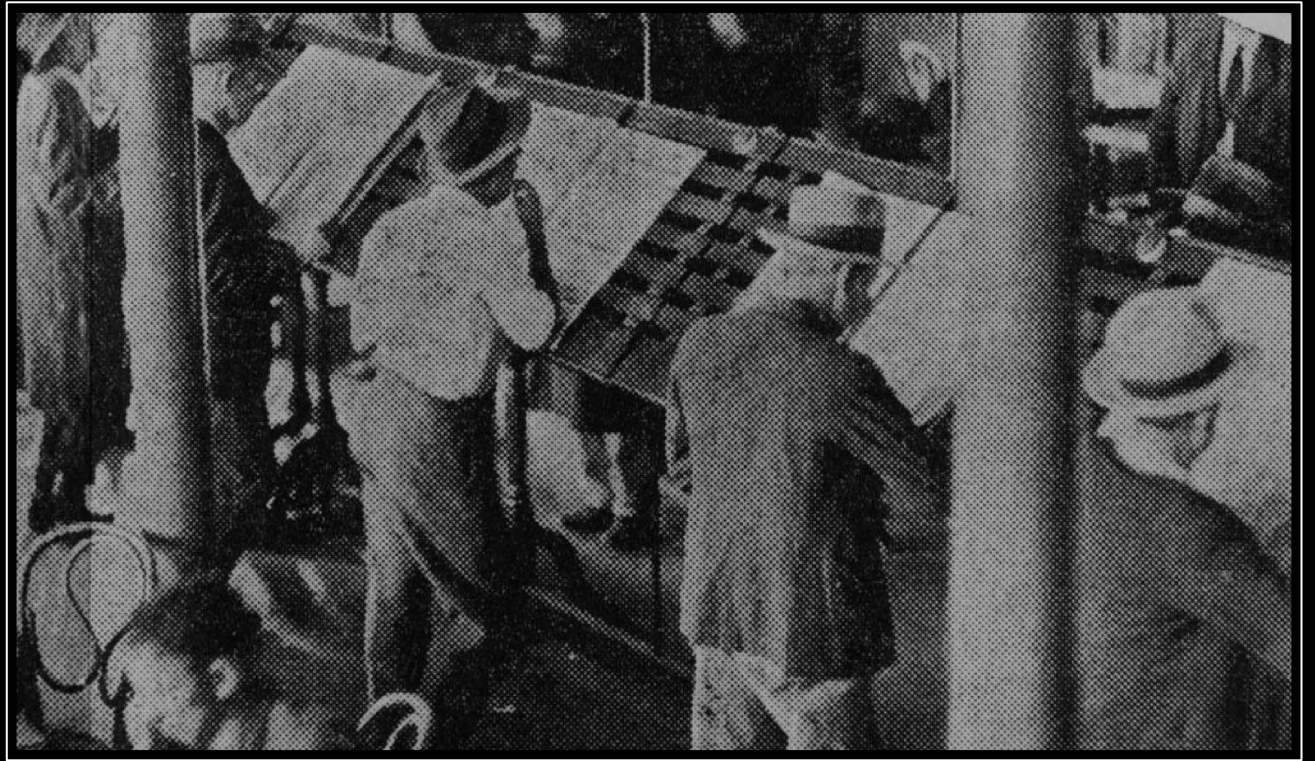


Historische Medien und Maschinelles Lernen

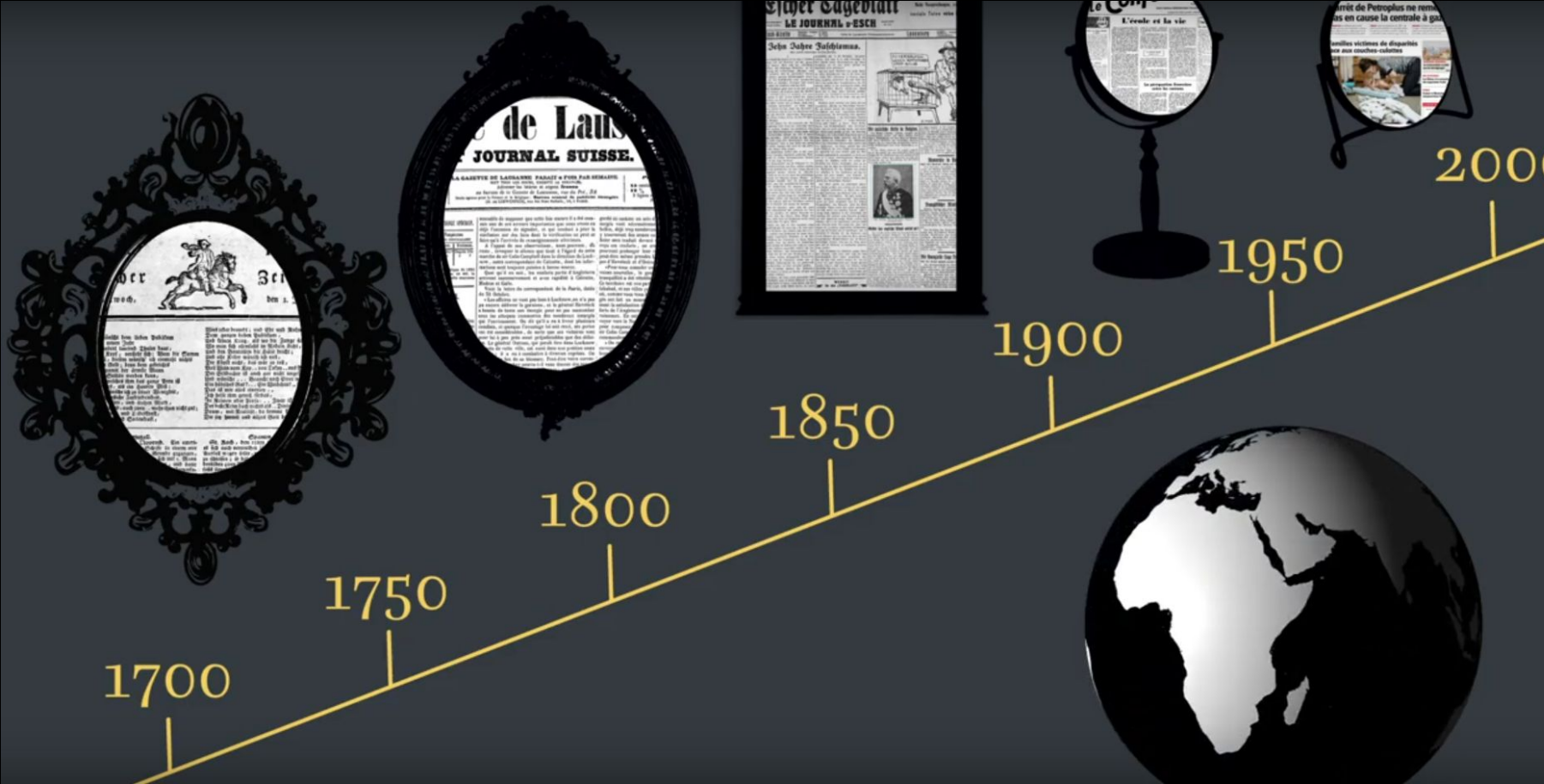
Vom Zusammenspiel von Infrastrukturen, Methoden und Akteuren



Marten Düring & impresso team



1. Ein paar Worte über historische Zeitungen



Erbes Blatt.
Erbliche Wäpfe.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Erbes Blatt.
Die Erbschaft des Erben ist ein Recht, welches dem Erben durch die Erbschaftsbescheinigung zufließt. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will. Er ist verpflichtet, die Erbschaft zu übernehmen, wenn er nicht zuvor erklärt hat, dass er die Erbschaft nicht übernehmen will.

Haute tension sur l'avenir électrique



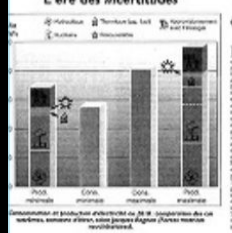
Haute Tension. A nouvelle tour sur le pont, au-dessus de 100 mètres de hauteur.

Le développement de l'énergie électrique est un des problèmes les plus importants de notre époque. La haute tension est la solution la plus économique et la plus efficace pour transporter l'énergie sur de longues distances. Les progrès réalisés dans ce domaine sont impressionnants et ouvrent de nouvelles perspectives pour l'avenir.

Les progrès réalisés dans ce domaine sont impressionnants et ouvrent de nouvelles perspectives pour l'avenir. La haute tension permet de transporter l'énergie sur de longues distances avec des pertes minimales. C'est une véritable révolution technologique qui change notre façon de produire et de consommer l'électricité.

La haute tension est la solution la plus économique et la plus efficace pour transporter l'énergie sur de longues distances. Les progrès réalisés dans ce domaine sont impressionnants et ouvrent de nouvelles perspectives pour l'avenir.

L'ère des incertitudes



Les progrès réalisés dans ce domaine sont impressionnants et ouvrent de nouvelles perspectives pour l'avenir. La haute tension permet de transporter l'énergie sur de longues distances avec des pertes minimales. C'est une véritable révolution technologique qui change notre façon de produire et de consommer l'électricité.

La haute tension est la solution la plus économique et la plus efficace pour transporter l'énergie sur de longues distances. Les progrès réalisés dans ce domaine sont impressionnants et ouvrent de nouvelles perspectives pour l'avenir.

Avis mortuaire.
Madame Alph. Dieleich
Madame Dieleich, née à Luxembourg le 15 Mars 1848, est décédée le 22 Mars 1912, à l'âge de 64 ans.

Volontés.
Le testament de Monsieur Dieleich, décédé le 22 Mars 1912, est enregistré au Greffe de la Cour d'Appel de Luxembourg.

Fell, Balleit- et Magerkochen.
Schneidwaren
Innstrickwaren
Brennholzherk
Bath-Brechholz

Herren- & Confection.
Volks-Kleider-Falle
Für ein oder drei Personen
Café Belge
Confection in gross
Bille-Monde
Bücher-Bibliothek
Buffet-Club
Engländer-Bier
Eisen- & Holz-Handlung

Waggon Kinderwagen!
Gebrüder Hiltig Magdel
Spezialwagen für Kinderwagen
Stollenwagen
3750
Schlehdorn-Wagen
4000
Stollenwagen 'Deenabur'

Automobil
Für 7-10 Liter Motoren
Stöckinger

Felix Fiedler-Neys.
Sargmanufaktur und Möbel-Lager
Sargmanufaktur
Felix Fiedler-Neys

Persil
Für Krankenwäsche
Das selbsttätige
Washmittel.
Das selbsttätige
Washmittel.
Das selbsttätige
Washmittel.

Erste Luxemburger
Katholik-Fabrik und Kattin-Spezialität
14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100

Chaussures Modernes
Pekels Steifen, 28 Pastör, 28
MARA STIEFEL
Überall gute Adresse
10 50
Einkaufspreise
Für Damen

Abt. Polstermöbel
Für Herren- & Damen
Für Herren- & Damen
Für Herren- & Damen

Reinende
Für Herren- & Damen
Für Herren- & Damen
Für Herren- & Damen

W. MICHEL - BRAUN
Kaufmann
Kaufmann
Kaufmann

Automobil
4 Sit., 1000 cc, 16 HP
komplett für 2900 Fr.
Vollständiger Wagen mit Produktion
Kaufmann
Kaufmann
Kaufmann

Wasch- und Bleichmittel
Lessive Kodu
Für Herren- & Damen
Für Herren- & Damen
Für Herren- & Damen

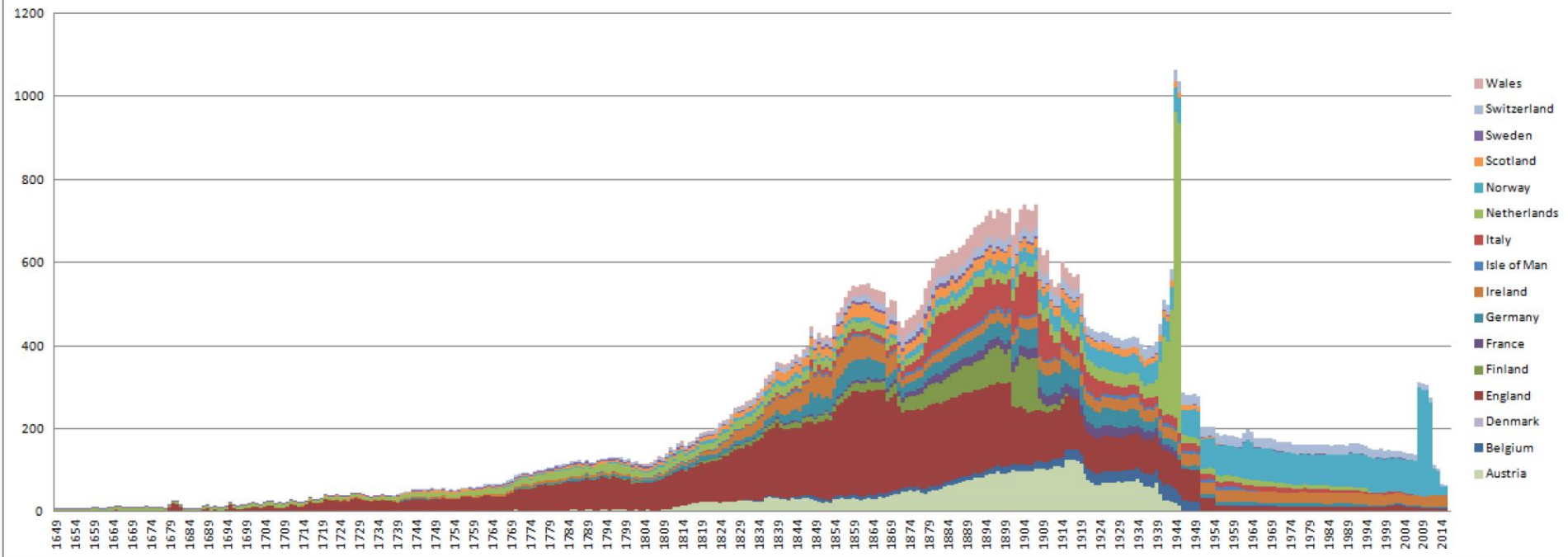
Automobil
4 Sit., 1000 cc, 16 HP
komplett für 2900 Fr.
Vollständiger Wagen mit Produktion
Kaufmann
Kaufmann
Kaufmann

Wasch- und Bleichmittel
Lessive Kodu
Für Herren- & Damen
Für Herren- & Damen
Für Herren- & Damen

Reinende
Für Herren- & Damen
Für Herren- & Damen
Für Herren- & Damen

Digitalisierte historische Zeitungs-Sammlungen

European Newspapers - Titles Per Year by Country



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/>

The screenshot shows the 'Bibliothèque nationale du Luxembourg Open Data' website. At the top, there is a navigation bar with 'HOME', 'DATA', 'TOOLS', and 'API'. Below the navigation bar, a message states: 'datasets contain XML (METS + ALTO), PDF, original TIFF and PNG files for every newspaper issue.' Three data packs are displayed:

- STARTER PACK (250MB):** 5 days of news, 5 newspaper issues, 22 pages, D'Wäschfra (1868). Includes Public Domain, CC0 (See copyright notice). Best for getting started & developing.
- DEV PACK (3GB):** 1 month of news, 26 newspaper issues, 112 pages, Luxemburger Wort (1877). Includes Public Domain, CC0 (See copyright notice). Best for getting started with Big Data.
- SAMPLE PACK (1GB):** 11 different newspaper titles, 1 issue per newspaper, News between 1845 and 1877. Includes Public Domain, CC0 (See copyright notice). Best for testing different newspapers and metadata.

<https://data.bnl.lu>



2. Show the total number of articles per year

In [another notebook](#), I look at different ways of visualising Trove newspaper searches over time. To 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))
```

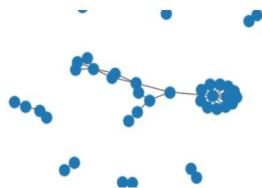
Ok, that's not all that useful. What would be more interesting is to show the total number of details in [this notebook](#) but, in short, we have to loop through the decades from 1800 to 201

These two functions do just that.

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



<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 IIIIF API

This [notebook](#) extracts a dataset from the [Europeana IIIIF API](#). It performs an automatic search, retrieving the manifests from the IIIIF server to create a dataset with the metadata as a CSV file.

[Binder](#)

[Preview](#)

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

Interdisziplinäre Forschungsprojekte



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

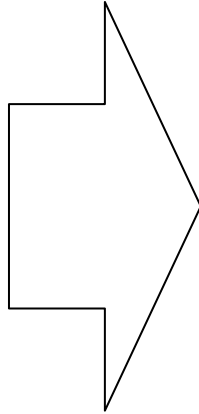
Gender studies

Kulturgeschichte

Sozialwissenschaften

Sozialgeschichte

...



Ziele

Layoutanalyse

Evolution von Genres

“Viralität”

Soziale Normen

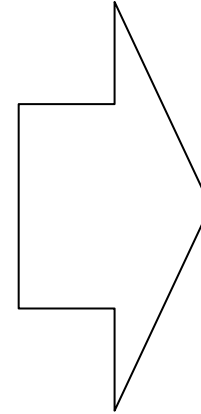
Meinungsbildung

Wissenshorizonte

Biographien

Ernährung

....



Zeitungselemente

Werbung

Kolumnen

Agenturnachrichten

Bilder

Kleinanzeigen

Radioprogramme

Todesanzeigen

....

EPFL



University of Zurich UZH



infoclio.ch

Unil



NZZ

LE TEMPS



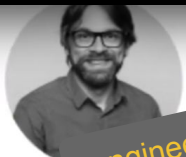
Critical content mining
of 200 years of
historical
newspapers

Impresso

*Inwieweit dienen semantische
Anreicherungen der Analyse und
Exploration historischer Zeitungen?*

Das Team

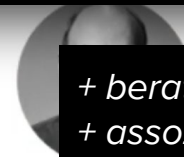
Estelle Bunout
Simon Clematide
Marten Duering
Maud Ehrmann
Andreas Fickers
Daniele Guido
Frédéric Kaplan
Peter Makarov
Matteo Romanello
Gerold Schneider
Paul Schroeder
Benoit Seguin
Phillip Stroëbel
Martin Volk
Thijs van Beek
Lars Wieneke



engineer



web dev



+ beratende HistorikerInnen
+ assoziierte WissenschaftlerInnen



designer /
web dev



NLP/DH



(digital)
historian



NLP/DH



NLP



NLP



(digital)
historian



designer /
web dev



(digital)
historian



web dev



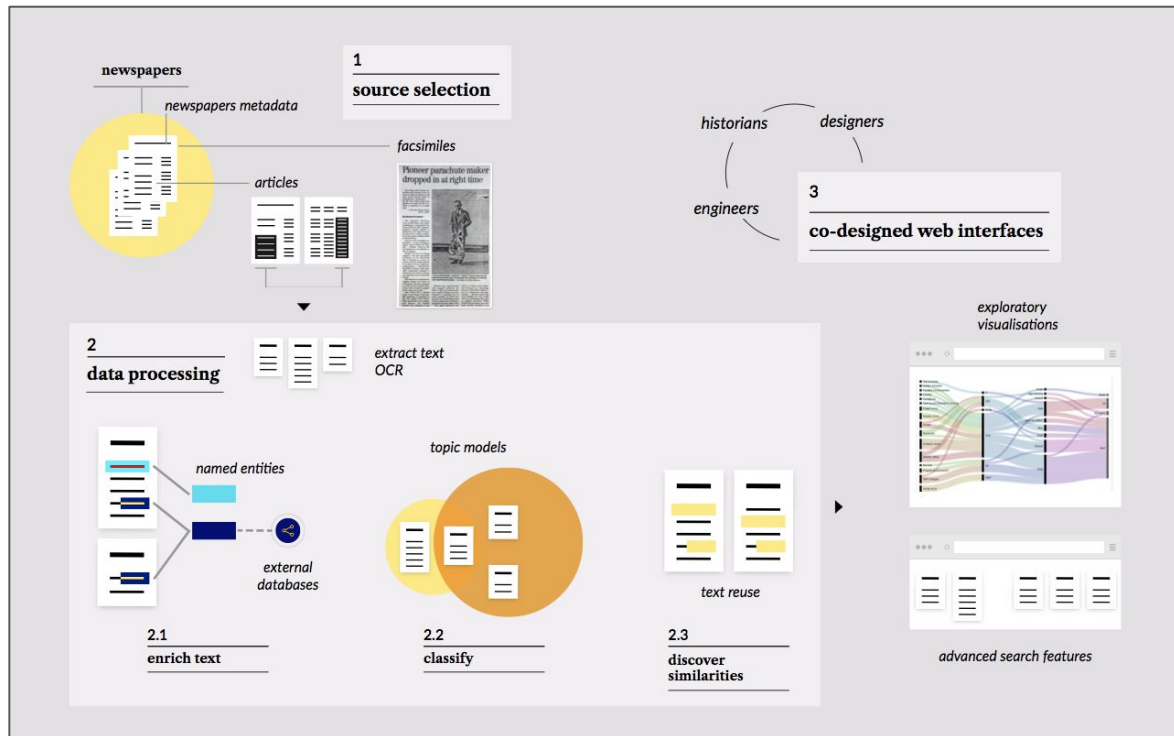
DH/robotics

Ziele und Forschungsfragen

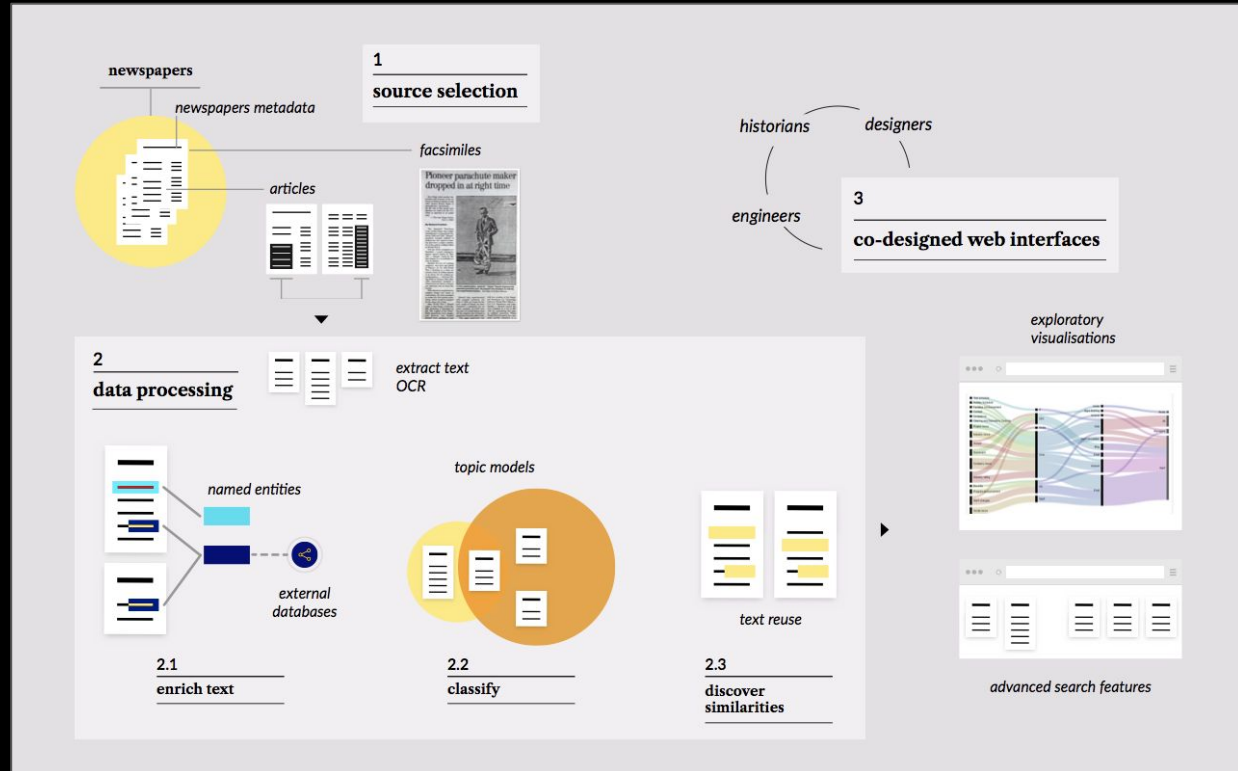
1. Wie passt man NLP-Werkzeuge an historischen Text an?

2. Wie erforscht man große und komplex Datenbestände?

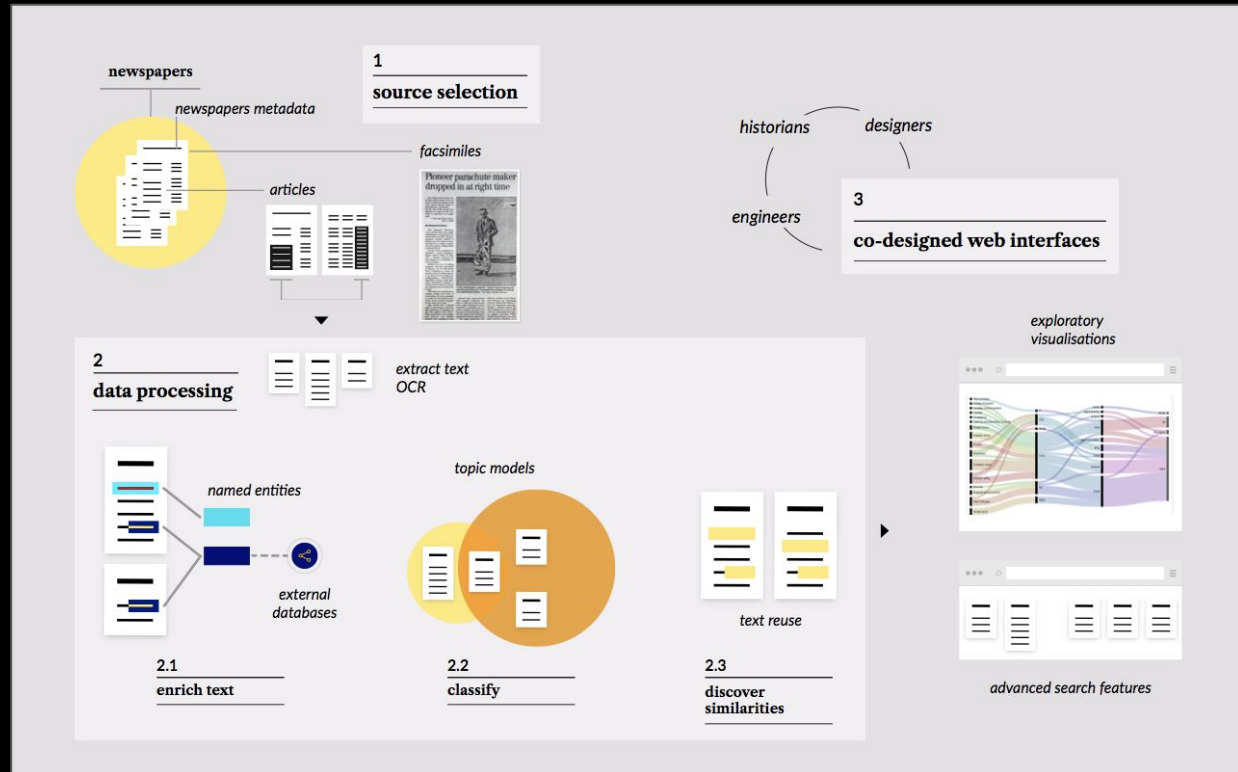
3. Welche Rückwirkungen hat dies auf historische Forschung?



1. Korpus-Aufbau



2. Semantische Anreicherung



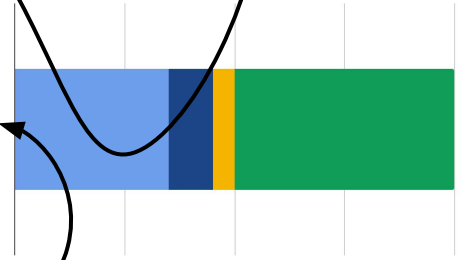
Topic Modeling



Topic distribution



Topic distribution



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 kamp london japanisch
havas amerikanisch feind alliiert
italienisch

zimmer lage vermieten
verkaufen villa hotel schön zürich
haus pension komfort preis see
telefon 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

lia crise britaaiiilque.

Gazette de Lausanne 📄 MONDAY, NOVEMBER 28, 1910 – P.2

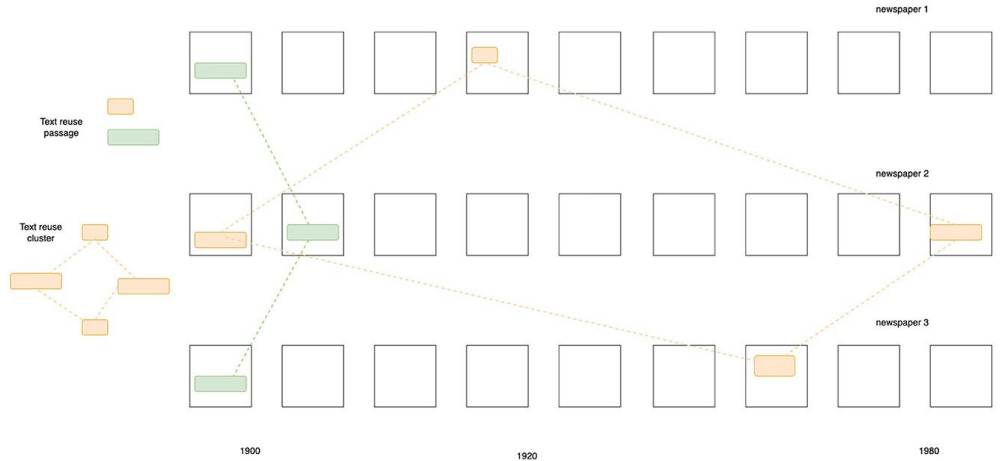
M. Winston Churchill, **ministre du commerce**, prononçait ven- dredi soir à Bradford a été coupé de plu- sieurs interruptions de suffragistes et de suffragettes qui ont été expulsés à tour de rôle. Comme l'orateur rentrait de Bradford à Londres, il fut attaqué dans le train par un individu, qui a essayé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



Ngrams

NGRAMS VIEWER

14,881 mentions of "titanic"; 2,299 mentions of "fukushima"; 12,664 mentions of "atomenergie" in 17,934 articles

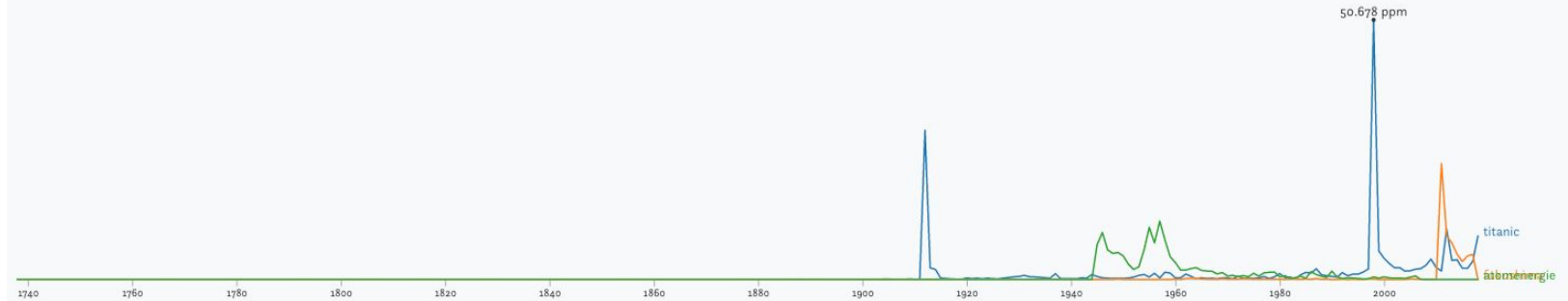
SEE ARTICLES

Enter unigram ⓘ

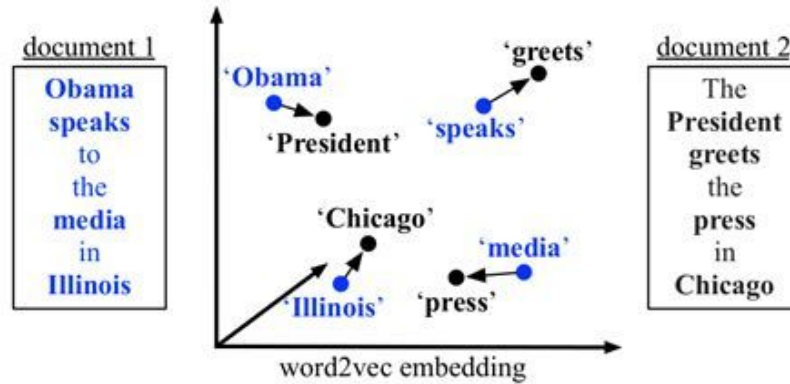
titanic x fukushima x atomenergie x

ADD SIMILAR ▾

YEARLY UNIGRAM MENTIONS (PER MILLION)



Word embeddings



Sprach-Erkennung

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

- French (32,770,496 results) ↗
- German (7,647,324 results) ↗
- English (65,207 results) ↗
- Luxembourgish (41,296 results) ↗

Inhaltstypen-Erkennung

A LOUER

24 juin 1933

100 m² à louer

100 m² à louer

100 m² à louer

Apprentis

100 m² à louer

100 m² à louer

100 m² à louer

AVIS DIVERS

100 m² à louer

100 m² à louer

100 m² à louer

PROBANDIER

100 m² à louer

100 m² à louer

100 m² à louer

CAUTIONNEUR

100 m² à louer

100 m² à louer

100 m² à louer

SWANDER

100 m² à louer

100 m² à louer

100 m² à louer

ZWIRBACHS HYGIENIQUES

100 m² à louer

100 m² à louer

100 m² à louer

SOCIÉTÉ DE MUSIQUE

100 m² à louer

100 m² à louer

100 m² à louer

QUATOUR DE VIENNE

100 m² à louer

100 m² à louer

100 m² à louer

SOIRÉE STRAUSS

100 m² à louer

100 m² à louer

100 m² à louer

Le docteur Mulliger

100 m² à louer

100 m² à louer

100 m² à louer

CORDONNERIE ROMANDE

100 m² à louer

100 m² à louer

100 m² à louer

VAL-DE-TRAVERS

BUTTES

Conseil général

(Corr.) Le Conseil général, assemblé sous la présidence de M. J. Jeanne, 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

RADIO

Radio Luxembourg Expérimental 1110 m - 200 kw.
Séjour Hollandaat: Vendredi, 23 avril 1933.
M. Concert de musique variée enregistrée: Die

Désignation des denrées		Moyens		Prix moyens					
		Net	Gross	des denrées vendues sur les marchés de					
Quantité	Unité	Net	Gross	Blé	Seigle	Orge	Avoine	Mais	Haricots
1000 kg	1000 kg	15 10	14 30	15	15 50	18 88	15 30	16 40	15 28
1000 kg	1000 kg	13 80	13 00	13 70	14 25	18 41	18 40	14 75	14 20
1000 kg	1000 kg	12 24	12	12	11	12	12	12	12 06
1000 kg	1000 kg	9 50	12	11	11	11	11	9 25	10 56
1000 kg	1000 kg	15 10	14 30	15	15 50	18 88	15 30	16 40	15 28
1000 kg	1000 kg	13 80	13 00	13 70	14 25	18 41	18 40	14 75	14 20
1000 kg	1000 kg	12 24	12	12	11	12	12	12	12 06
1000 kg	1000 kg	9 50	12	11	11	11	11	9 25	10 56

Monsieur Pierre LINSTER
vigueur

Président du Comité de la fabrique d'algues décerné de la médaille en bronze de l'Ordre de la Couronne de Danemark

leur bien-être et respecté, pour, beau-père, grand-père, beau-frère et oncle, récemment décédé à Bissen, le 27 avril, à 1 heure de matin, après une courte maladie, à l'âge de 80 ans, mari des Secours de notre Mère la sainte Église.

Enterré au cimetière de notre paroisse à Bissen, le samedi 29 avril, à 10 heures de matin.

Rosen, Lötzen, Tübingen, le 27 avril 1933.
Cet avis tient lieu de lettre de faire part.

relations qu'il entretenait avec les familles Linster, Glanzen, Linden, Hitz et les familles apparentées et la profonde douleur de leur part de la perte irréparable qu'ils viennent d'éprouver en la personne de

Monsieur Jean SCHANEN
leur bien-être et respecté, pour, beau-père, grand-père, beau-frère et oncle, récemment décédé à Bissen, le 27 avril, à 1 heure de matin, après une courte maladie, à l'âge de 80 ans, mari des Secours de notre Mère la sainte Église.

Enterré au cimetière de notre paroisse à Bissen, le samedi 29 avril, à 10 heures de matin.

Rosen, Lötzen, Tübingen, le 27 avril 1933.
Cet avis tient lieu de lettre de faire part.

AVIS MORTUAIRE.

Mme Jean Schanen, née Catherine Neirke; M. et Mme Pierre Schanen-Göttinger; M. et Mme; M. et Mme François Schanen-Kolisch et ses enfants; M. et Mme Hans-Frederik Schanen-Brunn et leurs enfants; M. et Mme Henri Schanen-Schönen et leur fils; M. et Mme Hans Bourne-Schönen et leur fils; M. et Mme Schanen les familles Schanen, Jansen, Heister, Staudt et les familles apparentées et la profonde douleur de leur part de la perte irréparable qu'ils viennent d'éprouver en la personne de

Monsieur Jean SCHANEN
leur bien-être et respecté, pour, beau-père, grand-père, beau-frère et oncle, récemment décédé à Bissen, le 27 avril, à 1 heure de matin, après une courte maladie, à l'âge de 80 ans, mari des Secours de notre Mère la sainte Église.

Enterré au cimetière de notre paroisse à Bissen, le samedi 29 avril, à 10 heures de matin.

Rosen, Lötzen, Tübingen, le 27 avril 1933.
Cet avis tient lieu de lettre de faire part.

Named entities

FILTER BY LOCATION (60,560 OPTIONS) ⓘ

- Lausanne (2,801,221 results) ↗
- Suisse, Moselle (2,473,678 results) ↗
- Switzerland (2,372,038 results) ↗
- Fribourg (2,150,280 results) ↗
- Paris (2,107,217 results) ↗
- France (2,057,462 results) ↗
- Gare de Cornavin (1,769,820 results) ↗
- Lake Neuchâtel (1,725,448 results) ↗
- La Chaux-de-Fonds (1,352,249 results) ↗
- Zürich (1,094,856 results) ↗

MORE... (60,550 MORE OPTIONS)

☰ "marburg" ▾ ✕

add keyword to search 🔍 

FILTER BY PERSON (7,444 OPTIONS) ⓘ

- Gottfried Keller (44 results) ↗
- Huldrych Zwingli (36 results) ↗
- Thomas Mann (36 results) ↗
- David Lloyd George (35 results) ↗
- Richard Wagner (34 results) ↗
- Richard Strauss (33 results) ↗
- Adolf Hitler (29 results) ↗
- Emil Adolf von Behring (29 results) ↗
- Gerhart Hauptmann (28 results) ↗
- Karl Barth (27 results) ↗

MORE... (7,434 MORE OPTIONS)

Angereicherte Daten

Topic modeling

Ngrams

Text reuse

Word embeddings

Language detection

Content type detection

Named entities

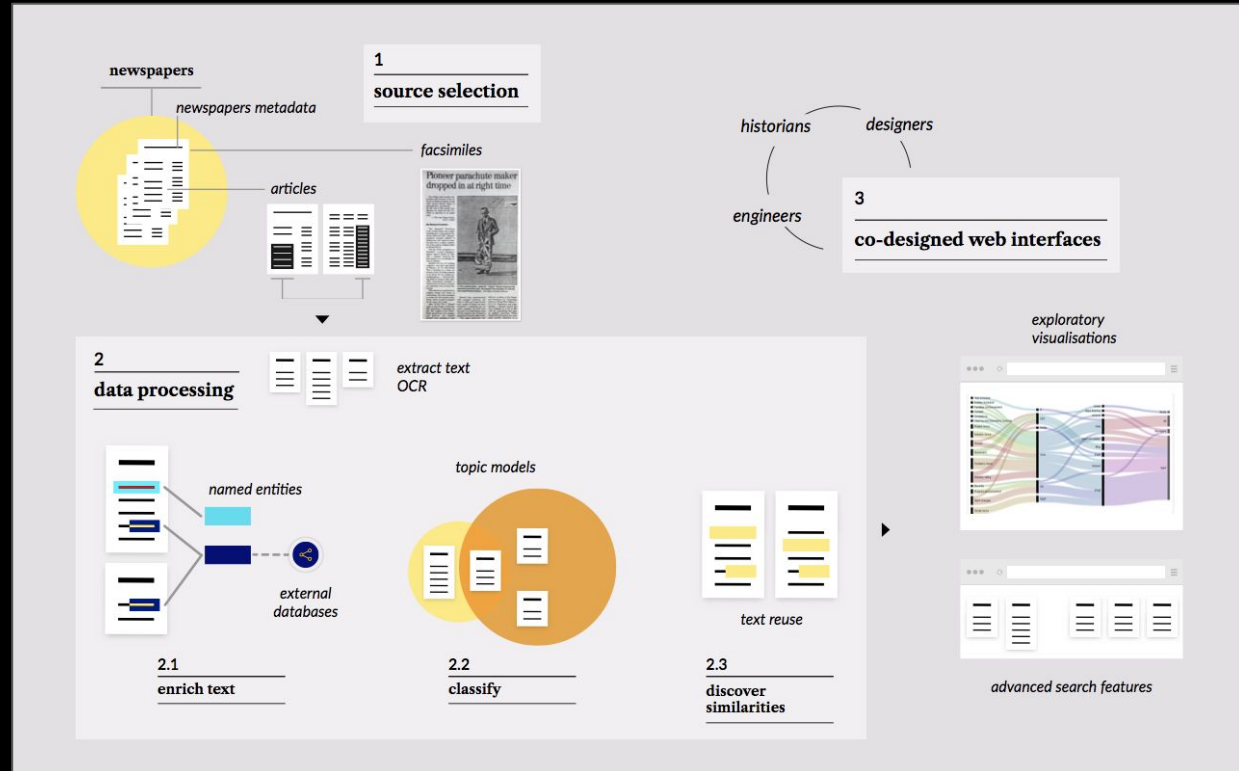
OCR quality assessment

Image similarity



The screenshot displays the Impresso website interface. At the top, there is a navigation bar with links for 'Search', 'Newspapers', 'Topics', 'Inspect & Compare', 'Text reuse', and 'Collections'. The main header area features the title 'Media Monitoring of the Past' in a large, white serif font, with 'Past' highlighted in yellow. Below the title, it says 'Mining 200 years of historical newspapers'. A sidebar on the left provides 'IMPRESSO DATA RUNDOWN' with statistics: 76 newspapers collected, 600,919 issues, 5,429,656 pages scanned, 47,738,468 content items identified, and 3,462,799 images. Below this is a search bar with the text 'search for ...'. At the bottom of the page, there is a section titled 'Impresso Challenges' with the subtitle 'How to explore the newspapers with persons or locations?'. A search bar at the bottom left shows a search for 'Robert Schumann'.

2. Co-design der impresso App



Workshops



Fallstudien

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



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

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 "séleénite" 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)
 - "alien" often mistaken for "(Woody) alien", "alien", "aliéné" and words with -alien/allien (312 585 results on Retronews.fr)
 - "alien": English words with numerous irrelevant meanings

3

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

Bridging the fields: Research scenarios

How do newspapers reflect 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

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)

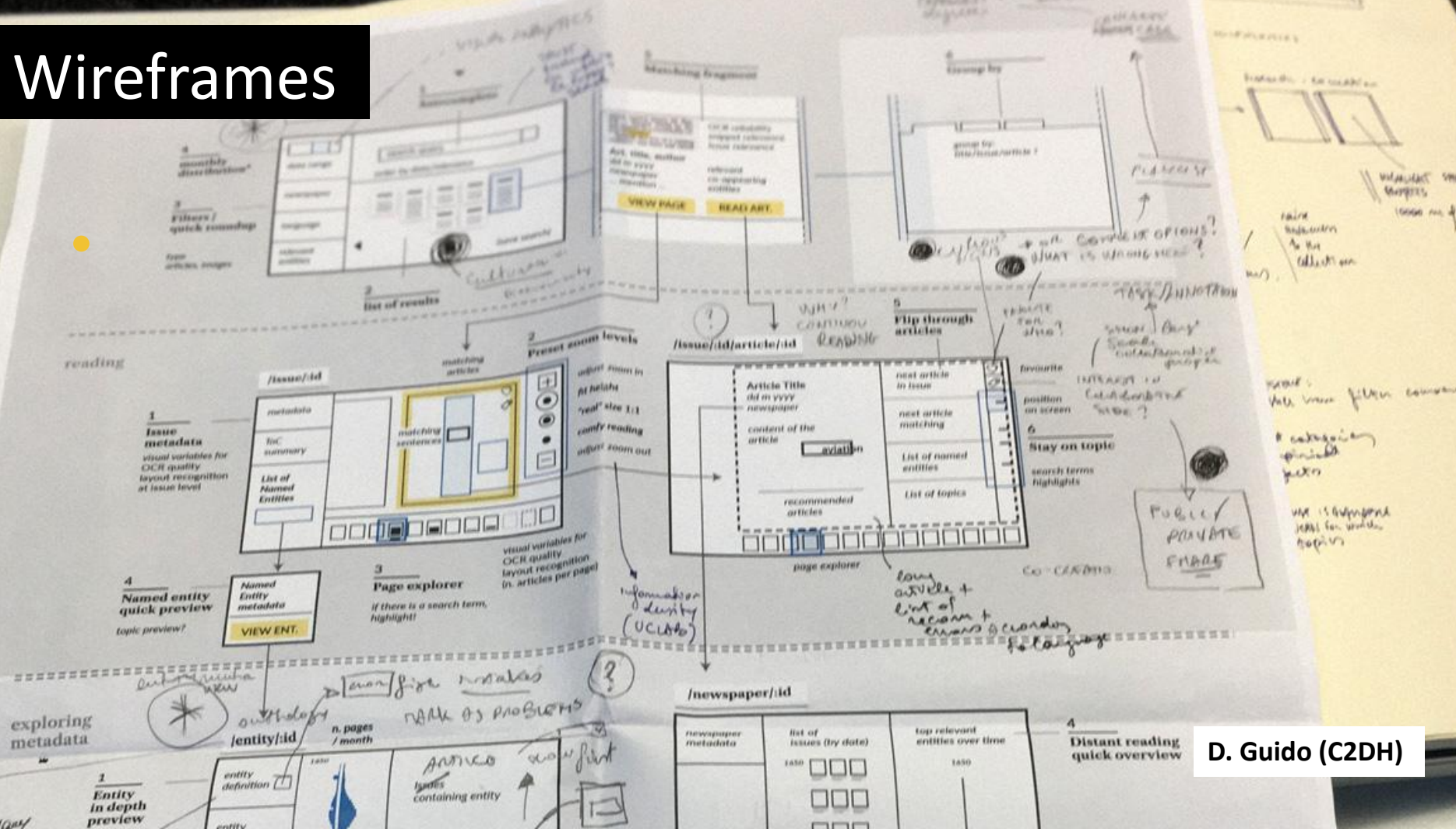
- Help me find the lengths of sections (German: *Rubriken*, French: *rubriques*) in both newspapers and the change of these lengths in time (Handout: 1.1.1)
- Help me find moments of emergence of sections in newspapers (Handout: 1.1.2)
- Help me find/collect words and collocations related to sections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

4

5

6

Wireframes



S search for articles C collecting
 R reading contents O organising
 M exploring metadata 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

The screenshot shows a search interface with the following sections:

- Navigation:** "explore" dropdown, "search" button.
- Filters:** "YOUR SEARCH" and "PAST SEARCHES" tabs.
- Search Input:** "FULL TEXT SEARCH" with a "REFINE ..." button and a search box containing "... type a text or a date or to start ...".
- Timeline:** "PUBLISHED IN (DATE RANGE) TIMELINE OF" with a chart showing results from 1745 to 1980. A note says "select a date range to show airticles published".
- Results:** "NEWSPAPER TITLES" section listing:
 - 53625 **La Gazette de Lausanne** - quotidien suisse de langue française édité à Lausanne
 - 43864 **Le Journal de Genève** - quotidien suisse qui a paru du 1826 au 28 février 1998
- Article Type:** 112 **partisans**, 50 **satirique**.
- Top Named Entities:** 103 **Napoleon**, 50 **Zurich, Suisse** location.
- Language:** 112 **French**.

Strength & opportunities

code

R.13 rest of the article and image regions side by side

R.14 add tags to the article; and add the article to your collections

R.15 List all the named entities of the article

R.16 Get the list of sources for the fed named entity in this article

R.17 named entity to a collection!

Or.2 selected (active) collection, with quantity of articles and pages

Or.3 default collections like "favorites" cannot be deleted or modified

Or.4 switch type of visualization (list vs graph, cf. Or.9)

Or.5 structured overview of your collection, export citations or metadata

Or.6 list of items included in your collection

Or.7 assign item to more than one collections

Or.8 "Format": newspaper or specific page regions

Or.9 "How did you find this item?" list of search results and link to the original query

Or.10 Explore also by tag (coming)

Or.11 suggest new tag

Or.12 suggest new "description" for the tag

Or.1 list of your "collections" sort by: test created, test modified, alphabetically

Or.11 Visual feedback C.7 of the data dimension chosen in [Or.11]

Or.11 Freely tune the data "dimensions" from the available metadata

component index

R.4 convenient and aesthetically pleasing summary of search - I also tested using the back button in the browser to go back to the previous page and that works as well (as a user, I am more inclined to use the back button rather than a "return to results button")

R.6 marginalia is great feature

Or.10 visualization of data can be very effective

S.1 ~~search bar~~ search bar draws on common practice of searching for something, which people are comfortable with

R.5 this feature is wonderful, saves time by being able to view articles without having to open them

when you select "download as," can you specifically download the marginalia?

albumic diagrams are difficult to parse, even for data viz professionals

perhaps, graph layout

how else can the front page introduce users to the collection? eg. featured collection, visual search? current search is designed for subject matter experts only

overall, iif server loads somewhat slowly

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!

find similar words

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

<input type="text"/>	German	⌵	50	⌵
----------------------	--------	---	----	---

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

find similar words



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

German

Click on one of the following words to update your search

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

Semantische Nachbarn

find similar words



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

German

Click on one of the following words to update your search

fahrrad motorrad velo fahrrades motorfahrrad
auto dreirad fahrrade lieferauto stahlroß
sidecar molorrad zweirad motorrade
leichtmotorrad motorrades lastauto iweirad
rasenmaschine motorvelo cozius personenauto

Synonyme

find similar words



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

French

Click on one of the following words to update your search

nucleaire nucieaire nucleaires nucléaire nueleaire
nucieaires nuoleaire thermonucleaire hydrogene
nuclôaires nuclei atorniques claire reacteur
thermonucleaires reacteurs nucliaire lhydrogene
cleaires meøatones telexuides lenergie

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 POUSSIÈRES 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 POUSSIÈRES
par les appareils portatifs

“ATOM”

NETTOYAGE des Tapis, Meubles, Rideaux, Portières,
Linceuls, Vitres, Fenêtres, Automobiles,
Bibliothèques, Billards, etc.

SIMPLICITÉ ! Maniement à la main !
HYGIÈNE !

Nettoyage par le vide à la portée de tous !
Indispensable dans les Hôtels, Restaurants,
Villas, Ménages, Hôpitaux, Bibliothèques, etc.

Nettoyage avec « ATOM »

DEMANDEZ PROSPECTUS :
F. & C. ZIEGLER, SCHAFFHOUSE

Démonstrations de l'appareil « ATOM » sont faites gratuitement et à domicile sur demande
par M. Fritz BONNET, 7, Quai des Bergues, 7, GENEVE.
S'adresser au magasin A l'Écluse, rue de la Tour-de-Tier.

SUPPRESSION RADICALE DES POUSSIÈRES
par les appareils portatifs,
Nettoyage sans « ATOM » « ATOM »
NETTOYAGE des Tapis, Meuble », Rideaux, Portières,
Literie, Vêtements, Fourrures, Automobiles,
,..-!Bibliothèques, Billards, etc...!
SIMPLICITÉ !, .. HYGIÈNE ! *
Maniement à la main. /
Nettoyage par le vide à la portée 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 Botschaf

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 übermitteln hätten




Codesign: Topic Modeling

Die Tat · Sunday, January 21, 1962

ADD TO COLLECTION ... ▾


FACSIMILE 


TRANSCRIPT 


Plasmaphysik

LOCATIONS [Switzerland](#), [Würenlingen](#), [Mosbach](#), [Baden](#), [Germany](#)

TOPICS

88.6% [DE wasser · betrieB · arbeit · energie · firma](#) 

3.9% [DE don · and · min · sage · uli](#) 

2.2% [DE schweiz · franke · bern · bereich · zukunft](#) 

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 

Plasmaphysik

 [REINHART FROSCH](#)

Aufbau der Materie


Das wichtigste Element aller Substanzen ist
das Atom , obschon es nicht aropos , das heisst

Codesign: Topic Modeling

VHTL-Zeitung · Friday, November 3, 1961


ADD TO COLLECTION ... ▾

FACSIMILE 

TRANSCRIPT 


Der Atommotor im Betrieb

TOPICS

94.2% DE wasser · betrieb · arbeit · energie · firma 

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

Der Atommotor im Betrieb

Die Zeit des Atommotors wird von derjenigen der Dampfmaschine und des Elektromotors so grundverschieden sein, daß die kühnste Phantasie nicht ausreicht, sie zu erfassen, geschweige denn zu gestalten!

Wie wird oder vielmehr wie muß der

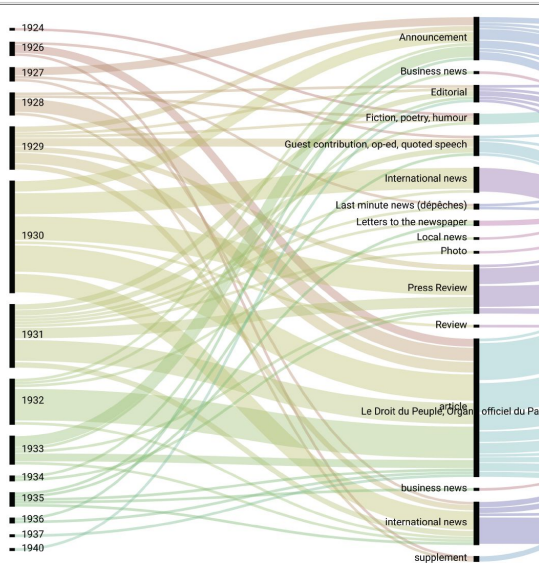
Codesign: Inspect & Compare

DISPLAY ITEMS AS LIST OF ITEMS VISUALIZE AS TIMELINE EXPLORE METADATA

SUMMARY

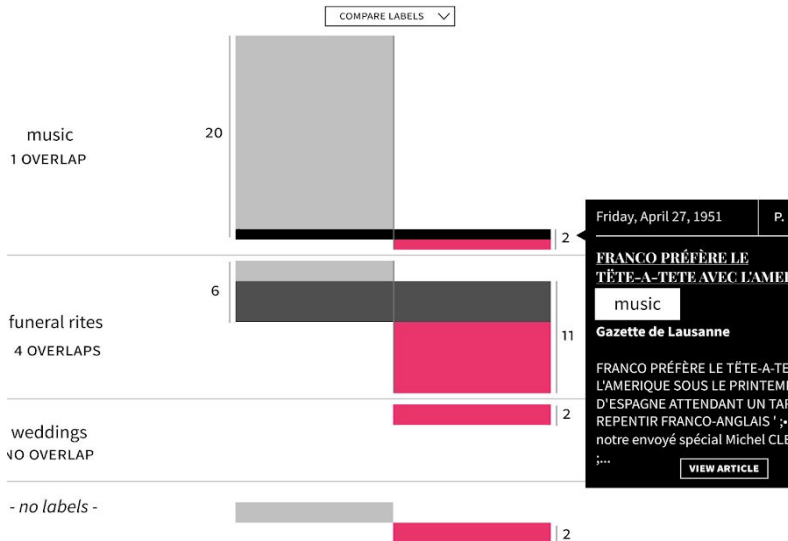
Alluvial diagram of year, article sections, newspaper titles based in the collection Panurope

PANEUROPE
29 JUNE 2018, 08H15 - collection of pan-europe and paneurope articles, manually curated



YEAR OF PUBLICATION	ARTICLE SECTIONS	NEWSPAPER
1 1924	2 Business news	23
5 1926	10 Editorial	23
5 1927	4 Fiction, poetry, humour	20
8 1928	9 Guest contribution, op-ed, quoted ...	15
15 1929	24 International news	14
39 1930	2 Last minute news (dépêches)	11

comparison of labels with the collection **On Catholicism**



LOCATION	ARTICLE SECTIONS	NEWSPAPER TITLES
1 1924	2 Business news	23 Gazette de Lausanne
5 1926	10 Editorial	23 Escher Tageblatt
5 1927	4 Fiction, poetry, humour	20 Feuille d'Avis de Neuchâtel et du Vignoble ...
8 1928	9 Guest contribution, op-ed, quoted ...	15 Journal de Genève
15 1929	24 International news	14 Obermosel-Zeitung

#design ☆

What do you think about intersection pattern @daniele, @Pau

Screen Shot 2020-04-01 at 15:22.09.png

Same with central Y axis

Screen Shot 2020-04-01 at 15:25.09.png

daniele 3:28 PM
fantastic! what if you invert the last two colors? so that the int

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

Add to Basket



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

Add to Basket



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

Add to Basket



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

Add to Basket



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

Customer Rating	★★★★★ (219)	★★★★☆ (338)	★★★★☆ (62)	★★★★☆ (2006)	★★★★☆ (13)
Price	€419.99	€999.00	€129.99	€326.99	€149.99
Sold By	Potensic-EU Official Store	Amazon.de	morlyrctoo	MondoTop	Potensic-EU Official Store

Codesign: Inspect & Compare

Media Monitoring of the Past
Search ...
Newspapers Topics Inspect & Compare Text reuse Collections
FAQ JOBS
Marten Düring STAFF

QUERY *

"titanic" film · cinéma · semaine · jean · joh REFINE ...

search for ...

OPEN IN SEARCH PAGE... (1,638 RESULTS)

INSPECT (A) + (B) COMPARE (A) & (B)

382 results in common

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

OPEN IN SEARCH PAGE... (382 RESULTS)

QUERY *

"titanic" bateau · mer · lac · bord · port REFINE ...

search for ...

OPEN IN SEARCH PAGE... (1,408 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

L'Express	656
Gazette de Lausanne	301
La Liberté	267
L'Impartial	240
Journal de Genève	91
Confédéré	49
La lutte syndicale	13
Le Peuple, La Sentinelle	10
d'Letzeburger Land	6
L'indépendance luxembourgeoise	4
D'Unio'n	1

COUNTRY

Switzerland	1,627
Luxembourg	11

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

L'Express	180
L'Impartial	65
Gazette de Lausanne	49
La Liberté	43
Confédéré	22
Journal de Genève	14
Le Peuple, La Sentinelle	3
L'indépendance luxembourgeoise	3
d'Letzeburger Land	2
La lutte syndicale	1

COUNTRY

Switzerland	377
Luxembourg	5

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

L'Express	533
La Liberté	227
Gazette de Lausanne	209
L'Impartial	193
Journal de Genève	82
L'indépendance luxembourgeoise	71
Confédéré	63
L'Essor	12
La lutte syndicale	8
Le Peuple, La Sentinelle	7
d'Letzeburger Land	2
Freiburger Nachrichten	1

COUNTRY

Switzerland	1,135
Luxembourg	

Codesign: Inspect & Compare

QUERY * COLLECTION

"arnhem OR arnhem(3 more)" match-équipe-figure-club-sais- REFIN

search for ... ADD FILTER...

OPEN IN SEARCH PAGE... (5,585 RESULTS)

181 results in common

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

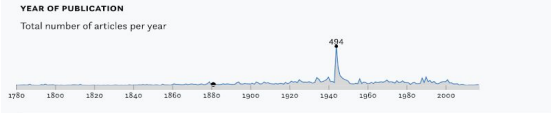
OPEN IN SEARCH PAGE... (181 RESULTS)

QUERY * COLLECTION

"arnhem OR arnhem(3 more)" "match" REFIN

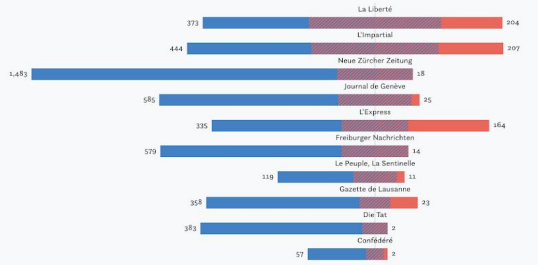
search for ... ADD FILTER...

OPEN IN SEARCH PAGE... (670 RESULTS)



SCALE: SQUARE ROOT SORT BY ABSOLUTE INTERSECTION

NEWSPAPER (10 OPTIONS)



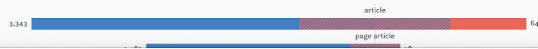
COUNTRY (1 OPTION)

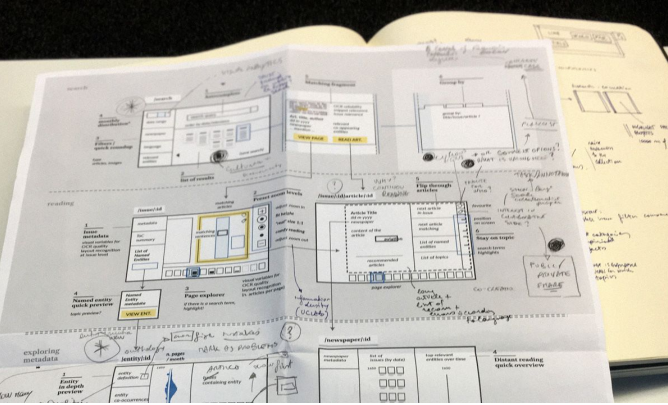


LANGUAGE (2 OPTIONS)



TYPE (3 OPTIONS)





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?



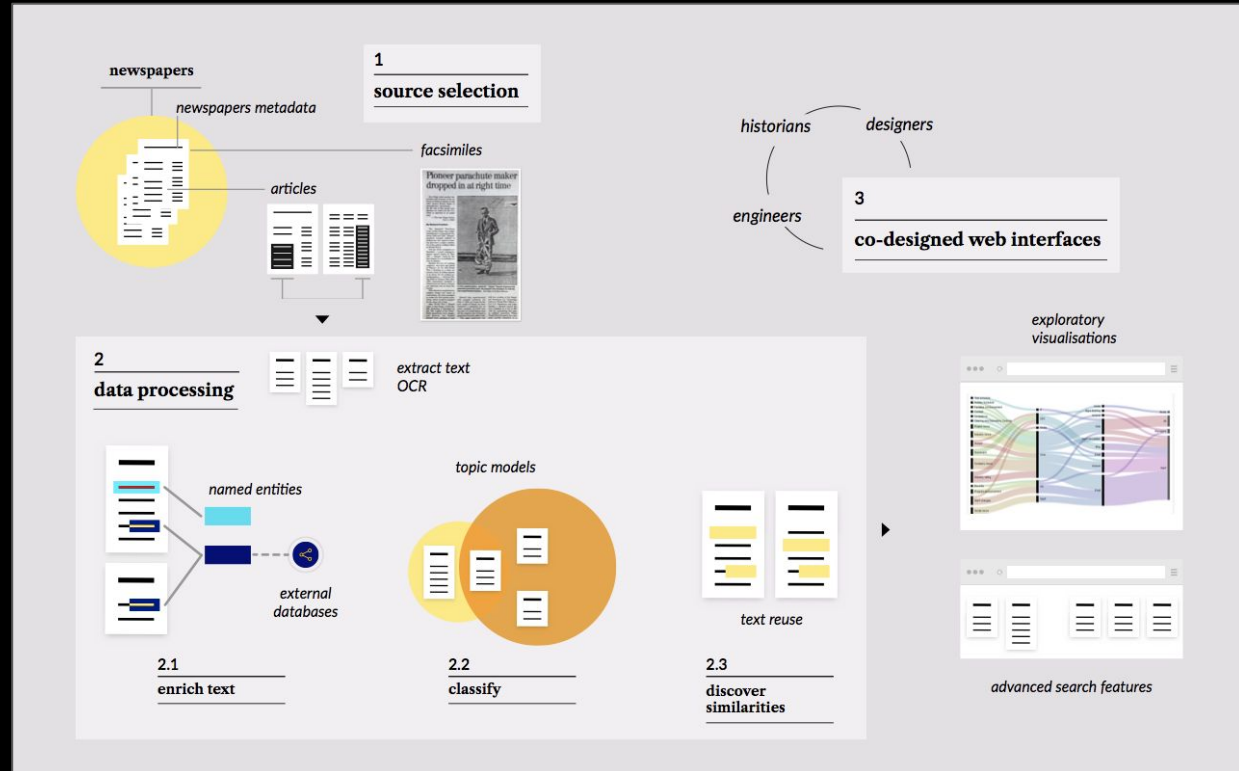
For legal reasons not all content is available.

To gain access to the **full impresso corpus** please [register](#) and sign our Non-Disclosure-Agreement.

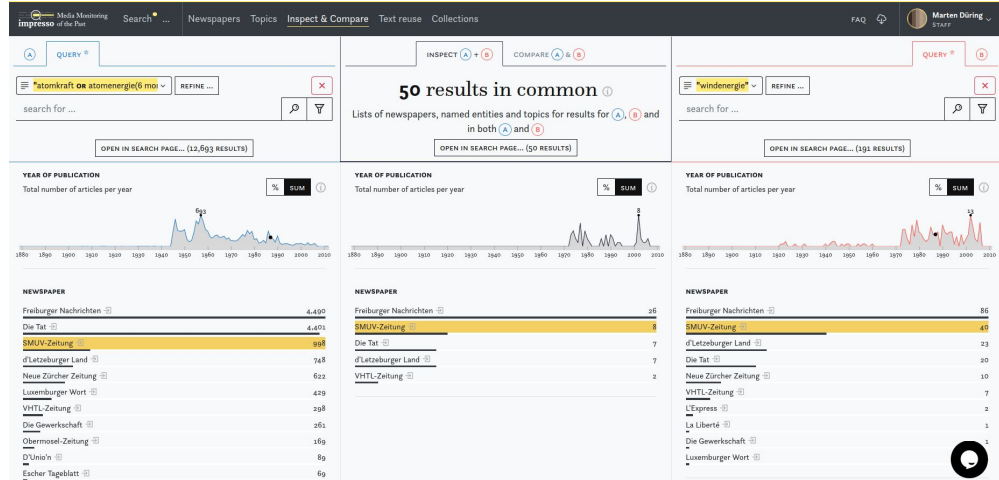
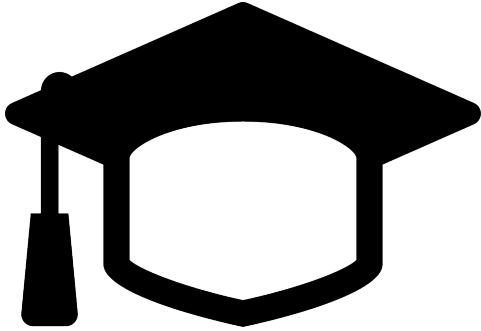
 [DOWNLOAD NON-DISCLOSURE-AGREEMENT FORM](#)

... 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**

NGRAMS VIEWER

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

SEE ARTICLES

* 2 search filters can't be applied. ⓘ

Enter unigram ⓘ

greve x streik x

ADD SIMILAR ▾

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

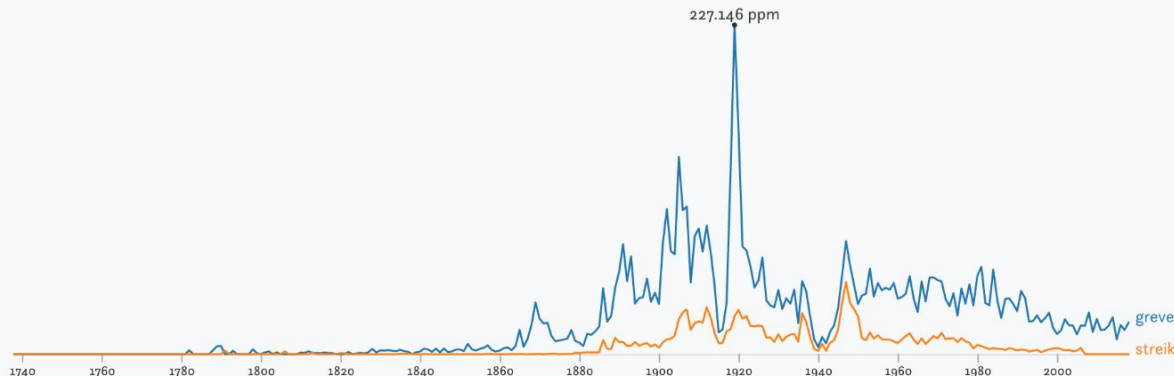
- French (250,490 results) ↗
- German (70,953 results) ↗
- Luxembourgish (628 results) ↗
- English (3 results) ↗

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) ↗

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 +



Fortgeschritten: Digitised newspapers as new artefacts

 **PARTHENOS** HOME TRAINING MODULES FOR TRAINERS FOR LEARN

DIGITISED NEWSPAPERS AS NEW ARTEFACTS

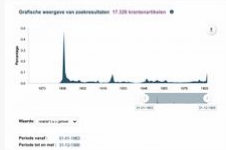

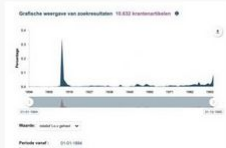
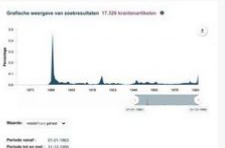
How are the analogue newspapers reconstructed in digital form?

Let's start at the beginning: what exactly are we accessing when we work with digitised newspapers, and how do we access it?

As Pelle Snickars explains, the process of digitisation creates many new layers of information on top of the original source, and the output of digitisation requires a dedicated interface to access it. Digitisation transforms the original source and opens it up for more uses and different means of access. Capturing the analogue source with a scanned image and processing the image with text recognition and layout analysis creates a new source format and changes the potential interaction with it. But raw digital output relies on many technical and institutional constraints, mainly copyright issues, and is not usable as such; many further steps are needed before it can be accessed by researchers. Users must be aware of the context in which it was produced, and also of the fact that the field is constantly changing, both in terms of technical improvement in the digitisation itself but also in the development of interfaces giving access to them. It is important not to lose sight of the fact that collections are being constantly enriched and the technical quality of newspaper digitisation is undergoing continuous improvement. It seems that the real challenge facing holders of digitised newspaper collections is how to provide users with facilitated access to a growing quantity of sources via a suitable interface. Behind seemingly stable institutional search interfaces, the digital newspapers collections are undergoing important and continued changes.

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

Comparison of Absolute and Relative Frequency

	Delpher Timeline on Titanic	Delpher Timeline on Eiffel Tower
URL	https://www.delpher.nl/nl/kranten/ngram?query=titanic&coll=ddd	https://www.delpher.nl/nl/kranten/ngram?query=Eiffeltoren&coll=ddd
Absolute Frequency		
Relative Frequency		

Click on an image to enlarge (when done, hit the 'back' button in your browser)

impresso data processing: Blog

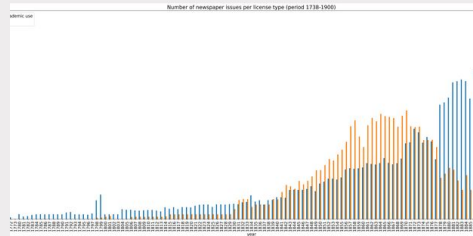
Blog

News and more

What's in our corpus?

Thu, 23.01 2020 — On the occasion of the first public release of the *impresso* interface, we wish to take stock of our newspaper corpus. More than a year has passed since the last corpus update...

[READ MORE](#)



Call for papers - Digitised newspapers - a new Eldorado for historians ?

Wed, 12.06 2019 — The large-scale digitisation of newspapers over the past decade has facilitated access to newspaper collections but also raised a series of issues for both libraries and users, and more specifically researchers: What does it mean to work in new ways with the traditional historical sources that are newspapers? How does the formal transformation of this source from analogue,



Named entity processing
Topic modeling
Text reuse
Word embeddings
Corpus composition

...

Komponenten und Anreicherungen: i-buttons und FAQs

FAQ

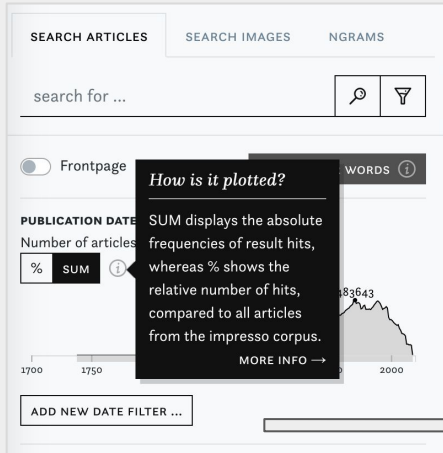
Use of the impresso interface

- + Where can I analyse the metadata of my collections or queries?
- + Limited search filters for this query
- + How does the image search work?
- + The search image page has limited search filters
- + How to make AND and OR queries?
- + What do ngrams show?
- + This ngram trends page has limited search filters
- How is it plotted?

- How is it plotted?

SUM displays the absolute frequencies of result hits, whereas % shows the relative number of hits, compared to all articles from the impresso corpus.

The highest result per year is displayed in number. The impresso interface groups the results per article, so the displayed frequency is the count of all articles containing at least one occurrence of the keyword. In other words, it is not the raw number of hits for the searched key-word, but the number of hits grouped by articles, as shown in the search summary. For instance, a query for 'Einstein' returns 8 967 articles. These 8 967 articles contain 13 535 individual hits of 'Einstein'. An item is included in the search when it contains at least one character, this means that the search is conducted also within advertisements and tables when they contain text.



impresso für Fortgeschrittene

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

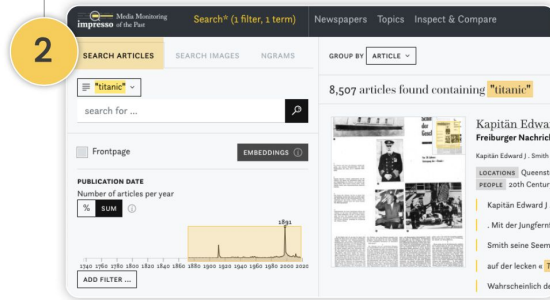
Impresso Challenges

- 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.

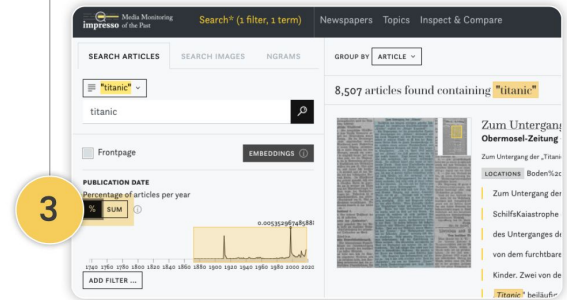
[DOWNLOAD CHALLENGES PDF](#)

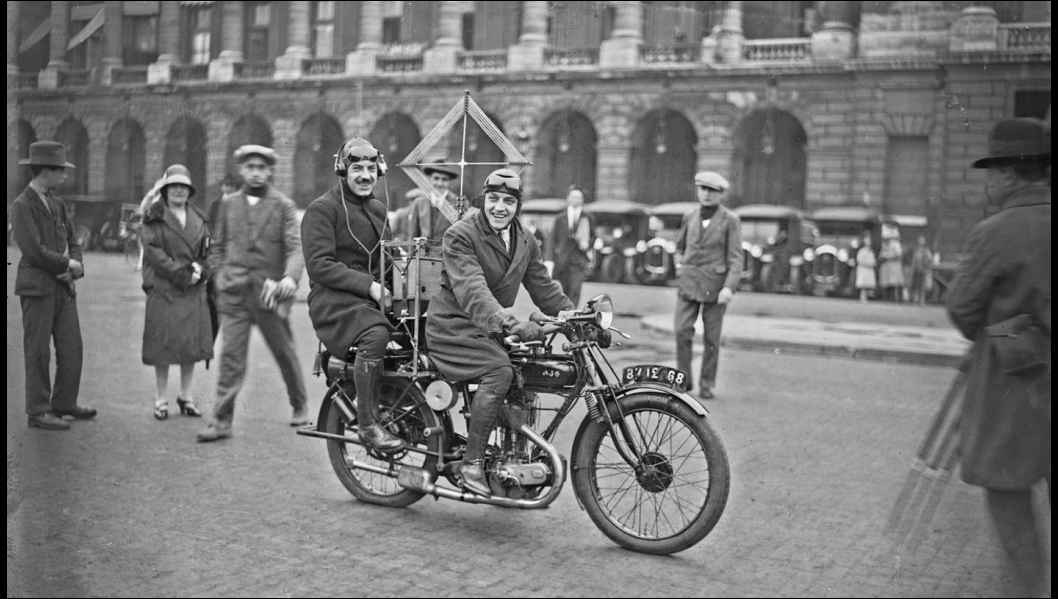
in contrast, search for the word titanic in the general search page: what is the difference between the number of hits shown in the ngram viewer and the search page?



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?





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;
- Erweiterung auf den ganzen *impresso* Korpus;
- Erste Analysen;
- Integration in die App.

The screenshot displays the Impresso search interface. At the top, the search bar contains the query "conference" and the filters "Reuters AND AFP". The search results are grouped by "ARTICLE" and ordered by "RELEVANCE". The first result is a personal use article from "Die Tat" dated Wednesday, May 1, 1968, titled "s-ffift Δff^^ J|L •dfjfr v: ... T mWk ...". The second result is another personal use article from "Freiburger Nachrichten" dated Wednesday, June 16, 1954, titled "Erdbeben registriert Pasadena (Kaliforni...". The third result is a personal use article from "Die Tat" dated Monday, June 2, 1958, titled "ten auf Ministerebene befassen. Diese Ko...". A dashed box highlights the "FILTER BY NEWS AGENCY" section, which shows the following options:

- Reuters(16,641 results)
- AFP(16,641 results)
- ATSSDA(3,361 results)
- UP UPI(1,430 results)
- AP(785 results)

Erwähnungen von Presseagenturen

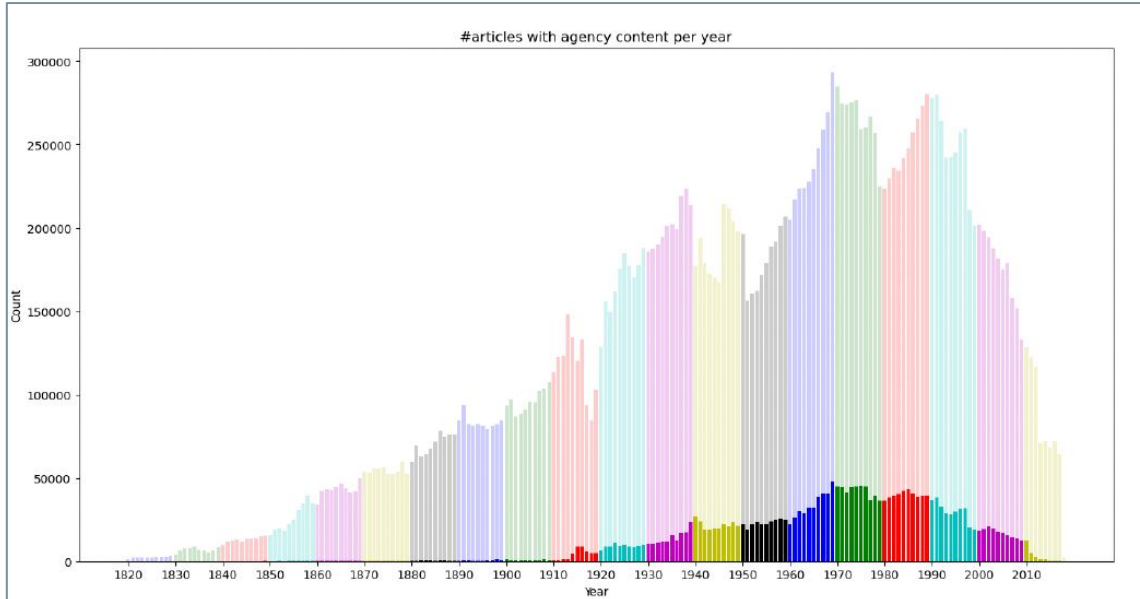


FIGURE 5.1

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

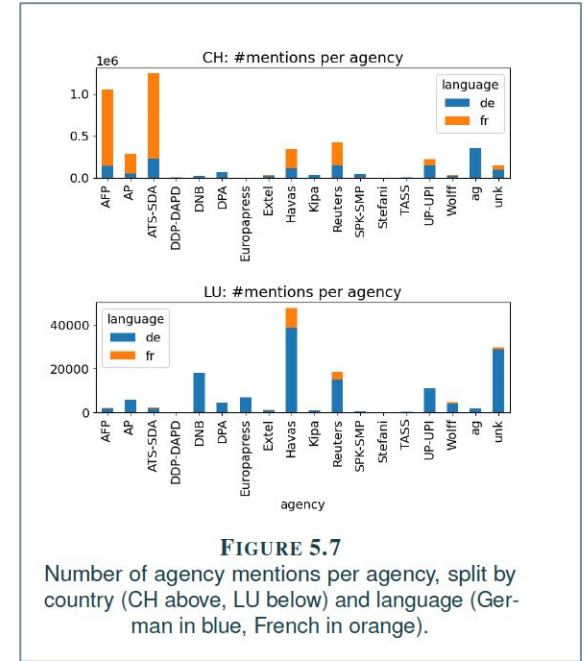
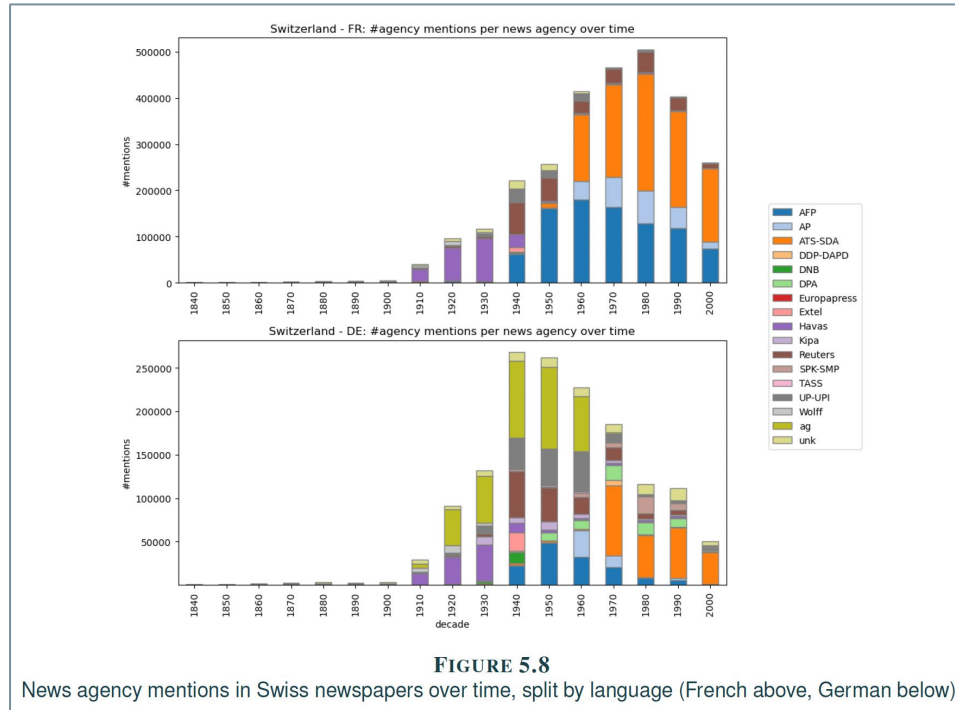


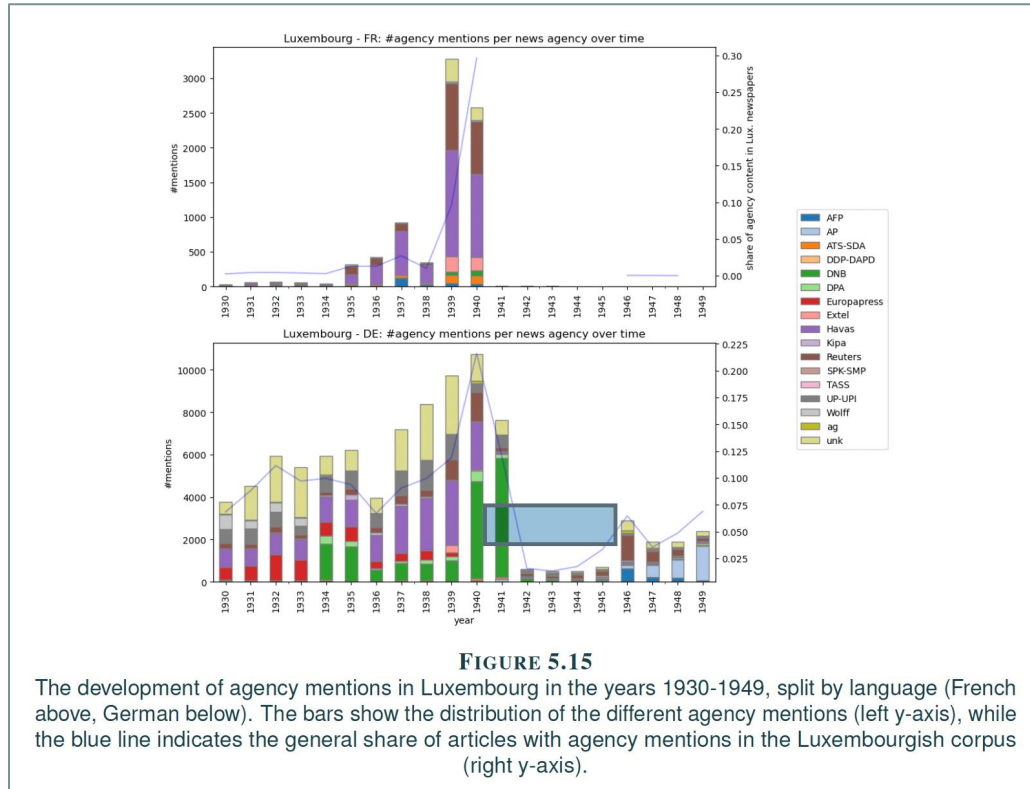
FIGURE 5.7

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

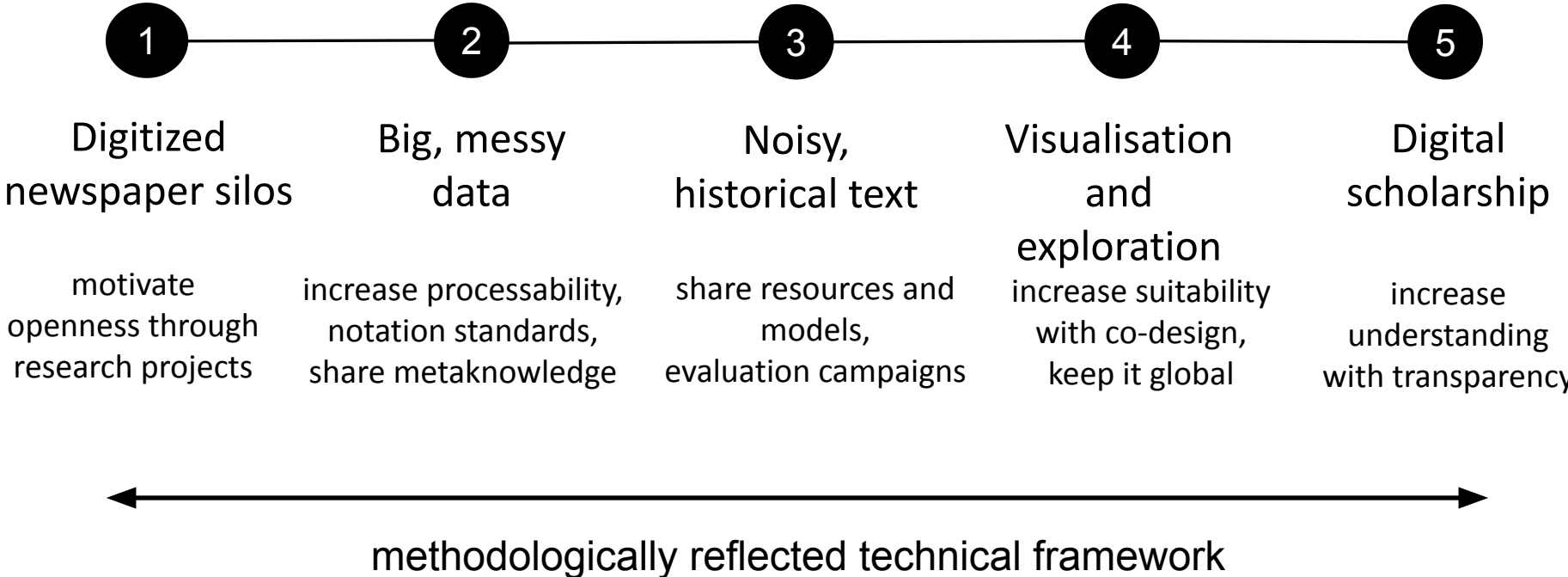
..in der schweizerischen Presse...



...in der luxemburgischen Presse (1940-1944)



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



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



Report from Dagstuhl Seminar 22292

Computational Approaches to Digitised Historical Newspapers

Edited by

Maud Ehrmann¹, Marten Düring², Clemens Neudecker³, and Antoine Doucet⁴

¹ EPFL - Lausanne, CH, maud.ehrmann@epfl.ch

² University of Luxembourg, LU, marten.during@uni.lu

³ Staatsbibliothek zu Berlin, DE, clemens.neudecker@bb.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



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

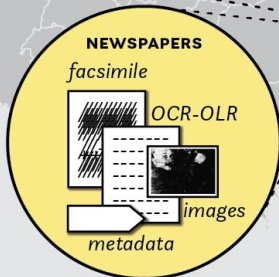
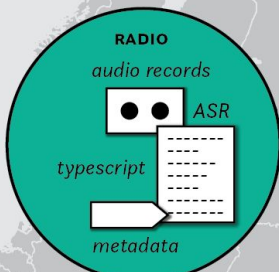
Research Networks

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

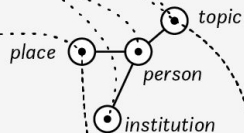
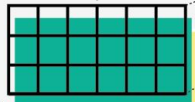


2 Media processing

Enriching & connecting

SEMANTIC ENRICHMENT
ACROSS LANGUAGES
ACROSS MEDIA

dense vector
representations



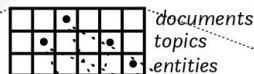
interview
advertisement
radio schedule

EXTERNAL
KNOWLEDGE

3 Media exploration

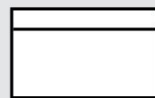
Connected and comparable enriched media sources

PROJECT DATA



API

IMPRESSO WEB APP



historians
designers
engineers

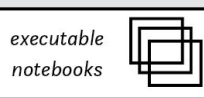
CO-DESIGNED INTERFACES

process data
using impresso API

USER-ORIENTED API

COMPATIBLE DATA

IMPRESSO DATA LAB



CO-DEVELOPED METHODS

EXTERNAL
DOCUMENTS



compatible
enrichments

community
of historians



For more information

impresso
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

JOIN OUR MAILING LIST

impresso website (update soon)

zenodo

Impresso - Media Monitoring of the Past

Recent uploads

- Extended Overview of CLEF HPE 2020: Named Entity Processing on Historical Newspapers**
@ Elrmeyri, @ Bonmalin, @ Fackler, @ Clentius
This paper presents an extended overview of the first edition of HPE (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. Steps to reproduction are hereby.
- CLEF HPE 2020 Named Entity Recognition and Linking on Historical Newspapers (slides)**
@ Elrmeyri, @ Bonmalin, @ Fackler, @ Clentius
Slides of the first 2020 task overview during the CLEF 2020 conference.
- Deep diving in NLP enhanced digitised newspapers. A hands-on session with the impresso interface**
@ Fackler, @ Bonmalin, @ Elrmeyri
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 of:
- Introducing the CLEF 2020 HPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers**

data on Zenodo

impresso-project

Media Monitoring of the Past

Repository

- fakenews-quiz**
- impresso-language-identification**
- epfl-aha-class**
- federal-gazette**
- impresso-data-sanitycheck**
- impresso-ahc-acquisition**
- CLEF-HPE-2020-internal**
- impresso-technical-cookbook**
- impresso-gycommons**

code on GitHub

impresso-project

5 subscribers

Media Monitoring of the Past

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

The objective of the impresso project 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.

Created playlists

- CLEF HPE 2020
- HPE 2020 - Binler Team
- CLEF HPE 2020 - NLP-UQAM Team
- CLEF HPE 2020 - ERTM Team

impresso youtube channel

impresso-project.ch

Media Monitoring of the Past

Mining 200 years of historical newspapers

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

78 newspapers collected,
600,919 issues,
5,429,856 pages scanned,
479,448 content items identified,
3,456,700 entities,
19,493,358,793 words.

2 countries of publication
230,086 named entities disambiguated

More? Check on our **blog**

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

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

Download our **DISCLOSURE AGREEMENT FORM**
signed form to info@impresso-project.ch

impresso-project.ch/app