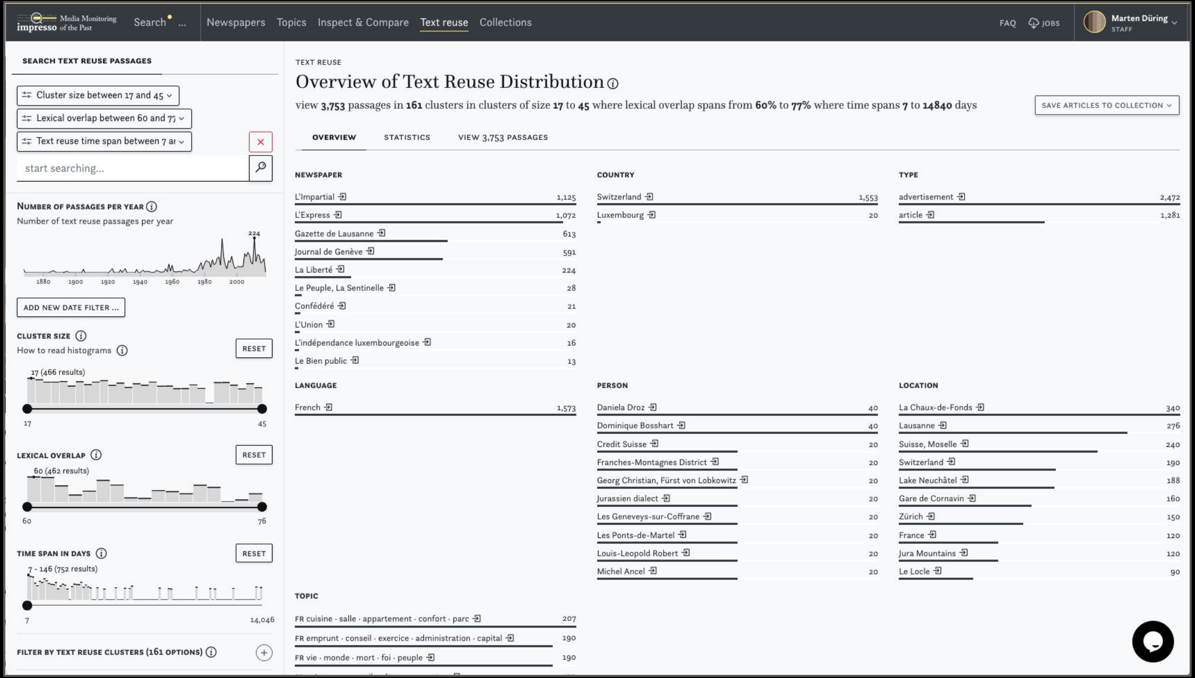


Type	Description	Measures	Examples
Duration	The time period which is covered by a cluster ranging from the earliest to the latest publication date of individual passages.	Publication date	Paju et al.'s notions of fast and slow text reuse fall into this category.
Virality	The speed (measured in days) 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.

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

3. Demo *impresso* Text reuse at Scale



SEARCH TEXT REUSE PASSAGES

Cluster size between 17 and 45

Lexical overlap between 60 and 77

Text reuse time span between 7 ar

start searching...

Number of passages per year

Number of text reuse passages per year

ADD NEW DATE FILTER ...

CLUSTER SIZE

How to read histograms

RESET

LEXICAL OVERLAP

RESET

TIME SPAN IN DAYS

RESET

FILTER BY TEXT REUSE CLUSTERS (161 OPTIONS)

TEXT REUSE

Overview of Text Reuse Distribution

view **3,753** passages in **161** clusters in clusters of size **17** to **45** where lexical overlap spans from **60%** to **77%** where time spans **7** to **14840** days

SAVE ARTICLES TO COLLECTION

OVERVIEW STATISTICS VIEW 3,753 PASSAGES

NEWSPAPER	
L'Impartial	1,125
L'Express	1,072
Gazette de Lausanne	613
Journal de Genève	591
La Liberté	224
Le Peuple, La Sentinelle	28
Confédéré	21
L'Union	20
L'indépendance luxembourgeoise	16
Le Bien public	13

COUNTRY	
Switzerland	1,553
Luxembourg	20

TYPE	
advertisement	2,472
article	1,281

LANGUAGE	
French	1,573

PERSON	
Daniela Droz	40
Dominique Bosshart	40
Credit Suisse	20
Franches-Montagnes District	20
Georg Christian, Fürst von Lobkowitz	20
Jurassien dialect	20
Les Geneveys-sur-Coffrane	20
Les Ponts-de-Martel	20
Louis-Leopold Robert	20
Michel Ancel	20

LOCATION	
La Chaux-de-Fonds	340
Lausanne	276
Suisse, Moselle	240
Switzerland	190
Lake Neuchâtel	188
Gare de Cornavin	160
Zürich	150
France	120
Jura Mountains	120
Le Locle	90

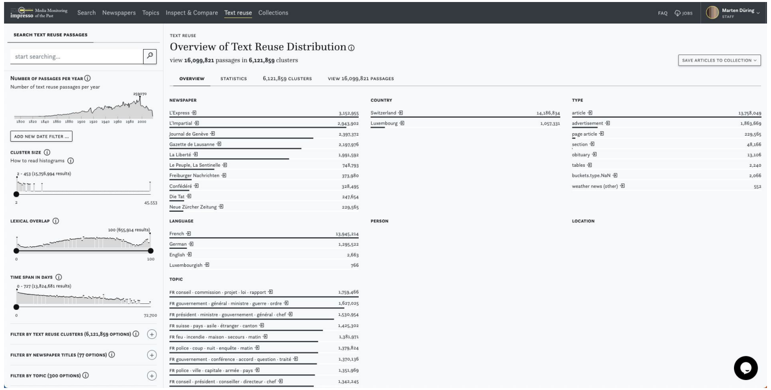
TOPIC	
FR cuisine · salle · appartement · confort · parc	207
FR emprunt · conseil · exercice · administration · capital	190
FR vie · monde · mort · foi · peuple	190

Types of text reuse in newspapers seen through filters

	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

Hands-on part and Challenge: Search and lexical overlap

1. **Go to** https://bit.ly/impresso-tr
2. **Go to** https://dev.impresso-project.ch/app/ and login
3. **Search** for “titanic”
4. **Exercise**: Familiarise yourself with all the different filters and views. Which questions emerge?
5. **Lexical overlap**: Go to Passages and experiment with the different sorting choices. What do you observe?



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)

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

FILTER BY NEWSPAPER TITLES (77 OPTIONS)

FILTER BY TOPIC (300 OPTIONS)

TEXT REUSE

Overview of Text Reuse Distribution

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

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

LOCATION

TOPIC	
FR conseil · commission · projet · loi · rapport	1,759,466
FR gouvernement · général · ministre · guerre · ordre	1,627,025
FR président · ministre · gouvernement · général · chef	1,530,954
FR suisse · pays · asile · étranger · canton	1,425,302
FR feu · incendie · maison · secours · matin	1,381,971
FR police · coup · nuit · enquête · matin	1,379,824
FR gouvernement · conférence · accord · question · traité	1,370,136
FR police · ville · capitale · armée · pays	1,351,969
FR conseil · président · conseiller · directeur · chef	1,342,245

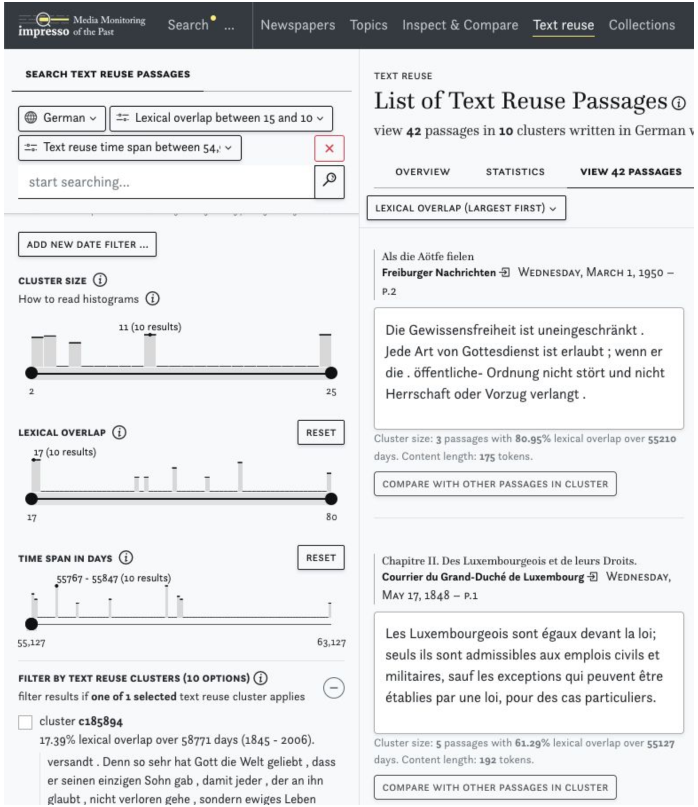
URL: https://dev.impresso-project.ch/app

Login: student@impresso-project.ch

Passwort: Marburg2023!

Hands-on part and Challenge: Time span

1. Start a new search by clicking the red “X”
2. Filter for “German”
3. Set Lexical overlap to 15% - 100%
4. Move to Passages view. What do you get?
5. Set time span to 55.00 - 67.000 days. What do you get?



FILTER BY TEXT REUSE CLUSTERS (10 OPTIONS)

filter results if **one of 1 selected** text reuse cluster applies

□ cluster **c185894**

17.39% lexical overlap over 58771 days (1845 - 2006).

versandt . Denn so sehr hat Gott die Welt geliebt , dass er seinen einzigen Sohn gab , damit jeder , der an ihn glaubt , nicht verloren gehe , sondern ewiges Leben

List of Text Reuse Passages

view **42** passages in **10** clusters written in German v

Als die Aötfe fielen

Freiburger Nachrichten Wednesday, March 1, 1950 – p.2

Die Gewissensfreiheit ist uneingeschränkt . Jede Art von Gottesdienst ist erlaubt ; wenn er die . öffentliche- Ordnung nicht stört und nicht Herrschaft oder Vorzug verlangt .

Cluster size: **3** passages with **80.95%** lexical overlap over **55210** days. Content length: **175** tokens.

Chapitre II. Des Luxembourgeois et de leurs Droits.

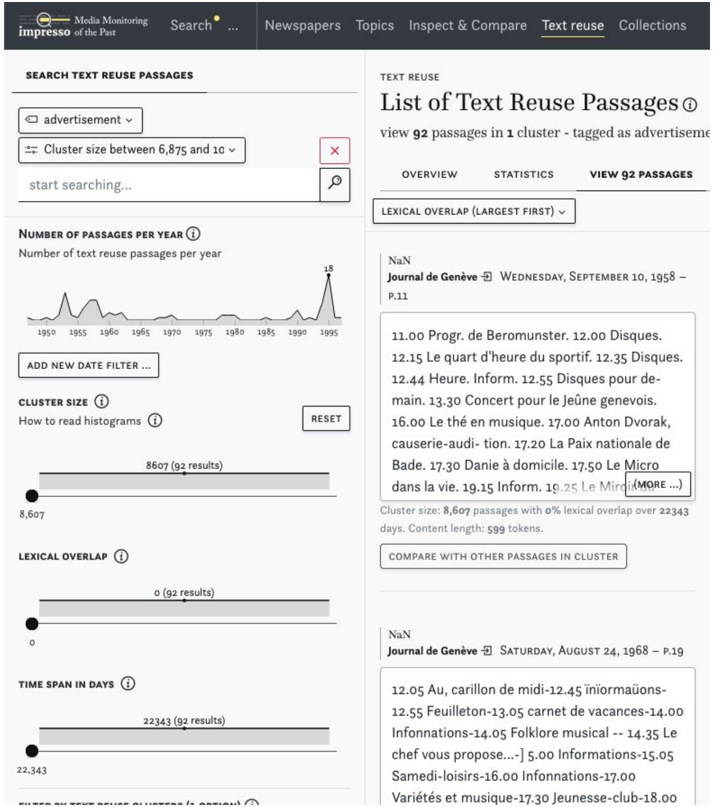
Courrier du Grand-Duché de Luxembourg Wednesday, May 17, 1848 – p.1

Les Luxembourgeois sont égaux devant la loi; seuls ils sont admissibles aux emplois civils et militaires, sauf les exceptions qui peuvent être établies par une loi, pour des cas particuliers.

Cluster size: **5** passages with **61.29%** lexical overlap over **55127** days. Content length: **192** tokens.

Hands-on part and Challenge: Cluster size

1. Start a new search by clicking the red “X”
2. Set filters for adverts
3. Set Cluster size between 90 - 100 passages.
4. What do you find? How can we explain this?



advertisement

Cluster size between 6,875 and 10

Number of text reuse passages per year

CLUSTER SIZE

8607 (92 results)

8,607

LEXICAL OVERLAP

0 (92 results)

0

TIME SPAN IN DAYS

22343 (92 results)

22,343

TEXT REUSE

List of Text Reuse Passages

view **92** passages in **1** cluster - tagged as advertiseme

NaN

Journal de Genève WEDNESDAY, SEPTEMBER 10, 1958 – P.11

11.00 Progr. de Beromunster. 12.00 Disques. 12.15 Le quart d'heure du sportif. 12.35 Disques. 12.44 Heure. Inform. 12.55 Disques pour de- main. 13.30 Concert pour le Jeûne genevois. 16.00 Le thé en musique. 17.00 Anton Dvorak, causerie-audi- tion. 17.20 La Paix nationale de Bade. 17.30 Danie à domicile. 17.50 Le Micro dans la vie. 19.15 Inform. 19.25 Le Miroir du

Cluster size: **8,607** passages with **0%** lexical overlap over **22343** days. Content length: **599** tokens.

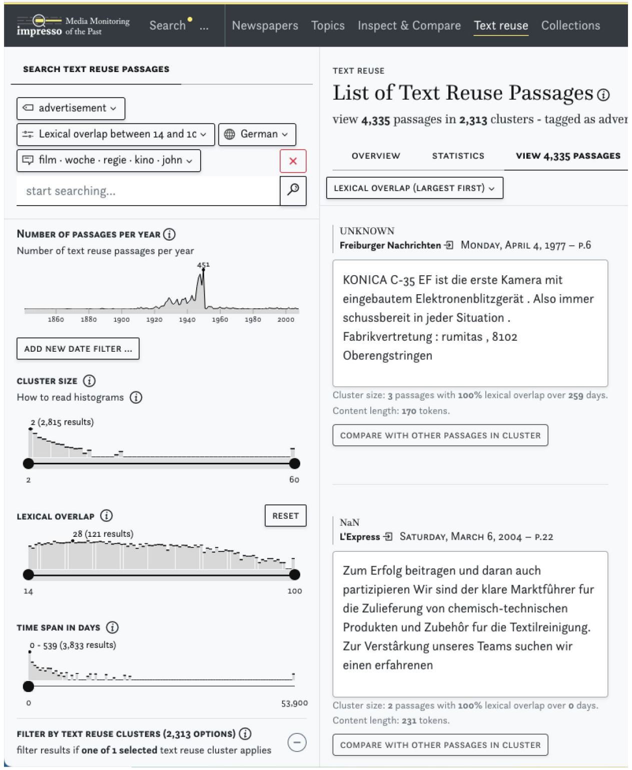
NaN

Journal de Genève SATURDAY, AUGUST 24, 1968 – P.19

12.05 Au, carillon de midi-12.45 ïnïormaüons-12.55 Feuilleton-13.05 carnet de vacances-14.00 Infonnations-14.05 Folklore musical -- 14.35 Le chef vous propose...-] 5.00 Informations-15.05 Samedi-loisirs-16.00 Infonnations-17.00 Variétés et musique-17.30 Jeunesse-club-18.00

Hands-on part and Challenge: Topics etc

1. Start a new search by clicking the red “X”
2. Set filters to “German” and adverts
3. Experiment with different topics: medical, classified ads, media…
4. Use the Passages tab and experiment with different sorting operations: e.g. largest cluster size vs. lexical overlap, passage size etc.



TEXT REUSE

List of Text Reuse Passages

view **4,335** passages in **2,313** clusters - tagged as adver

NUMBER OF PASSAGES PER YEAR

Number of text reuse passages per year

CLUSTER SIZE

How to read histograms

LEXICAL OVERLAP

TIME SPAN IN DAYS

FILTER BY TEXT REUSE CLUSTERS (2,313 OPTIONS)

filter results if **one of 1 selected** text reuse cluster applies

UNKNOWN

Freiburger Nachrichten MONDAY, APRIL 4, 1977 – P.6

KONICA C-35 EF ist die erste Kamera mit eingebautem Elektronenblitzgerät . Also immer schussbereit in jeder Situation . Fabrikvertretung : rumitas , 8102 Oberengstringen

Cluster size: **3** passages with **100%** lexical overlap over **259** days.

Content length: **170** tokens.

NaN

L'Express SATURDAY, MARCH 6, 2004 – P.22

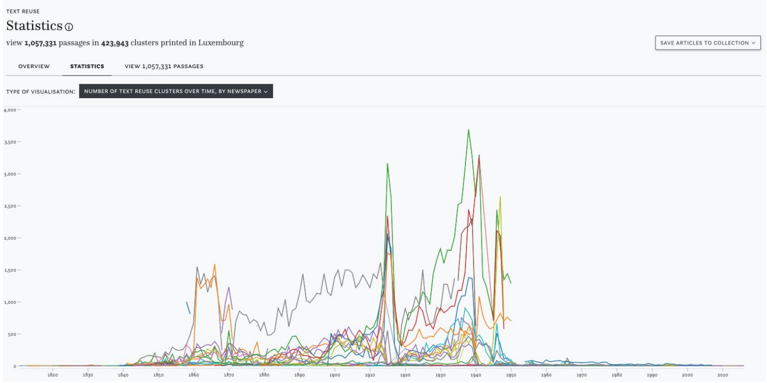
Zum Erfolg beitragen und daran auch partizipieren Wir sind der klare Marktfûhrer fur die Zulieferung von chemisch-technischen Produkten und Zubehôr fur die Textilreinigung. Zur Verstârkung unseres Teams suchen wir einen erfahrenen

Cluster size: **2** passages with **100%** lexical overlap over **0** days.

Content length: **231** tokens.

Hands-on part and Challenge: Topics etc

1. Start a new search by clicking the red “X”
2. Set the **country** filters to “Luxembourg”
3. Select the Statistics view + number of text reuse clusters over time
4. How can we explain this graph?



TEXT REUSE

Statistics ⓘ

view **1,057,331** passages in **423,943** clusters printed in Luxembourg

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

Your turn…find instances of:

1. Build you own queries, combine different keywords and filters. Which impact do they have?
2. Explore the Statistics view. What can it tell you about the data?
3. Open https://bit.ly/impresso-tr and go to slides 44ff. Pick one of the project ideas or duplicate slides to add your own.

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

Your names here

Your idea here

Observations:

<SCREENSHOT(S)>

URL to query	
Number of passages	
Lexical overlap	
Time span	

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Standardized content (weather, radio programme…)

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Your names here

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

The longest running advertisement

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Press agency reports

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Your names here

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

The two newspapers that co-publish the m

Observations:

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

A new form of text reuse!?

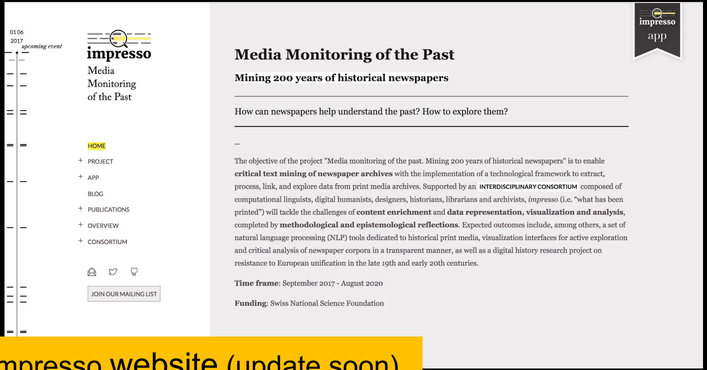
<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

For more information



Media
Monitoring
of the Past

Media Monitoring of the Past

Mining 200 years of historical newspapers

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

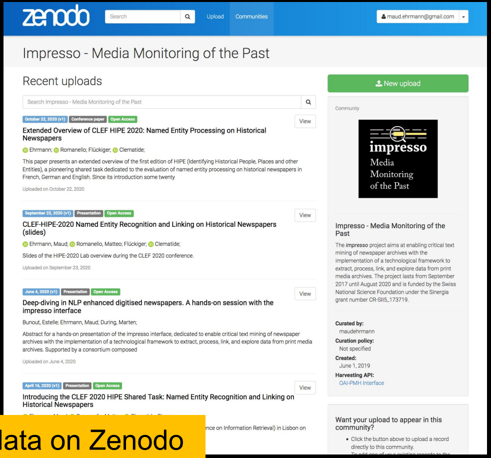
—

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)



Impresso - Media Monitoring of the Past

Recent uploads

October 22, 2020 (v1) Conference paper Open Access

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

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

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

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

…nce on Information Retrieval) in Lisbon on

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

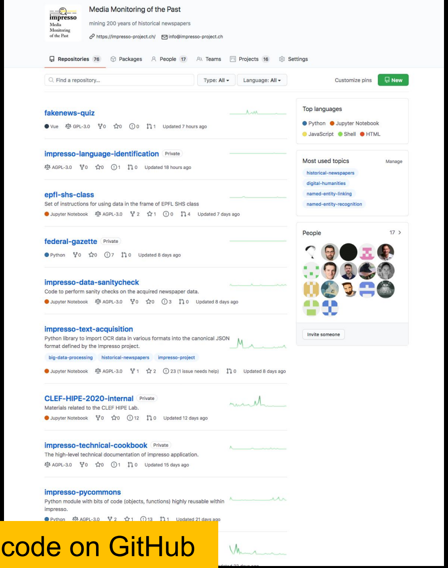
Created: June 1, 2019

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

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

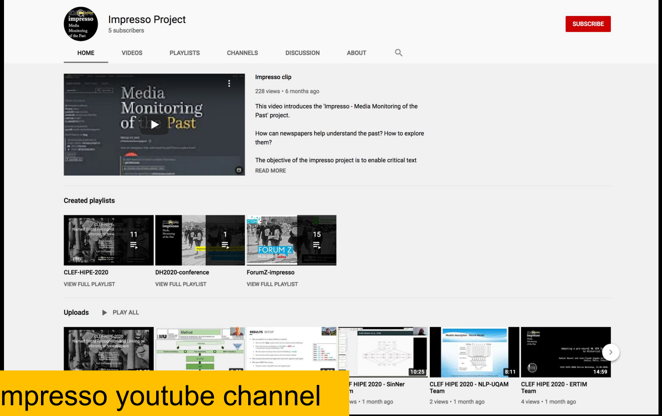
impresso-pycommons
Python module with bits of code (objects, functions) highly reusable within impresso.
Python AGPL-3.0 2 1 13 1 Updated 21 days ago

Top languages
Python Jupyter Notebook JavaScript Shell HTML

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

People 17

code on GitHub



Impresso Project
5 subscribers

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
VIEW FULL PLAYLIST

DH2020-conference
VIEW FULL PLAYLIST

ForumZ-impresso
VIEW FULL PLAYLIST

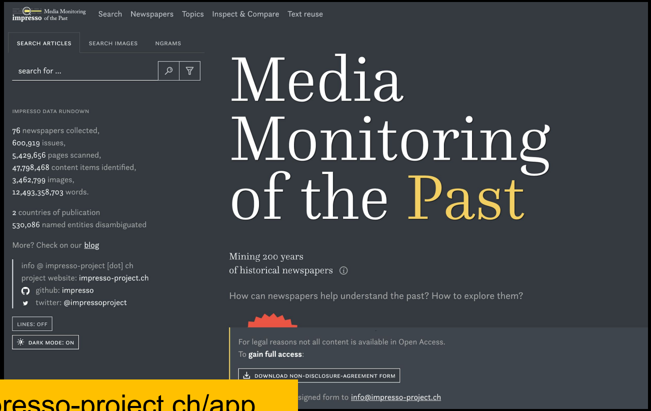
Uploads

…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 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

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**:

…signed form to info@impresso-project.ch

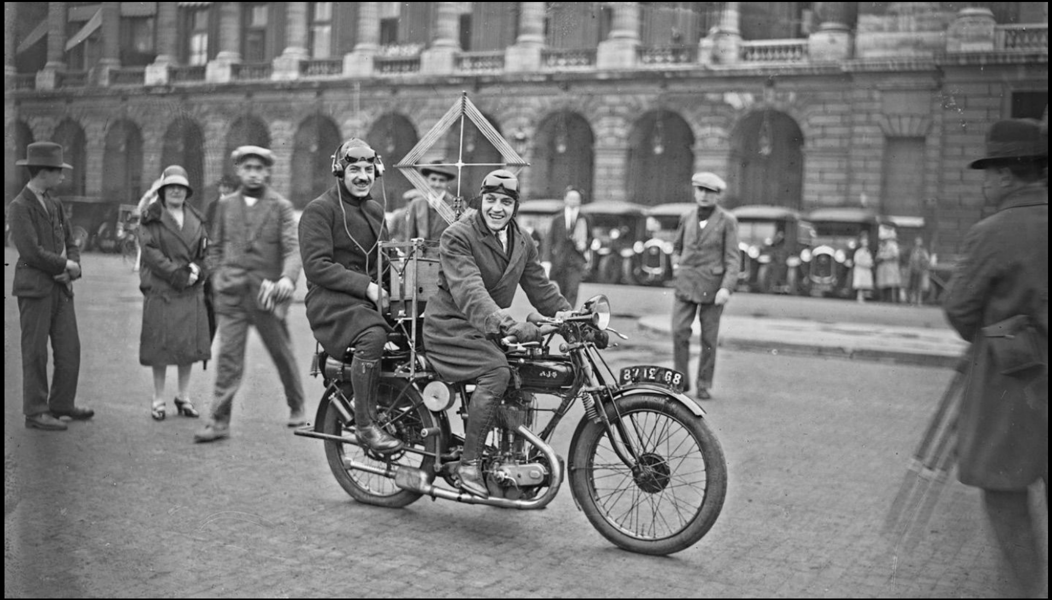
impresso-project.ch/app

Temporary account

URL: https://dev.impresso-project.ch/app

Login: student@impresso-project.ch

Passwort: Marburg2023!

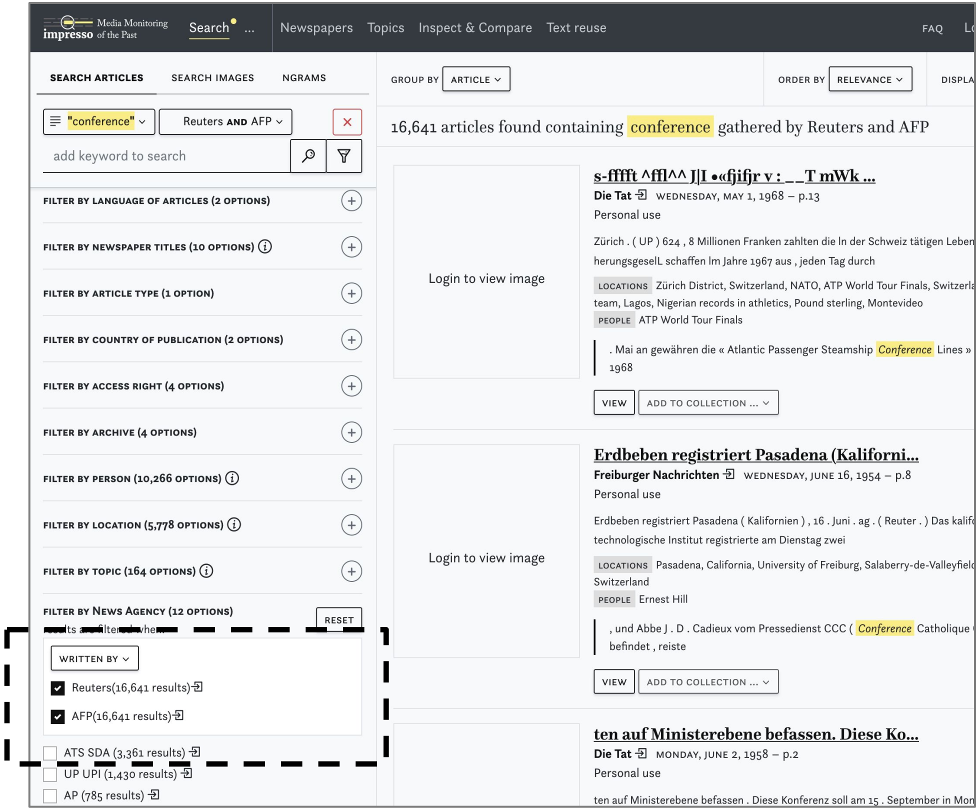


Ausblick: Was kommt als Nächstes?

Neue Perspektiven auf Agenturnachrichten

Erkennung von Presseagenturen

- Auf Basis eines Trainingskorpus (27 Agenturen, ca 2000 Artikel, fr und de);
- Training und Evaluierung von untersch. Sprachmodellen;
- Eweiterung auf den ganzen *impresso* Korpus;
- Erste Analysen;
- Integration in die App.



"conference" | 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

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

Zürich . (UP) 624 , 8 Millionen Franken zahlten die In 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

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

Erwähnungen von Presseagenturen

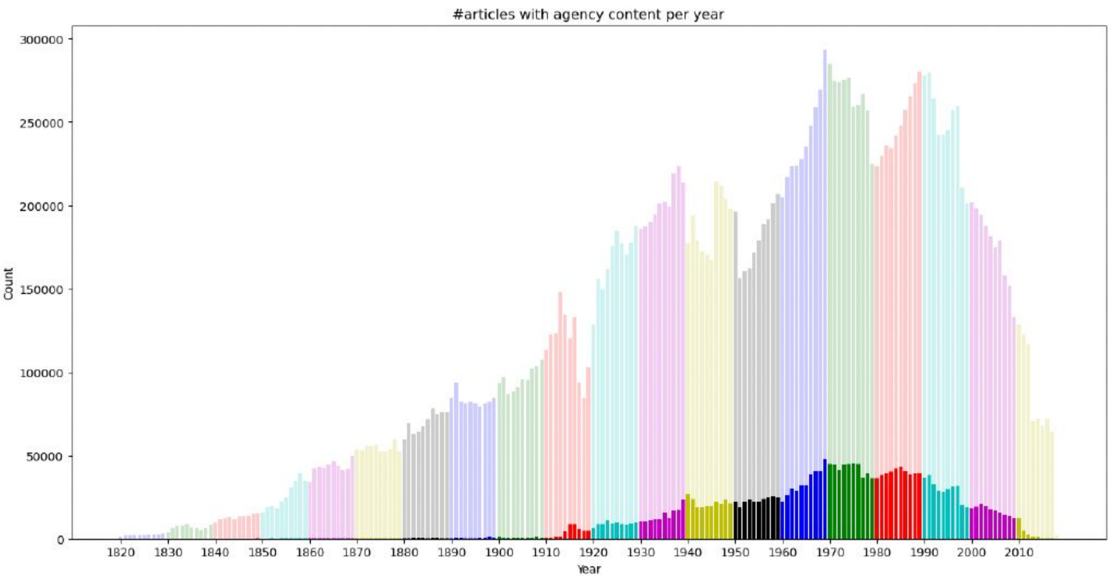


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

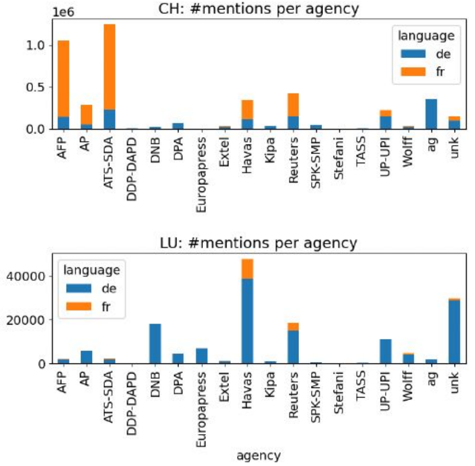


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

..in der schweizerischen Presse…

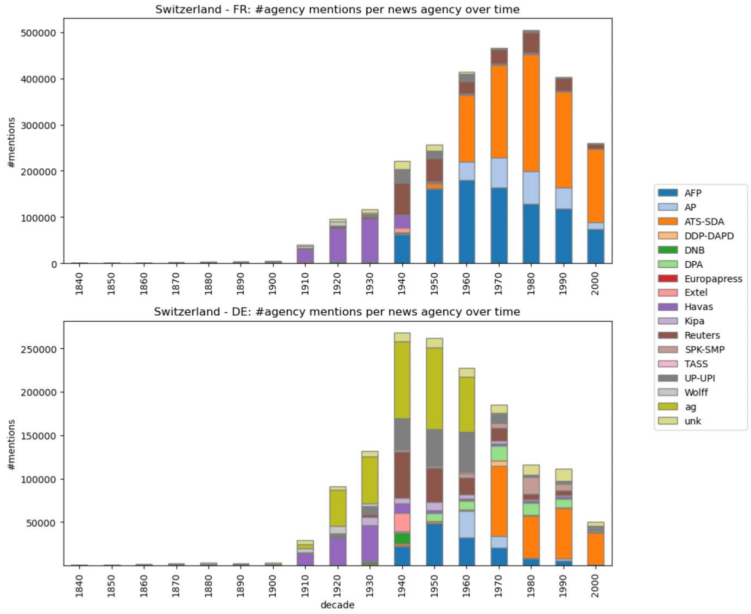


FIGURE 5.8

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

…in der luxemburgischen Presse (1940-1944)

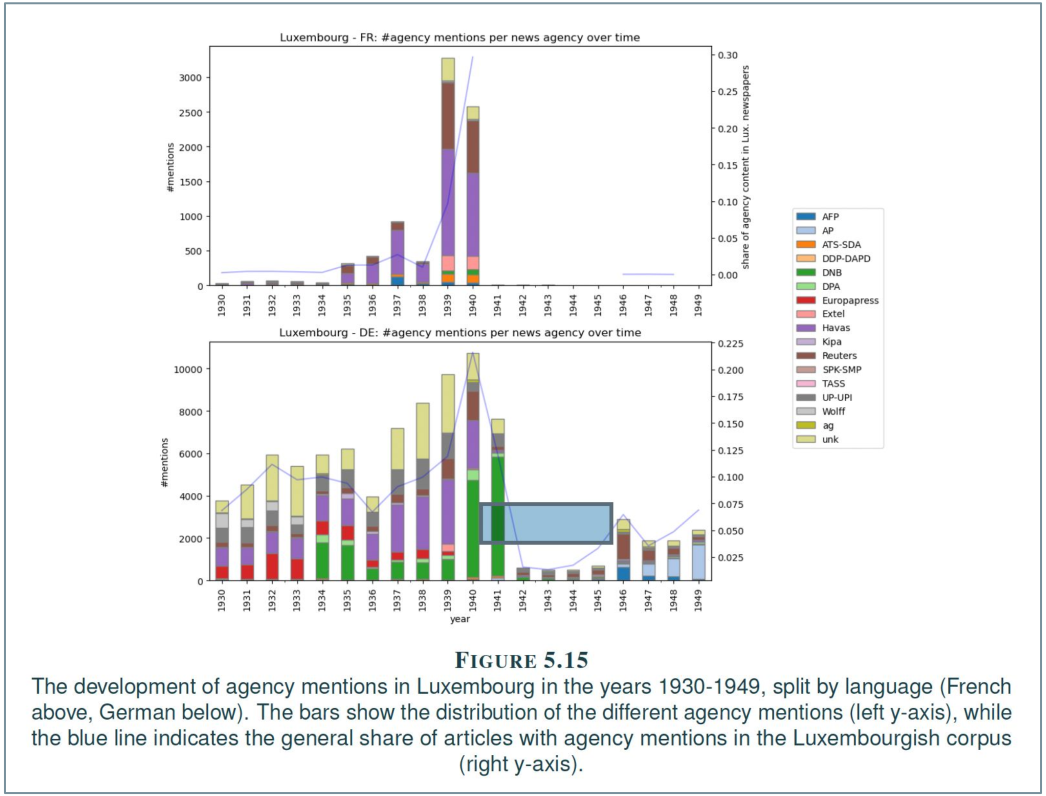
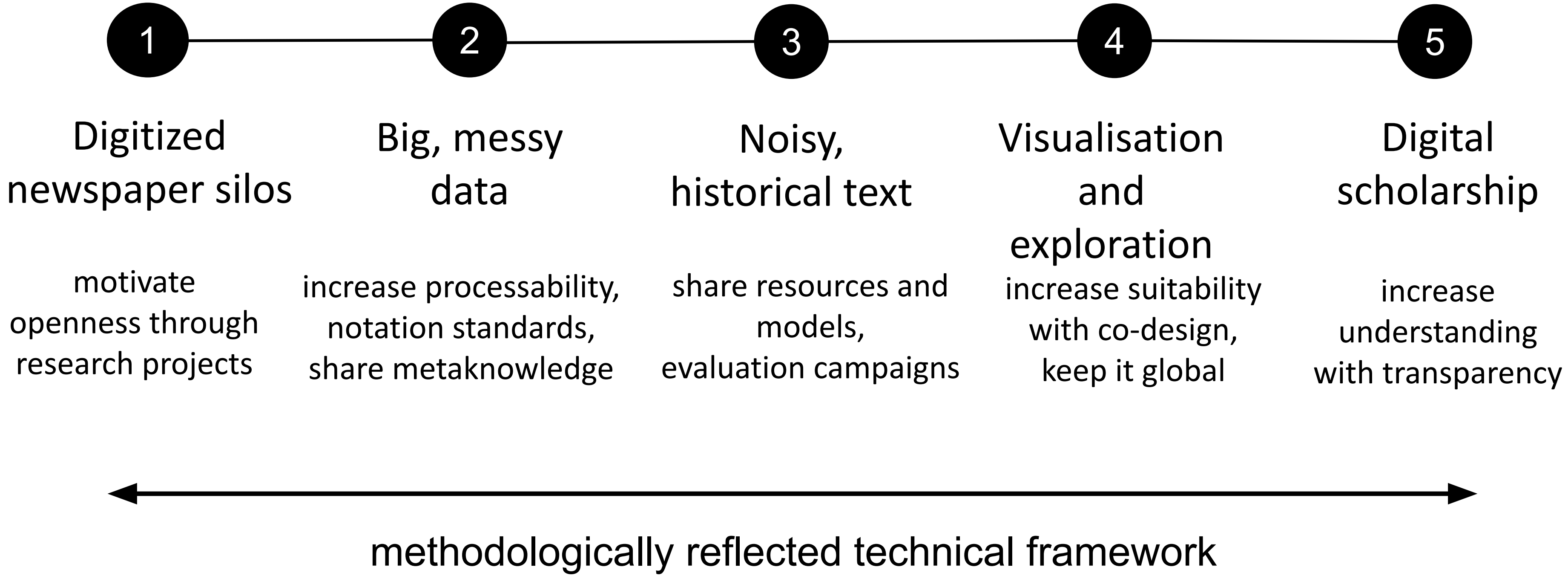


FIGURE 5.15

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

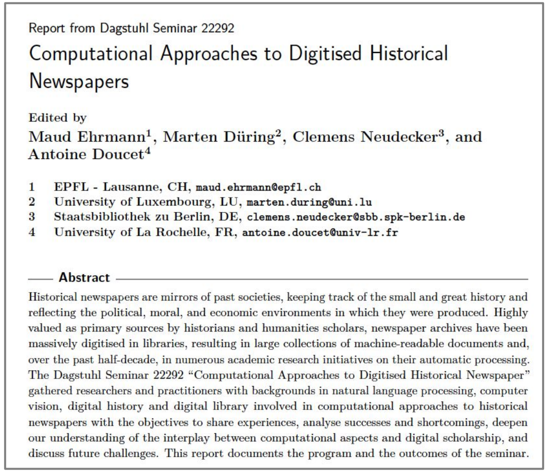
methodologically reflected technical framework

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

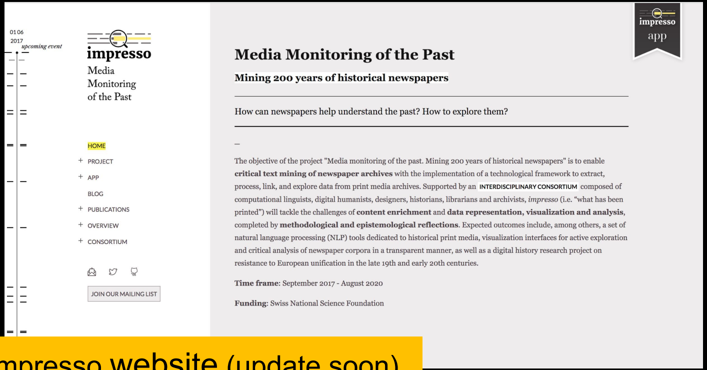
Edited by
Maud Ehrmann[1], Marten Düring[2], Clemens Neudecker[3], and Antoine Doucet[4]

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.

For more information



Media Monitoring of the Past

Mining 200 years of historical newspapers

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

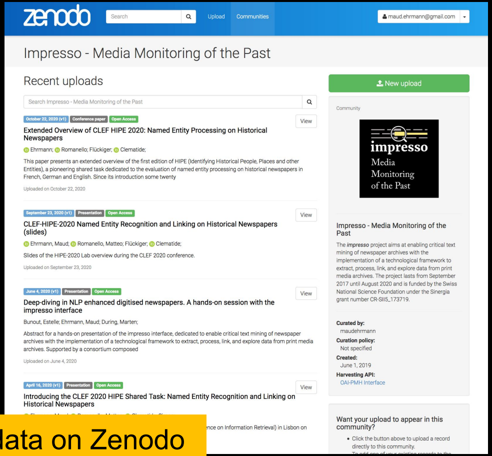
—

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)



Impresso - Media Monitoring of the Past

Recent uploads

Search Impresso - Media Monitoring of the Past

October 22, 2020 (v1) Conference paper Open Access

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

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

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

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

ence on Information Retrieval) in Lisbon on

Community

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

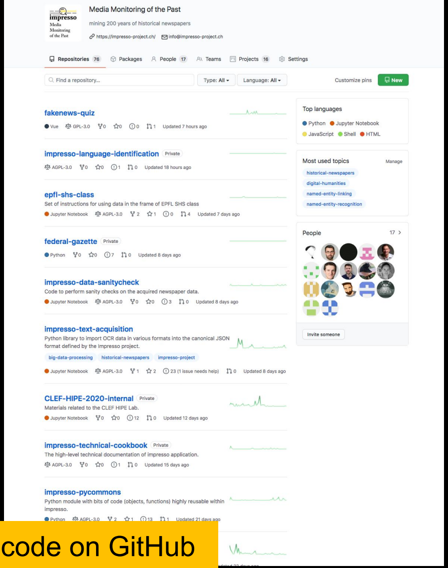
Created: June 1, 2019

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

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

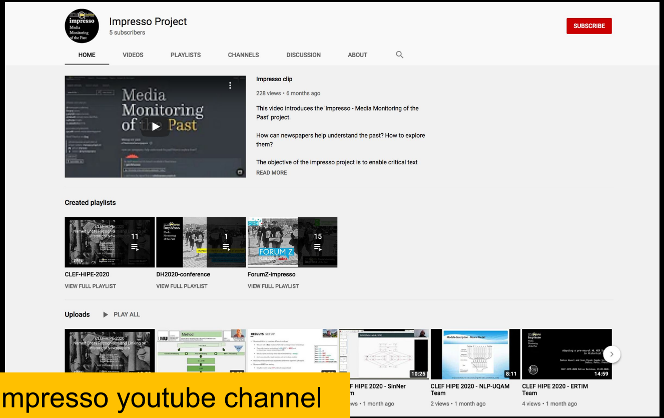
impresso-pycommons
Python module with bits of code (objects, functions) highly reusable within impresso.
Python AGPL-3.0 2 1 13 1 Updated 21 days ago

Top languages
Python Jupyter Notebook JavaScript Shell HTML

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

People 17

code on GitHub



Impresso Project
5 subscribers

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
VIEW FULL PLAYLIST

DH2020-conference
VIEW FULL PLAYLIST

ForumZ-impresso
VIEW FULL PLAYLIST

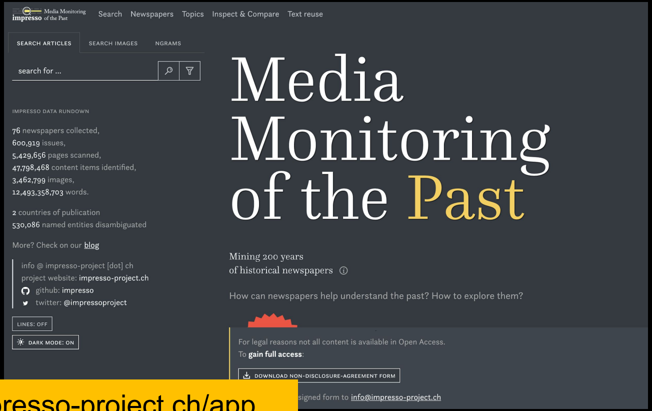
Uploads

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 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

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**:

signed form to info@impresso-project.ch

impresso-project.ch/app

Text reuse and historical research objectives

1. (Trans-) national media ecosystems
2. Newspaper content as *bricolage*
3. Historicising virality
4. Tracing historical events
5. Capturing historical Zeitgeist

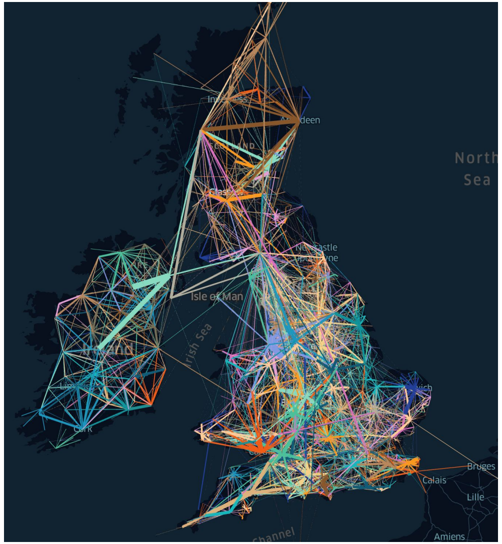
(Trans-) nationale Medien-Ökosysteme

Goal: Highlight the connectivity of historical media across borders and languages. Which ideological, commercial, and financial structures made this work and how did they shape media?

TR: (How) does content flow through the international network of newspapers?

Focus e.g. on cut-paste practices, the relevance of information infrastructure, geography, cross-border cultural affinities.

- Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational methods for uncovering reprinted texts in antebellum newspapers. *Am. Liter. Hist*. 27, E1–E15. doi: 10.1093/alh/ajv029
- Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news: a computational and multilingual approach to digitized newspaper collections. *Media History* 29, 287–304. doi: 10.1080/13688804.2022.2146905
- Salmi, H., Rantala, H., Vesanto, A., and Ginter, F. (2019). “The long-term reuse of text in the finnish press, 1771–1920,” in *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, eds. C. Navarretta, M. Agirrezabal, and B. Maegaard (Copenhagen, Denmark: CEUR Workshop Proceedings), 253–273.
- Paju, P., Salmi, H., Rantala, H., Lundell, P., and Marjanen, Vesanto, A. (2022). “Textual migration across the baltic sea: Creating a database of text reuse between Finland and Sweden,” in *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR Workshop Proceedings*, eds. K. Berglund, M. La Mela, and I. Zwart (Aachen: CEUR-WS.org), 361–369.
- Beelen, K. Digitising newspapers press directories to understand the landscape of historical newspapers, 2023, https://livingwithmachines.ac.uk/digitising-newspapers-press-directories-to-understand-the-landscape-of-historical-newspapers/.



Beelen, K. Digitising newspapers press directories to understand the landscape of historical newspapers

Newspaper content as bricolage

From which textual resources were newspapers compiled?

TR: How did formulaic ways of representing information emerge and spread (e.g. weather reports)? Can we determine the origins of content? Which content was influential?

Focuses on parallel developments across titles and the emergence of common practices.

- Walma, L. W. B. (2015). Filtering the “news:” Uncovering morphine's multiple meanings on delpher's dutch newspapers and the need to distinguish more article types. *Tijdschrift voor Tijdschriftstudies*. 38, 61–78. doi: 10.18352/ts.345
- Thèrenty, M.-E., and Venayre, S. (2021). *Le monde à la une. Une histoire de la presse par ses rubriques*. Anamosa, illustrated èdition edition. doi: 10.3917/anamo.there.2021.02

Historicising virality

Which conditions (geography, type of information, infrastructure) enable rapid dissemination?

Focus is on the breadth and speed with which content spreads. Pioneering work of Paju et al. who define a virality score based on the number of titles within a cluster, the number of unique printing locations, and the distance in days between the first and last passage publication date.

Paju, P., Salmi, H., Rantala, H., Lundell, P., and Marjanen, Vesanto, A. (2022). “Textual migration across the baltic sea: Creating a database of text reuse between Finland and Sweden,” in *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR Workshop Proceedings*, eds. K. Berglund, M. La Mela, and I. Zwart (Aachen: CEUR-WS.org), 361–369.

Textual Migration Across the Baltic Sea: Creating a Database of Text Reuse Between Finland and Sweden

Petri Paju[1], Hannu Salmi[1], Heli Rantala[1], Patrik Lundell[2], Jani Marjanen[3] and Aleksi Vesanto[1]

[1] *University of Turku, Department of Cultural History, Turku, FI-20014, Finland*
[2] *Örebro University, School of Humanities, Örebro, SE-70182, Sweden*
[3] *University of Helsinki, Department of Digital Humanities, Helsinki, FI-00014, Finland*

Abstract

In this paper, we present a database and an interface on text reuse between newspapers and journals published in the Swedish language in Sweden and Finland during the 1645–1918 time frame. Using two national, digital newspaper collections, we detected their textual similarities with a computational method to study the textual migration, i.e., information flows, between the two countries. For purposes of this project, we developed a database of detected clusters of text reuse and an online interface to search, examine and analyse the transnational movement of information. The database, *Text Reuse in the Swedish-language Press, 1645–1918,* is accessible online and includes texts from over 1,100 newspapers and journals published at approximately 150 locations at various times during the 274-year time frame.

Keywords

computational history, text reuse, historical newspaper, digital collections, database construction, Finland, Sweden, transnational history, information flow

1. Introduction

This short paper presents a database and an interface on text reuse among Swedish-language newspapers and journals during the 1645–1918 time frame. The database and interface were built as part of the project, *Information flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–1918*. The database and the accompanying project aim to study information flows, particularly between Sweden and Finland from the period when present-day Finland was part of the Swedish kingdom to the establishment of Finland as a Grand Duchy in the Russian Empire after 1809 and until the Independence of Finland in 1917 and Civil War in 1918. Even after the 1809 separation, news and other texts circulated because of the common cultural heritage and shared language, i.e., Swedish. The border was relatively easy to cross, and newspapers circulated between Sweden and Finland regularly. However, because their national histories eventually diverged, these press materials have been preserved, processed, and siloed in two national libraries. Still, print media digitisation makes it possible to study overlaps in large collections of texts and see how information was spread across the Baltic Sea.

In our project, textual migration was traced using a method based on the software BLAST, which can be applied to text-reuse detection. With the method, we detected every text passage with 300 or more characters of similarity and combined these passages into reuse clusters. We included Swedish-language papers published in Sweden and Finland, but excluded the Finnish-language press in Finland, as textual migration within Finland has been studied in previous publications [1, 2]. To strengthen the

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: petpaju@utu.fi (A. 1); hansalmi@utu.fi (A. 2); hemara@utu.fi (A. 3); Patrik.Lundell@oru.se (A. 4); Jani.Marjanen@helsinki.fi (A. 5); avjves@gmail.com (A. 6)
ORCID: 0000-0002-2486-2364 (A. 1); 0000-0001-8607-6126 (A. 2); 0000-0002-4108-4904 (A. 3); 0000-0002-6221-9089 (A. 4); 0000-0002-3085-4862 (A. 5)

CEUR Workshop Proceedings (CEUR-WS.org)

Tracing historical events

How do individual media titles situate events in the political, economic, social, and cultural context relevant to them? How does this affect the perception of these events by their audiences?

Focus is on the appropriation of often global news on the local scale.

- Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., et al. (2020). Spreading news in 1904. *Media History* 26, 391–407. doi: 10.1080/13688804.2019.1652090
- Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news: a computational and multilingual approach to digitized newspaper collections. *Media History* 29, 287–304. doi: 10.1080/13688804.2022.2146905

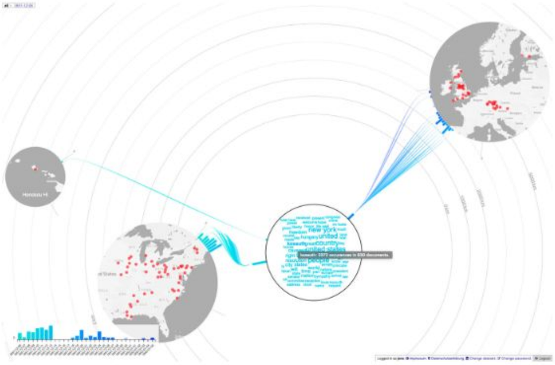
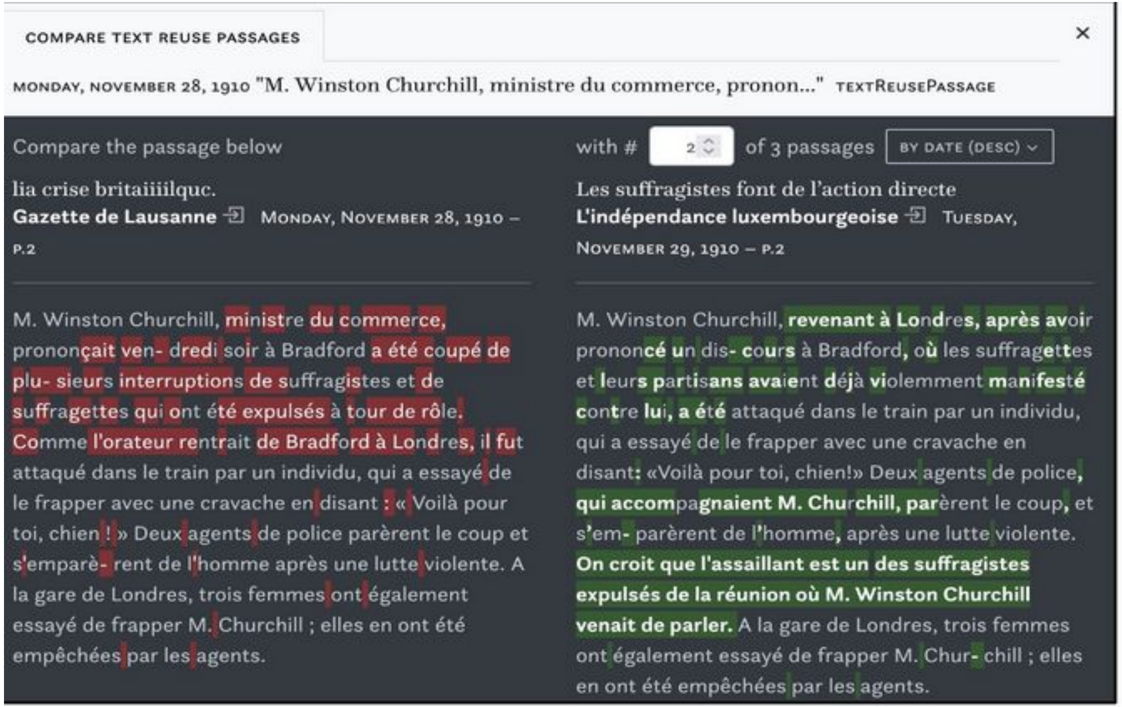


FIGURE 2.
The bar plot of the number of reprinted articles mentioning Kossuth in OcEx corpus shows that the American tour established Kossuth as an international celebrity.

Keck, J., Oiva, M., and Fyfe, P. (2022). Lajos kossuth and the transnational news

Tracing historical events



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, 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.

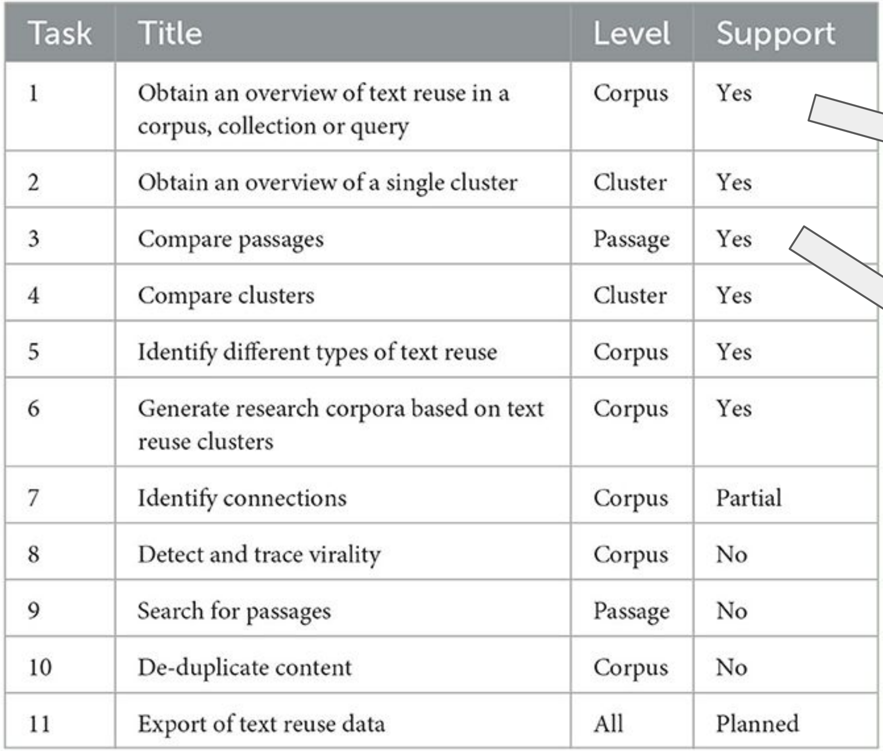
Capturing historical Zeitgeist

Historical media partially capture the attitudes, norms, beliefs, moods and feelings of past generations, or Zeitgeist.

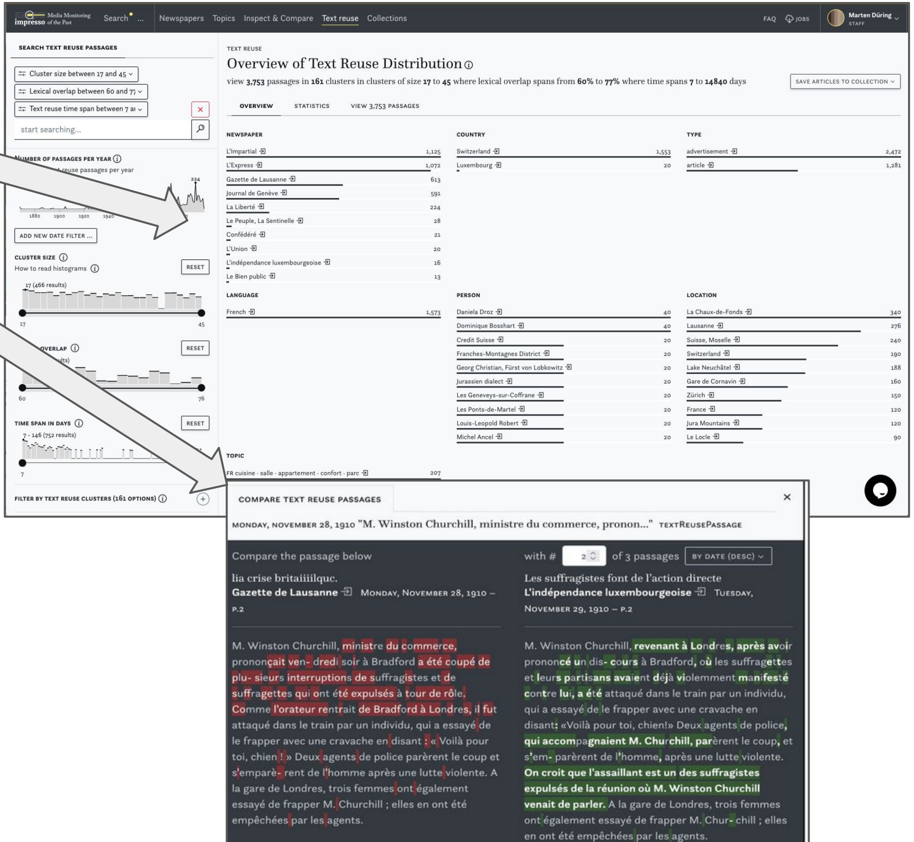
Focus is on texts which were produced independently but still share certain characteristics.

- Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., et al. (2022). Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914. *Dig. Human. Quart*. 16, 7445. Available online at: https://www.digitalhumanities.org/dhq/vol/16/2/000550/000550.html
- Paasikivi, S., Salmi, H., Vesanto, A., and Ginter, F. (2022). Infectious media: Cholera and the circulation of texts in the finnish press, 1860–1920. *Media Hist*. 29, 17–38. doi: 10.1080/13688804.2022.2054408

Tasks to support research objectives



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	Yes
3	Compare passages	Passage	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	Passage	No
10	De-duplicate content	Corpus	No
11	Export of text reuse data	All	Planned



Cluster size between 17 and 45

Lexical overlap between 60 and 77

Text reuse time span between 7 a

TEXT REUSE

Overview of Text Reuse Distribution

view 3,753 passages in 161 clusters in clusters of size 17 to 45 where lexical overlap spans from 60% to 77% where time spans 7 to 14840 days

NEWSPAPER	
L'Impartial	1,125
L'Express	1,072
Gazette de Lausanne	613
Journal de Genève	591
La Liberté	224
Le Peuple, La Sentinelle	28
Confédéré	21
L'Union	20
L'indépendance luxembourgeoise	16
Le Bien public	13

COUNTRY	
Switzerland	1,553
Luxembourg	20

TYPE	
advertisement	2,472
article	1,281

LANGUAGE	
French	1,573

PERSON	
Daniela Droz	40
Dominique Bosshart	40
Credit Suisse	20
Franches-Montagnes District	20
Georg Christian, Fürst von Lobkowitz	20
Jurassien dialect	20
Les Geneveys-sur-Coffrane	20
Les Ponts-de-Martel	20
Louis-Leopold Robert	20
Michel Ancel	20

LOCATION	
La Chaux-de-Fonds	340
Lausanne	276
Suisse, Moselle	240
Switzerland	190
Lake Neuchâtel	188
Gare de Cornavin	160
Zürich	150
France	120
Jura Mountains	120
Le Locle	90

TOPIC	
FR cuisine · salle · appartement · confort · parc	207

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

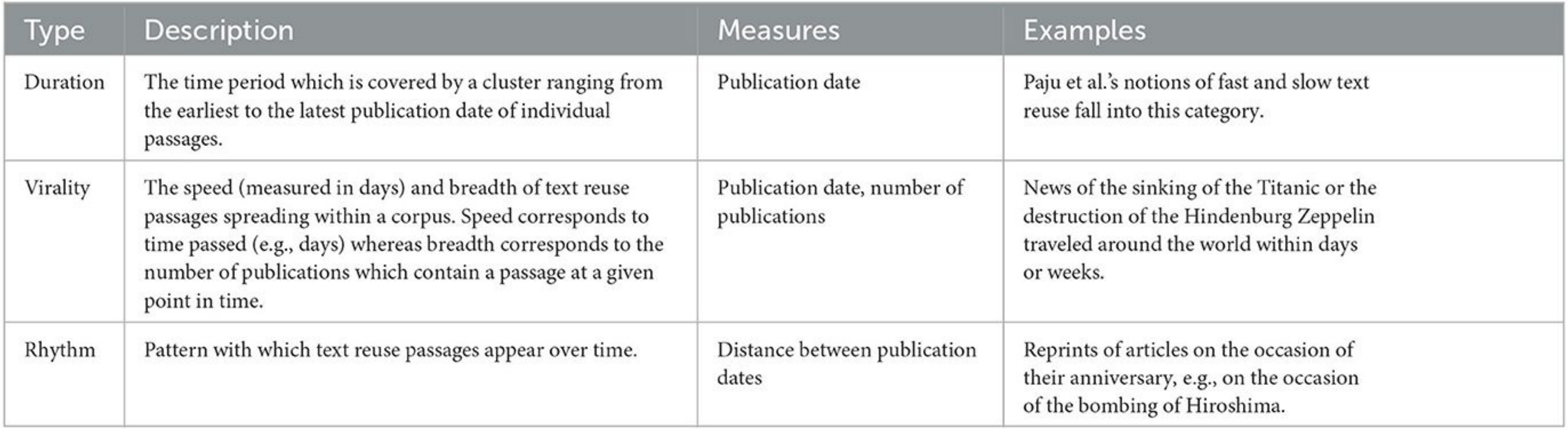
lia crise britaiiiilquc.
Gazette de Lausanne MONDAY, NOVEMBER 28, 1910 – P.2

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

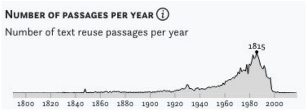
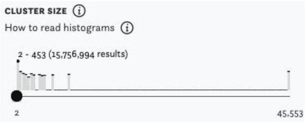
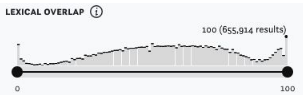
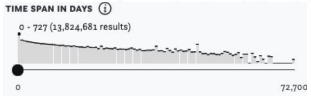
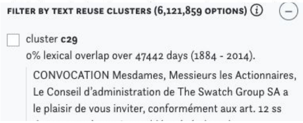
M. Winston Churchill, revenant à Londres, après avoir prononcé un dis- cours à Bradford, où les suffragettes et leurs partisans avaient déjà violemment manifesté contre lui, a été attaqué dans le train par un individu, qui a essayé de le frapper avec une cravache en disant: «Voilà pour toi, chien!» Deux 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.

Temporalities in text reuse data

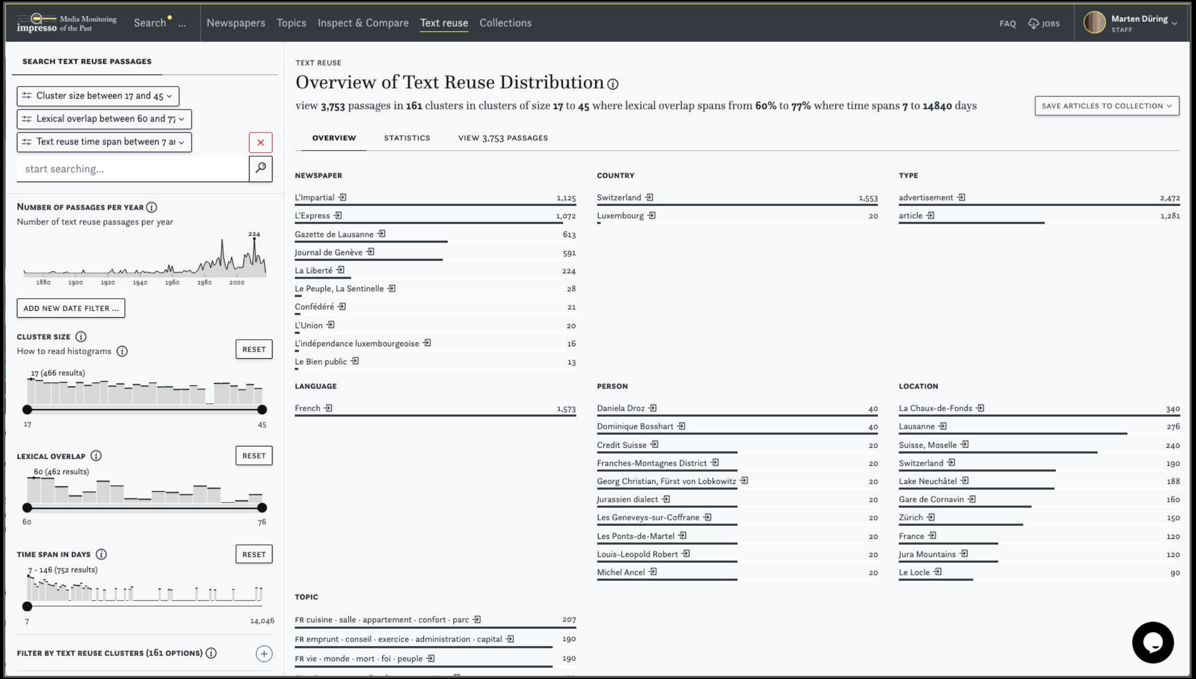


Type	Description	Measures	Examples
Duration	The time period which is covered by a cluster ranging from the earliest to the latest publication date of individual passages.	Publication date	Paju et al.'s notions of fast and slow text reuse fall into this category.
Virality	The speed (measured in days) 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.

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.  FILTER BY TEXT REUSE CLUSTERS (6,121,859 OPTIONS) ⓘ ⊖ □ 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

3. Demo *impresso* Text reuse at Scale



Cluster size between 17 and 45

Lexical overlap between 60 and 77

Text reuse time span between 7 a

Number of passages per year

Number of text reuse passages per year

Cluster size

How to read histograms

Lexical overlap

Time span in days

Filter by text reuse clusters (161 options)

TEXT REUSE

Overview of Text Reuse Distribution

view **3,753** passages in **161** clusters in clusters of size **17** to **45** where lexical overlap spans from **60%** to **77%** where time spans **7** to **14840** days

NEWSPAPER	
L'Impartial	1,125
L'Express	1,072
Gazette de Lausanne	613
Journal de Genève	591
La Liberté	224
Le Peuple, La Sentinelle	28
Confédéré	21
L'Union	20
L'indépendance luxembourgeoise	16
Le Bien public	13

COUNTRY	
Switzerland	1,553
Luxembourg	20

TYPE	
advertisement	2,472
article	1,281

LANGUAGE	
French	1,573

PERSON	
Daniela Droz	40
Dominique Bosshart	40
Credit Suisse	20
Franches-Montagnes District	20
Georg Christian, Fürst von Lobkowitz	20
Jurassien dialect	20
Les Geneveys-sur-Coffrane	20
Les Ponts-de-Martel	20
Louis-Leopold Robert	20
Michel Ancel	20

LOCATION	
La Chaux-de-Fonds	340
Lausanne	276
Suisse, Moselle	240
Switzerland	190
Lake Neuchâtel	188
Gare de Cornavin	160
Zürich	150
France	120
Jura Mountains	120
Le Locle	90

TOPIC	
FR cuisine · salle · appartement · confort · parc	207
FR emprunt · conseil · exercice · administration · capital	190
FR vie · monde · mort · foi · peuple	190

Types of text reuse in newspapers seen through filters

	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

Hands-on part and Challenge: Search and lexical overlap

1. **Go to** https://dev.impresso-project.ch/app/ and login
2. **Go to** https://bit.ly/impresso-tr
3. **Search** for “titanic”
4. **Exercise**: Familiarise yourself with all the different filters and views. Which questions emerge?
5. **Lexical overlap**: Go to Passages and experiment with the different sorting choices. What do you observe?

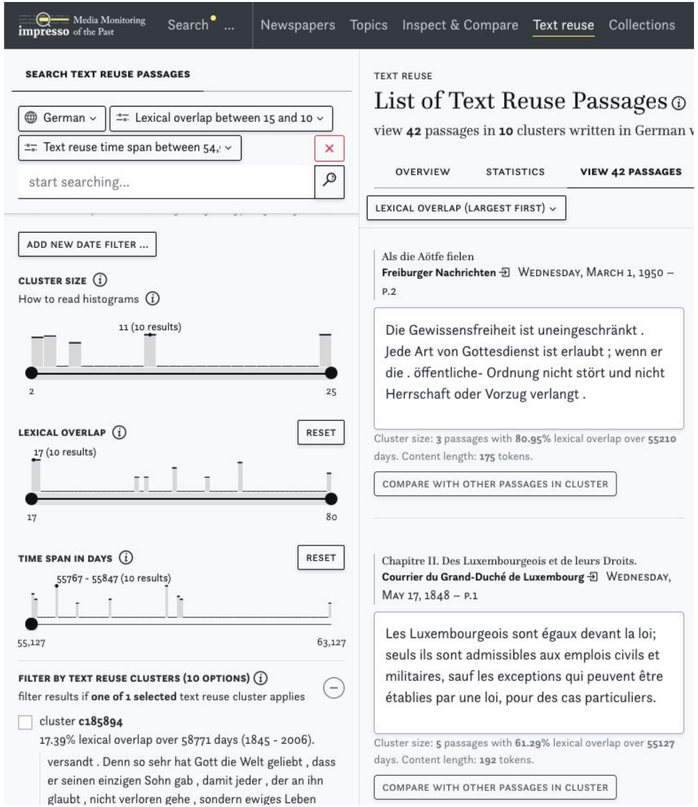
URL: https://dev.impresso-project.ch/app

Login: student@impresso-project.ch

Passwort: Marburg2023!

Hands-on part and Challenge: Time span

1. Start a new search by clicking the red “X”
2. Filter for “German”
3. Set Lexical overlap to 15% - 100%
4. Move to Passages view. What do you get?
5. Set time span to 55.00 - 67.000 days. What do you get?



German

Lexical overlap between 15 and 10

Text reuse time span between 54,

CLUSTER SIZE

How to read histograms

11 (10 results)

2 – 25

LEXICAL OVERLAP

17 (10 results)

17 – 80

TIME SPAN IN DAYS

55767 - 55847 (10 results)

55,127 – 63,127

FILTER BY TEXT REUSE CLUSTERS (10 OPTIONS)

filter results if **one of 1 selected** text reuse cluster applies

□ cluster **c185894**
17.39% lexical overlap over 58771 days (1845 - 2006).

versandt . Denn so sehr hat Gott die Welt geliebt , dass er seinen einzigen Sohn gab , damit jeder , der an ihn glaubt , nicht verloren gehe , sondern ewiges Leben

TEXT REUSE

List of Text Reuse Passages

view **42** passages in **10** clusters written in German

LEXICAL OVERLAP (LARGEST FIRST)

Als die Aötfe fielen
Freiburger Nachrichten WEDNESDAY, MARCH 1, 1950 – P.2

Die Gewissensfreiheit ist uneingeschränkt . Jede Art von Gottesdienst ist erlaubt ; wenn er die . öffentliche- Ordnung nicht stört und nicht Herrschaft oder Vorzug verlangt .

Cluster size: **3** passages with **80.95%** lexical overlap over **55210** days. Content length: **175** tokens.

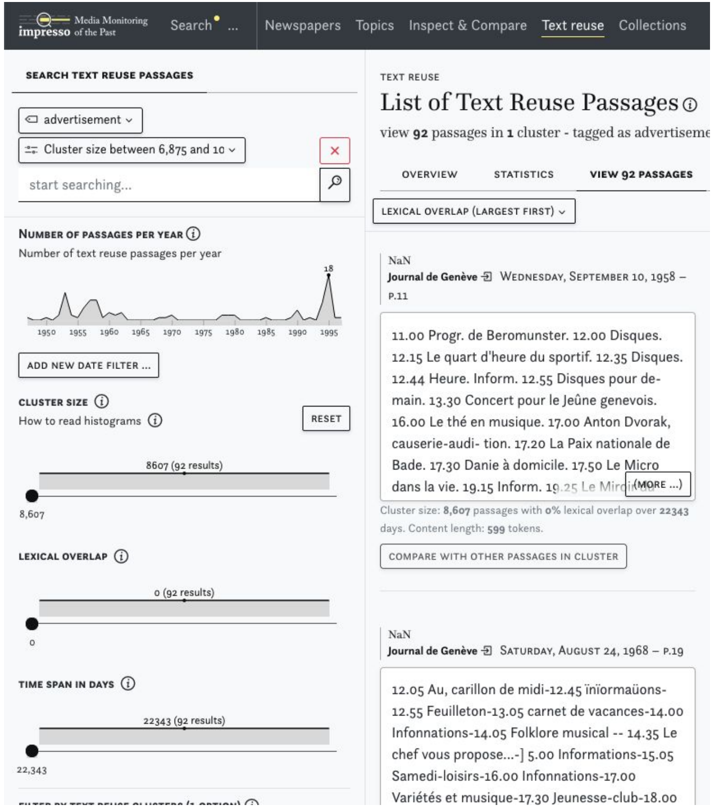
Chapitre II. Des Luxembourgeois et de leurs Droits.
Courrier du Grand-Duché de Luxembourg WEDNESDAY, MAY 17, 1848 – P.1

Les Luxembourgeois sont égaux devant la loi; seuls ils sont admissibles aux emplois civils et militaires, sauf les exceptions qui peuvent être établies par une loi, pour des cas particuliers.

Cluster size: **5** passages with **61.29%** lexical overlap over **55127** days. Content length: **192** tokens.

Hands-on part and Challenge: Cluster size

1. Start a new search by clicking the red “X”
2. Set filters for adverts
3. Set Cluster size between 90 - 100 passages.
4. What do you find? How can we explain this?



SEARCH TEXT REUSE PASSAGES

advertisement

Cluster size between 6,875 and 10

start searching...

NUMBER OF PASSAGES PER YEAR

Number of text reuse passages per year

CLUSTER SIZE

How to read histograms

8607 (92 results)

8,607

LEXICAL OVERLAP

0 (92 results)

0

TIME SPAN IN DAYS

22343 (92 results)

22,343

FILTER BY TEXT REUSE CLUSTERS (1 OPTION)

TEXT REUSE

List of Text Reuse Passages

view **92** passages in **1** cluster - tagged as advertiseme

LEXICAL OVERLAP (LARGEST FIRST)

NaN

Journal de Genève WEDNESDAY, SEPTEMBER 10, 1958 – P.11

11.00 Progr. de Beromunster. 12.00 Disques. 12.15 Le quart d'heure du sportif. 12.35 Disques. 12.44 Heure. Inform. 12.55 Disques pour demain. 13.30 Concert pour le Jeûne genevois. 16.00 Le thé en musique. 17.00 Anton Dvorak, causerie-audi- tion. 17.20 La Paix nationale de Bade. 17.30 Danie à domicile. 17.50 Le Micro dans la vie. 19.15 Inform. 19.25 Le Miroir du

Cluster size: **8,607** passages with **0%** lexical overlap over **22343** days. Content length: **599** tokens.

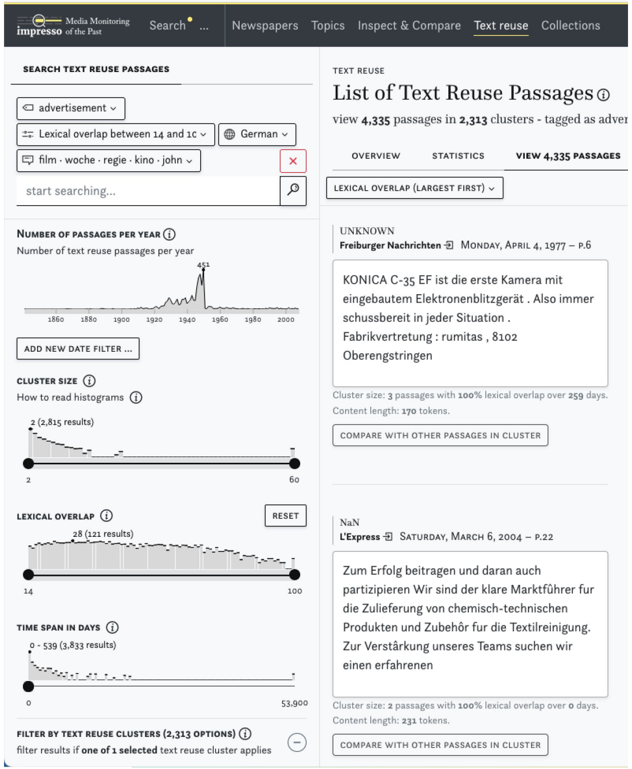
NaN

Journal de Genève SATURDAY, AUGUST 24, 1968 – P.19

12.05 Au, carillon de midi-12.45 ïnïormaüons-12.55 Feuilleton-13.05 carnet de vacances-14.00 Infonnations-14.05 Folklore musical -- 14.35 Le chef vous propose...-] 5.00 Informations-15.05 Samedi-loisirs-16.00 Infonnations-17.00 Variétés et musique-17.30 Jeunesse-club-18.00

Hands-on part and Challenge: Topics etc

1. Start a new search by clicking the red “X”
2. Set filters to “German” and adverts
3. Experiment with different topics: medical, classified ads, media…
4. Use the Passages tab and experiment with different sorting operations: e.g. largest cluster size vs. lexical overlap, passage size etc.



SEARCH TEXT REUSE PASSAGES

advertisement

Lexical overlap between 14 and 1c

German

film · woche · regie · kino · john

NUMBER OF PASSAGES PER YEAR

Number of text reuse passages per year

CLUSTER SIZE

How to read histograms

2 (2,815 results)

LEXICAL OVERLAP

28 (121 results)

TIME SPAN IN DAYS

0 - 539 (3,833 results)

FILTER BY TEXT REUSE CLUSTERS (2,313 OPTIONS)

filter results if **one of 1 selected** text reuse cluster applies

TEXT REUSE

List of Text Reuse Passages

view **4,335** passages in **2,313** clusters - tagged as adver

OVERVIEW STATISTICS **VIEW 4,335 PASSAGES**

LEXICAL OVERLAP (LARGEST FIRST)

UNKNOWN
Freiburger Nachrichten MONDAY, APRIL 4, 1977 – P.6

KONICA C-35 EF ist die erste Kamera mit eingebautem Elektronenblitzgerät . Also immer schussbereit in jeder Situation . Fabrikvertretung : rumitas , 8102 Oberengstringen

Cluster size: **3** passages with **100%** lexical overlap over **259** days.
Content length: **170** tokens.

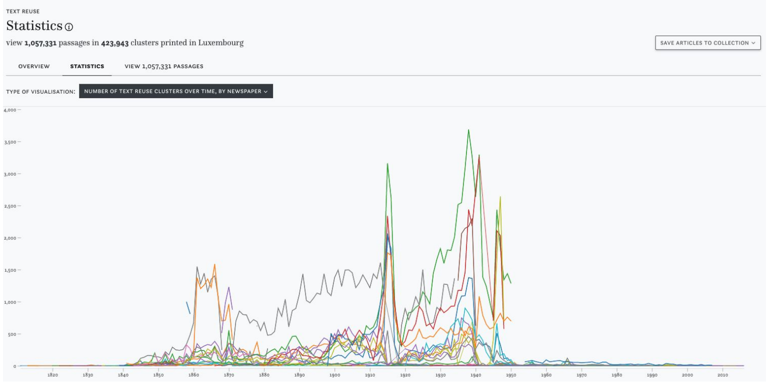
NaN
L'Express SATURDAY, MARCH 6, 2004 – P.22

Zum Erfolg beitragen und daran auch partizipieren Wir sind der klare Marktfûhrer fur die Zulieferung von chemisch-technischen Produkten und Zubehôr fur die Textilreinigung. Zur Verstârkung unseres Teams suchen wir einen erfahrenen

Cluster size: **2** passages with **100%** lexical overlap over **0** days.
Content length: **231** tokens.

Hands-on part and Challenge: Topics etc

1. Start a new search by clicking the red “X”
2. Set the **country** filters to “Luxembourg”
3. Select the Statistics view + number of text reuse clusters over time
4. How can we explain this graph?



TEXT REUSE

Statistics

view **1,057,331** passages in **423,943** clusters printed in Luxembourg

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

Your turn…find instances of:

1. Build you own queries, combine different keywords and filters. Which impact do they have?
2. Explore the Statistics view. What can it tell you about the data?
3. Open https://bit.ly/impresso-tr and go to slides 44ff. Pick one of the project ideas or duplicate slides to add your own.

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

Your names here

Your idea here

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

Standardized content (weather, radio programme…)

Observations:

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

URL to query	
Number of passages	
Lexical overlap	
Time span	

The longest running advertisement

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Your names here

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

Press agency reports

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Observations:

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

The two newspapers that co-publish the m

Observations:

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

URL to query	
Number of passages	
Lexical overlap	
Time span	

Your names here

A new form of text reuse!?

<SCREENSHOT(S)>

Insert here a brief description of what type of text reuse it is and what we can learn from it.

Observations:

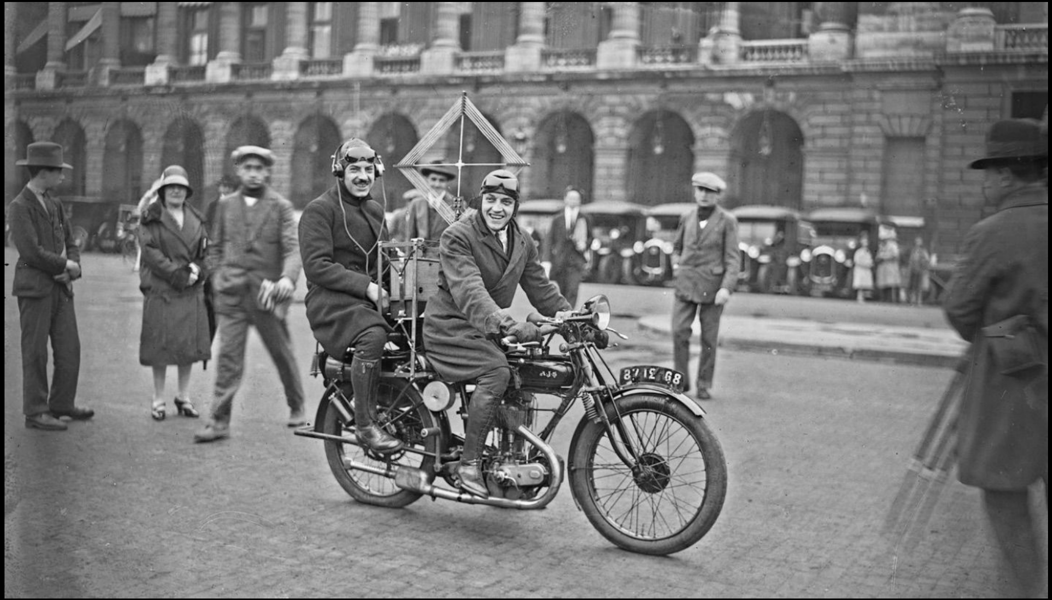
URL to query	
Number of passages	
Lexical overlap	
Time span	

Temporary account

URL: https://dev.impresso-project.ch/app

Login: student@impresso-project.ch

Passwort: Marburg2023!

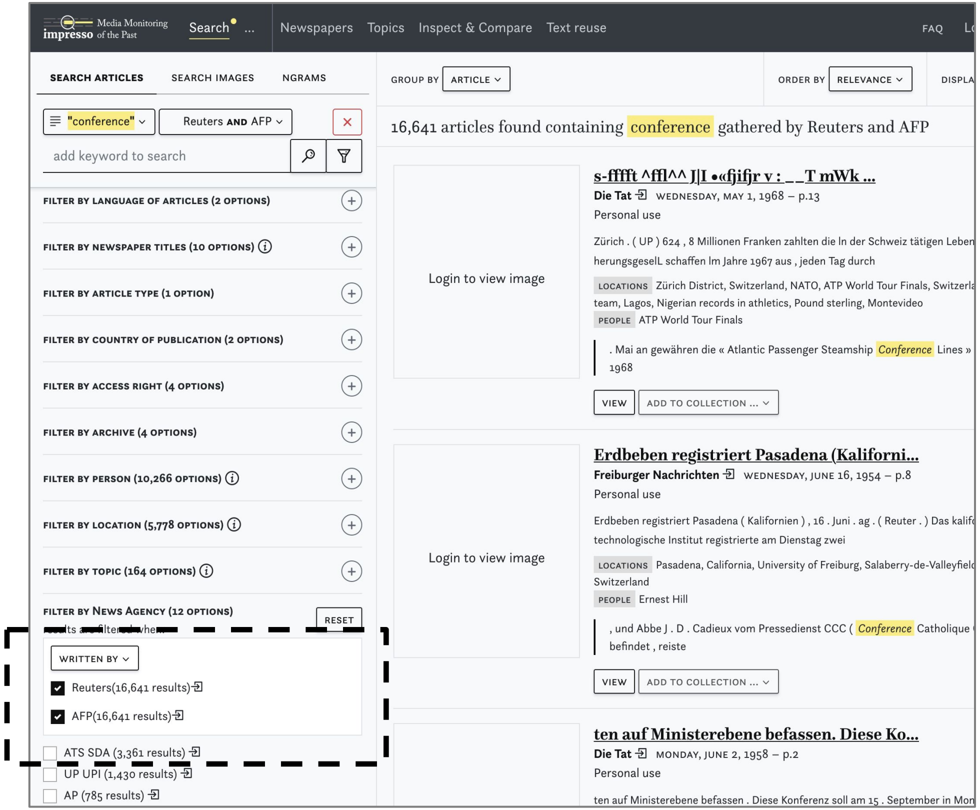


Ausblick: Was kommt als Nächstes?

Neue Perspektiven auf Agenturnachrichten

Erkennung von Presseagenturen

- Auf Basis eines Trainingskorpus (27 Agenturen, ca 2000 Artikel, fr und de);
- Training und Evaluierung von untersch. Sprachmodellen;
- Eweiterung auf den ganzen *impresso* Korpus;
- Erste Analysen;
- Integration in die App.



"conference" Reuters AND AFP

add keyword to search

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) RESET

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)

Login to view image

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 In 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

VIEW ADD TO COLLECTION ...

Login to view image

Erdbeben registriert Pasadena (Kaliforni...

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

Personal use

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

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

PEOPLE Ernest Hill

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

VIEW ADD TO COLLECTION ...

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

Erwähnungen von Presseagenturen

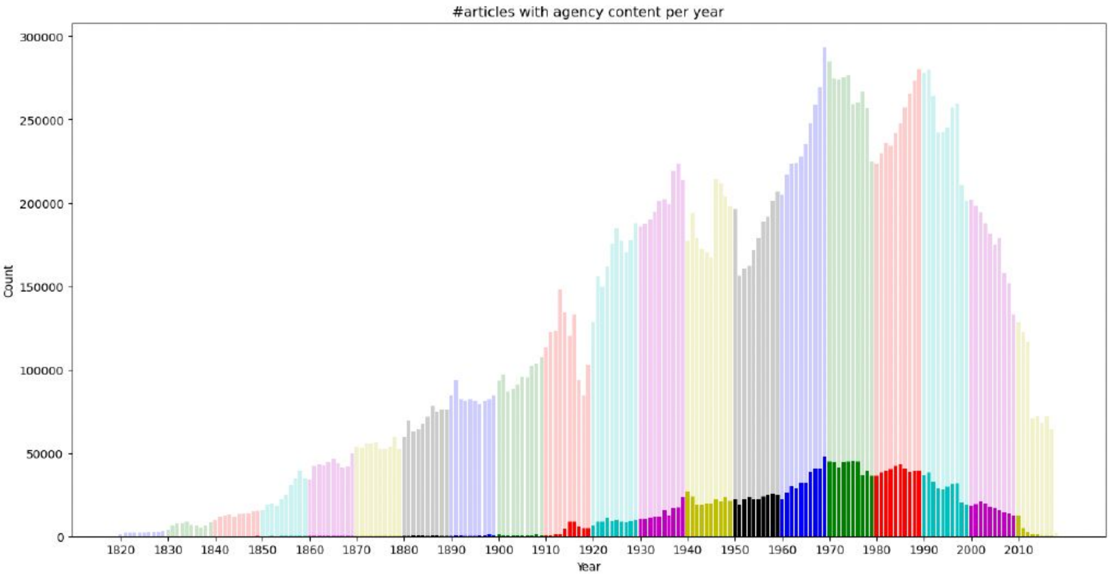


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

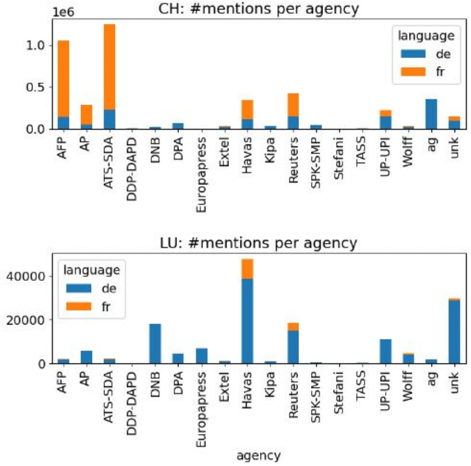


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

..in der schweizerischen Presse…

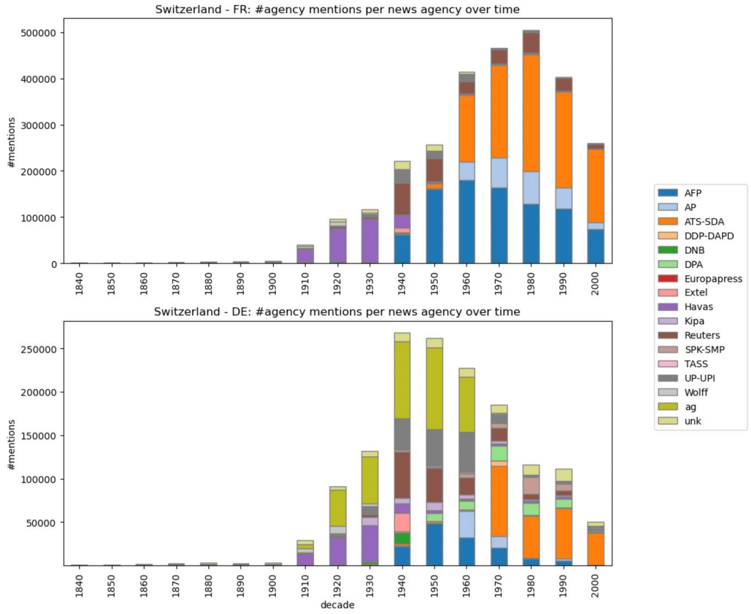


FIGURE 5.8
News agency mentions in Swiss newspapers over time, split by language (French above, German below).

…in der luxemburgischen Presse (1940-1944)

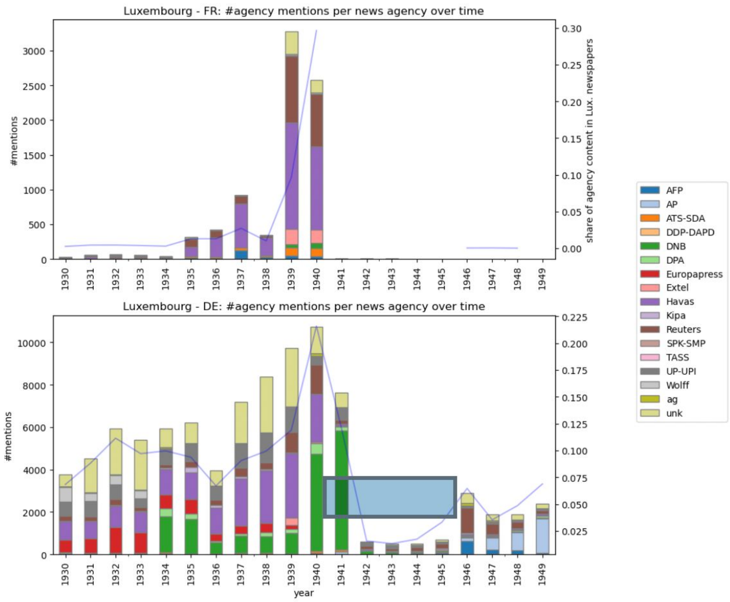
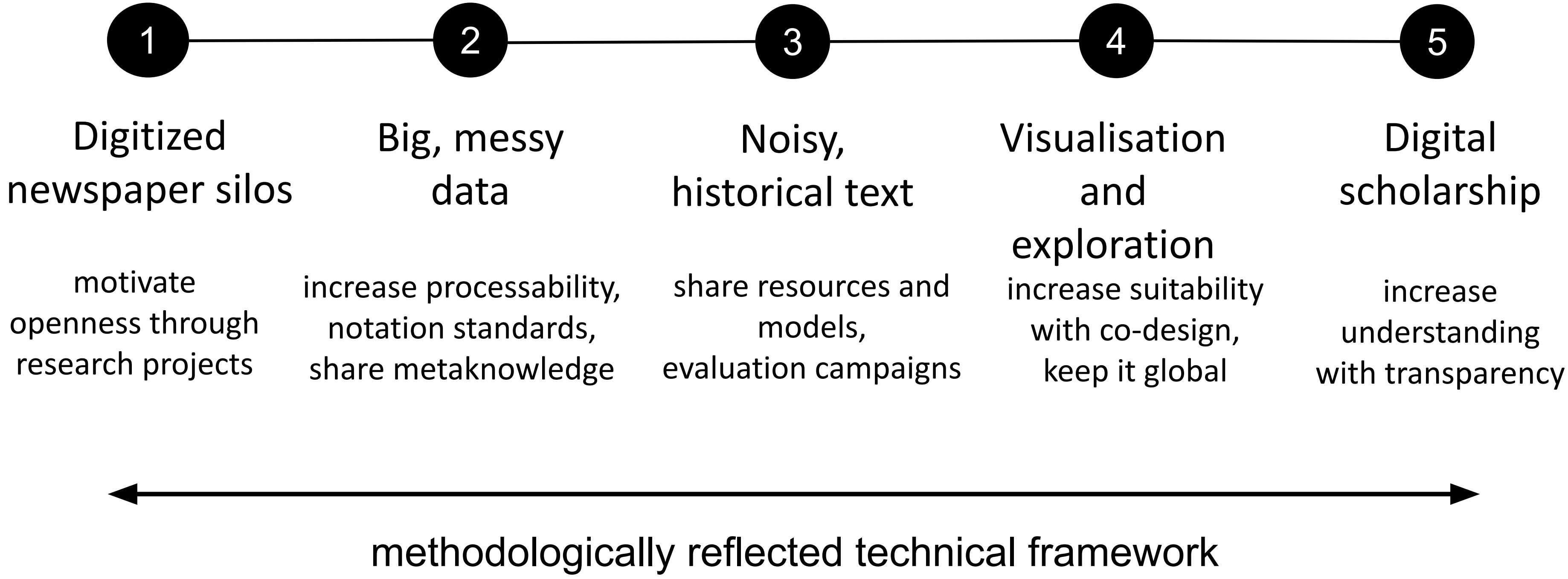


FIGURE 5.15

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

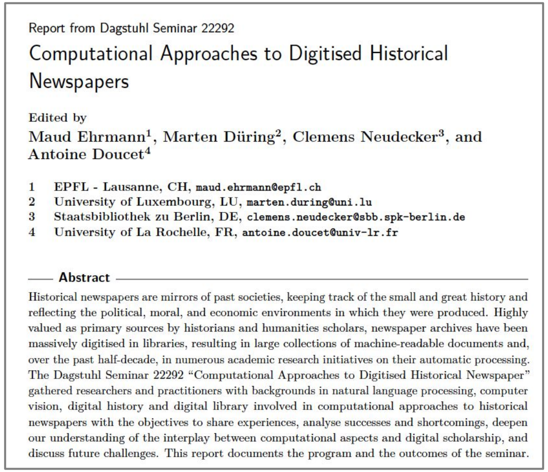
SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

What’s next ? (from the community)

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

© SCHLOSS DAGSTUHL – LZI GMBH
licensed under Creative Commons License CC BY-NC-ND



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.