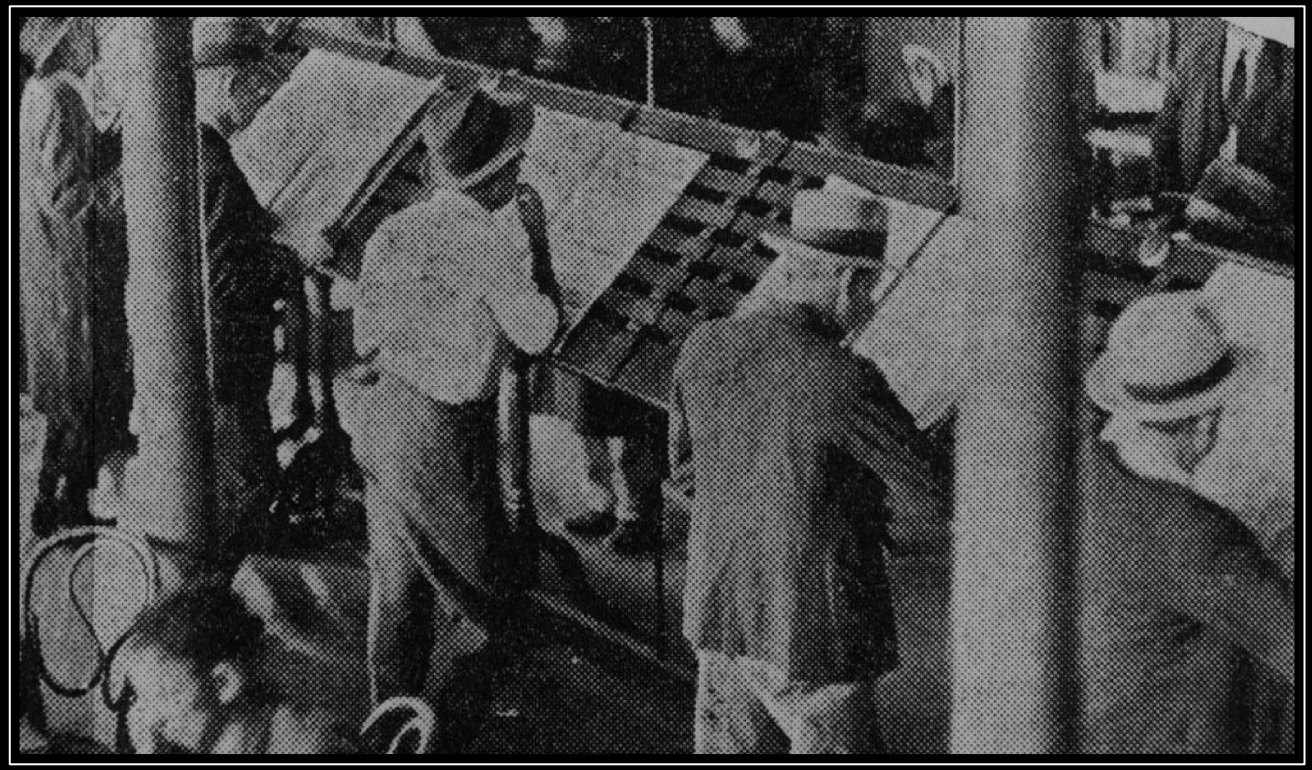
Historische Medien und Maschinelles Lernen

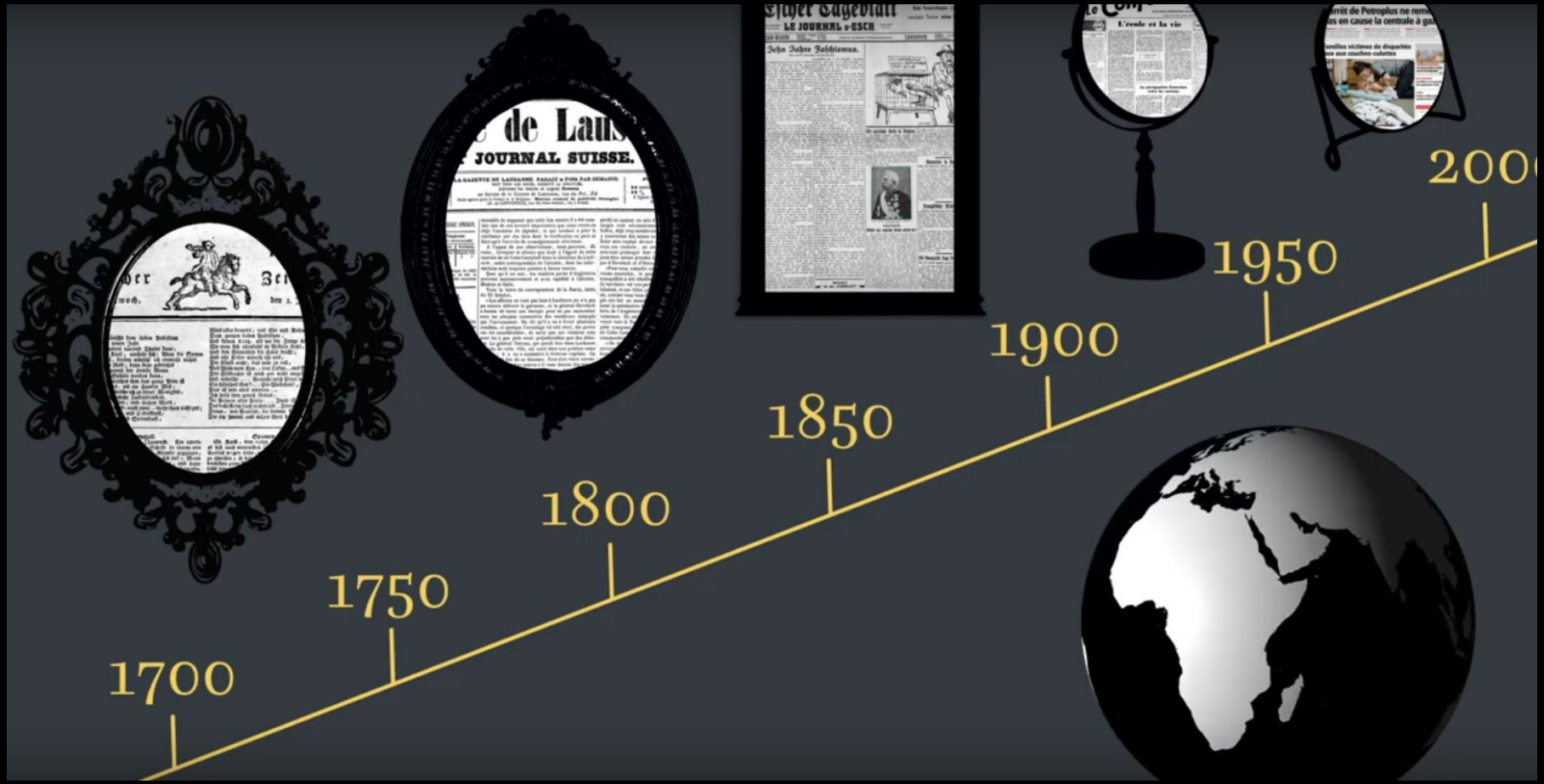
Vom Zusammenspiel von Infrastrukturen, Methoden und Akteuren

Marten Düring & impresso team

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



1. Ein paar Worte über historische Zeitungen


JOURNAL SUISSE.
LE JOURNAL D'ESCH
Zehn Jahre Falschmünz
L'école et la vie
1700
1750
1800
1850
1900
1950
200

Luxemburger Wort

für Wahrheit und Recht

Mittwoch, 24. April 1912

Nr. 115

65. Jahrgang.

Erstes Blatt.

Serbische Wahlen.

Schloß Altenstein.

Türkisch-italienischer Krieg

Erzbischof Dr. v. Albert †

Zum Untergang der Titanic.

Haute tension sur l'avenir électrique

ENERGIE DE L'OUEST SUISSE (EOS)

L'ère des incertitudes

Christophe Babaiantz: le prix du renoncement à Kaiseraugst

Avis mortuaire.

Madame Alph. Diederich

Waggon Kinderwagen!

Chaussures Modernes

Packels Steffen, 28 Pastorstr. 28.

MARA STIEFEL

Persil

Für Krankenwäsche

Das selbsttätige Waschmittel.

Volks-Kleider-Halle

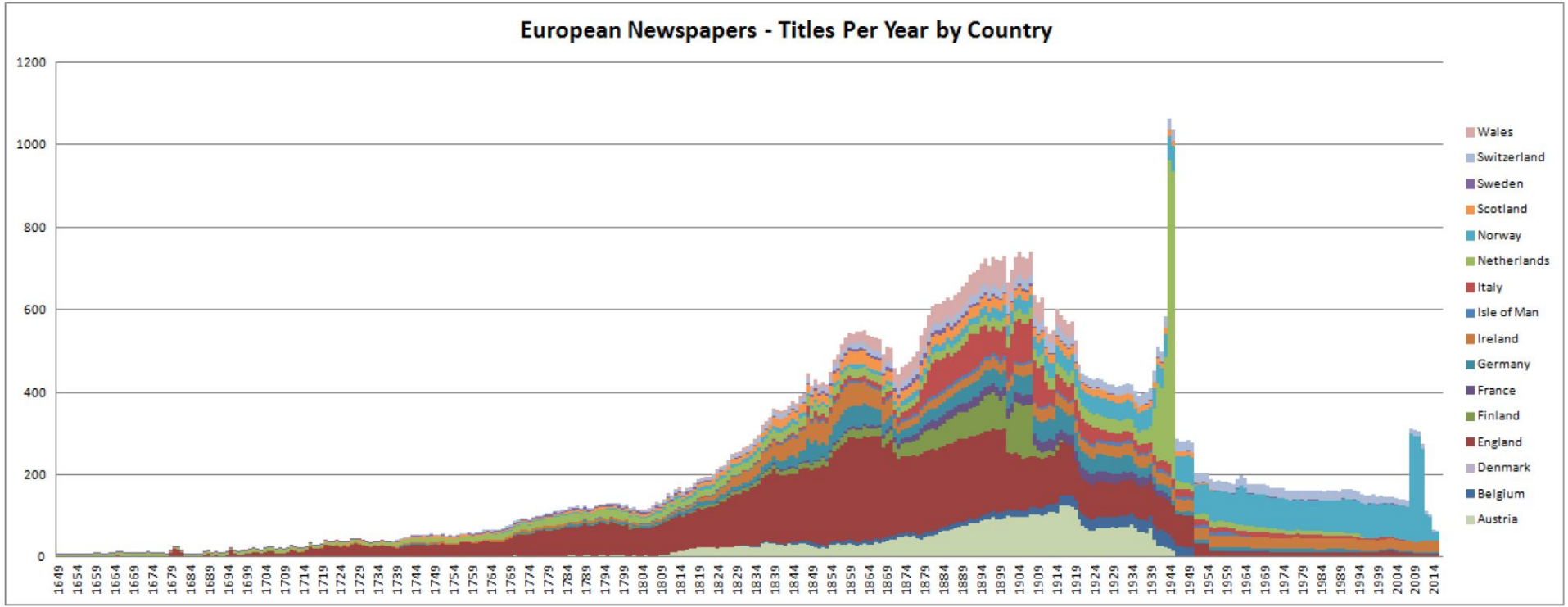
Herren-Konfektion

Abteilung Möbel

Lessive Body

Spar-Einlagen

Digitalisierte historische Zeitungs-Sammlungen



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

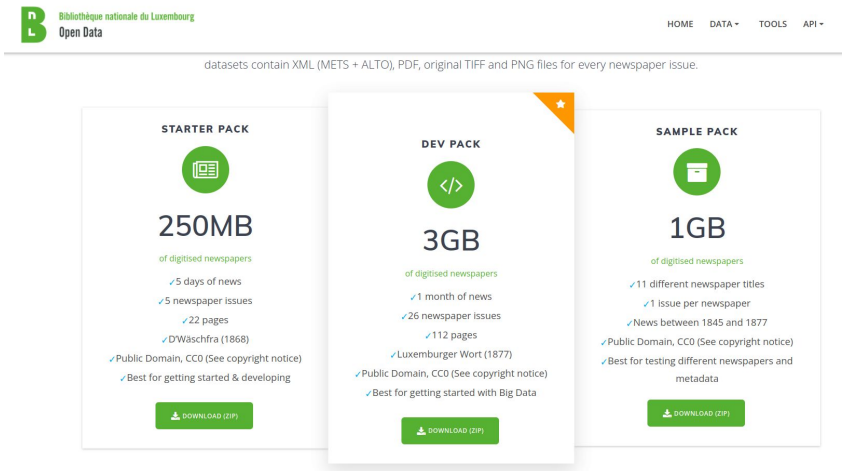
Historische Zeitungen als Herausforderung

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

Zeitungen als Daten

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

https://collectionsasdata.github.io/



https://data.bnl.lu

TROVE

2. Show the total number of articles per year

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

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

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

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

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

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

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

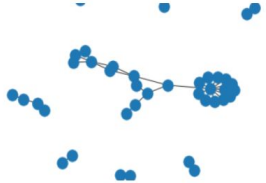
Ok, that's not all that useful. What would be more interesting is to show the total number of a
details in this notebook but, in short, we have to loop through the decades from 1800 to 201

These two functions do just that.

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

OPEN A GLAM LAB

https://glamlabs.io/



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

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

Binder Preview

COMPUTER VISION APPLIED TO SMITHSONIAN OPEN ACCESS

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

Binder Preview

ACCESSING EUROPEANA IIIF API

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

Binder Preview

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

Interdisziplinäre Forschungsprojekte

NEWS EYE

LIVING WITH MACHINES

impresso
Media
Monitoring
of the Past

NUMA PRESSE

Beispiele für datengetriebene historische Forschung

Verbundenheiten

> “Reconstruct the cultural connectivity between 19thC Sweden and Finland using text reuse detection.”

Zeitgeist

> “Capture anti-modernist thought in the Swiss press using naive Bayes classifiers.”

Medienformate

> “Reconstruct the evolution of newspaper genres using topic modeling.”

Rekompilationen

> “Understand the interplay between historical cut and paste journalism and information infrastructure.”

Forschungsinteressen: Geschichte der Medien und Medien als Quellen

Fachdisziplinen

Mediengeschichte

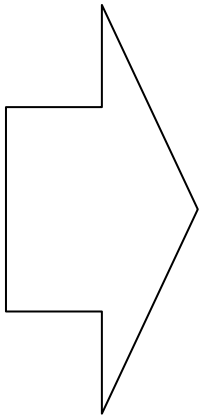
Gender studies

Kulturgeschichte

Sozialwissenschaften

Sozialgeschichte

…



Ziele

Layoutanalyse

Evolution von Genres

“Viralität”

Soziale Normen

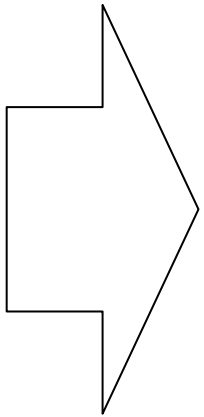
Meinungsbildung

Wissenshorizonte

Biographien

Ernährung

….



Zeitungselemente

Werbung

Kolumnen

Agenturnachrichten

Bilder

Kleinanzeigen

Radioprogramme

Todesanzeigen

….

EPFL

University of Zurich[UZH]

UNIVERSITÉ DU LUXEMBOURG

C²DH LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Département fédéral de l'intérieur DFI
Office fédéral de la culture OFC
Bibliothèque nationale suisse BN

Bibliothèque nationale de Luxembourg

infoclio.ch

Unil

University of Basel

archives de l'état du valais staatsarchiv wallis

NZZ

LE TEMPS

Impresso

Critical content mining of 200 years of historical newspapers

Inwieweit dienen semantische Anreicherungen der Analyse und Exploration historischer Zeitungen?

Das Team

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

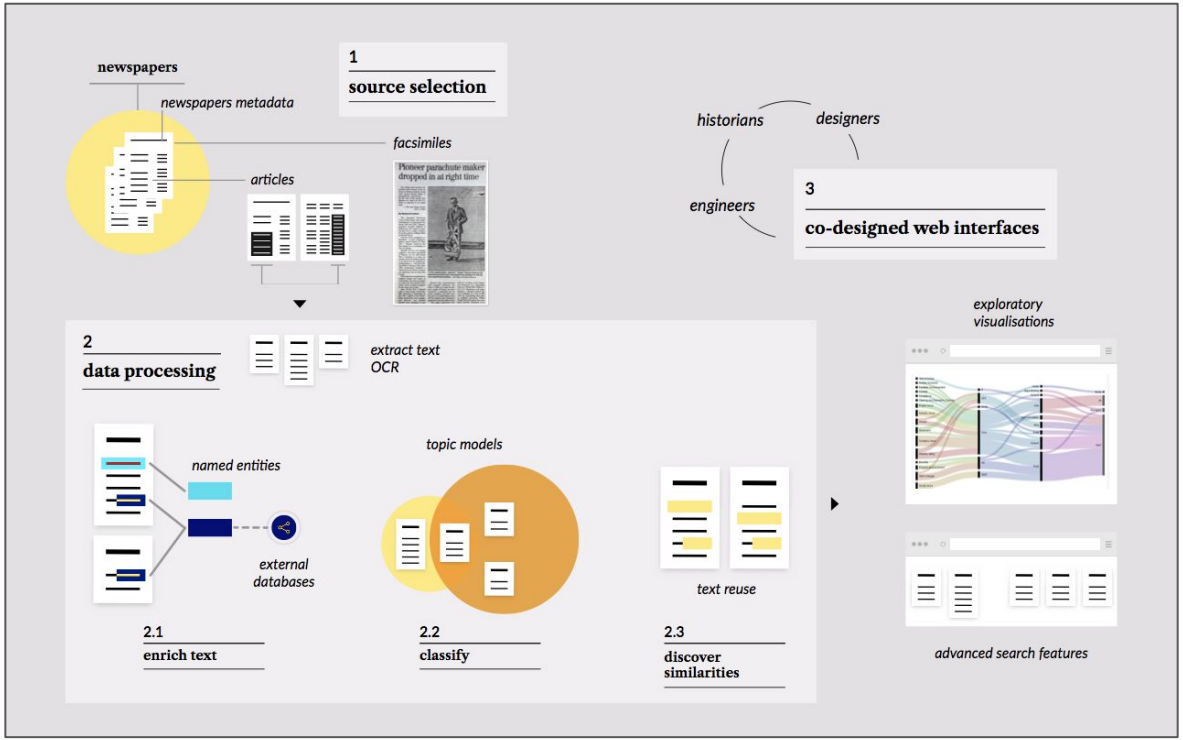
+ beratende HistorikerInnen
+ assoziierte WissenschaftlerInnen

Ziele und Forschungsfragen

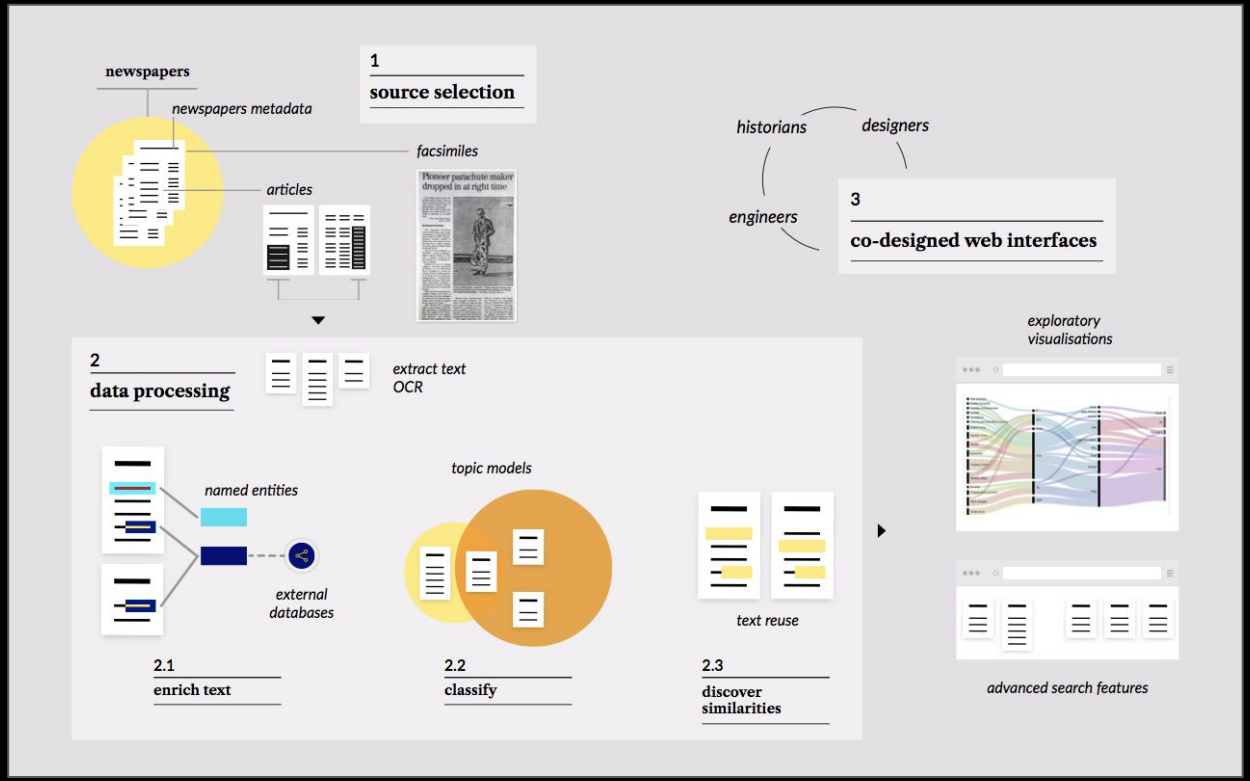
1. Wie passt man NLP-Werkzeuge an historischen Text an?

2. Wie erforscht man große und komplex Datenbestände?

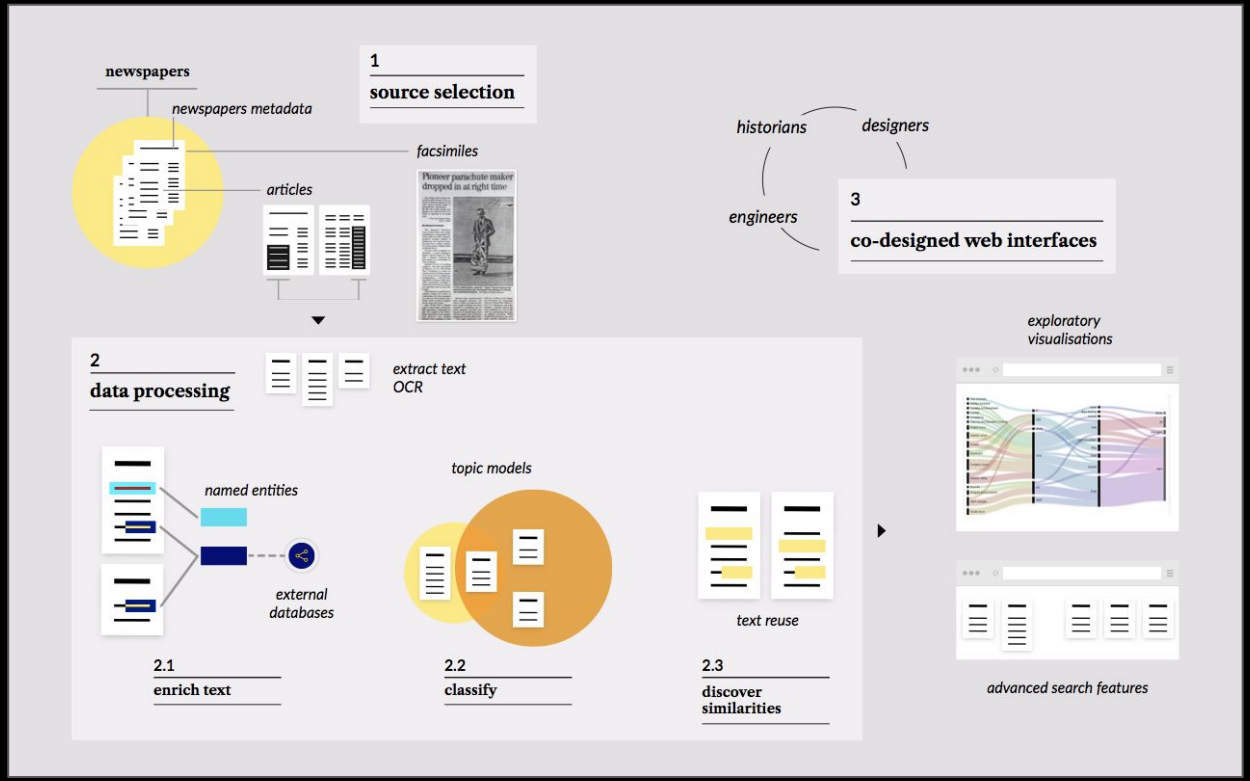
3. Welche Rückwirkungen hat dies auf historische Forschung?



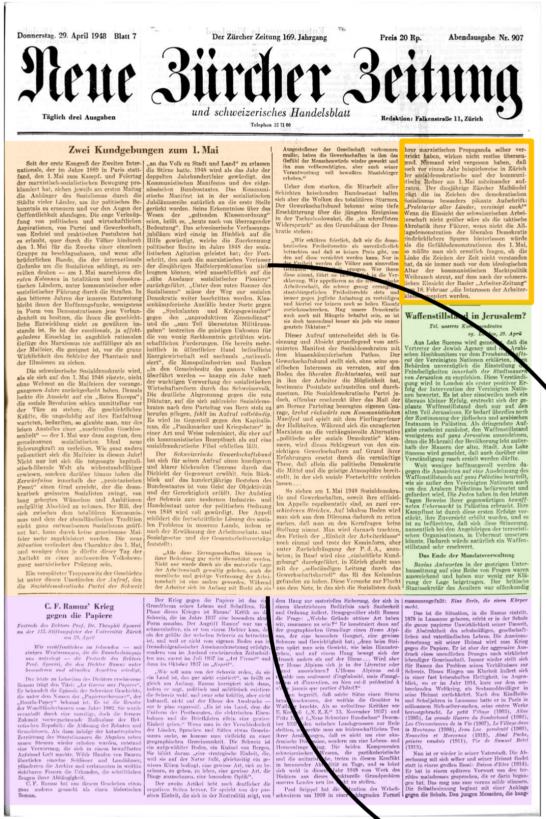
1. Korpus-Aufbau

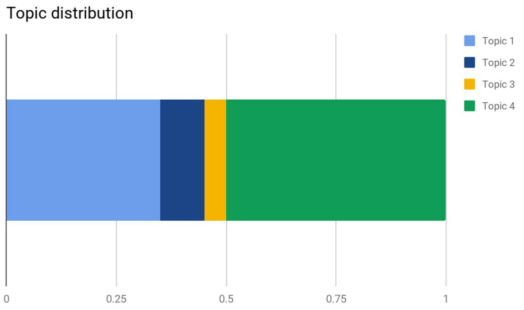


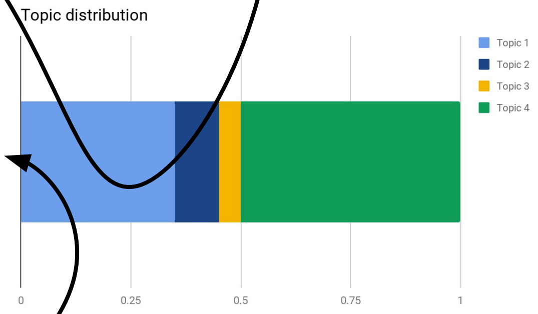
2. Semantische Anreicherung



Topic Modeling







kunst konzert bild theater künstler oper musikalisch spielen werk künstlerisch orchester aufführung ausstellung lied musik vortrag stück chor musil programm

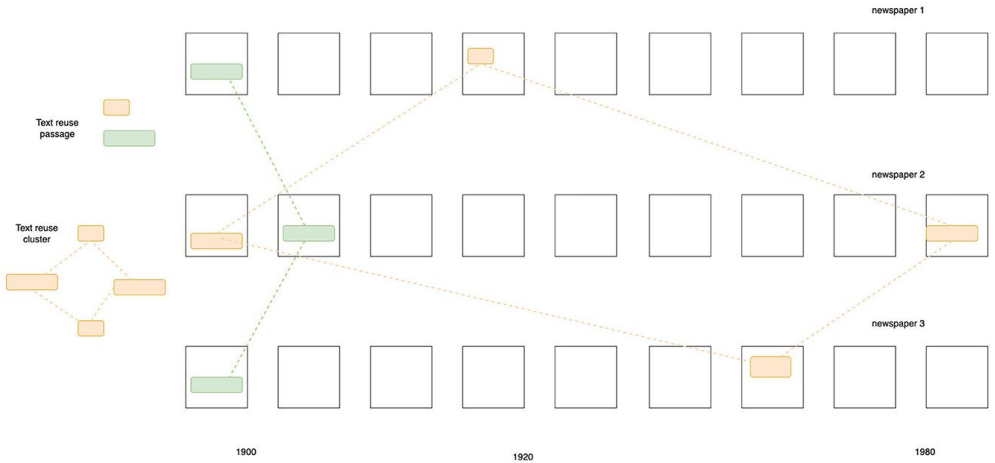
deutschen britisch deutsch truppe krieg russisch angriff flugzeug feindlich schwer front armee kampf london japanisch havas amerikanisch feind alliiert italienisch

zimmer lage vermieten verkaufen wohnung garten groß auskunft villa hotel schön zürich haus pension komfort preis see telephon sonnig modern

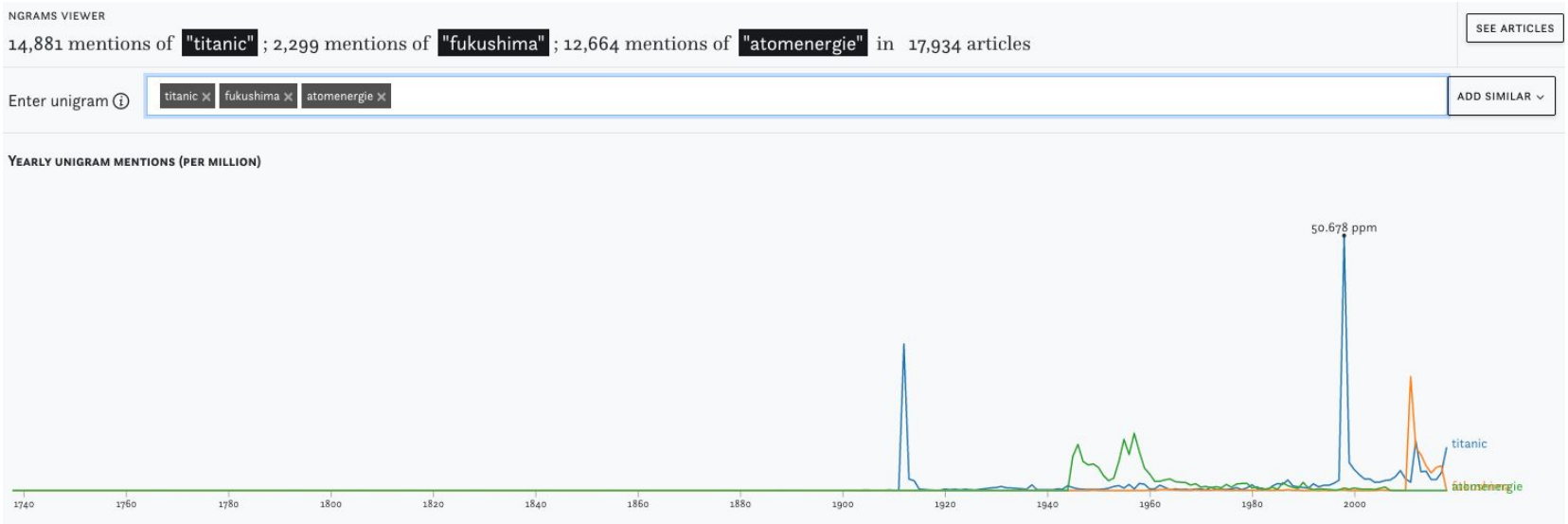
P. Stroëbel and S. Clematide (UZH)

Text reuse

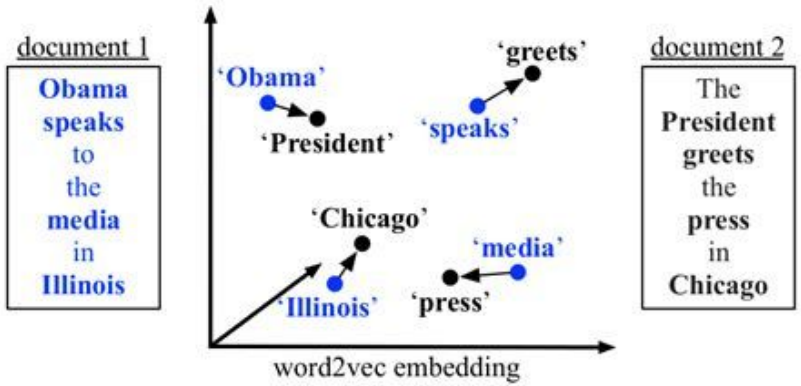

COMPARE TEXT REUSE PASSAGES
MONDAY, NOVEMBER 28, 1910 "M. Winston Churchill, ministre du commerce, pronon..." TEXTREUSEPASSAGE
Compare the passage below
with # 2 of 3 passages
BY DATE (DESC)
lia crise britaiiiilquc.
Gazette de Lausanne MONDAY, NOVEMBER 28, 1910 – P.2
Les suffragistes font de l'action directe
L'indépendance luxembourgeoise TUESDAY, NOVEMBER 29, 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.
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.


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

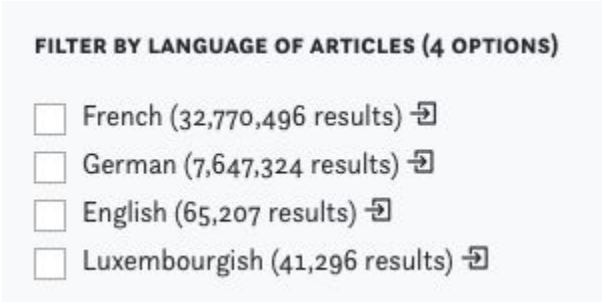
Ngrams


NGRAMS VIEWER
14,881 mentions of "titanic" ; 2,299 mentions of "fukushima" ; 12,664 mentions of "atomenergie" in 17,934 articles
SEE ARTICLES
Enter unigram
titanic
fukushima
atomenergie
ADD SIMILAR
YEARLY UNIGRAM MENTIONS (PER MILLION)
50.678 ppm
titanic
atomenergie
1740
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000

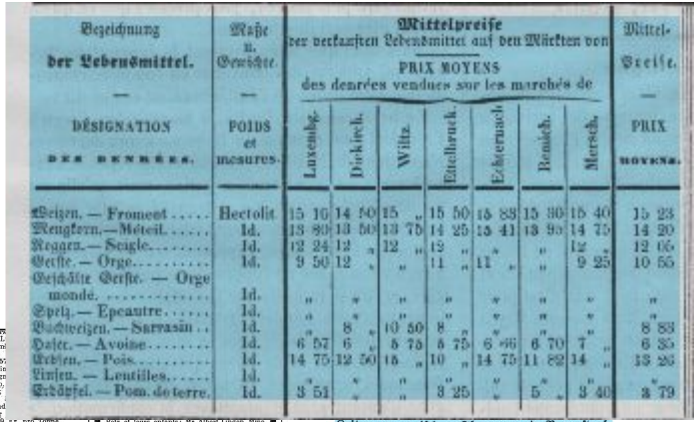
Word embeddings



Sprach-Erkennung



Inhaltstypen-Erkennung



Bezeichnung der Lebensmittel. — DÉSIGNATION DES DENRÉES.	Maße u. Gewichte. — POIDS et mesures.	Mittelpreise der verkauften Lebensmittel auf den Märkten von — PRIX MOYENS des denrées vendues sur les marchés de: Luxembg.	Diekirch.	Wiltz.	Ettelbruck.	Echternach.	Remich.	Mersch.	Mittel-Preise. — PRIX MOYENS.
Weizen. — Froment	Hectolit.	15 16	14 50	15 „	15 50	15 88	15 30	15 40	15 23
Mengkorn. — Méteil	Id.	13 80	13 50	13 75	14 25	15 41	13 95	14 75	14 20
Roggen. — Seigle	Id.	12 24	12 „	12 „	12 „	„	„	12 „	12 06
Gerste. — Orge	Id.	9 50	12 „	„	11 „	11 „	„	9 25	10 55
Geschälte Gerste. — Orge mondé	Id.	„	„	„	„	„	„	„	„
Spelz. — Epeautre	Id.	„	„	„	„	„	„	„	„
Buchweizen. — Sarrasin	Id.	„	8 „	10 50	8 „	„	„	„	8 83
Hafer. — Avoine	Id.	6 57	6 „	5 75	5 75	6 46	6 70	7 „	6 35
Erbsen. — Pois	Id.	14 75	12 50	15 „	10 „	14 75	11 82	14 „	13 26
Linsen. — Lentilles	Id.	„	„	„	„	„	„	„	„
Erdäpfel. — Pom. de terre	Id.	3 51	„	„	3 25	„	5 „	3 40	3 79

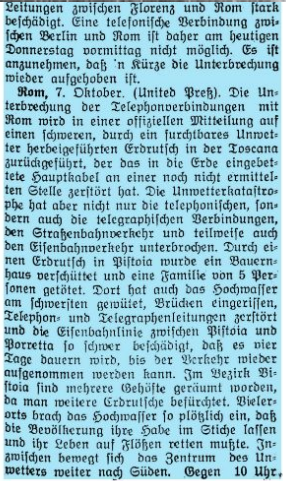
VAL-DE-TRAVERS

BUTTES

Conseil général

(Corr.) Le Conseil général, assemblé sous la présidence de M. J. Jeannet, vote les deux arrêtés ci-dessous: Le Conseil communal est autorisé à verser aux chômeurs affiliés à une caisse de chômage une allocation d'hiver pour autant qu'ils remplissent les conditions prévues dans le projet d'arrêté du Conseil d'Etat du 24 décembre 1932 concernant le versement d'une allocation d'hiver. L'allocation sera versée sitôt l'arrêté du Conseil d'Etat en vigueur.

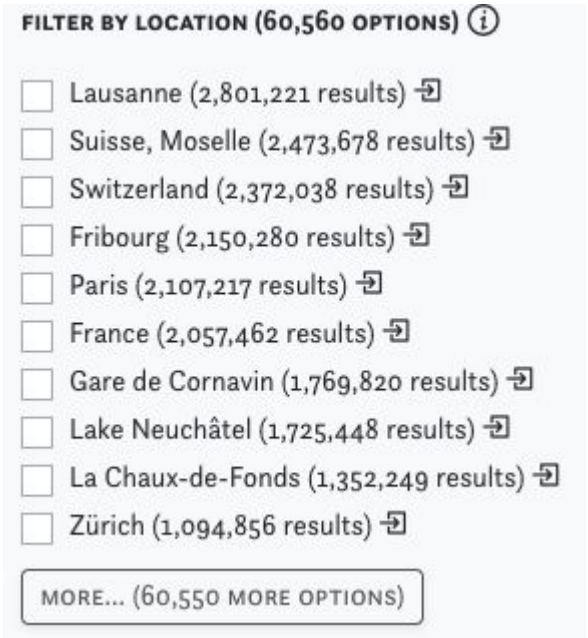
Le Conseil communal est autorisé à verser aux soutiens de famille qui reçoivent encore l'allocation de crise

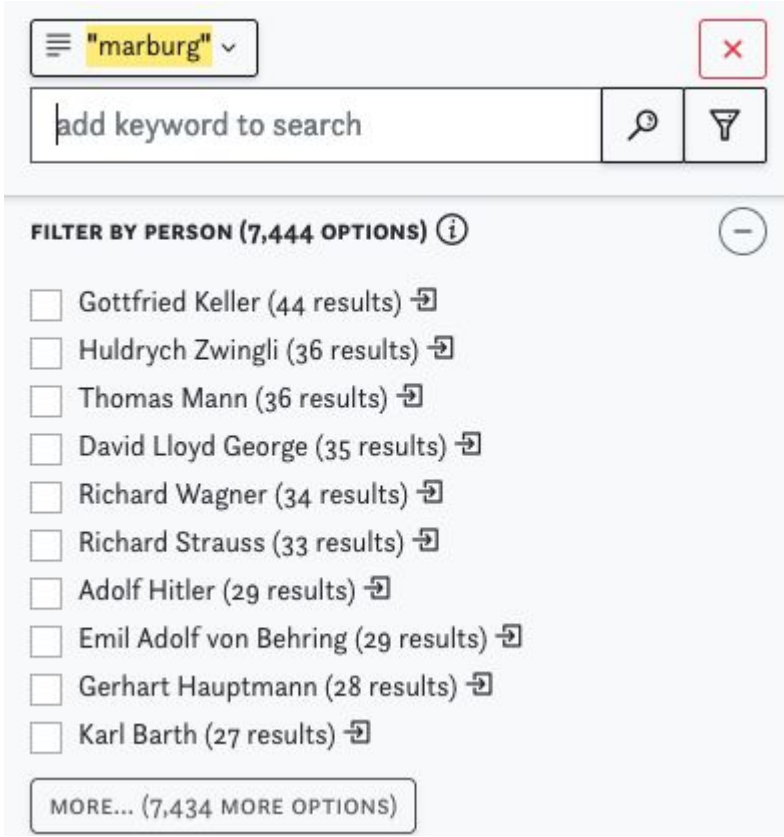


Leitungen zwischen Florenz und Rom stark beschädigt. Eine telefonische Verbindung zwischen Berlin und Rom ist daher am heutigen Donnerstag vormittag nicht möglich. Es ist anzunehmen, daß 'n Kürze die Unterbrechung wieder aufgehoben ist.

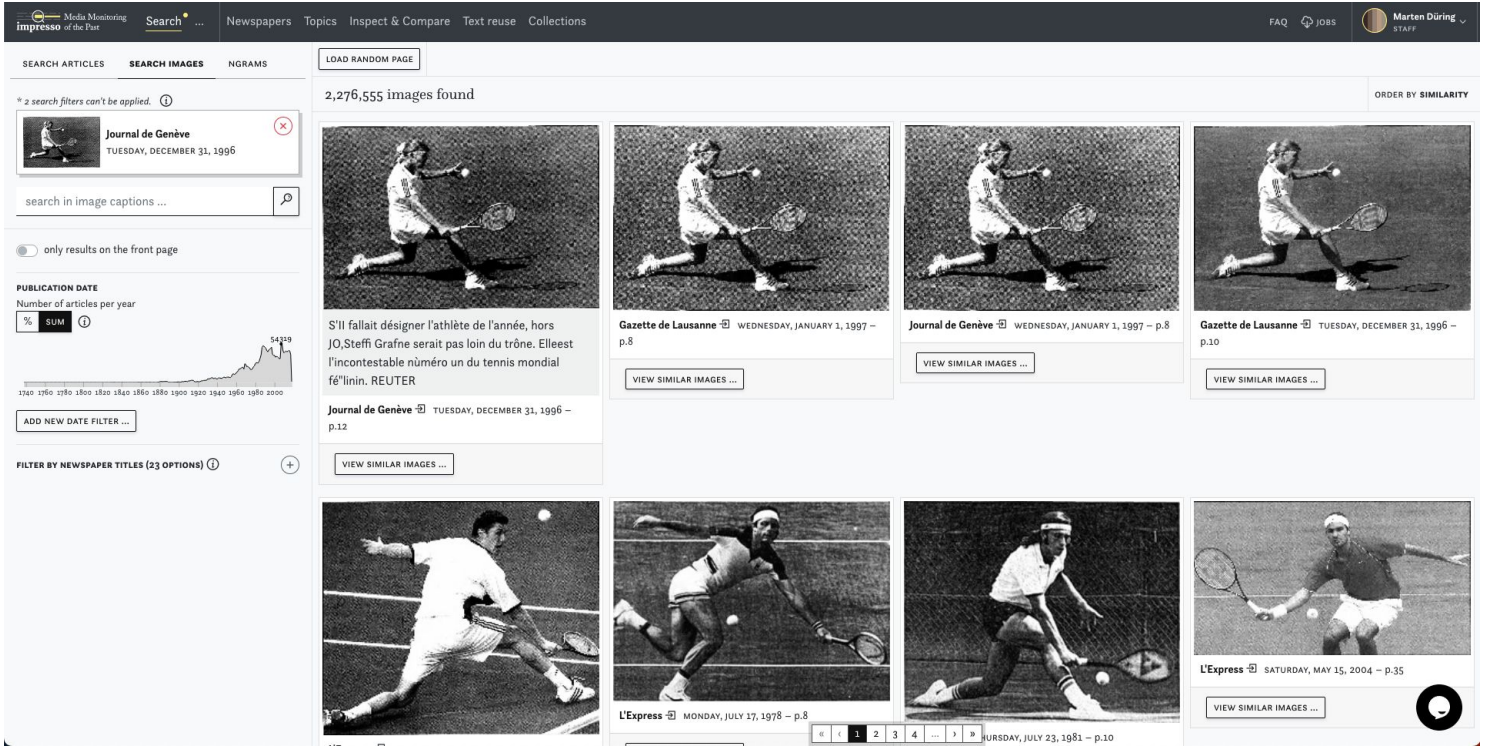
Rom, 7. Oktober. (United Preß). Die Unterbrechung der Telephonverbindungen mit Rom wird in einer offiziellen Mitteilung auf einen schweren, durch ein furchtbares Unwetter herbeigeführten Erdrutsch in der Toscana zurückgeführt, der das in die Erde eingebettete Hauptkabel an einer noch nicht ermittelten Stelle zerstört hat. Die Unwetterkatastrophe hat aber nicht nur die telephonischen, sondern auch die telegraphischen Verbindungen, den Straßenbahnverkehr und teilweise auch den Eisenbahnverkehr unterbrochen. Durch einen Erdrutsch in Pistoia wurde ein Bauernhaus verschüttet und eine Familie von 5 Personen getötet. Dort hat auch das Hochwasser am schwersten gewütet, Brücken eingerissen, Telephon- und Telegraphenleitungen zerstört und die Eisenbahnlinie zwischen Pistoia und Porretta so schwer beschädigt, daß es vier Tage dauern wird, bis der Verkehr wieder aufgenommen werden kann. Im Bezirk Pistoia sind mehrere Gehöfte geräumt worden, da man weitere Erdrutsche befürchtet. Vielerorts brach das Hochwasser so plötzlich ein, daß die Bevölkerung ihre Habe im Stiche lassen und ihr Leben auf Flößen retten mußte. Inzwischen bewegt sich das Zentrum des Unwetters weiter nach Süden. Gegen 10 Uhr

Named entities





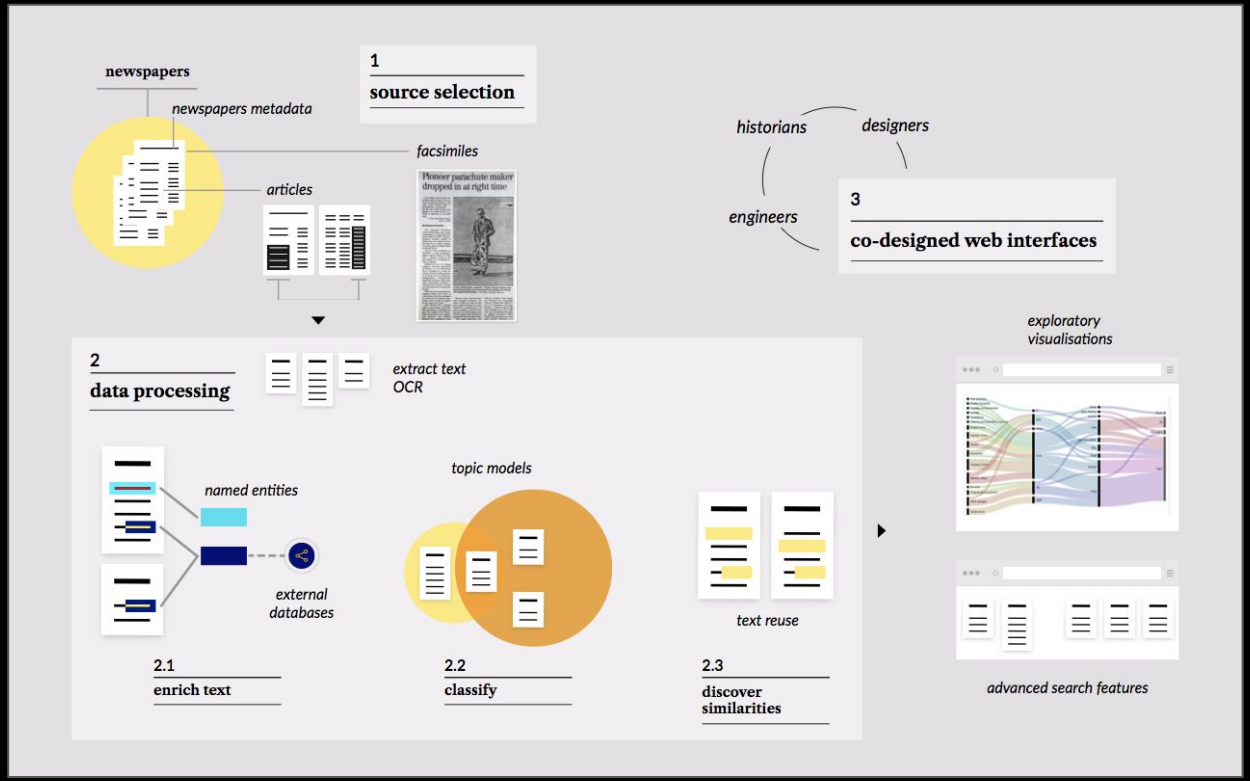
Bildähnlichkeits-Erkennung



Angereicherte Daten

Topic modeling
Ngrams
Text reuse
Word embeddings
Language detection
Content type detection
Named entities
OCR quality assessment
Image similarity

2. Co-design der impresso App



Workshops

Fallstudien

Media Monitoring impresso of the Past

Bridging the fields: Research scenarios

Tracking the anti-European posture in the public debate in Switzerland and Luxembourg (1848-1945)

1. Goal: Identify the postures in the debates around the European idea
2. Using people and slogans to collect a broad corpus of articles, across several newspaper titles / cluster them according to manual classification and automatically generated information
3. Quantity and diversity of the collected results / types of incarnation of the European idea.

Estelle >> Marten >> Julien >> Solenn >> Enrico >> Tobias >> Gerold

1

1

The coverage of The Battle of Arnhem in European Newspapers (1944-present)

1. Help me find all articles which cover the events at Arnhem: When does coverage start?
2. Help me discover similarities and differences in articles about the battle 1944 to present
3. Help me understand how "popularity" of the battle is changing over time. How could that be measured?

Estelle >> **Marten** >> Julien >> Solenn >> Enrico >> Tobias >> Gerold

2

2

Media Monitoring impresso of the Past

ALIENS IN NEWSPAPERS

What we want to look for:

- Concomitance with an **event** (astronomical discovery, UFO, release of a movie, publication of a book…)
- "Keyword in **context**"(–history of representations)
  - **Characterization of these non-humans** (judgement, anthropomorphism…)
  - Links with **semantic fields**: war, scare, space, colonization, creatures, intelligence…
  - Mentions of **persons** (e.g. authors, directors, astronomers…)
- A popular subject? (origin of these newspapers, type of the content–fiction, interviews…–size of the articles…)
- Evolutions of representations through time & possible differences between countries

French keywords: *extraterrestre, martien, sélénite, alien…*

First results on *e-newspaperarchives.ch*:

- 1289 results for "extraterrestre" from 1896 to 2012
- 26 results for "sélénite" from 1866 to 2010
- 1097 results for "martien" from 1863 to 2009
- 8725 results for "alien" from 1872 to 2014

Difficulties:

- **Substantivized adjectives:** "extraterrestre" can apply to phenomena and things too
- **Homonyms and similar words:**
  - "martien" often mistaken for "**Martigny**" (Swiss city)
  - "allen" often mistaken for "[Woody] allen", "alleń", "alléné" and words with -allen/allen (312 585 results on Retronews !!)
  - "alien": English words with numerous irrelevant meanings

3

3

Media Monitoring impresso of the Past

Bridging the fields: Research scenarios

Funding secondary education in nineteenth century Europe: structuration of a public debate

1. Research question : how educational policies were addressed by journalists and which data were deemed necessary to support the given point of view;
2. Help me find: articles dealing with educational policies, their "type" and the media associated to them ; linked named entities to "follow them";
3. Expected results/difficulties

Estelle >> Marten >> Julien >> **Solenn** >> Enrico >> Tobias >> Gerold

4

4

Media Monitoring impresso of the Past

Bridging the fields: Research scenarios

How do newspapers reflects the history of computing in Switzerland ?

5 mins

1. Research question
2. Operationalisation/tools
3. Expected results/difficulties

Estelle >> Marten >> Julien >> Solenn >> **Enrico** >> Tobias >> Gerold

5

5

Journalistic Cultures during the 19th Century: a Comparison between the *NZZ* and *le JdG*

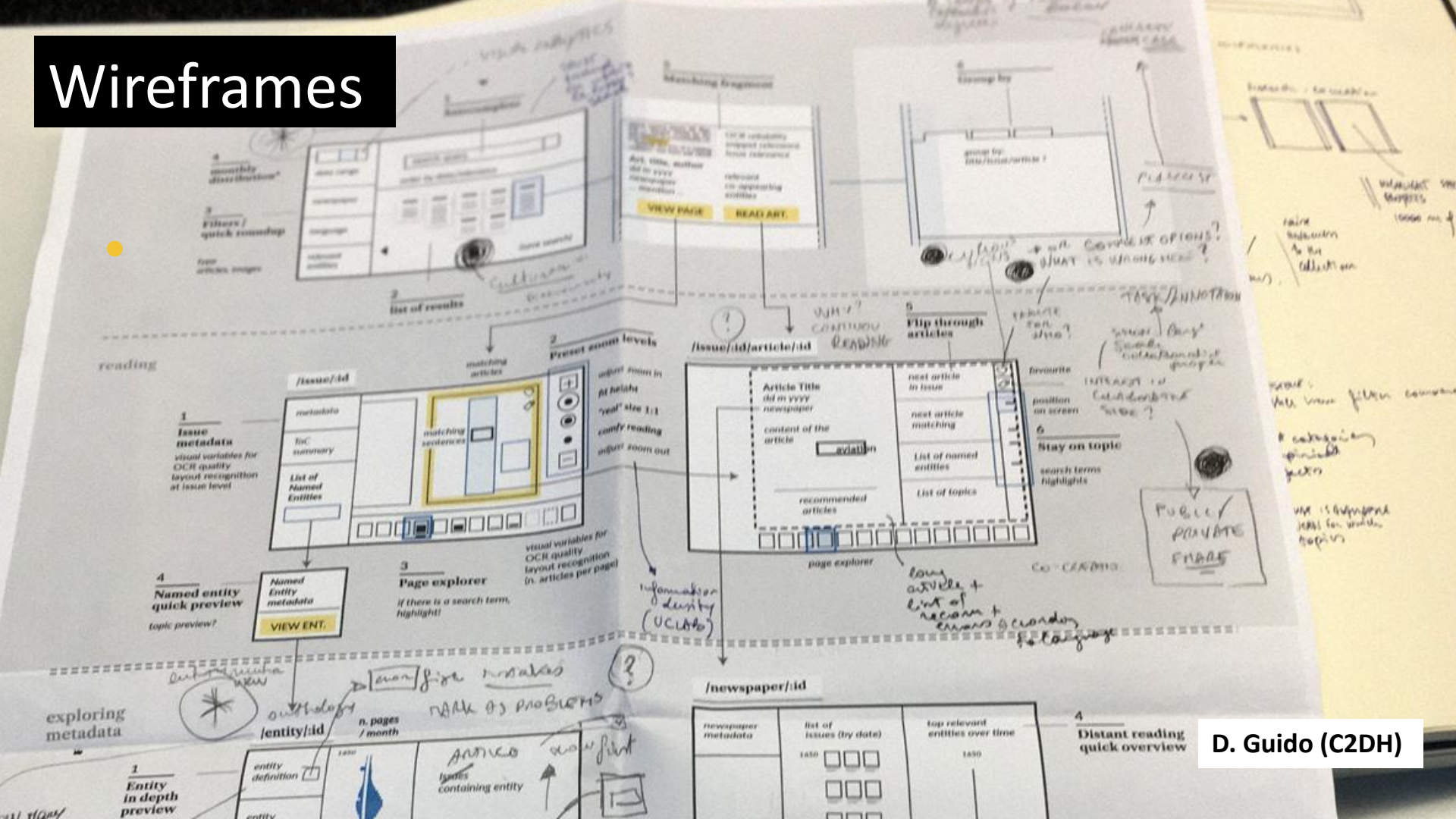
Main focuses (among others): genre change and meta-discourses

The three most important operationalisations (cf. handout)

- Help me find the lengths of sections (German: *Rubriken*, French: *rubriques*) in both newspapers and the change of these lengths in time
  (Handout: 1.1.1.)
- Help me find moments of emergence of sections in newspapers
  (Handout: 1.1.2.)
- Help me find/collect words and collocations related to sections
  (Handout: 1.2.1.)

6

Wireframes



D. Guido (C2DH)

S search for articles **C** collecting **M** exploring metadata
R reading contents **O** organising **E** visual experiments

Search

code for the SWOT

component description

S.1

search autocomplete: based on input, suggests words, date ranges, named entities or article categories

red text: uncompleted features

S.2

timeline n. results per year. This acts as "date range" filter

S.3

metadata filters / quick roundup how many search results per newspapers titles, languages, page content tags, format tags

explore

search

YOUR SEARCH | PAST SEARCHES

GROUP RESULTS BY: ISSUE | PAGE | ARTICLES | SENTENCE

FULL TEXT SEARCH — REFINE ...

... type a text or a date or to start ...

PUBLISHED IN (DATE RANGE)

TIMELINE OF: N. ARTICLES | N. ISSUES

1745 1800 1850 1900 1980

select a date range to show alrticles published

NEWSPAPER TITLES

53625 **La Gazette de Lausanne** quotidien suisse de langue française édité à Lausanne

43864 **Le Journal de Gèneve** quotidien suisse qui a paru du 1826 au 28 février 1998

MORE ...

ARTICLE TYPE

112 **partisans**

50 **satirique**

MORE ...

TOP NAMED ENTITIES

103 **Napoleon**

50 **Zurich, Suisse** location

MORE ...

LANGUAGE

112 **French**

SEARCH SUMMARY

Find **53625987** articles in our collection.

1 of 112

Les recherches d'

L'INFORMATION

INTERNATIONAL NEWS

Gazette de Lausan

Les recherches d'une possibilités JEAN-PIE de multiples...

2 of 112

Les USA et l'AFN

INTERNATIONAL NEWS

Gazette de Lausan

Les USA et l'AFN Un la terrasse de café o contre l assaut...

3 of 112

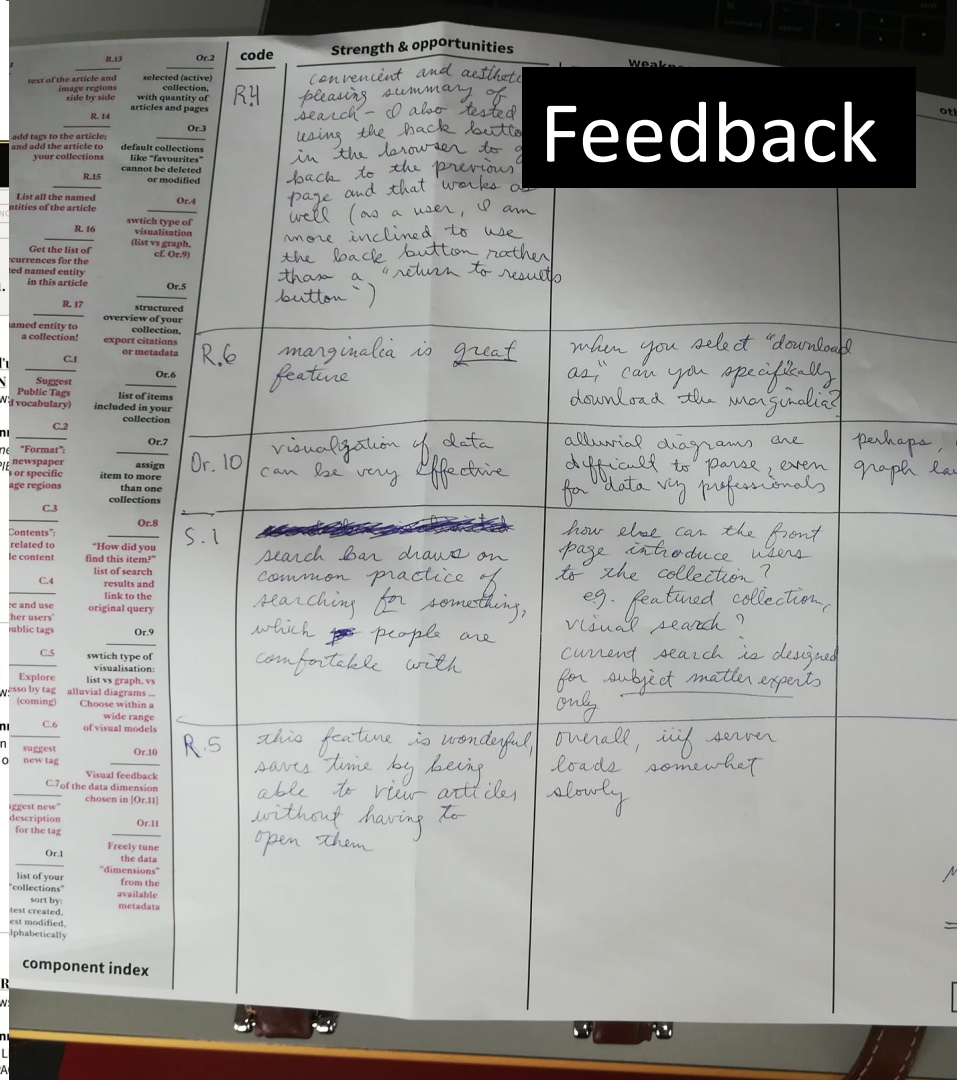
FRANCO PRÉFÈR

INTERNATIONAL NEWS

Gazette de Lausan

FRANCO PRÉFÈRE L PRINTEMPS D'ESPA

Feedback

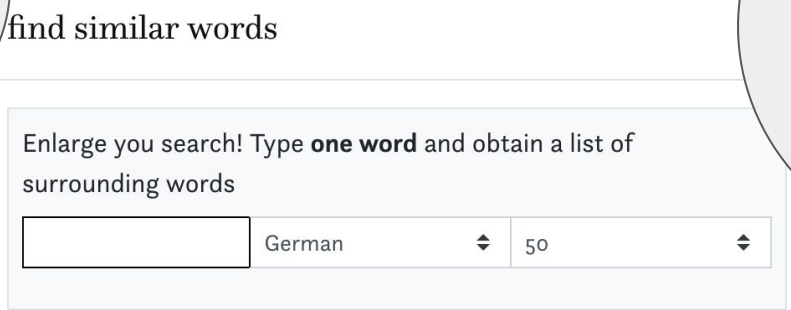


Codesign: Word embeddings

You really need to do something about OCR quality!

Keywords alone do not suffice to find everything I need.

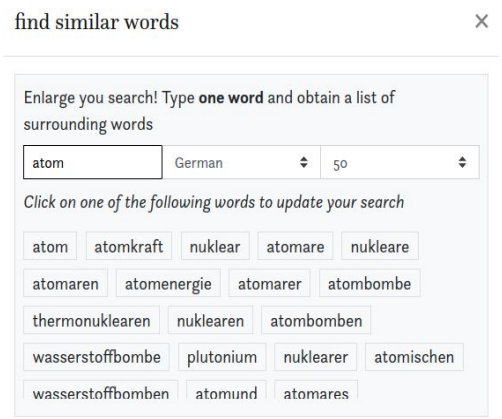
Different keywords yield different results. I get confused!



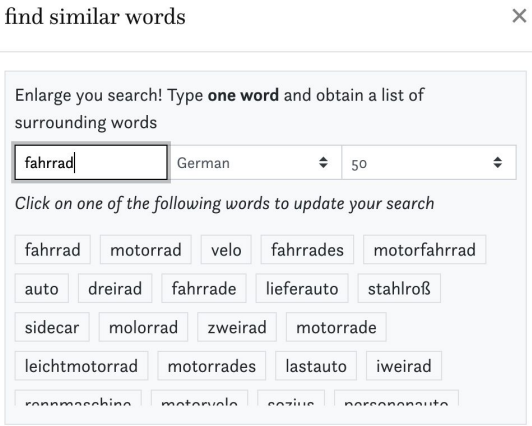
Let's compute word embeddings as well and see how they help with data-driven analyses of history!

Let's combine Search and Advanced Search in one!

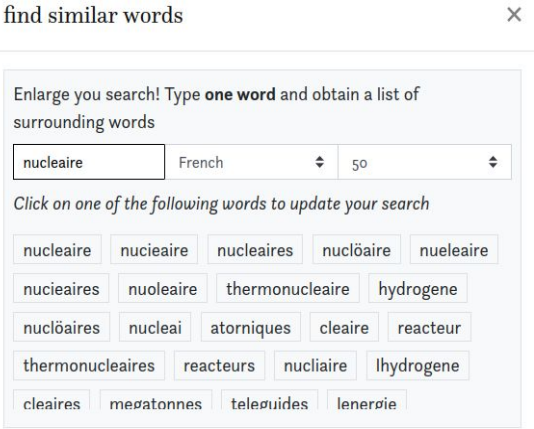
Codesign: Word embeddings



Semantische Nachbarn



Synonyme



OCR Fehler

Codesign: Topic Modeling

Facetten von “atom”

Codesign: Topic Modeling

Journal de Genève · Friday, August 3, 1906

ADD TO COLLECTION ... | FACSIMILE | TRANSCRIPT

SUPPRESSION RADICALE DES POUSSIERES par

LOCATIONS Schaffhausen, Gare de Cornavin

TOPICS

12.6% FR maison · vente · choix · magasin · qualité
10.6% FR avion · vol · appareil · aviation · pilote
4.5% FR tir · gauche · droite · page · main
3.5% FR bureau · place · suite · famille · ménage
3.5% FR état · lit · bureau · table · bois
2.5% FR district · lieu · tribunal · ville · paix
2.5% FR enfant · famille · père · vie · mère
2.5% FR cas · santé · médecin · maladie · docteur
2.5% FR main · tête · temps · coup · air
2.5% FR service · garde · domicile · cas · mois
2.5% FR mois · numéro · carte · adresse · poste
2.5% FR loi · droit · conseil · cas · article

ADD TO COLLECTION ...

NO TEXT REUSE PASSAGES AVAILABLE

SUPPRESSION RADICALE DES POUSSIERES
par les appareils portatifs,
Nettoyage sans « ATOM » « ATOM "
NETTOYAGE des Tapis, Meuble », Rideaux, Portières,
Literie, Vêtements, Fourrures,. Automobiles,
',.-.'Bibliothèques, Billards, etc...,'
SIMPLICITÉ !, .. HYGIÈNEI *
Maniement a la main. /
Nettoyage par le vide à laportée de tous !
Indispensable dans les Hôtels ; Restaurants,
Villas, Ménages, Hôpitaux, Bibliothèques, etc. Nettoyage avec « ATOM »

Codesign: Topic Modeling

Die Tat · Wednesday, June 18, 1958

ADD TO COLLECTION ... | FACSIMILE | TRANSCRIPT

Chruschtschow spielt den Ungeduldigen

LOCATIONS Lage, Xmal Deutschland, Ned Washington

TOPICS

89.7% DE regierung · afp · sowjetunion · reuter · moskau

2.8% DE juli · juni · august · mai · samstag

ADD TO COLLECTION ...

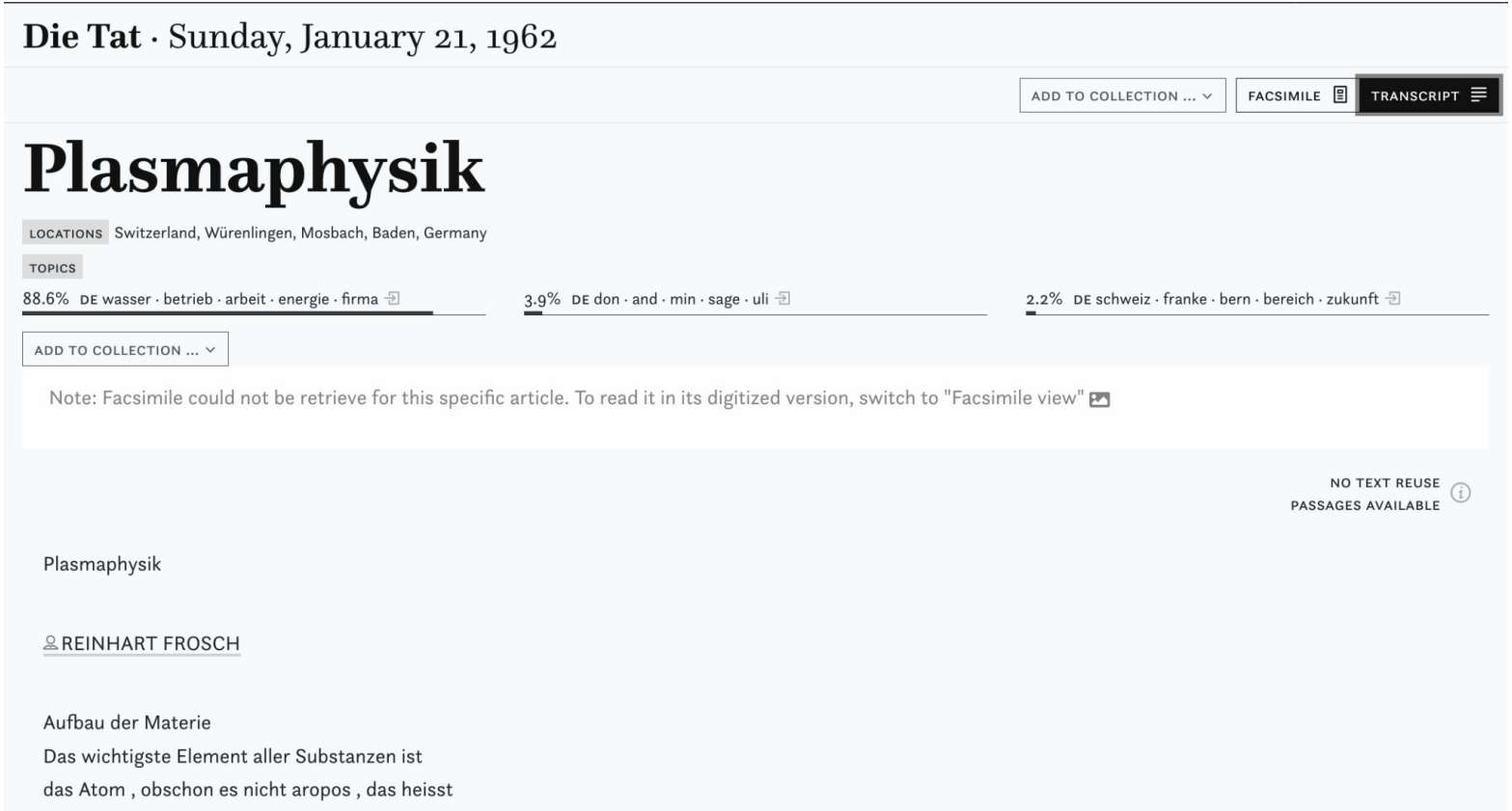
Note: Facsimile could not be retrieve for this specific article. To read it in its digitized version, switch to "Facsimile view"

NO TEXT REUSE PASSAGES AVAILABLE

Chruschtschow spielt den Ungeduldigen

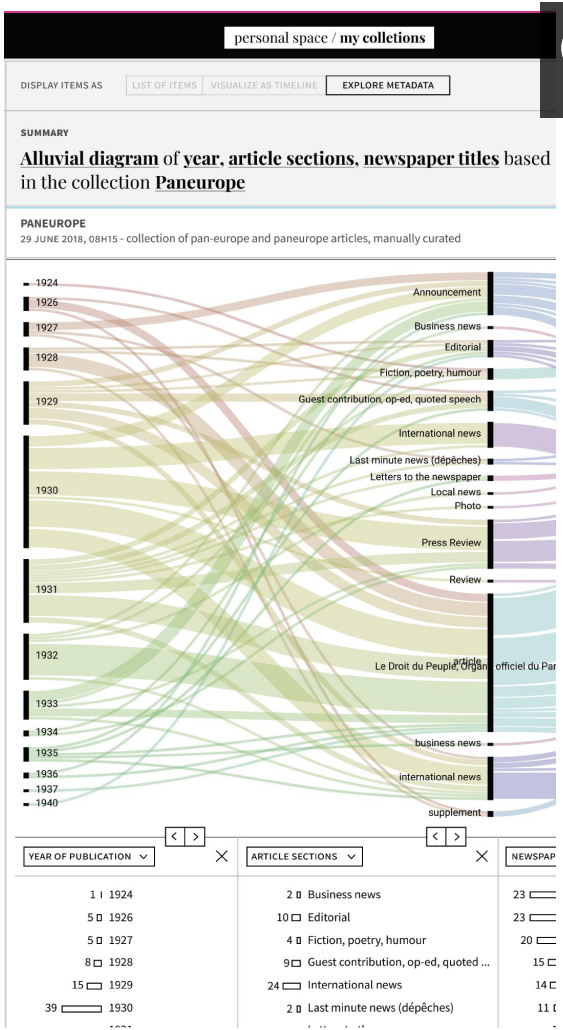
Veröffentlichung der letzten Botsohaf
Moskau , 17 . Juni . (Reuter) Die Agentur Tass
verbreitete am Montag eine Zusammenfassung
des Inhalts des Schreibens , das der sowjetische
Ministerpräsident Chruschtschow Präsident
Eisenhower am 11 . Juni überreichen ließ .
Der Augenblick sei gekommen , heißt es u . a .
in der Botschaft , um ehrlich festzustellen , ob alle
Weltmächte tatsächlich eine Gipfelkonferenz
wünschten . Die westlichen Vorschläge betreffend
die Tagesordnung einer solchen Konferenz sowie
andere Dokumente , welche die Westmächte der
Sowjetunion übermittelt hätten

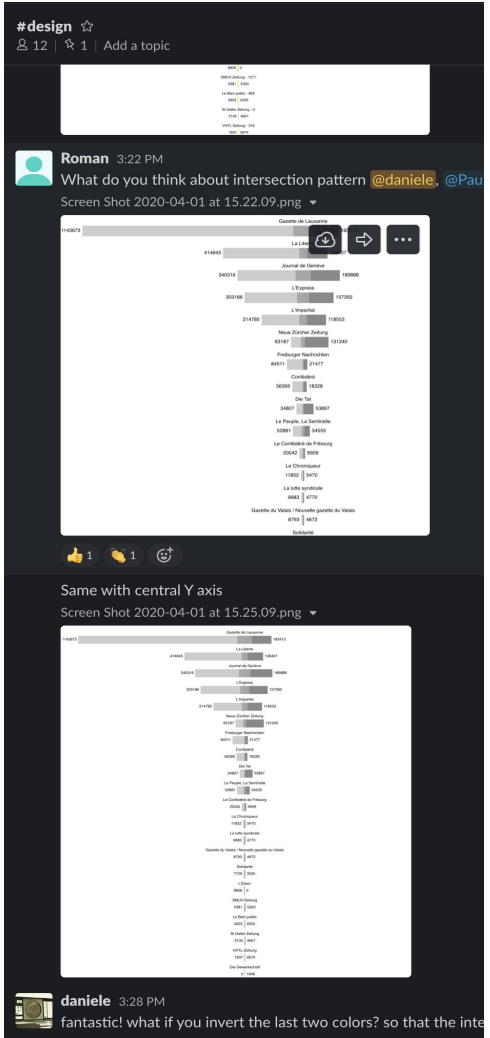
Codesign: Topic Modeling



Codesign: Topic Modeling

Codesign: Inspect & Compare



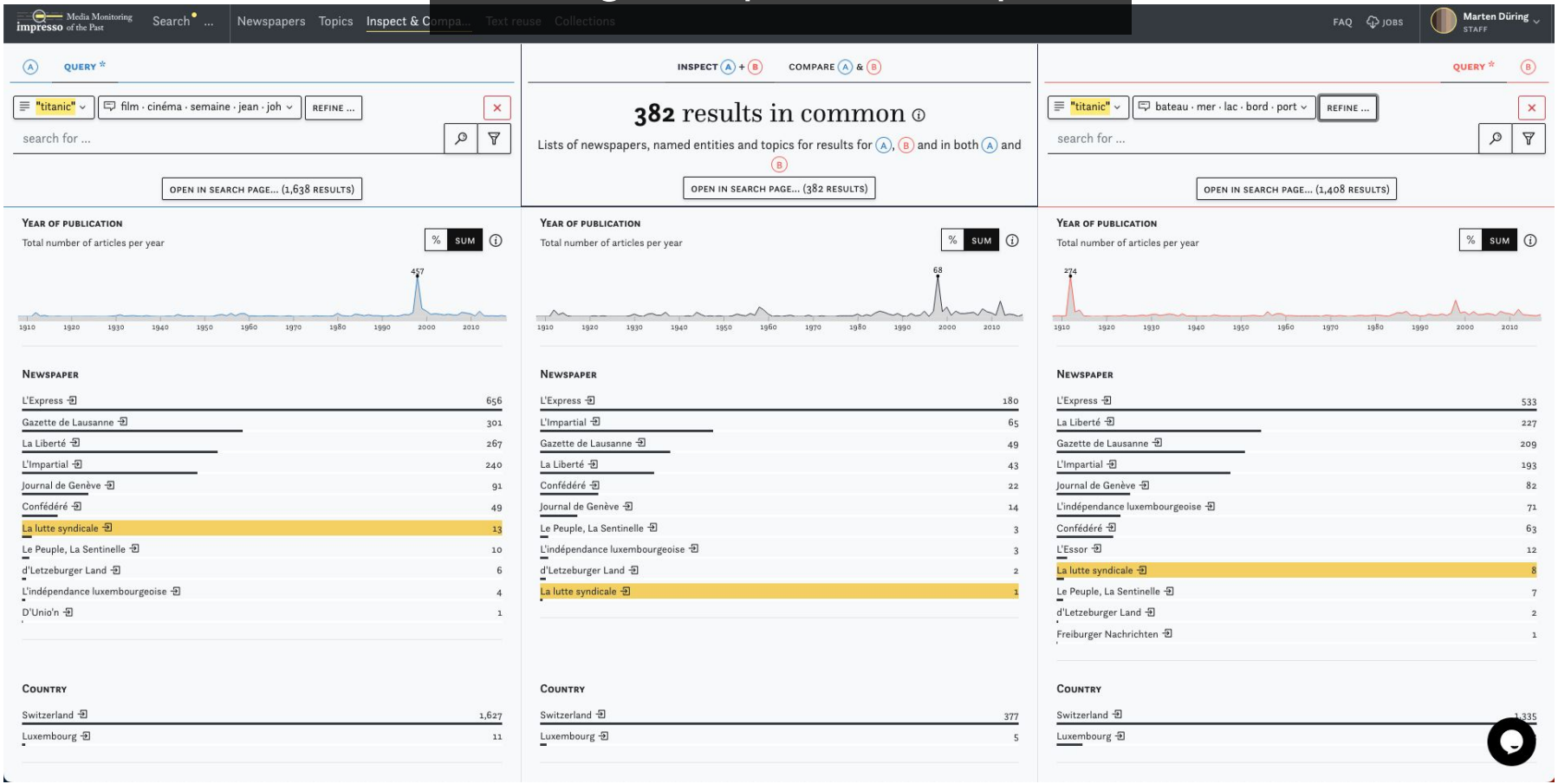


Produktvergleich als Inspiration

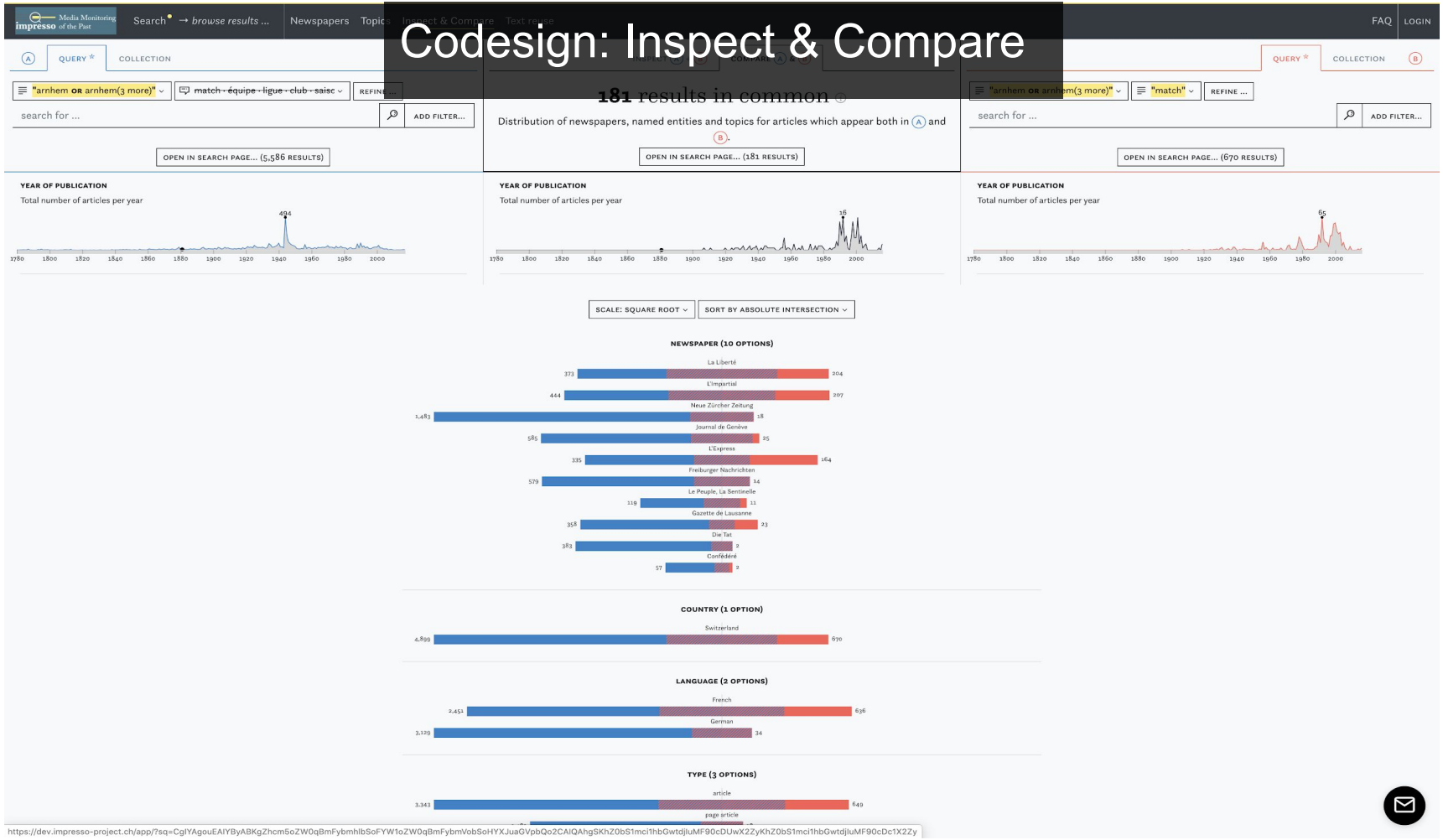
Compare with similar items

					
	This item Potensic ATOM 4K GPS Drone with 3-Axis Gimbal, 96 Min. Long Flight Time, 6KM FPV Transmission, Visual Following/QuickShots/RTH, Under 249g, Wind Force 5, 12MP Photos Camera Drone for Beginners, Adults	DJI Mini 4 Pro (DJI RC 2 Remote Control), Foldable Mini Drone with 4K Camera for Adults, Under 249g, 34 Minutes Flight Time, 20km Video Transmission, Omnidirectional Image Recognition, Class C0	12PRO Drone with Camera Brushless Motor Drone for Beginners and Adults with Motorised Adjustable 135° Camera 1080P HD 2 Cameras Drone 5G WiFi FPV RC Foldable Quadcopter Altitude Hold 2 Batteries	DJI Mini 2 SE, Lightweight and Foldable Mini Camera Drone with 2.7K Video, Intelligent Modes, 10km Video Transmission, 31 min Flight Time, Under 249g, Easy to Use, Photo Shooting, Street Shooting	Potensic ATOM SE Drone Replacement with Camera Includes a Set of Propellers and All Electronic Components, Without Battery/Remote Control, Only for Repair and Replacement, Can't Fly Alone
	Add to Basket	Add to Basket	Add to Basket	Add to Basket	Add to Basket
Customer Rating	★★★★★ (219)	★★★★½ (338)	★★★★½ (62)	★★★★½ (2006)	★★★★½ (13)
Price	€419.99	€999.00	€129.99	€326.99	€149.99
Sold By	Potensic-EU Official Store	Amazon.de	morlyrctooy	MondoTop	Potensic-EU Official Store

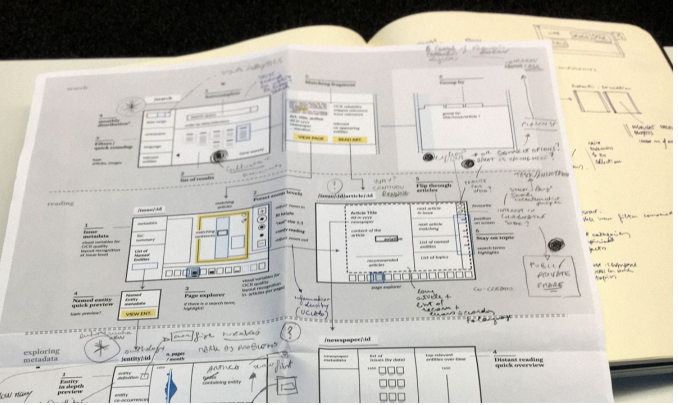
Codesign: Inspect & Compare



Codesign: Inspect & Compare



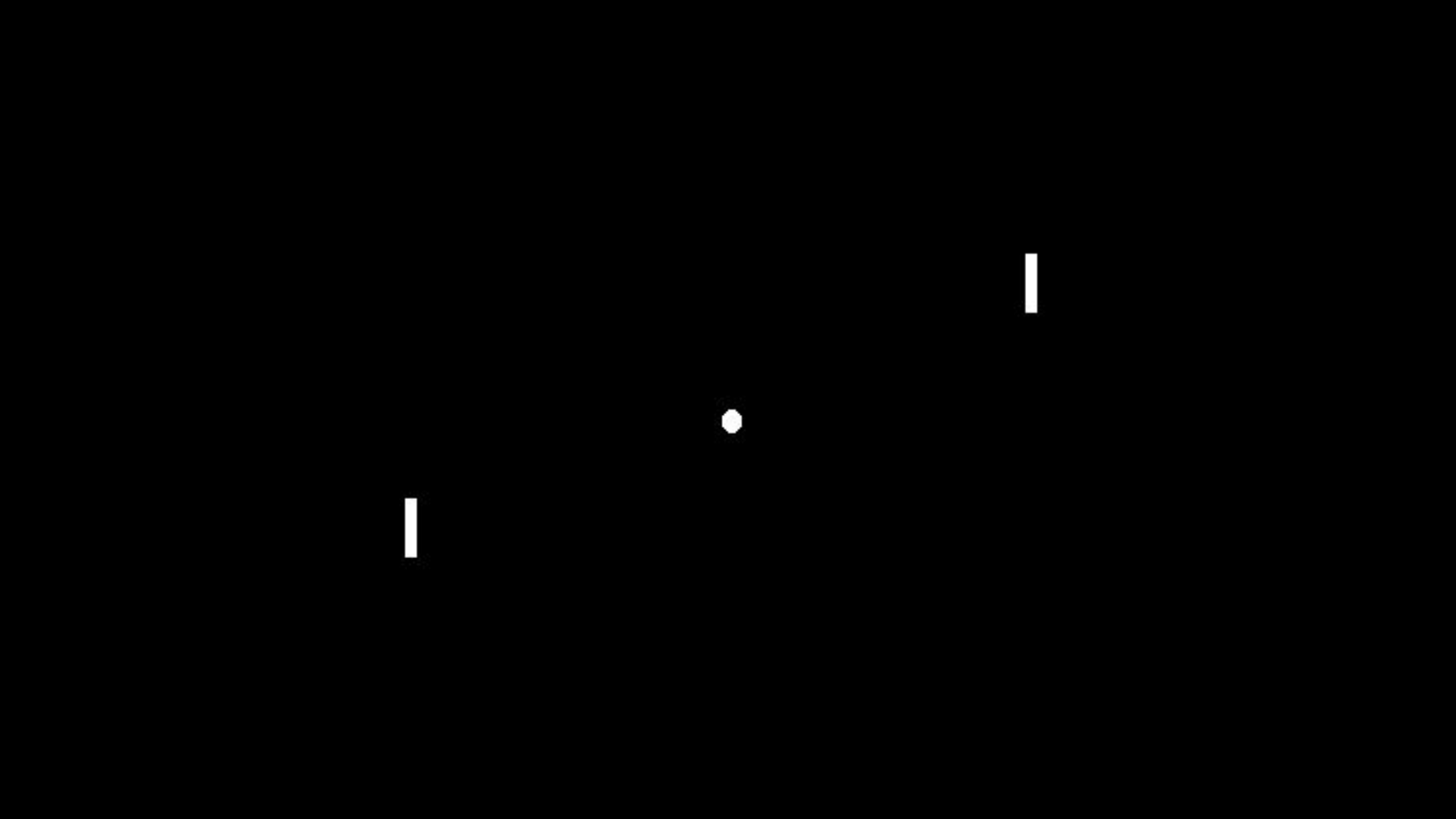
https://dev.impresso-project.ch/app/?sq=CgIYAgouEAIYByABKgZhcm5oZW0qBmFybmhlbSoFYW1oZW0qBmFybmVobSoHYXJuaGVpbQo2CAIQAhgSKhZ0bS1mci1hbGwtdjluMF90cDUwX2ZyKhZ0bS1mci1hbGwtdjluMF90cDc1X2Zy



Codesign Fazit

Voraussetzungen sind **Zeit, Prototypen und Praxis** (Henne-Ei-Problem).

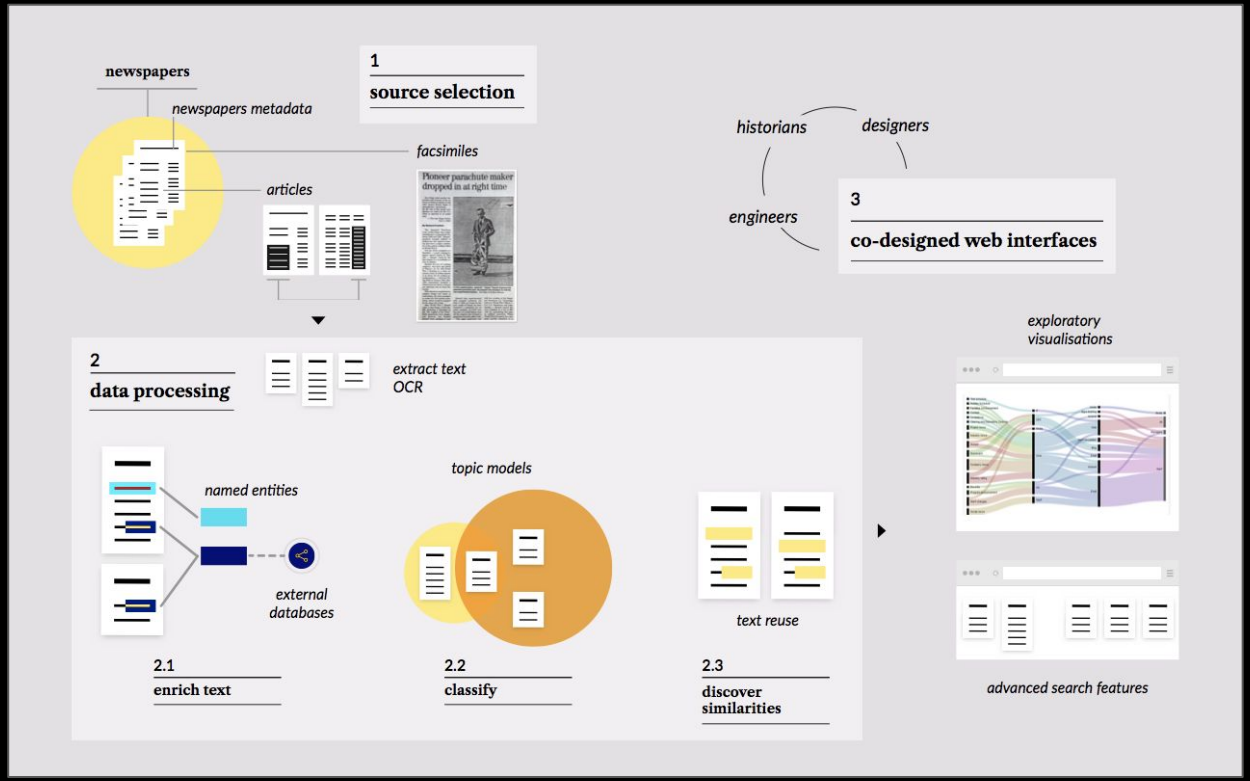
Experimente sind essentiell.



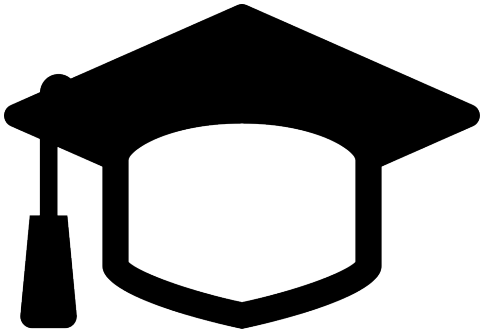
3. Demo

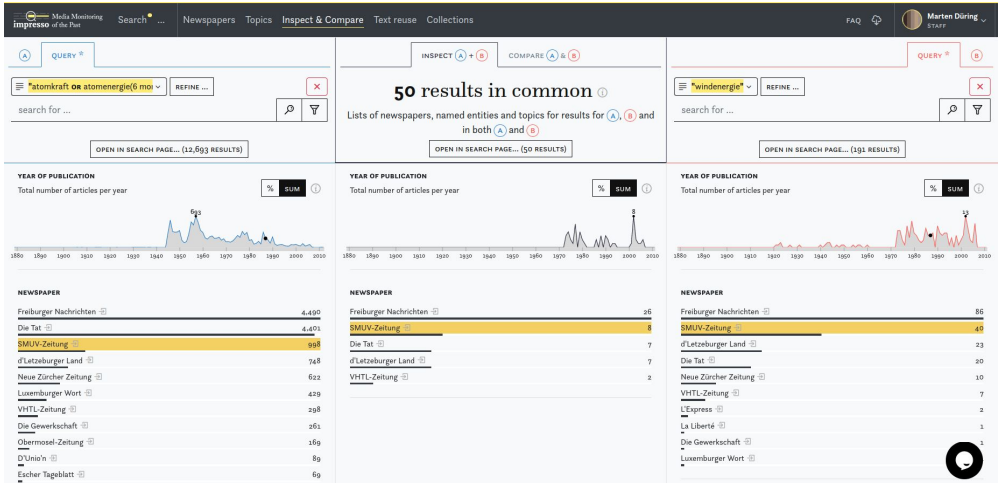

Media Monitoring of the Past
Mining 200 years of historical newspapers
How can newspapers help understand the past? How to explore them?
For legal reasons not all content is available.
To gain access to the full impresso corpus please register and sign our Non-Disclosure-Agreement.
DOWNLOAD NON-DISCLOSURE-AGREEMENT FORM
... and return the signed form to info@impresso-project.ch

4. Historische Forschung

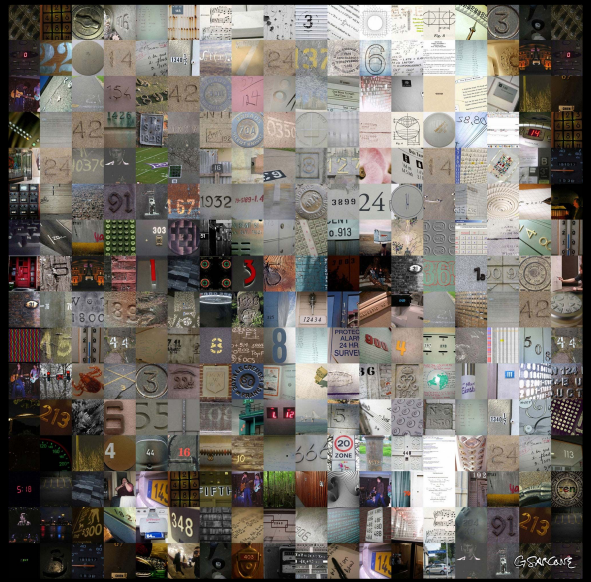


Target audience





HistorikerInnen ohne besondere digitale 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)

SEARCH ARTICLES | SEARCH IMAGES | NGRAMS

* 2 search filters can't be applied.

only results on the front page

PUBLICATION DATE

Number of articles per year

% SUM

ADD NEW DATE FILTER ...

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results

- [ ] French (250,490 results)
- [ ] German (70,953 results)
- [ ] Luxembourgish (628 results)
- [ ] English (3 results)

FILTER BY NEWSPAPER TITLES (68 OPTIONS)

check one or more newspaper to filter results

- [ ] Journal de Genève (51,937 results)
- [ ] L'Express (39,486 results)
- [ ] Gazette de Lausanne (38,025 results)
- [ ] La Liberté (37,995 results)
- [ ] L'Impartial (33,138 results)

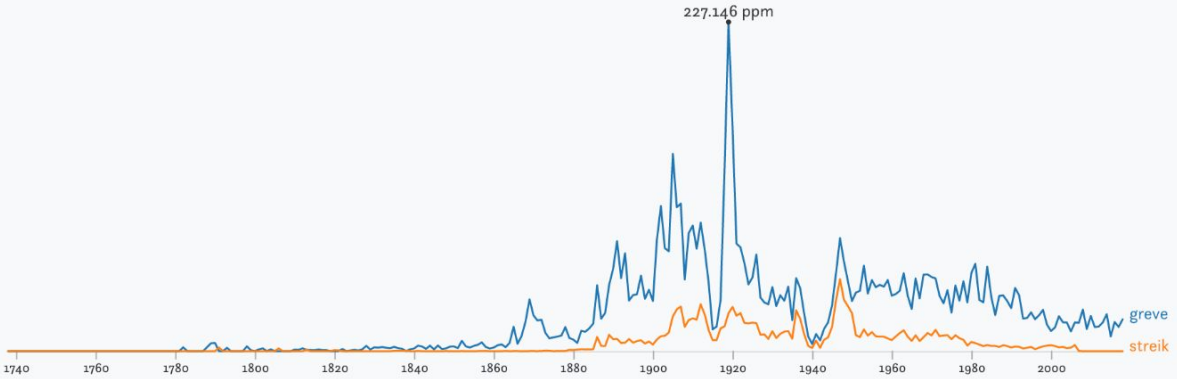
NGRAMS VIEWER

497,781 mentions of "greve" ; 123,490 mentions of "streik" in 322,074 articles

SEE ARTICLES

Enter unigram greve × streik × ADD SIMILAR

YEARLY UNIGRAM MENTIONS (PER MILLION)



DOWNLOAD DATA IN JSON

Krösus und das Orakel von Delphi

Heinrich Leutemann's The Oracle of Delphi Entranced; Cresus portrait

impresso als Orakel

Welche Frage genau wird an die App gestellt?

Vorsicht vor *confirmation bias*.

Was müssen wir wissen, um angemessene Fragen zu stellen und Ergebnisse zu interpretieren?

impresso als Orakel

Welche Frage genau wird an die App gestellt?

Vorsicht vor *confirmation bias*.

Was müssen wir wissen, um angemessene Fragen zu stellen und Ergebnisse zu interpretieren?

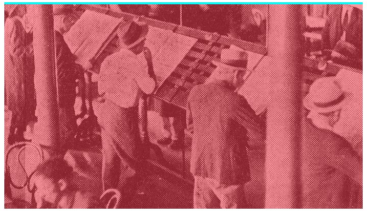
Generelle Herausforderung für die digitalen Geisteswissenschaften:

Der kritische Umgang mit:

- Digitalisierung
- Daten
- Metadaten
- KI/ hier: Text mining etc
- Visualisierung
- UX

Einstieg: Ranke2 - From the shelf to the web

Ranke.2 ~~Source criticism in the digital age



From the shelf to the web, exploring historical newspapers in the digital age.

NEWSPAPERS

Impresso.

TAGS: digitisation of newspapers , digital source criticism , optical character recognition , search engines , interface

go to lesson

4 out of 4 — Looking for Robert Schuman(n) ¶

Names of people and locations are more prone to OCR mistakes, as they cannot be found in the dictionaries that are used to recognise printed words. This means that if the software used for digitisation has not integrated a dictionary of names, an accurate match for a query of a name relies completely on the correct identification of each single letter of the name.

Instructions +

4.a How can we identify articles about "Robert Schuman"? 20 min _

- Read key information that can be useful to distinguish the two Robert Schuman(n)s:

Robert Schuman, the politician, is most famous for the eponymous declaration

Robert Schumann, the composer

- What kind of search filters could be used to distinguish the two personalities?
- Which period is relevant for each of them?
- What kind of keywords could be helpful?
- What other names would they potentially be associated with?

4.b Collecting articles on Robert Schuman(n) +

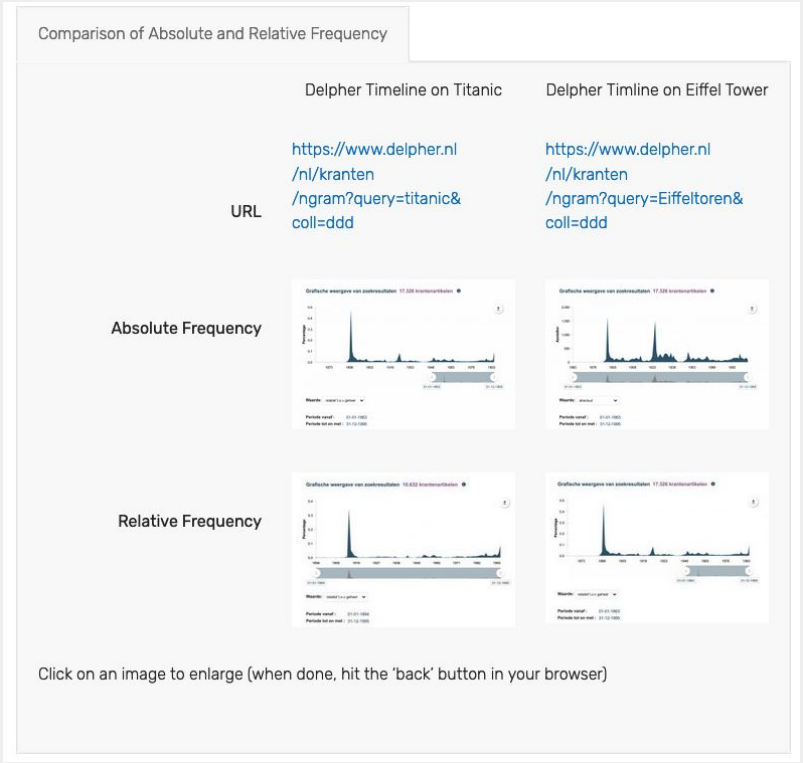
4c. Looking for Robert Schuman in Luxembourg +

Reading/viewing suggestions +

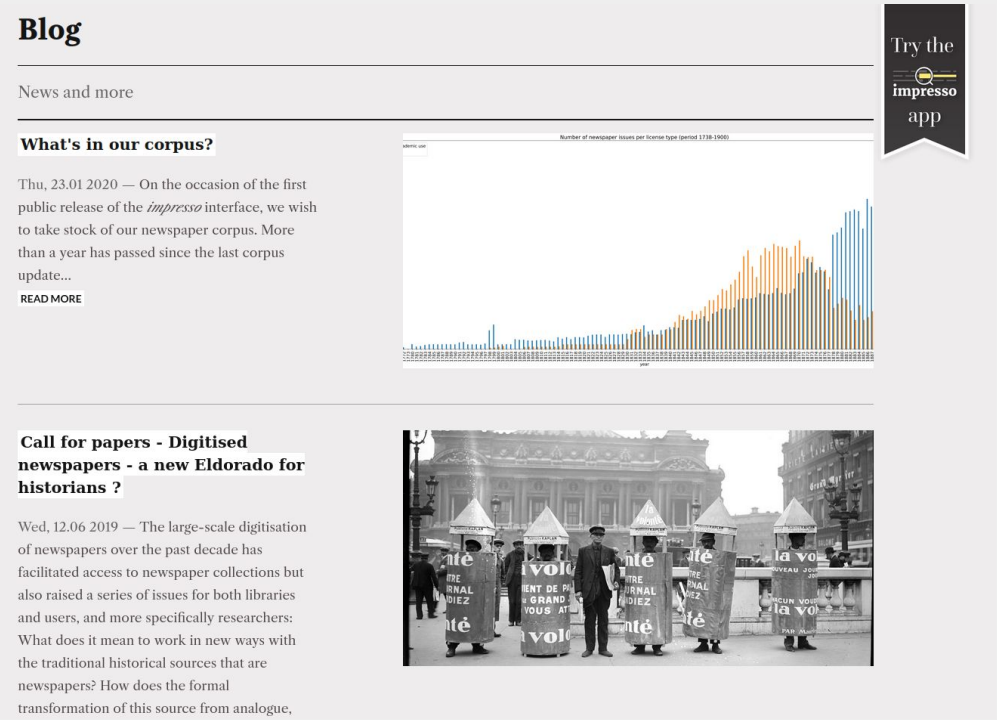
https://ranke2.uni.lu/

Fortgeschritten: Digitised newspapers as new artefacts

https://training.parthenos-project.eu/



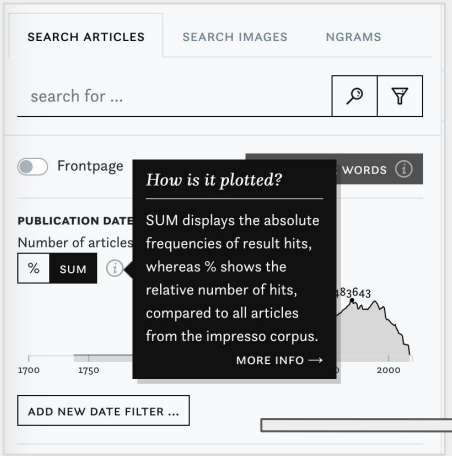
impresso data processing: Blog

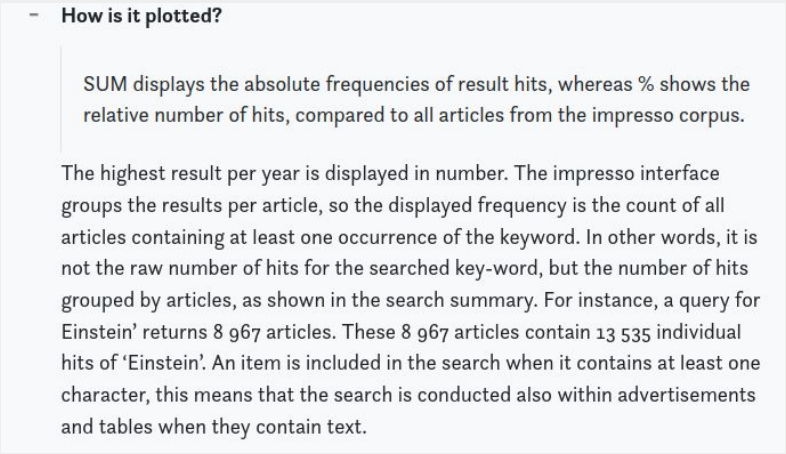


Named entity processing
Topic modeling
Text reuse
Word embeddings
Corpus composition
…

https://impresso-project.ch/blog/

Komponenten und Anreicherungen: i-buttons und FAQs





https://impresso-project.ch/app/faq

impresso für Fortgeschrittene

Structured in three skill levels, teaches advanced usage of the interface using case studies.

Impresso Challenges

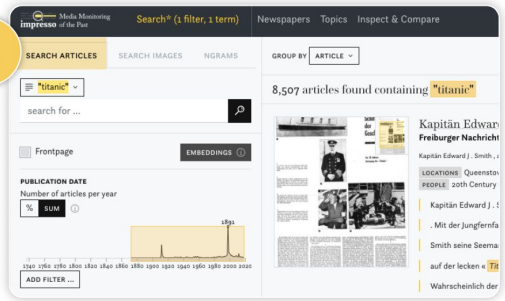
How to explore the newspapers with persons or locations?
What are topics good for?
What elements can be compared?

Get a better understanding of this interfaces' features and how they can interact with 3 challenges, starting with an initiation and leading to an expert level use of the interface.

DOWNLOAD CHALLENGES PDF

in contrast, search for the word titanic in the general search page: what is the difference between the number of hits shown in the ngram viewer and the search page?

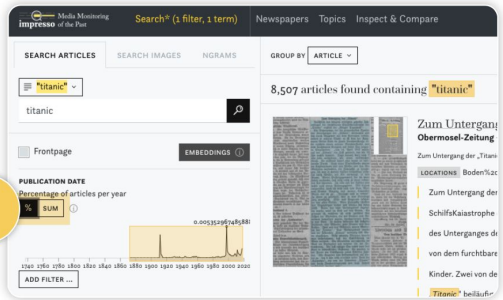
2



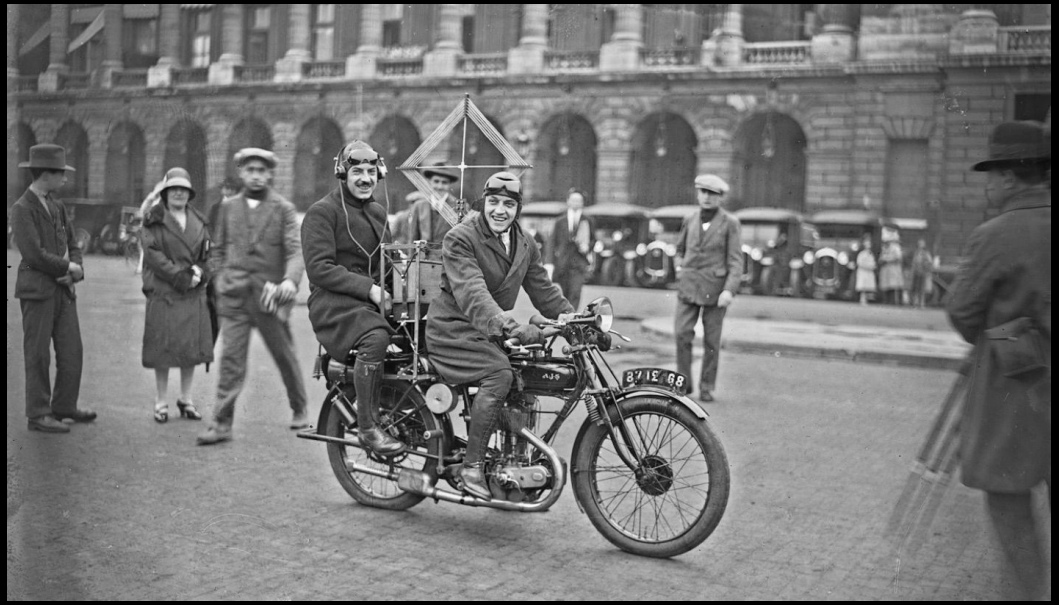
describe what you see:
when are the highest hits for the word "Titanic" and how can you explain it?

change the frequency line from SUM to %
what changes?

3



https://impresso-project.ch/assets/impresso-challenges-1.2.3.pdf

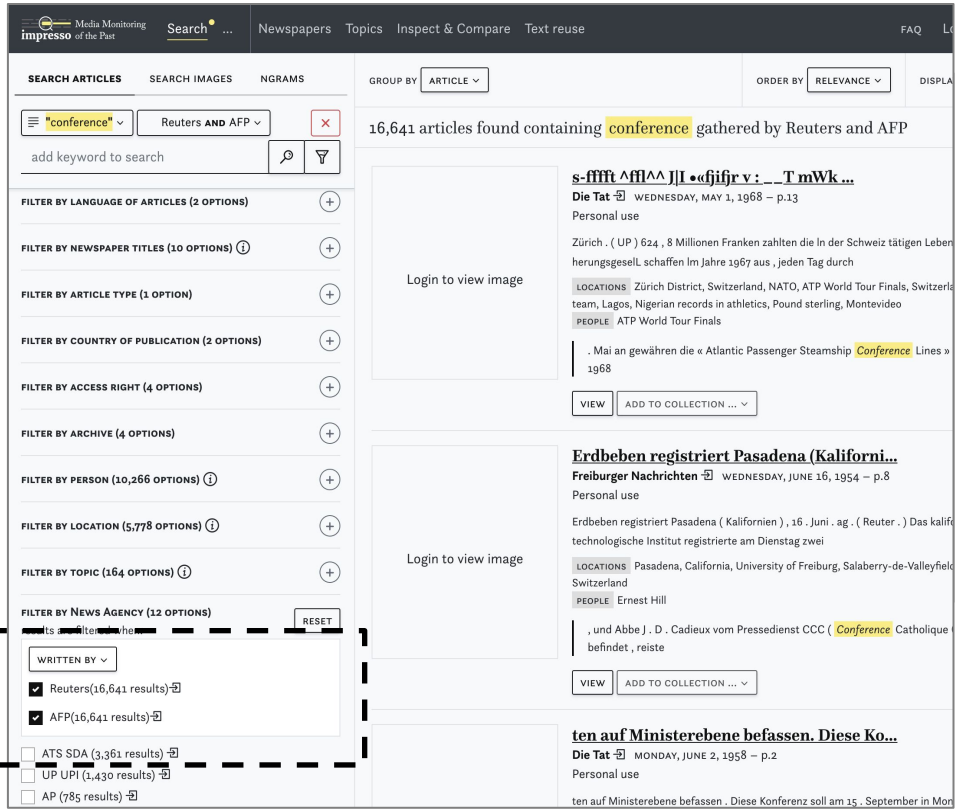


Ausblick: Was kommt als Nächstes?

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.



Erwähnungen von Presseagenturen

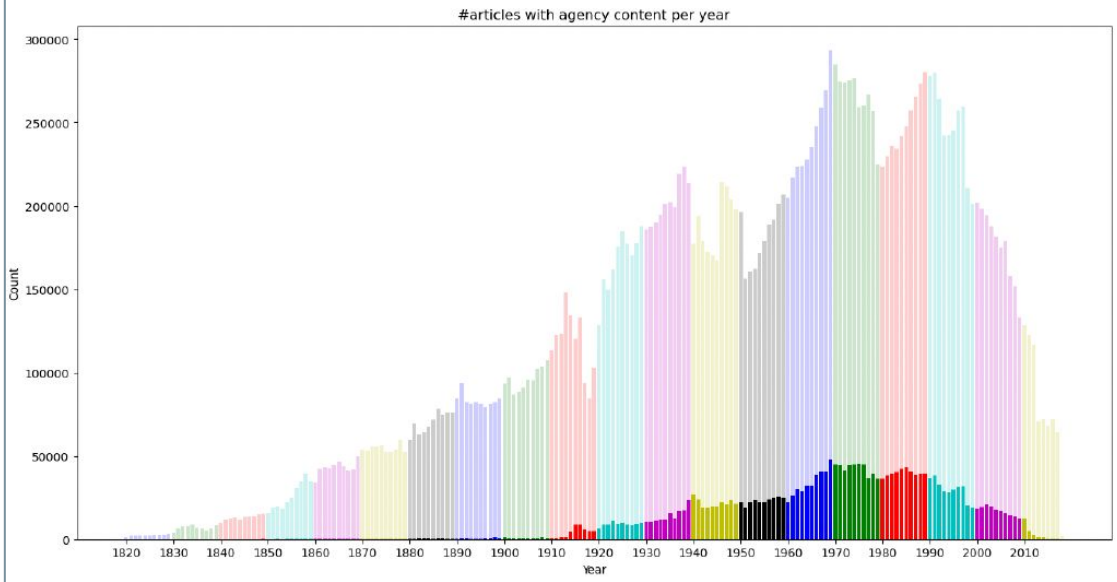


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

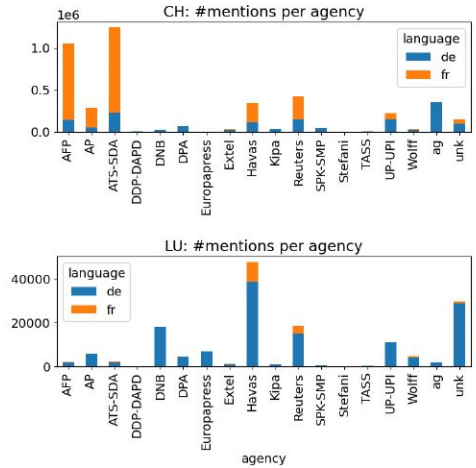


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

..in der schweizerischen Presse…

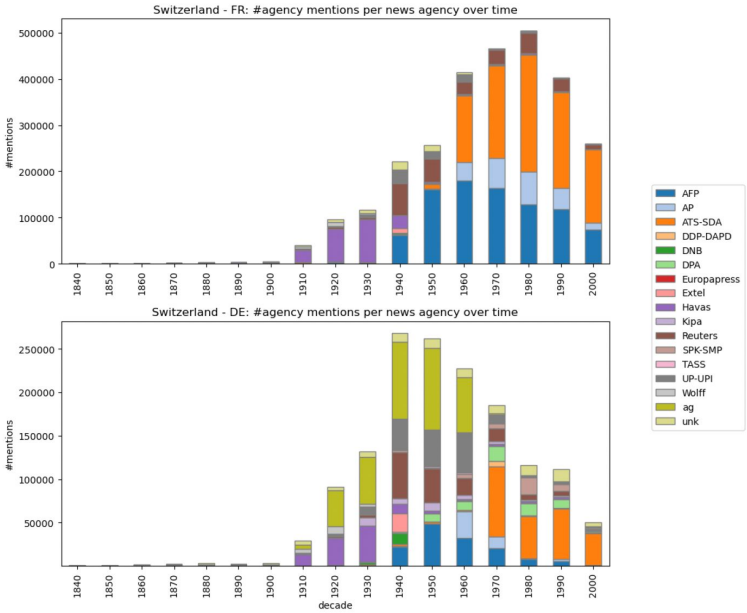


FIGURE 5.8
News agency mentions in Swiss newspapers over time, split by language (French above, German below).

…in der luxemburgischen Presse (1940-1944)

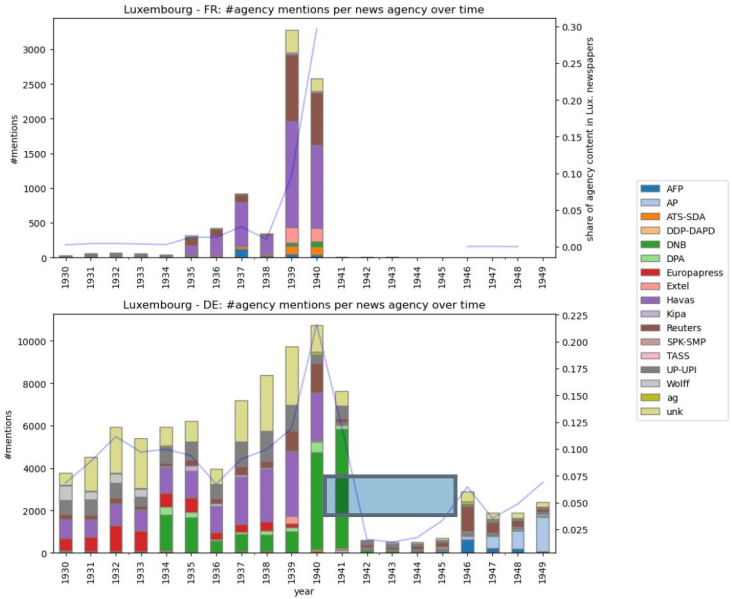
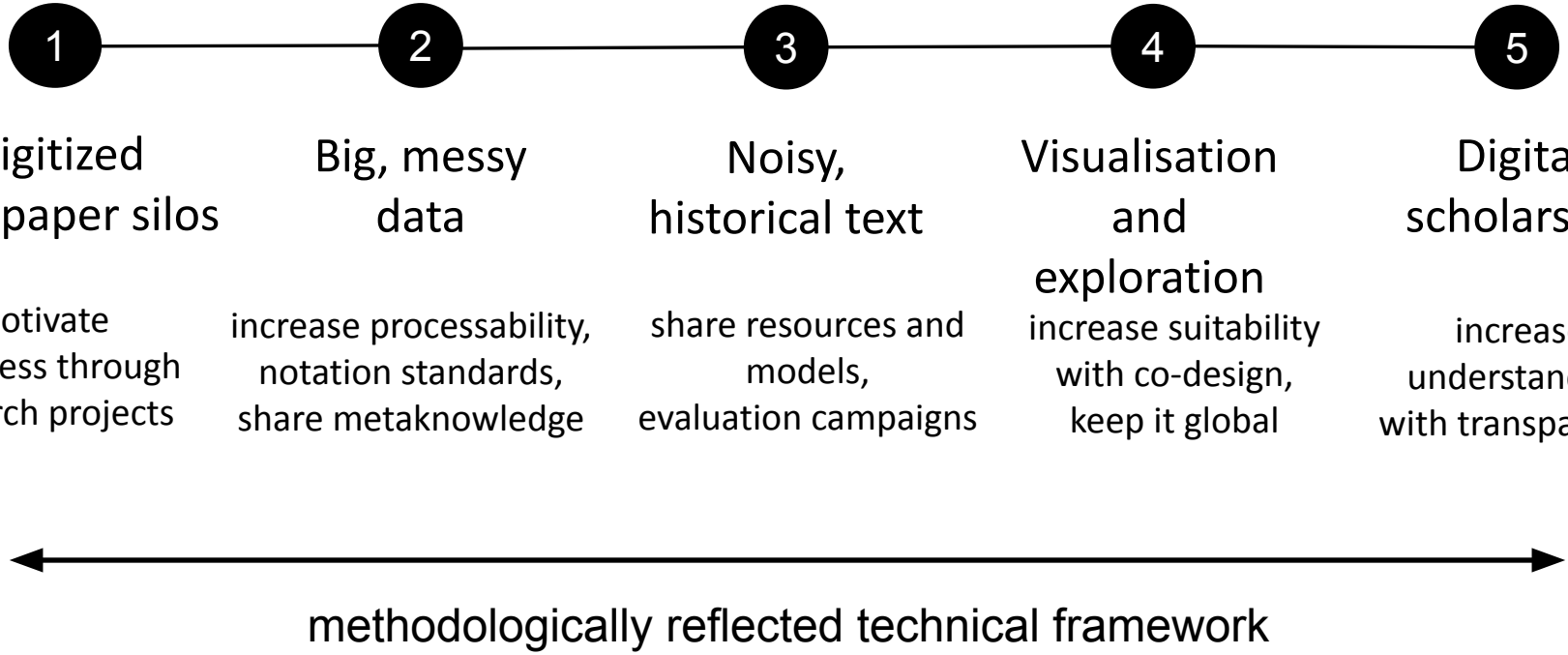


FIGURE 5.15

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?



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.

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

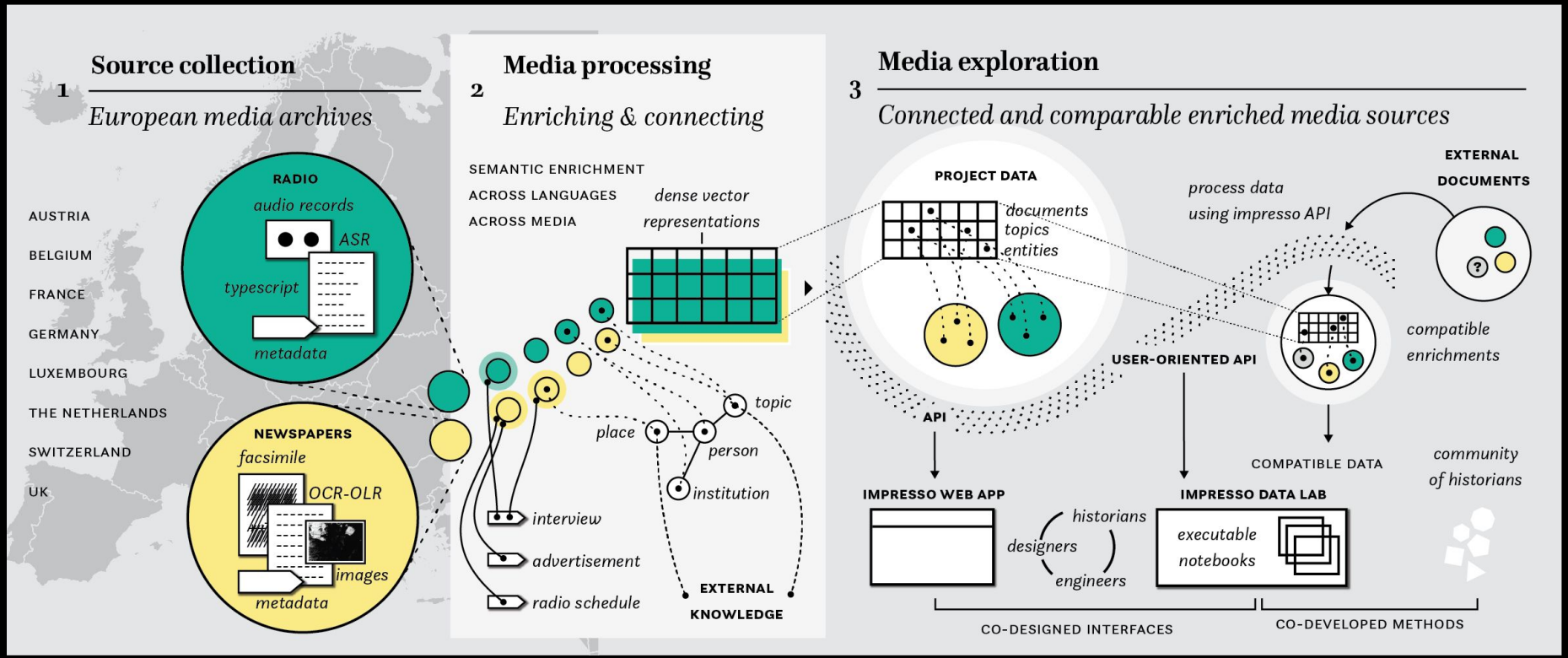
Le Temps
Neue Zürcher Zeitung

Audiovisual heritage institutions and archives (holding digitised radio collections)

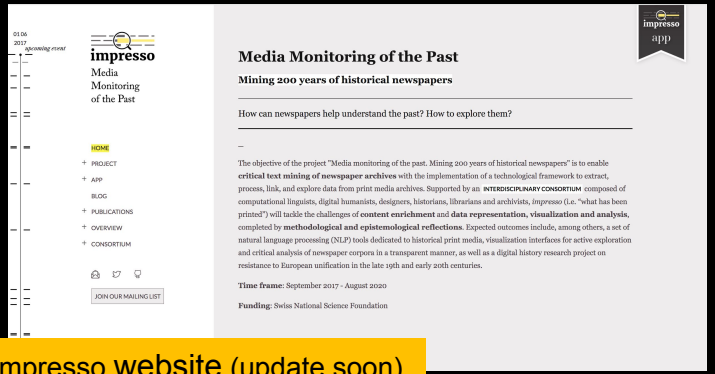
Radio Television Suisse (RTS)
Österreichischer Rundfunk, ORF
British Broadcasting Corporation (BBC)
DeutschlandRadio
Institut National de l'Audiovisuel, INA
Nederlands Instituut voor Beeld en Geluid, NISV

Research Networks

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

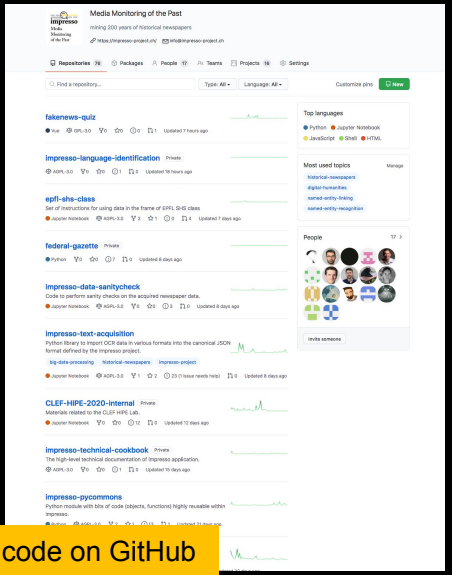

1
Source collection
European media archives
AUSTRIA
BELGIUM
FRANCE
GERMANY
LUXEMBOURG
THE NETHERLANDS
SWITZERLAND
UK
RADIO
audio records
ASR
typescript
metadata
NEWSPAPERS
facsimile
OCR-OLR
images
metadata
2
Media processing
Enriching & connecting
SEMANTIC ENRICHMENT
ACROSS LANGUAGES
ACROSS MEDIA
dense 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

For more information

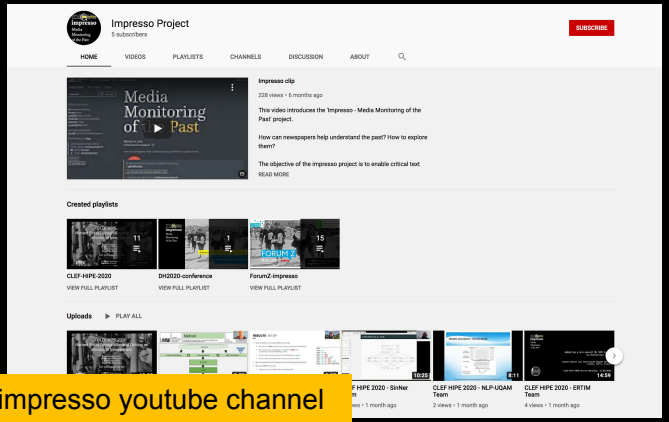


impresso website (update soon)

data on Zenodo



code on GitHub



impresso youtube channel

impresso-project.ch/app