



Historical Newspaper Content Mining: findings from the impresso project



SPEAKERS



Estelle Bunout
Research associate, Leibniz Centre for
Contemporary History Potsdam



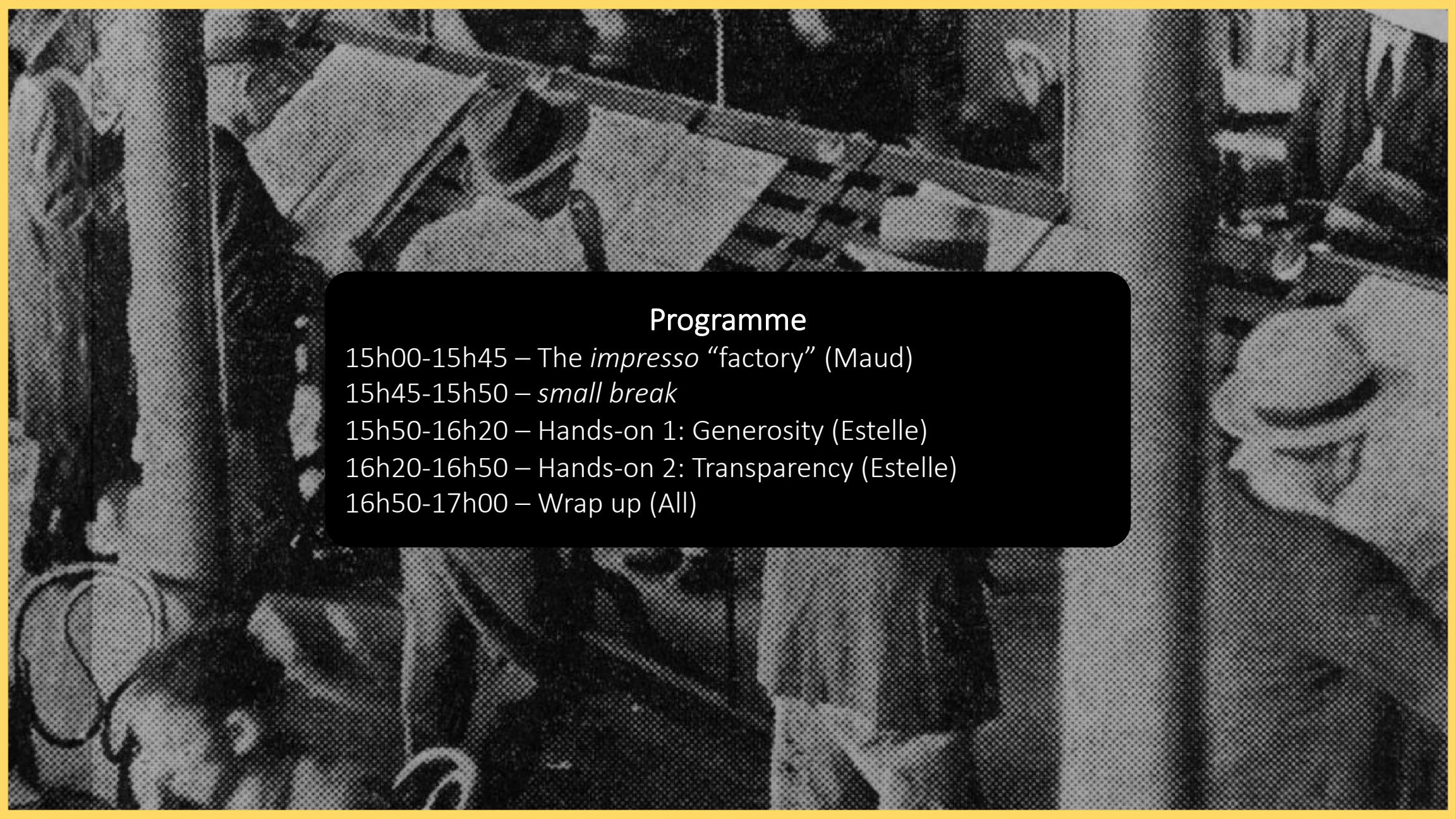
Maud Ehrmann
Research scientist, École Polytechnique Fédérale
de Lausanne (EPFL)

NOTES

- **Please turn off your camera and microphone.** You can unmute your microphone when asked.
- **Questions?** Put them in the chat box. We'll put questions to the speakers during the Q&A time and the hands-on sessions.
- **The workshop is being recorded.** All participants will receive a link to the recording shortly.
- **Slides will be uploaded on Zenodo after the workshop**
All participants will receive a link to the slides shortly.

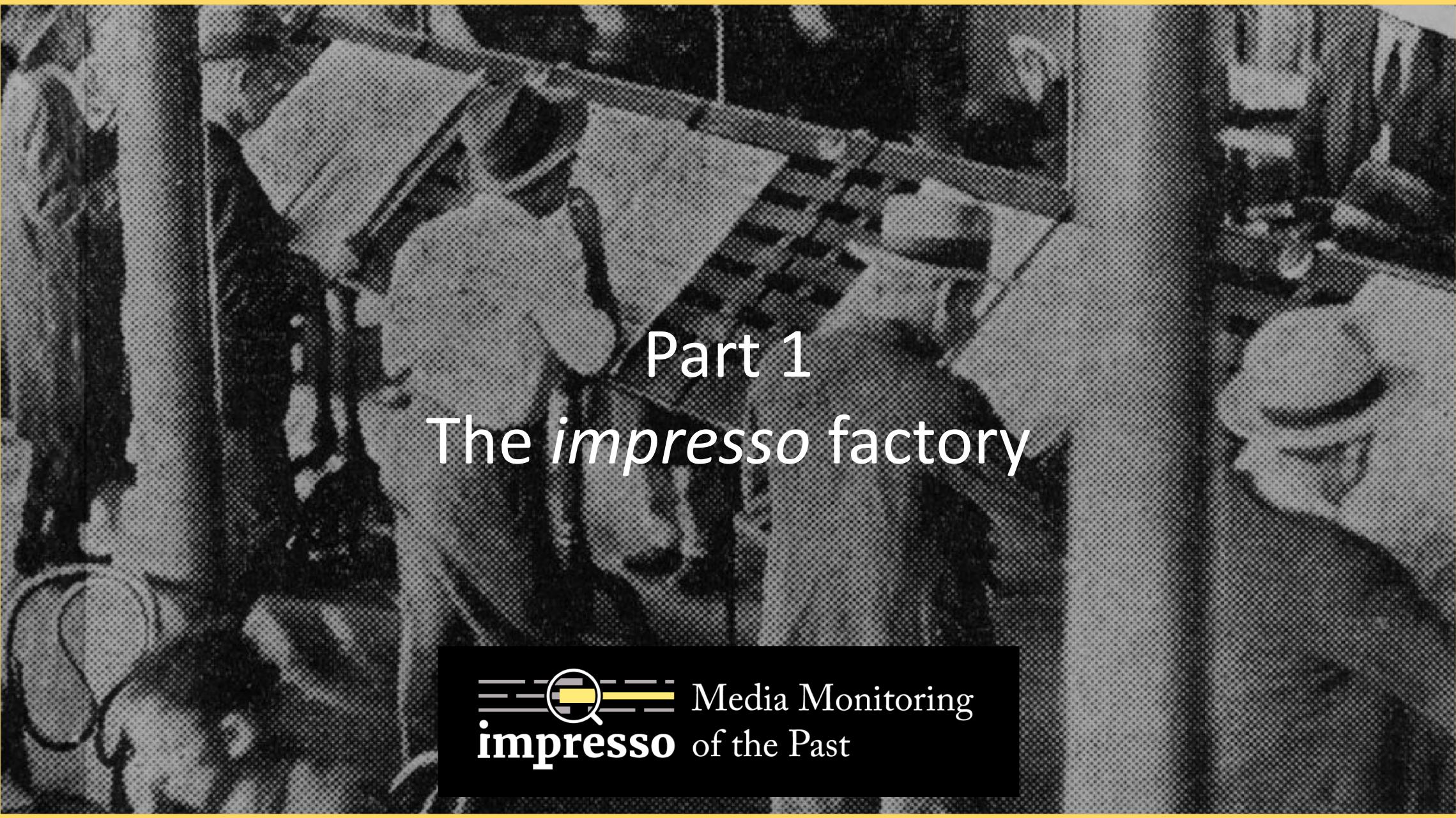


Estelle Bunout, Maud Ehrmann & the *impresso* team
Workshop AI4LAM - 31.02.2021



Programme

- 15h00-15h45 – The *impresso* “factory” (Maud)
- 15h45-15h50 – *small break*
- 15h50-16h20 – Hands-on 1: Generosity (Estelle)
- 16h20-16h50 – Hands-on 2: Transparency (Estelle)
- 16h50-17h00 – Wrap up (All)



Part 1

The *impresso* factory



Media Monitoring
impresso of the Past

NEWSPAPERS





British Library National Newspaper Archive - Photo by Christopher Furlong / Getty images

This screenshot shows the homepage of the Colorado Historic Newspapers Collection. It features a search bar at the top with options for 'Search the Collection' and 'Browse by'. Below this is a 'Top Journeymen Editors' section with a list of names and their counts. A 'About Our Collection' section provides an overview of the CHNC's mission and history. On the right, there are sections for 'Topics in History' (including 'Bushell', 'Nikola Tesla', and 'Rocky Mountain National Park'), 'Listen to the Past' (with a link to a video), and 'Learn more about CHNC'.

This screenshot shows the homepage of the Bibliothèque nationale de Luxembourg. It includes a search bar for 'Chercher un article' and 'Chercher un titre'. A sidebar on the left lists various news titles like 'Pardate', 'Par'lire', and 'movement'. A central column displays a news item from 'Zur Zukunftsgestaltung zum Land a Riffkin-Prozess'.

This screenshot shows the homepage of LE TEMPS. It features a search bar for 'Rechercher dans les archives de la Gazette de Lausanne et du Journal de Genève'. Below the search bar, there is a snippet of text: 'Ex: Rumus, Confédération helvétique, taillerie, g3.22.07.2016'. The page also includes a 'Nos partenaires' section.

This screenshot shows the homepage of Chronicling America, part of the Library of Congress. It features a search bar for 'Search Pages' and 'Advanced Search'. Below the search bar, there is a grid of historical newspaper pages from various publications like 'The New York Tribune' and 'The New-York Daily Review'.

This screenshot shows the homepage of ANNO - Austria! Newspapers Online. It features a search bar and a sidebar with links for 'Über ANNO', 'Suchen in ANNO', 'Drucken aus ANNO', 'Kontakt', 'Impressum', and 'Service'. The main content area shows a historical newspaper page from 'SPORTSALON'.

This screenshot shows the homepage of ZEFTYS Zeitungsinformationsystem. It features a search bar and a sidebar with links for 'Start', 'Suchen', 'Über uns', and 'Aktuelles'. The main content area shows a historical newspaper page from 'Der Reichsführer hat für Seine Majestät den Throne in einzutreten'.

This screenshot shows the homepage of POLONA. It features a search bar and a sidebar with links for 'Suchen', 'Feuilleter', 'Tags', and 'Aide'. The main content area shows a grid of historical newspaper pages from various publications.

This screenshot shows the homepage of E-NewsPaper Archives.ch. It features a search bar and a sidebar with links for 'Faites une recherche dans la collection', 'SUR LA COLLECTION', and 'ETENDUE DE LA COLLECTION'. The main content area shows a grid of historical newspaper pages from various publications.

This screenshot shows the homepage of Georgia Historic Newspapers. It features a search bar and an interactive map of Georgia. The main content area shows a grid of historical newspaper pages from various publications.

This screenshot shows the homepage of The European Library. It features a search bar and a sidebar with links for 'About', 'Membership', 'Discover', and 'Access our Data'. The main content area shows a grid of historical newspaper pages from various publications.

This screenshot shows the homepage of Delfpher. It features a search bar and a sidebar with links for 'Explore newspapers', 'Explore newsbooks', and 'Explore newsletters'. The main content area shows a 3D visualization of many historical newspaper pages floating in space.

This screenshot shows the homepage of Trove, the digital library of Australia. It features a search bar and a sidebar with links for 'Search articles by', 'Browse articles by', and 'Top test correctors'. The main content area shows a grid of historical newspaper pages from various publications.

Help me

... “expand my query based on common OCR mistakes and spelling variants” + multilingual synonyms



Eegierung
Begierung
Regieruug
Kegierung
Regiexung

similar words
+ edit distance

Behorden
Finanzdirektion
Staatsrath
Eegierungen
Behörde
Reichregierung
Staatsrat
Staatsrathes

similar words

→ lexical processing

Help me

editorial

... “find the frequency of word X in
articles of certain type or thematic”

... “distinguish ‘real’ occurrences
from false positives emerging from
ads, stock market or meteo pages”

feuilleton

→ segmentation (CV) and classification (NLP)

title banner and 'ears'



A detailed image of a historical newspaper page from 'Gazette de Lausanne ET JOURNAL SUISSE' dated November 26, 1926. The page is filled with columns of French text. At the top left, there's a box for 'EDITION DE MIDI' and 'LIBERTÉ ET PATRIE' at the top right. The main title 'Gazette de Lausanne' is prominently displayed in the center. Several arrows point from the surrounding text to specific parts of the page, highlighting different sections like 'bulletin politique', 'lettre de Berlin', and 'feuilleton'. Other annotations include 'culture' pointing to the letter from Berlin, and 'local news' pointing to the feuilleton section.



culture

local news

A detailed image of a historical newspaper page from 'Gazette de Lausanne ET JOURNAL SUISSE' dated November 26, 1926. The page is filled with columns of French text. At the top left, there's a box for 'EDITION DE MIDI' and 'LIBERTÉ ET PATRIE' at the top right. The main title 'Gazette de Lausanne' is prominently displayed in the center. Several arrows point from the surrounding text to specific parts of the page, highlighting different sections like 'bulletin politique', 'lettre de Berlin', and 'feuilleton'. Other annotations include 'culture' pointing to the letter from Berlin, and 'local news' pointing to the feuilleton section.



culture

local news

Help me

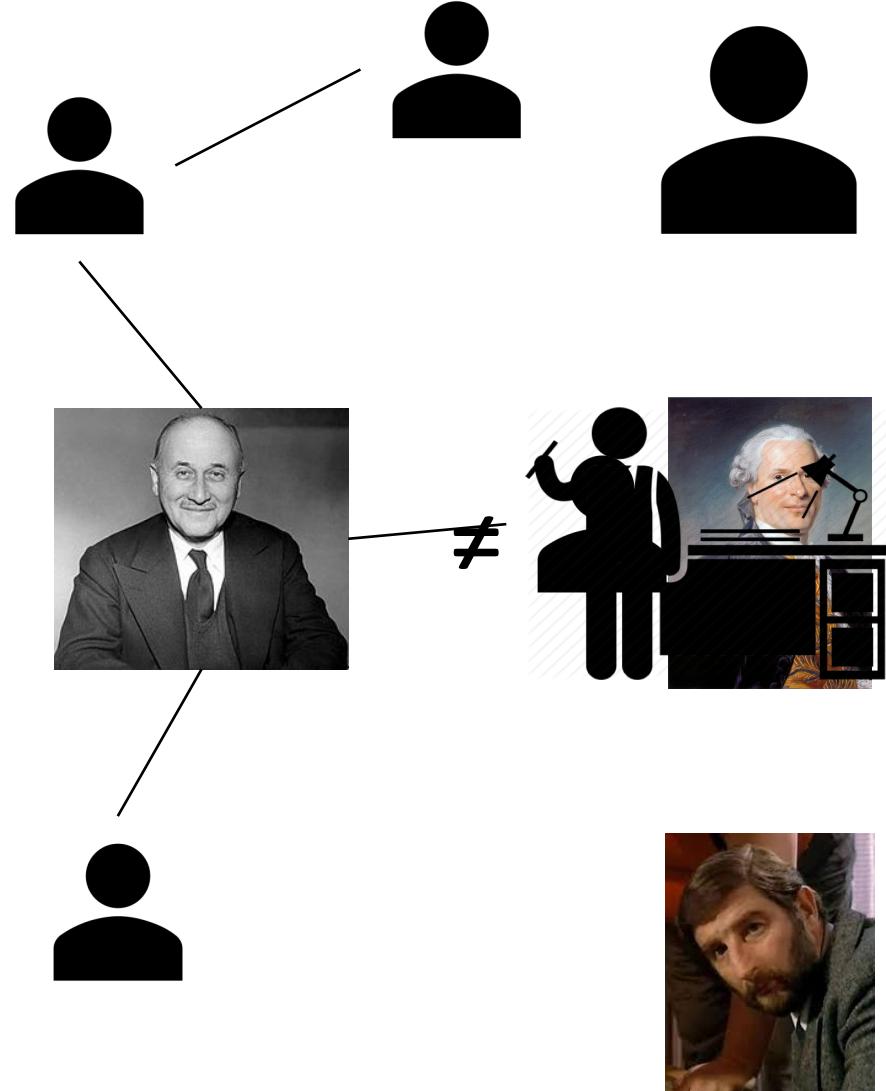
... “ find all articles mentioning Jean Monnet.

And the right one!”

... “ know with whom the name of X is associated.”

... “ find articles mentioning person X as **author** vs. as **main subject** of the article”

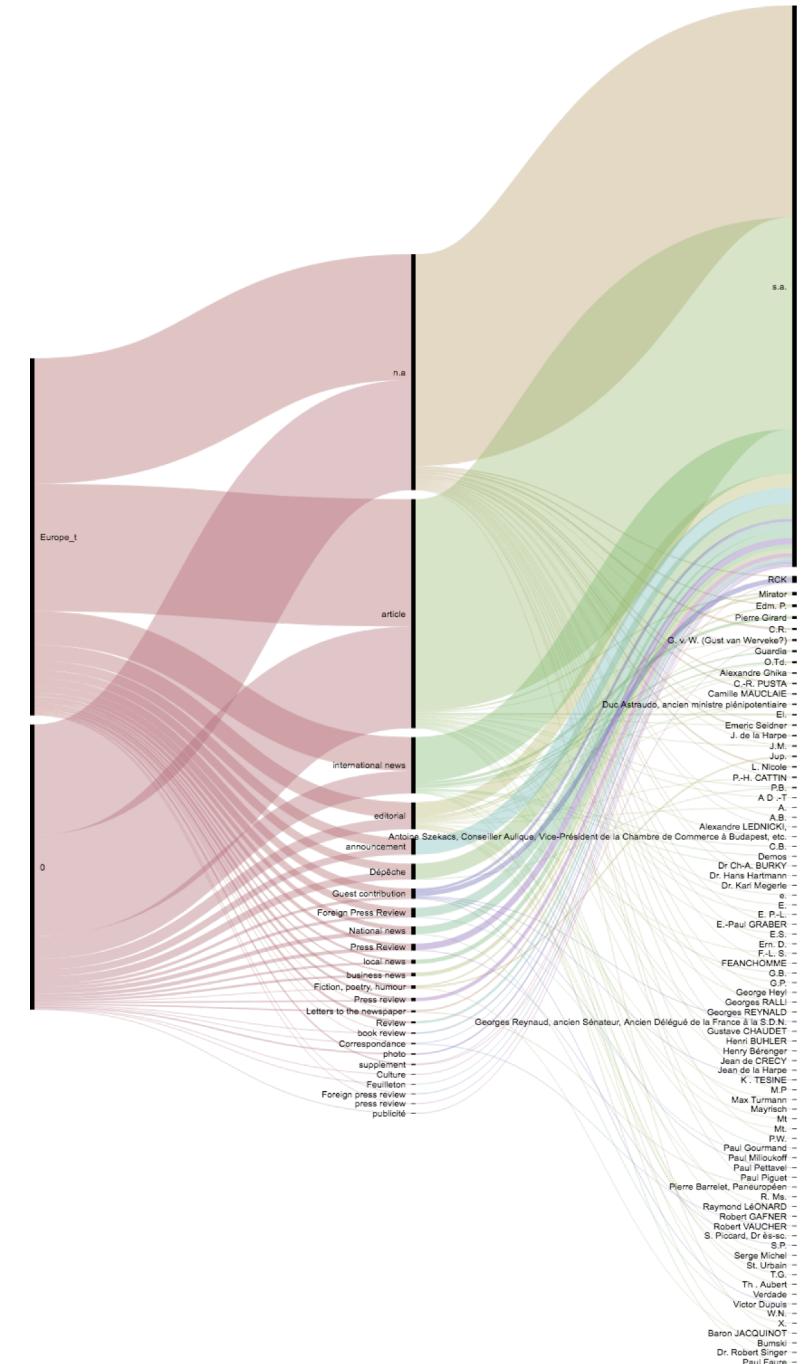
→ named entity processing



Help me

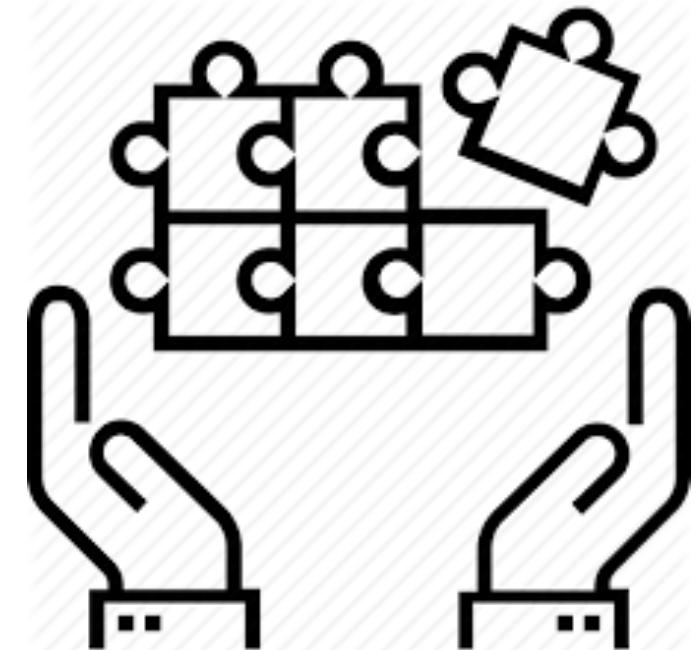
... “ collect and compare sub-collections based on metadata (time, titles, political orientation, etc) but also based on content”

→ interface, metadata



Help me

- ... “discover similar items I might be interested in”
- ... “appreciate the high-level thematic(s) of the corpus”
- ... “build and interact with my own collection”
- ...



How to enable semantic indexing and
exploration of large collections of historic
newspapers?



Media Monitoring
impresso of the Past

EPFL

C²DH



University of
Zurich UZH

[FNSNF]

SWISS NATIONAL SCIENCE FOUNDATION



Swiss National Library, SNL



National Library of Luxembourg,
BnL



State Archives of Valais, AEV.



Swiss Economic Archives, SWA.

LE TEMPS

NZZ

Le Temps

Neue Zürcher Zeitung, NZZ.

Unil

History department, University
of Lausanne, UNIL.

infoclio.ch

infoclio

+ many others institutions



Critical content mining
of 200 years of
historical
newspapers

The team

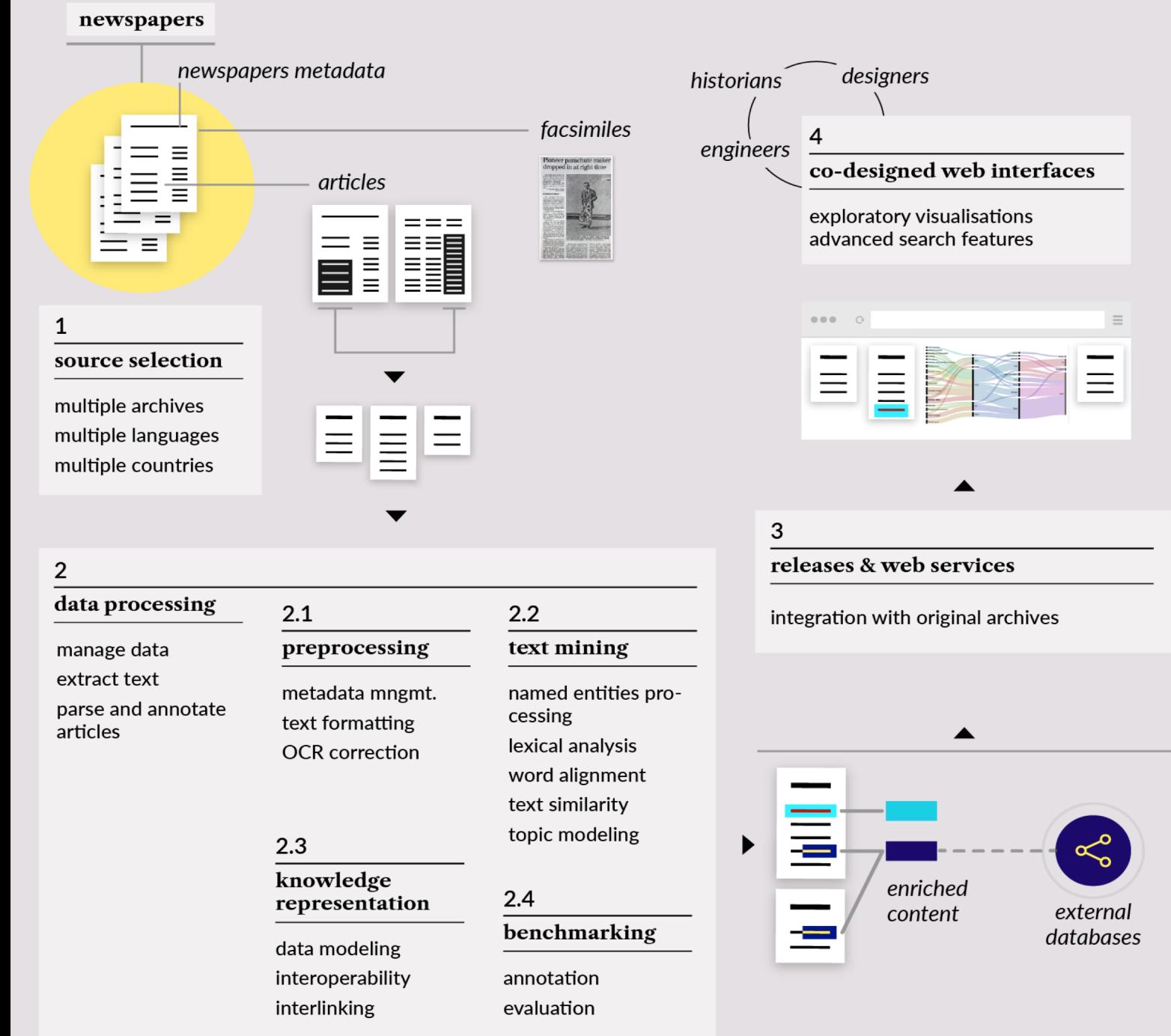
*Estelle Bunout
Simon Clematide
Marten Duering
Maud Ehrmann
Andreas Fickers
Daniele Guido
Frédéric Kaplan
Peter Makarov
Matteo Romanello
Gerold Schneider
Paul Schroeder
Benoit Seguin
Phillip Stroëbel
Martin Volk
Thijs van Beek
Lars Wieneke*



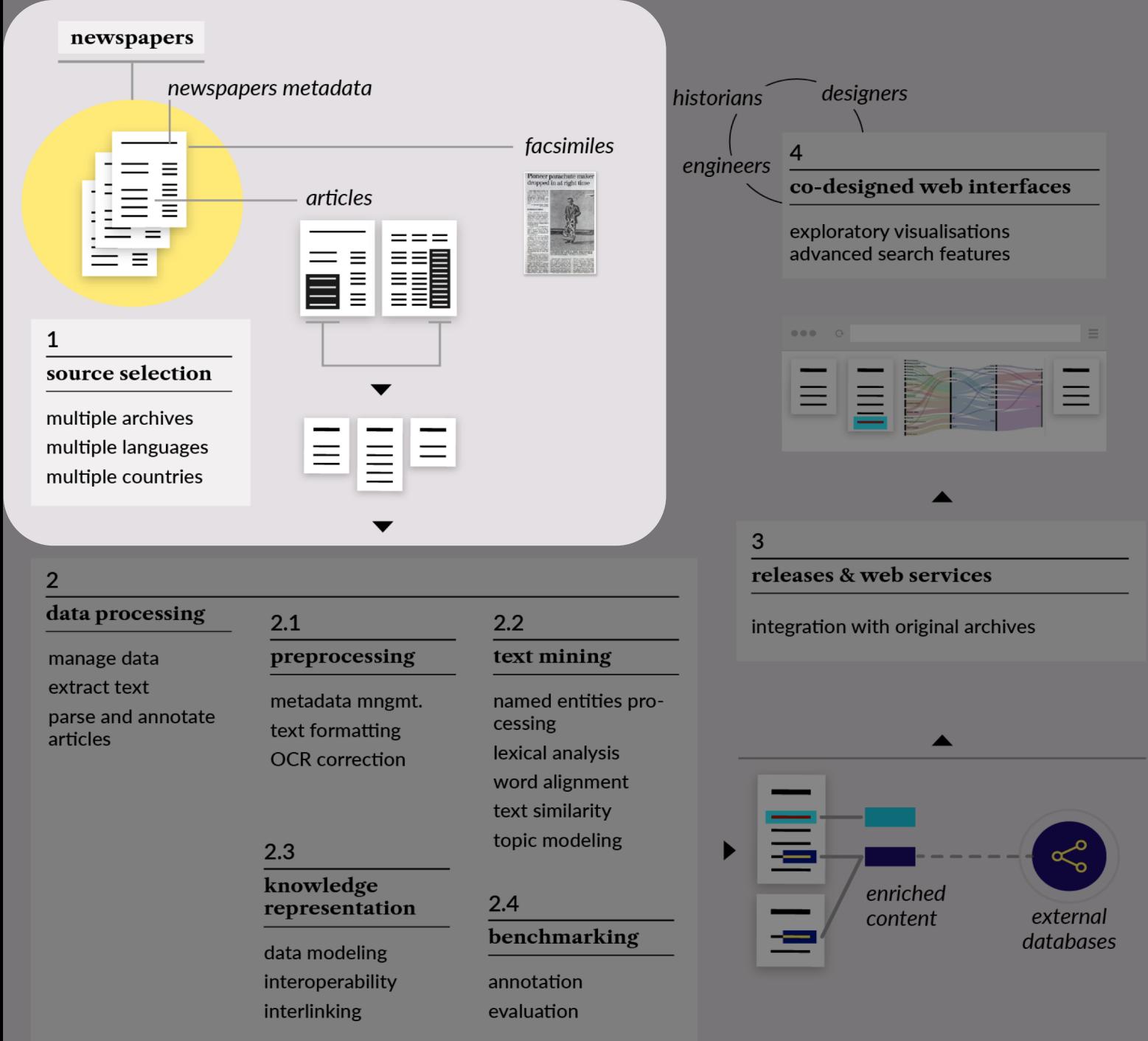
*+ a team of historical advisors
+ a team of associated historians*



...and our plan:

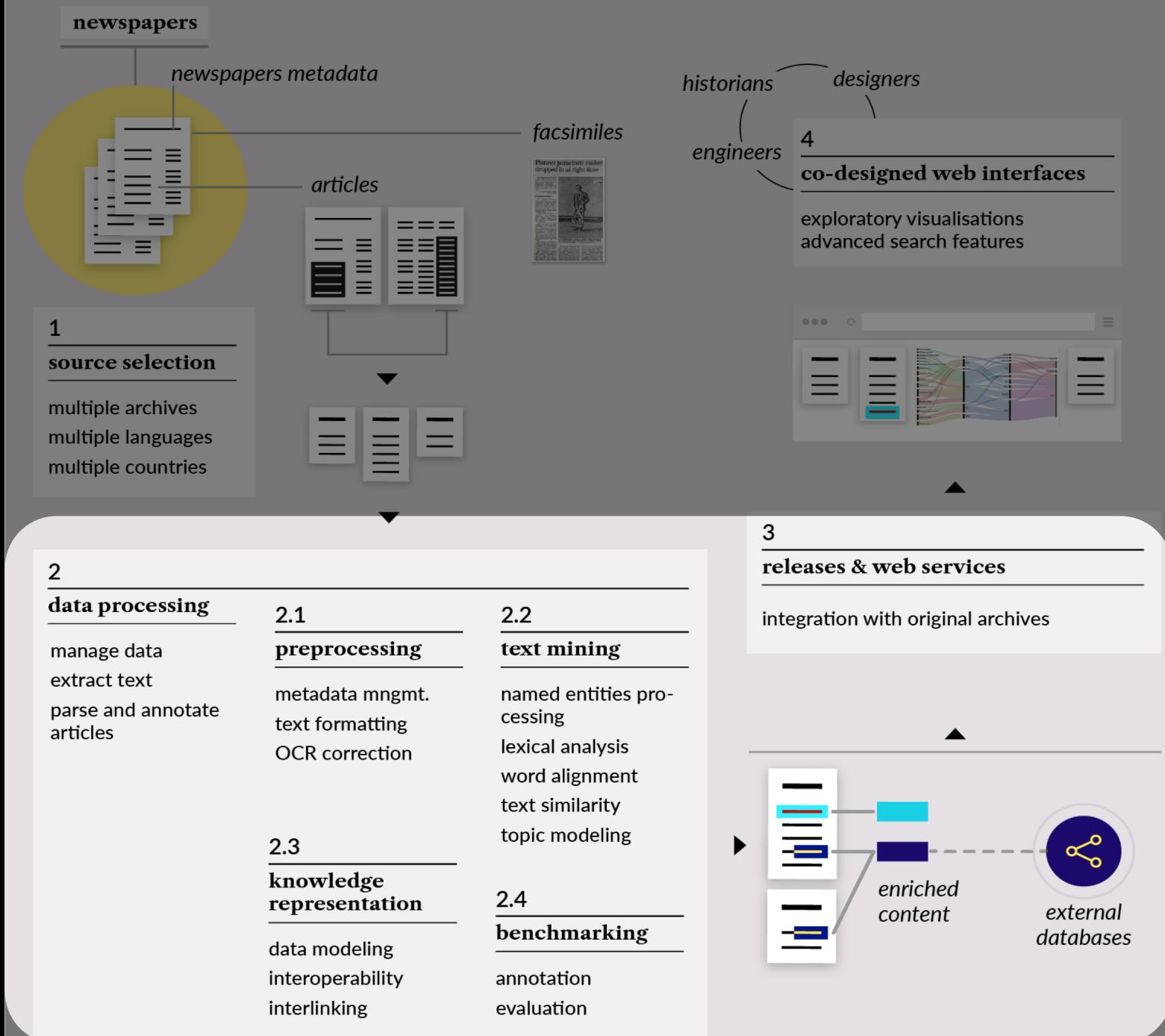


...and our plan:



...and our plan:

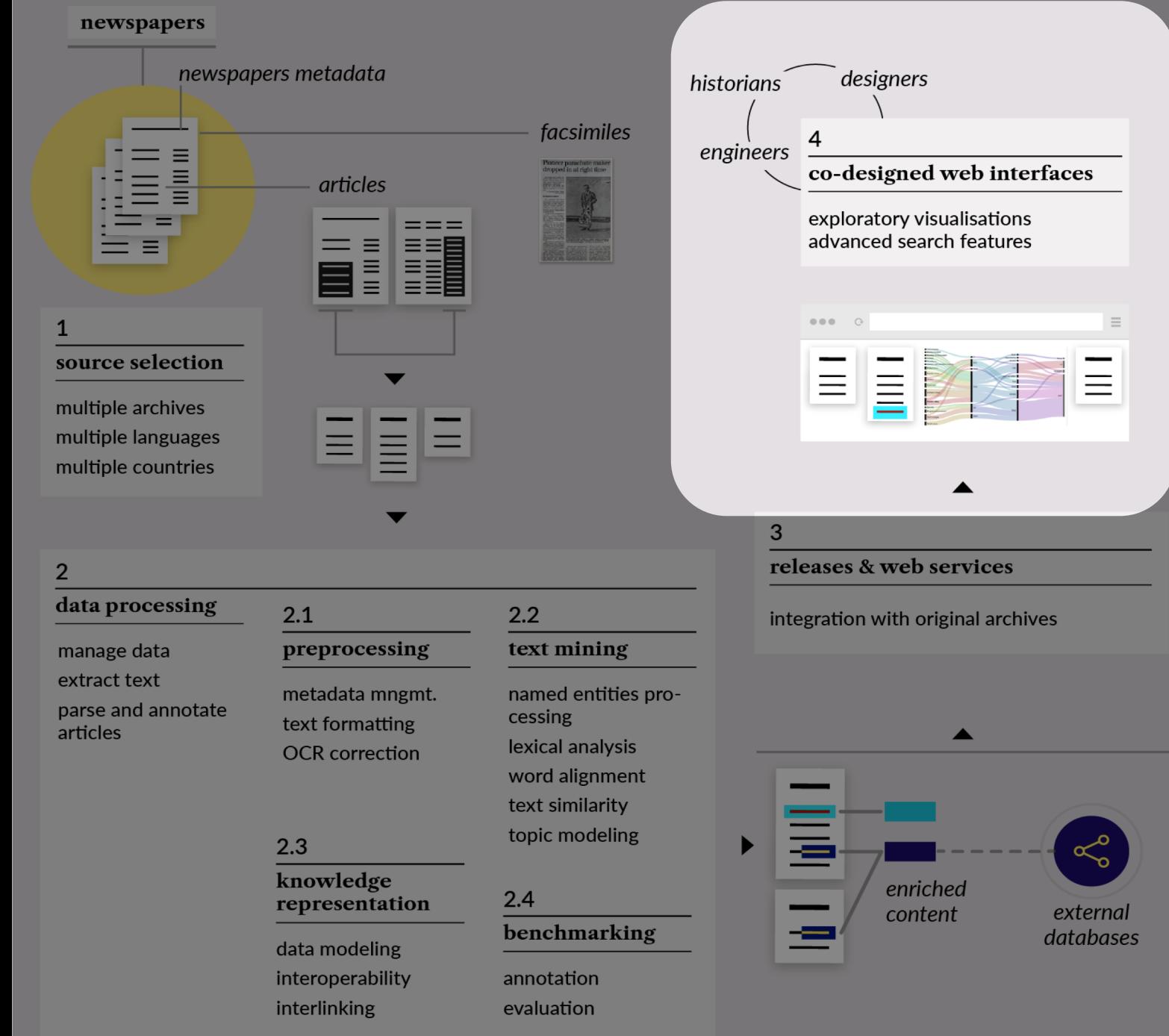
1. How to adapt NLP tools to historical texts?



...and our plan:

1. How to adapt NLP tools to historical texts?

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

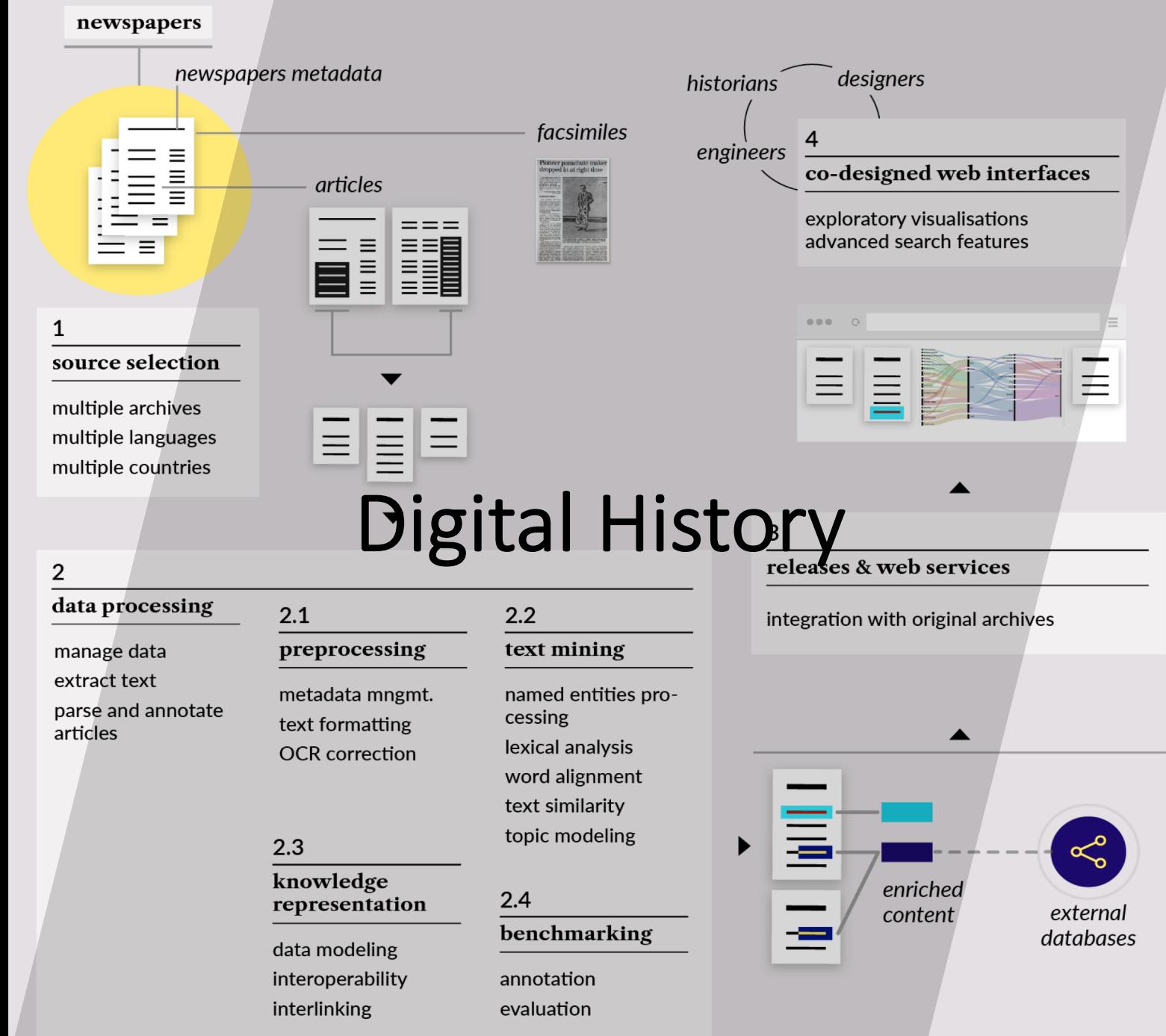


...and our plan:

1. How to adapt NLP tools to historical texts?

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

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



Guiding principles

- Co-design, pull-push
- Experimentation
- Transparency
- Generosity
- Continuum

The screenshot shows the homepage of the Media Monitoring of the Past website. At the top, there's a navigation bar with links for 'Search', 'Newspapers', 'Topics', 'Inspect & Compare', and 'Text reuse'. On the far right are 'FAQ' and 'LOGIN' buttons. The main title 'Media Monitoring of the Past' is prominently displayed in large, white, serif font. Below the title is a search bar with tabs for 'SEARCH ARTICLES', 'SEARCH IMAGES', and 'NGRAMS', followed by a search input field and two icons. A section titled 'IMPRESSO DATA RUNDOWN' provides statistical information: 76 newspapers collected, 600,919 issues, 5,429,656 pages scanned, 47,798,468 content items identified, 3,462,799 images, and 12,493,358,703 words. It also mentions 2 countries of publication and 530,086 named entities disambiguated. Below this, there's a link to the 'blog'. On the left side of the main content area, there are social media links for GitHub (@impresso) and Twitter (@impressoproject). At the bottom left, there are buttons for 'LINES: OFF' and 'DARK MODE: ON'. On the right side, there's a section about mining historical newspapers, a link to the 'Non-Disclosure-Agreement Form', and a contact email address. A small circular icon with a number '3' is in the bottom right corner.

keyword suggestion

The screenshot shows a search interface for historical news articles. A yellow arrow points from the text "keyword suggestion" to the search bar where "arnhem" is typed. The search results page displays several news items, with one article from "Le Peuple, La Sentinel" on April 17, 1945, titled "Arnhem nettoyé". Another arrow points from the text "named entities" to the bottom of the page, specifically to the section "Communication aux porteurs d'obligations de la Société de la Banque de Luxembourg".

marginalia

A yellow arrow points from the text "marginalia" to a historical newspaper clipping from "La Sentinel" dated April 17, 1945. The clipping discusses the liberation of Arnhem. Marginalia are handwritten notes in the margins of the document, such as "Ils font le jeu de la réaction" and "LA PAIX APPROCHE Sera-t-il la « Paix du Peuple » solennellement promise ?". Another arrow points from the text "named entities" to the same newspaper clipping.

named entities

A yellow arrow points from the text "named entities" to a historical newspaper clipping from "La Sentinel" dated October 6, 1944, with the headline "LA PAIX APPROCHE". The clipping discusses the approaching peace. Entity recognition annotations include "Jules Humbert-Droz PERSON" with a timeline chart and "Swiss communist (1891-1971)". Another arrow points from the text "named entities" to the same newspaper clipping.

*How to enable semantic indexing and exploration
of large collections of historic newspapers?*

5 challenges

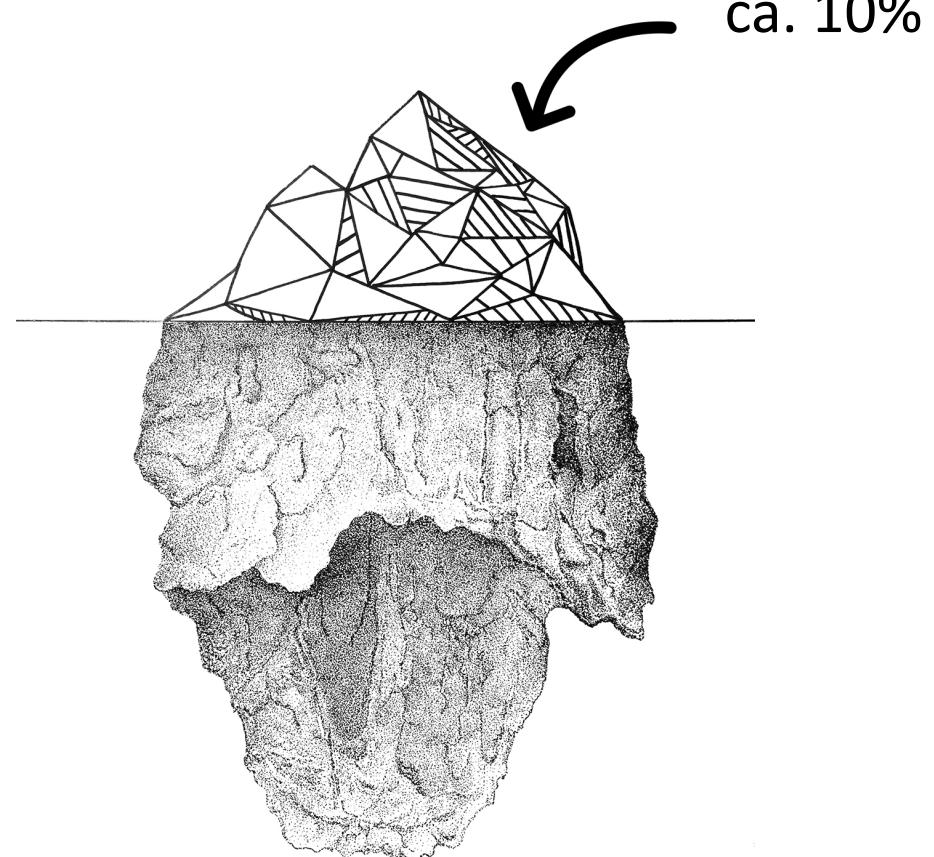
How to enable semantic indexing and exploration of large collections of historic newspapers?

1

Digitized
newspaper silos

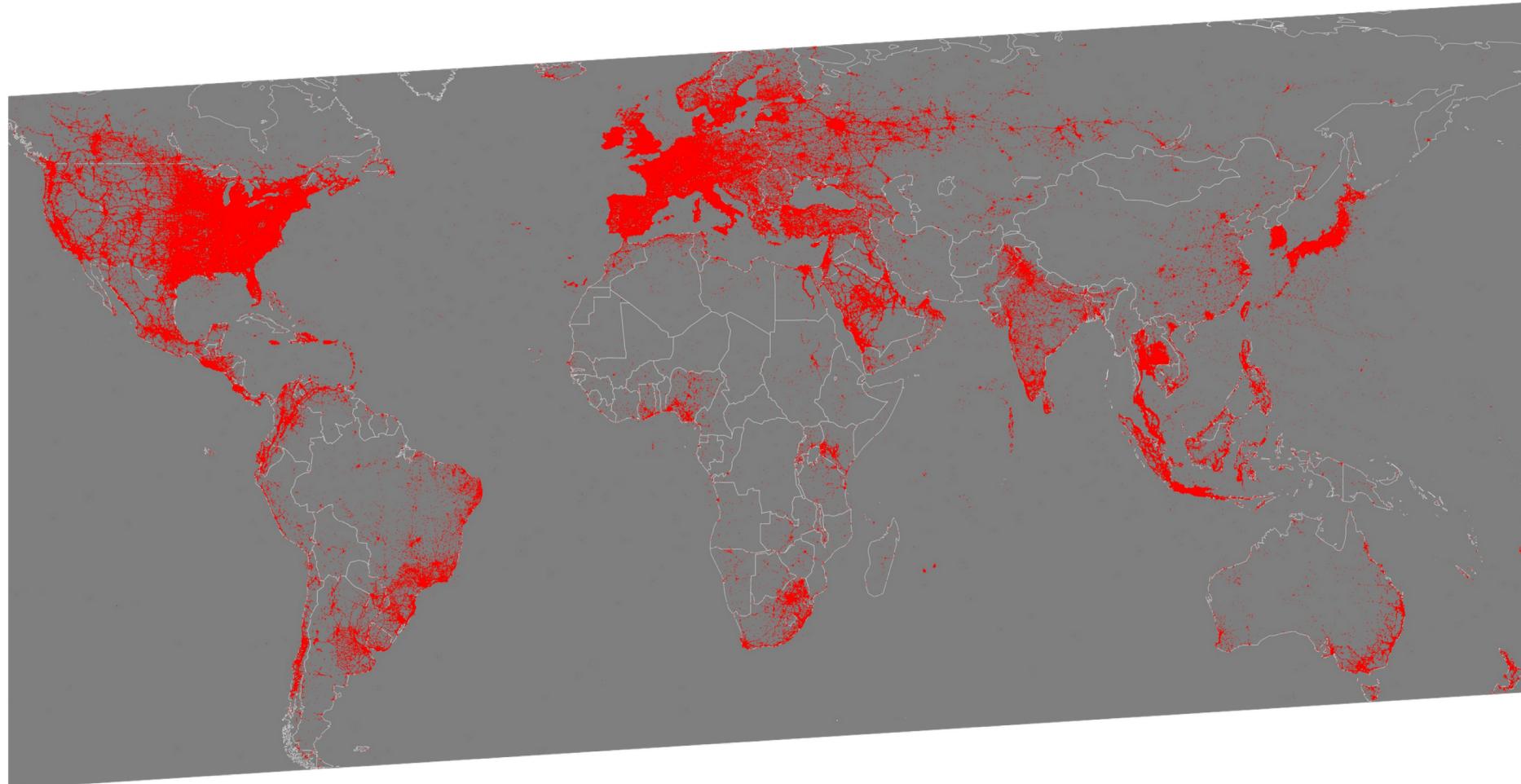
Shape of the (newspaper) digital landscape

- Partial



Shape of the (newspaper) digital landscape

- Partial
- Imbalanced



Cf: L. Manovich, *Cultural Data: Possibilities and Limitations of Digitized Archives*, <http://manovich.net/index.php/projects/cultural-data>

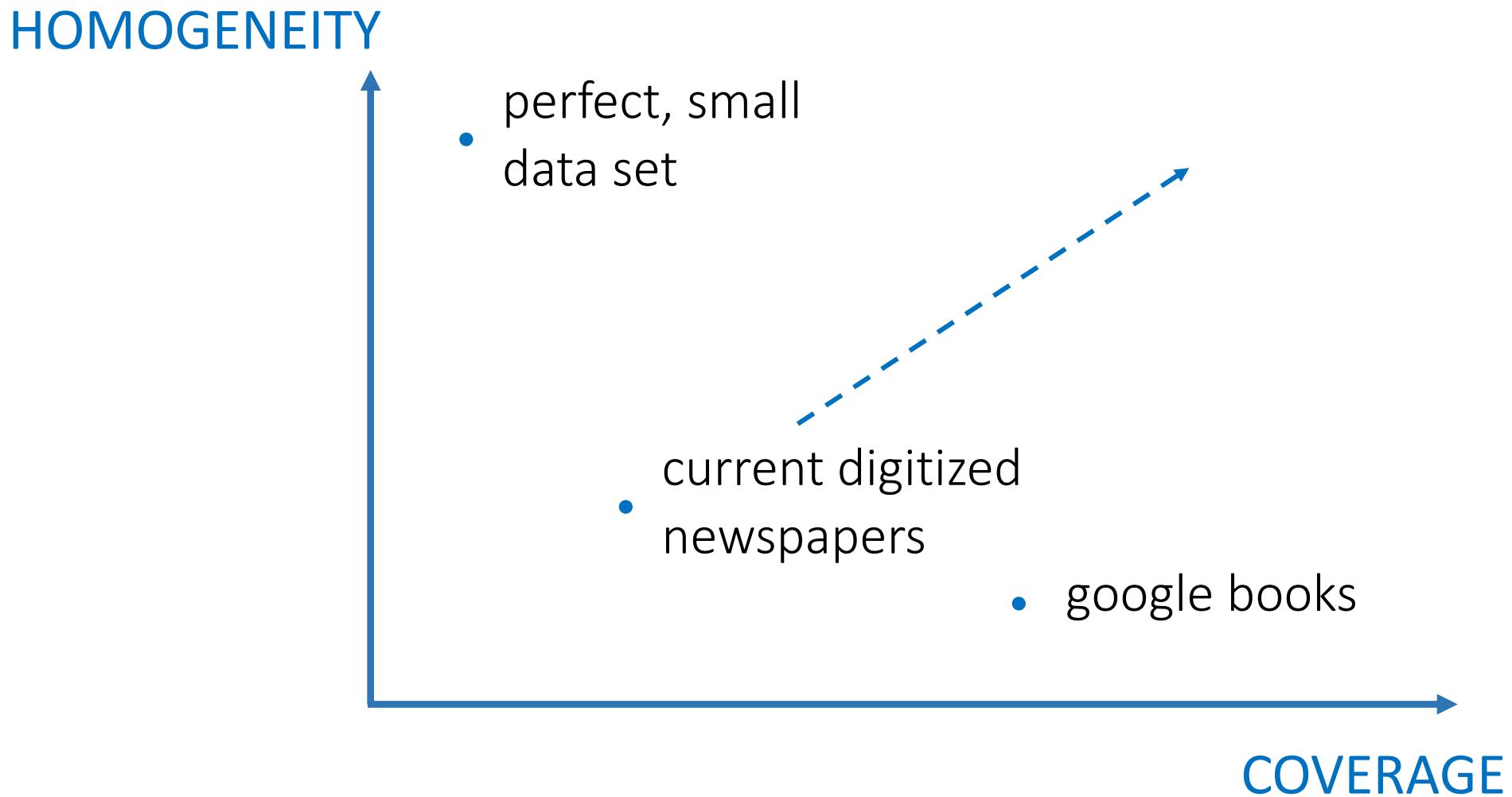
Cf. L. Putnam, *The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast*, *The American Historical Review* (2016)

Shape of the (newspaper) digital landscape

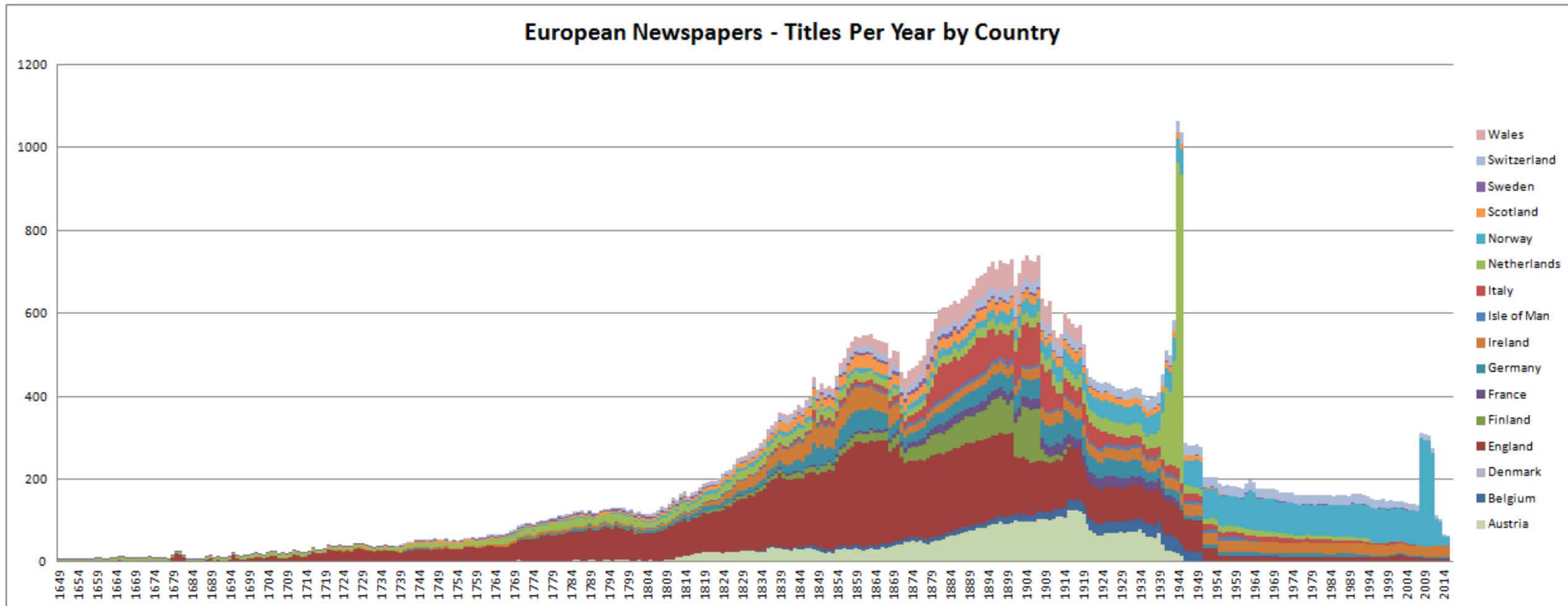
- Partial
- Imbalanced
- Heterogeneous



Shape of the (newspaper) digital landscape



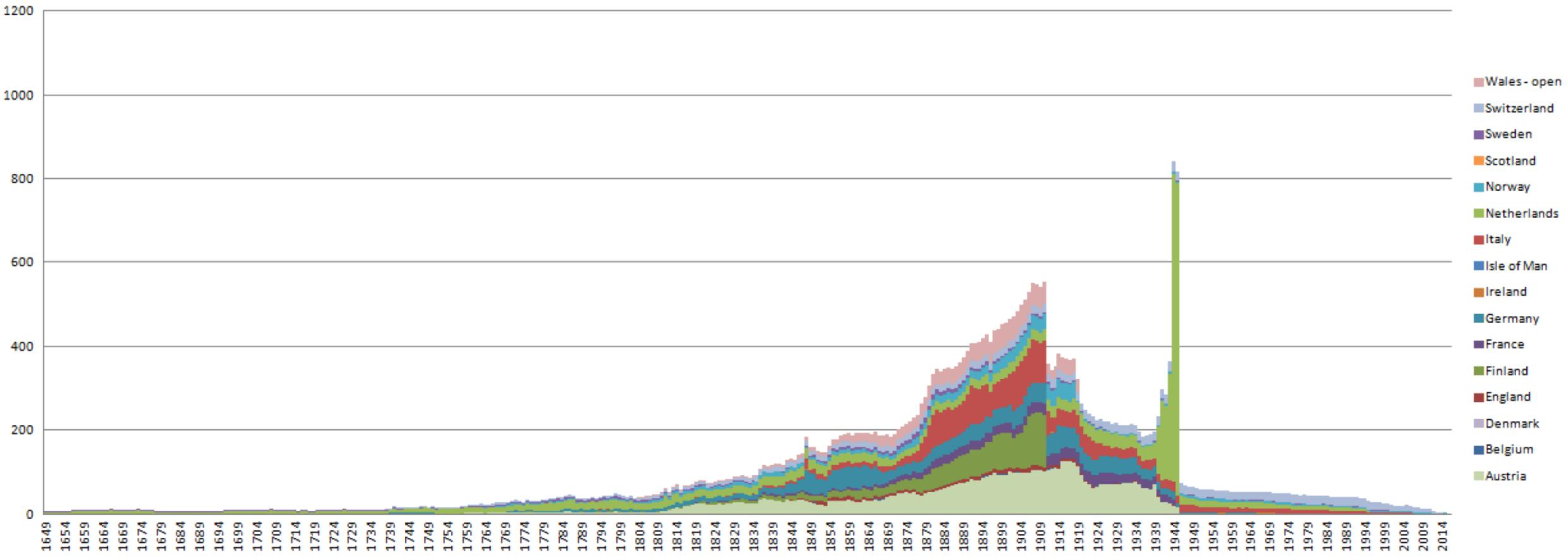
Access policies



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

Access policies

European Newspapers - Open Access Titles Per Year by Country



“Data acquisition crusade”

Tedious and time consuming

- source inventory
- conventions with right holders
- physical data acquisition

The screenshot shows a Google Sheets document titled "MediaSources-SwissNationalLibrary". The spreadsheet contains a large amount of data across numerous columns. The columns include: Number, Source, Title, Format, Access Method, Access Method Description, Digital Access Method, Digital Access Method Description, Language, Date of Production, Date of First Broadcast, Date of First Publication, Date of First Release, Medium, Medium Description, Size, Size Description, Notes, Notes Description, Status, Status Description, Last Update, and Last Update Description. The data is presented in a grid format with many rows of entries.

Bridging silos?

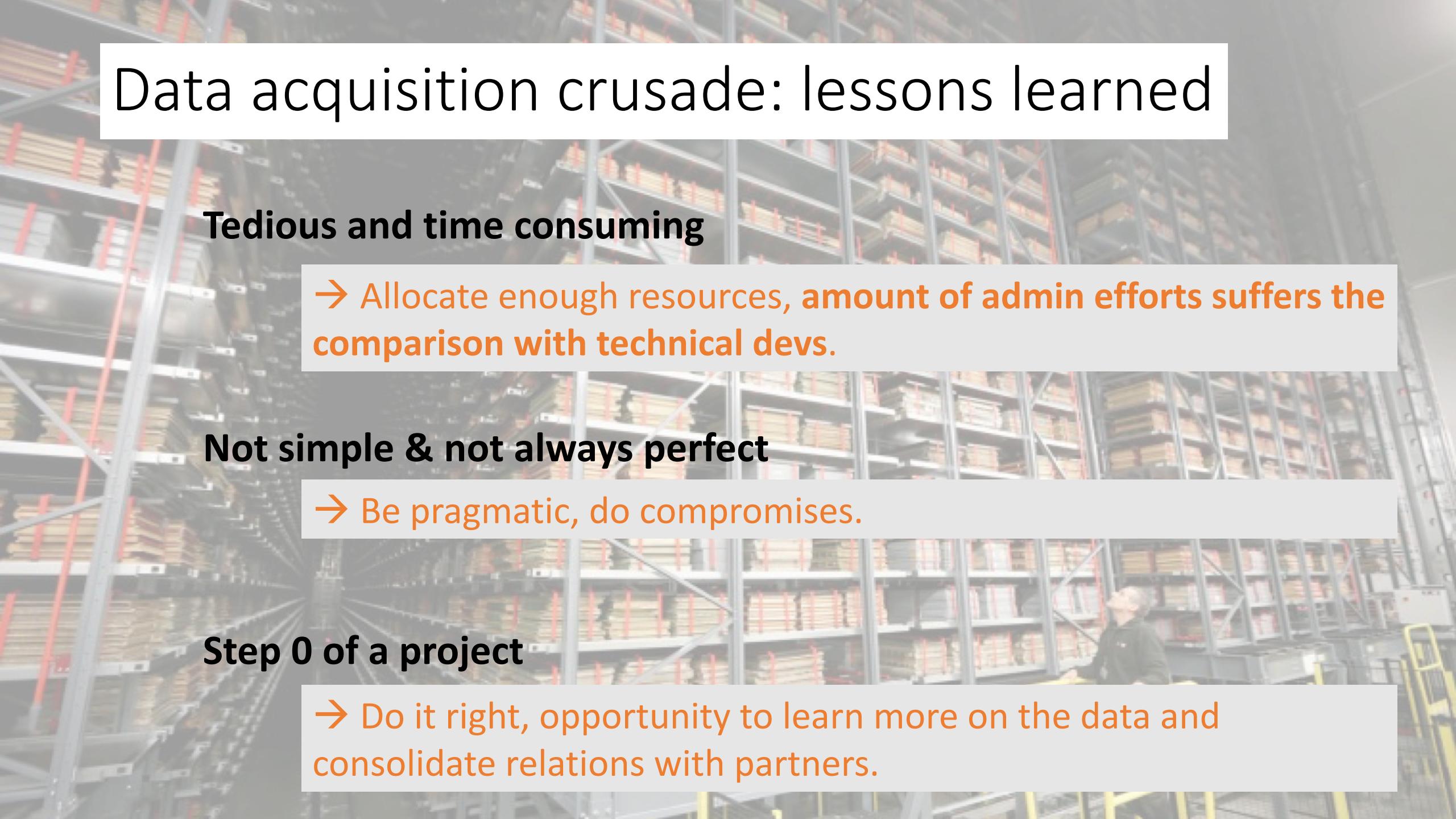
- robotics and politics
- hybrid search strategy
- best argument: demonstrating the potential of digitized newspapers

N E W S
E  E



Living with machines

Data acquisition crusade: lessons learned

A large warehouse with tall metal shelving units filled with boxes and books. A person is visible in the background, looking up at the shelves.

Tedious and time consuming

→ Allocate enough resources, amount of admin efforts suffers the comparison with technical devs.

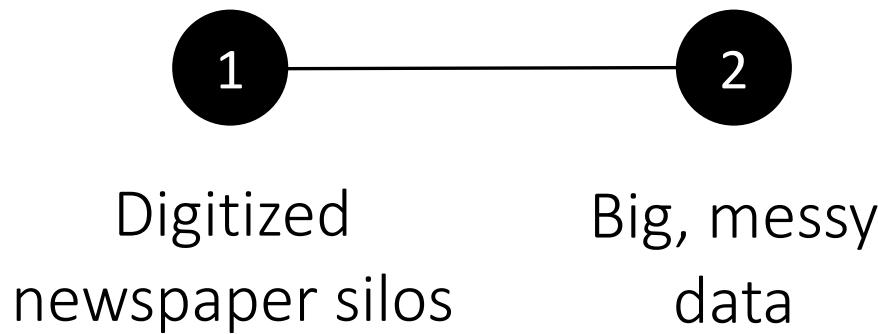
Not simple & not always perfect

→ Be pragmatic, do compromises.

Step 0 of a project

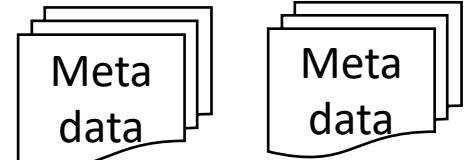
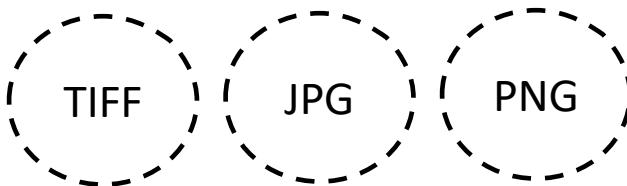
→ Do it right, opportunity to learn more on the data and consolidate relations with partners.

How to enable semantic indexing and exploration of large collections of historic newspapers?



What happens to digitized items?

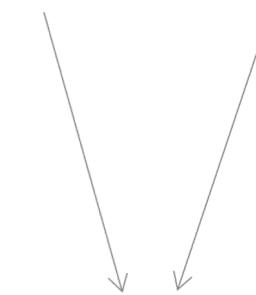
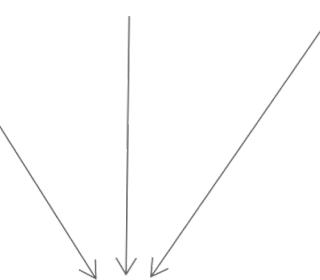
*digital
archives*



*processable
data*

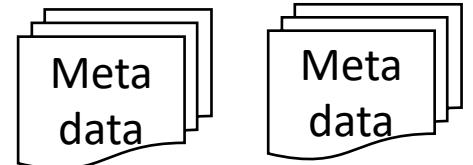
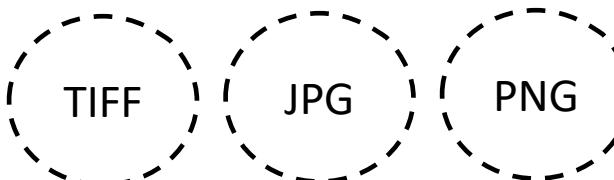
{ JSON }

JPEG2000



What happens to digitized items?

*digital
archives*



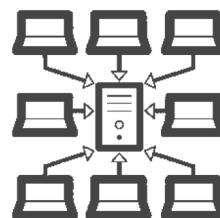
1 year: 25GB

*processable
data*



{ JSON }

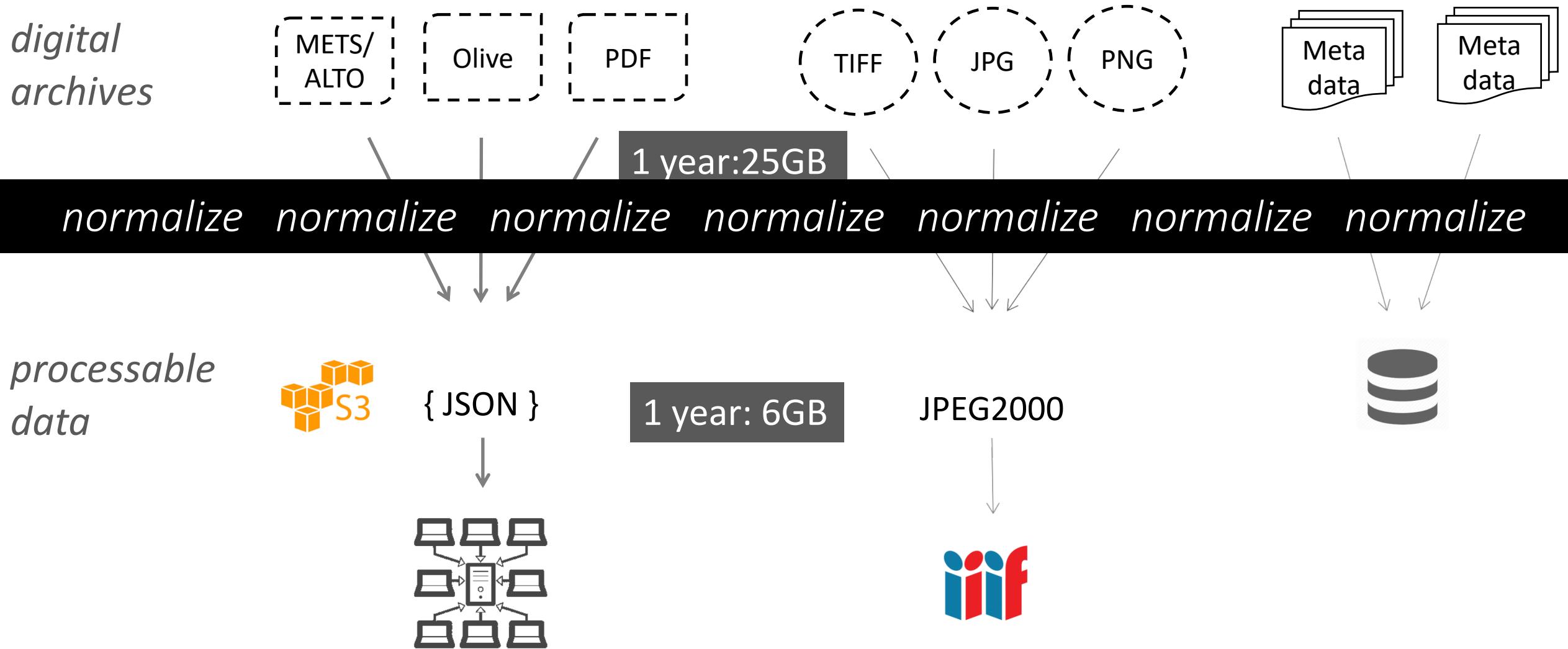
1 year: 6GB



JPEG2000



What happens to digitized items?



Lost in translation?

*digital
archives*



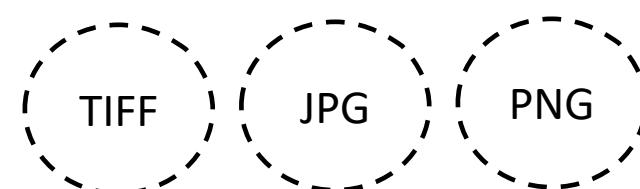
empty folders

corrupted (zip) archives

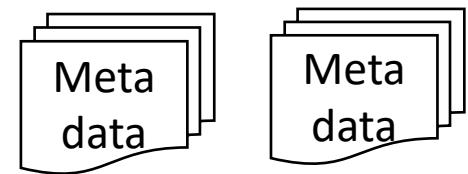
inconsistencies

{ JSON }

*processable
data*



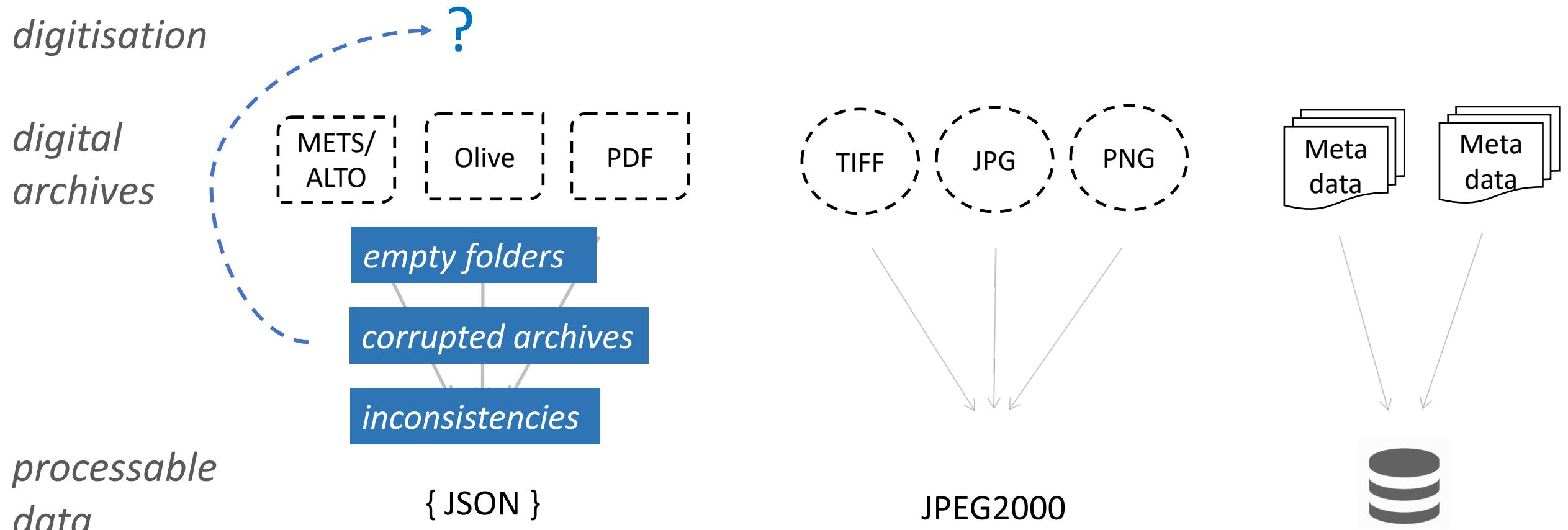
JPEG2000



Inconsistencies

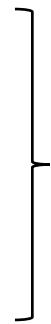
- mixed up page order
- empty text field (ca. 3,3%)
- duplicated text
- wrong language tag (ca 0.1%)

Inputs



Archives holes

1. publication periodicity



⇒ historical knowledge, often encoded as metadata

2. publication interruption

3. paper copies lost

⇒ preservation history knowledge, rarely encoded as metadata

4. digital copies lost or damaged

⇒ requires digital checks, almost never encoded as metadata

Actually, holes and imperfections matter

```
# get zip archive and check it
archive = os.path.join(issue_dir.path, "Document.zip")
if not os.path.isfile(archive):
    logger.info(f"No Document.zip in issue {issue_dir.path}. Skipping it.")
    return
else:
    try:
        working_archive = zipfile.ZipFile(archive)
    except zipfile.BadZipfile as e:
        logger.info(f"Corrupted Document.zip in issue {issue_dir.path}. Skipping it.")
    return
```

Count, Store, Compare!

Lost in translation?

*digital
archives*



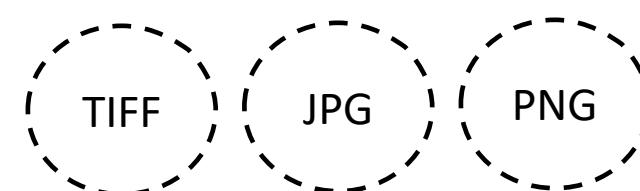
empty folders

corrupted archives

inconsistencies

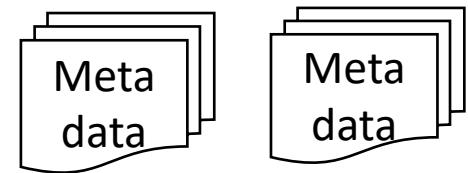
{ JSON }

*processable
data*



box coordinates

JPEG2000



Nightmare

Lost in translation?

*digital
archives*



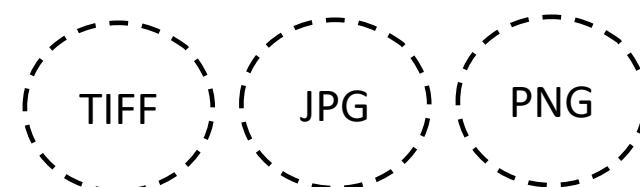
empty folders

corrupted archives

inconsistencies

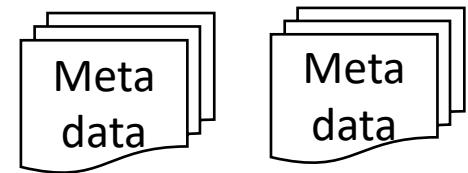
{ JSON }

*processable
data*



box coordinates

JPEG2000

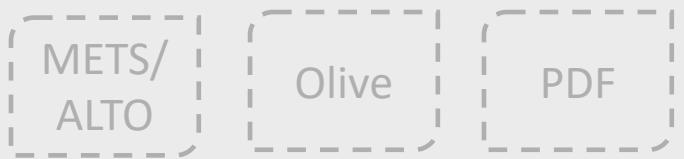


*partial, non
normalized info*



Lost in translation?

*digital
archives*



empty folders

corrupted archives

inconsistencies

*processable
data*

Harmonization and enrichment needed



box coordinates

Quotidien: 17
Wöchentl.: 10
Hebdomadaire: 2
2x par semaine: 4
3x par semaine: 1
2x wöchentl.: 6
Tägl.: 6
Settimanale: 4
3x par semaine,: 2
6x par semaine: 3
Bi-hebdomadaire: 1
quotidien,: 1
trois fois par semaine,: 2
Quotidien,: 1
quotidien: 2
Wöchentl.,: 1
Mensile: 1
2x all'anno,: 1
Quotidiano,: 2
Quotidiano: 1
12x all'anno: 1
3x la settimana: 1
Halbwöchentl.: 4
3x wöchentl.: 1
Ca. alle 1-2 Jahre: 1
2x monatl.: 1
2-3x wöchentl.: 1
6x wöchentl.: 2
2x wöchentl.: 1
Annuel: 2
Ca. 6x jährl.: 1
Jährl.: 2



*, non
normalized info*



Overview of the *impresso* corpus

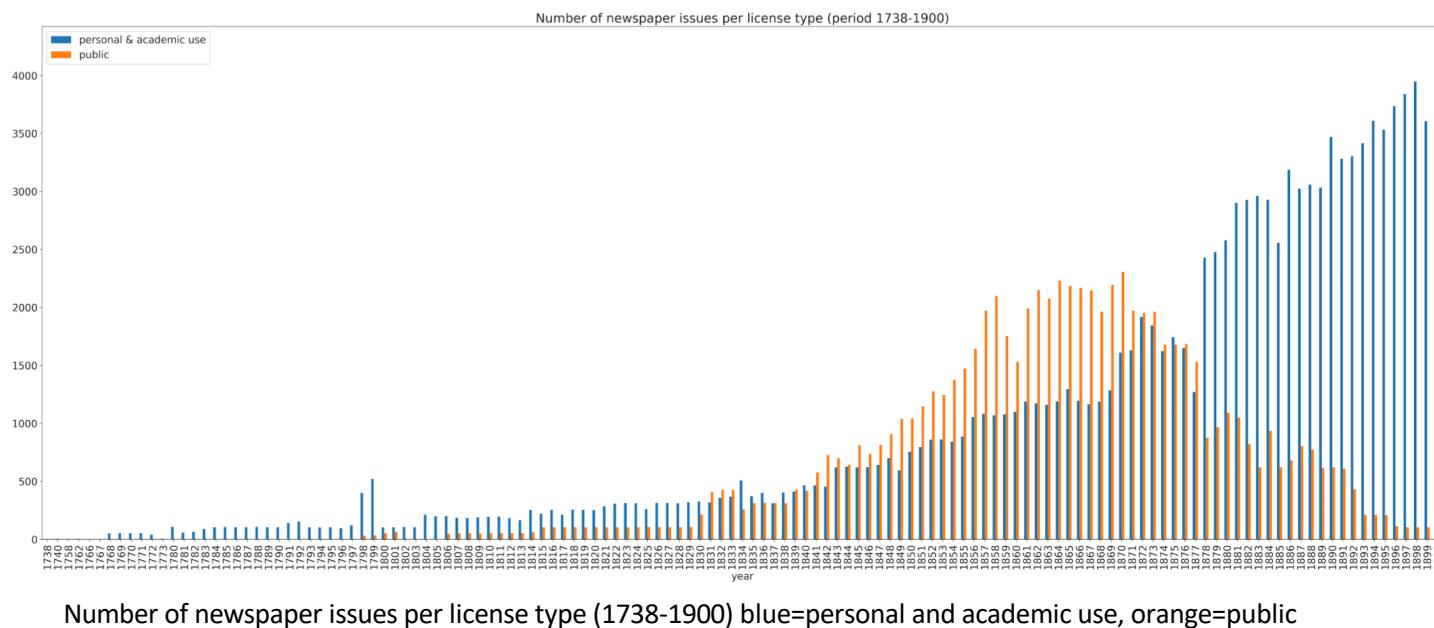
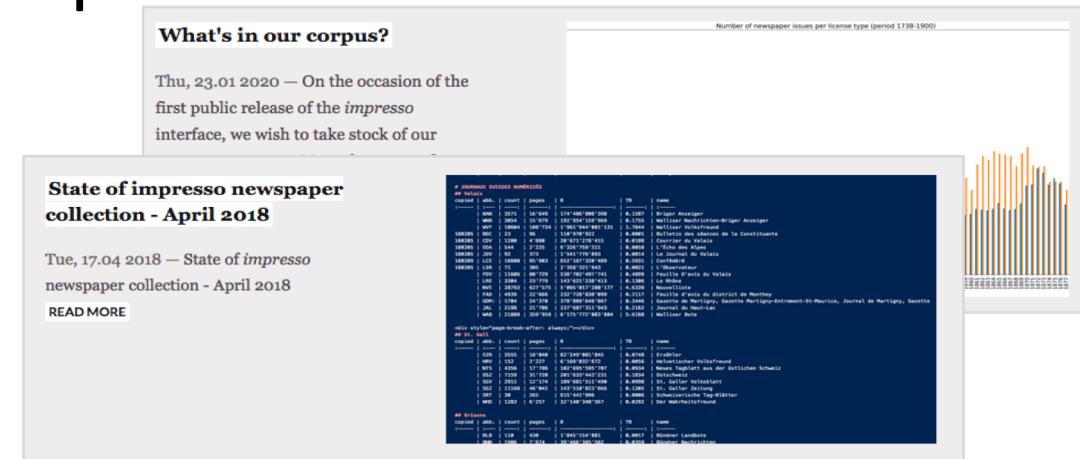
Growing throughout the project:
51 (2018), 76 (2019), 93 (2021) titles

Upcoming release:

- o CH, LU & FR newspapers
- o 58M ‘content items’
- o 15 billion words (token)

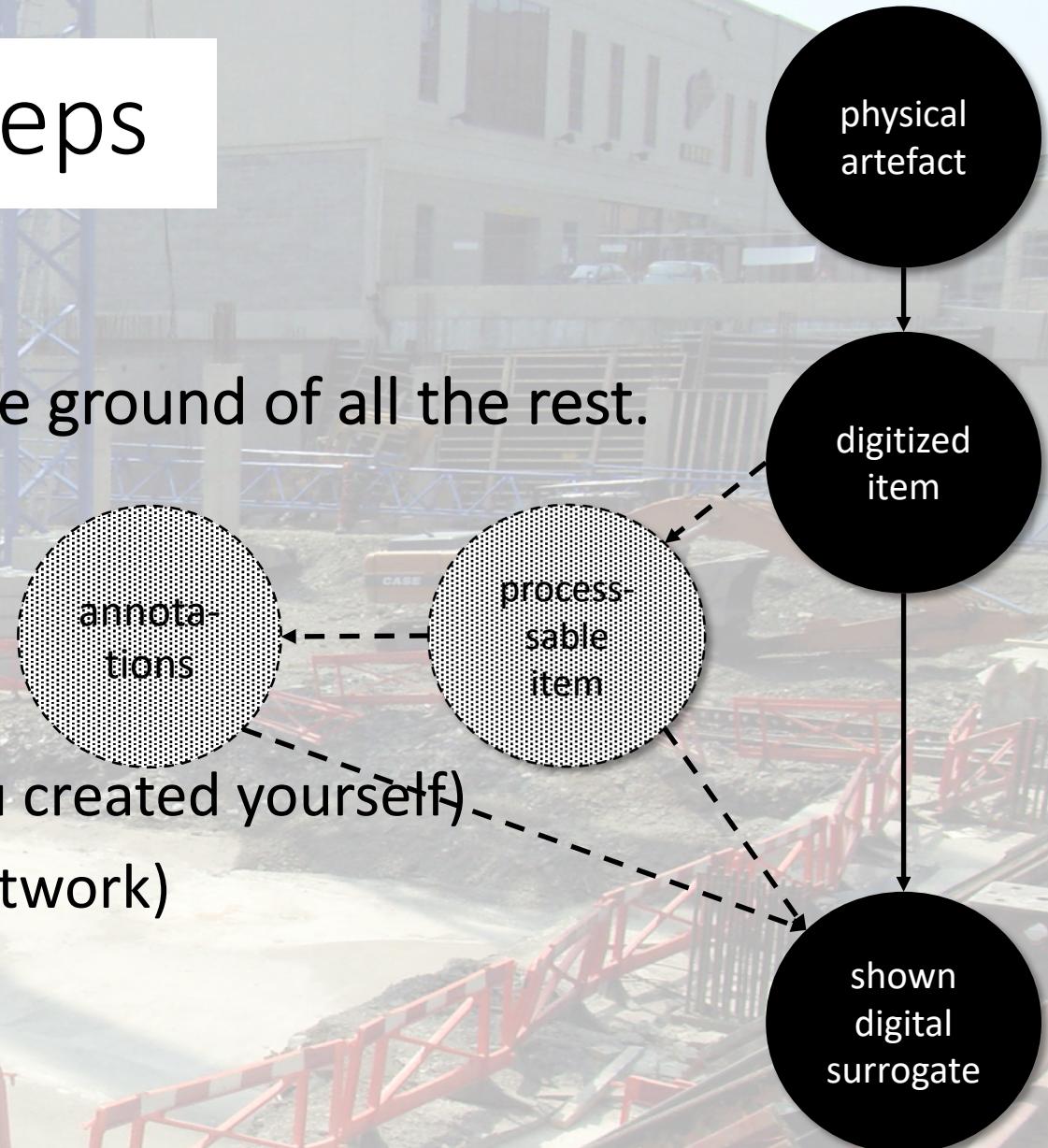
Different rights statements regimens:

- o public domain
- o private and academic use



A whole world of hidden steps

- Laborious, usually invisible work which lay the ground of all the rest.
- Caring about quality in big data context is not easy:
 - do not trust any data (not even the ones you created yourself)
 - do not trust any infrastructure (S3 pb, IO, network)
 - forget about all your assumptions
 - think big, but triple-check the tiny details

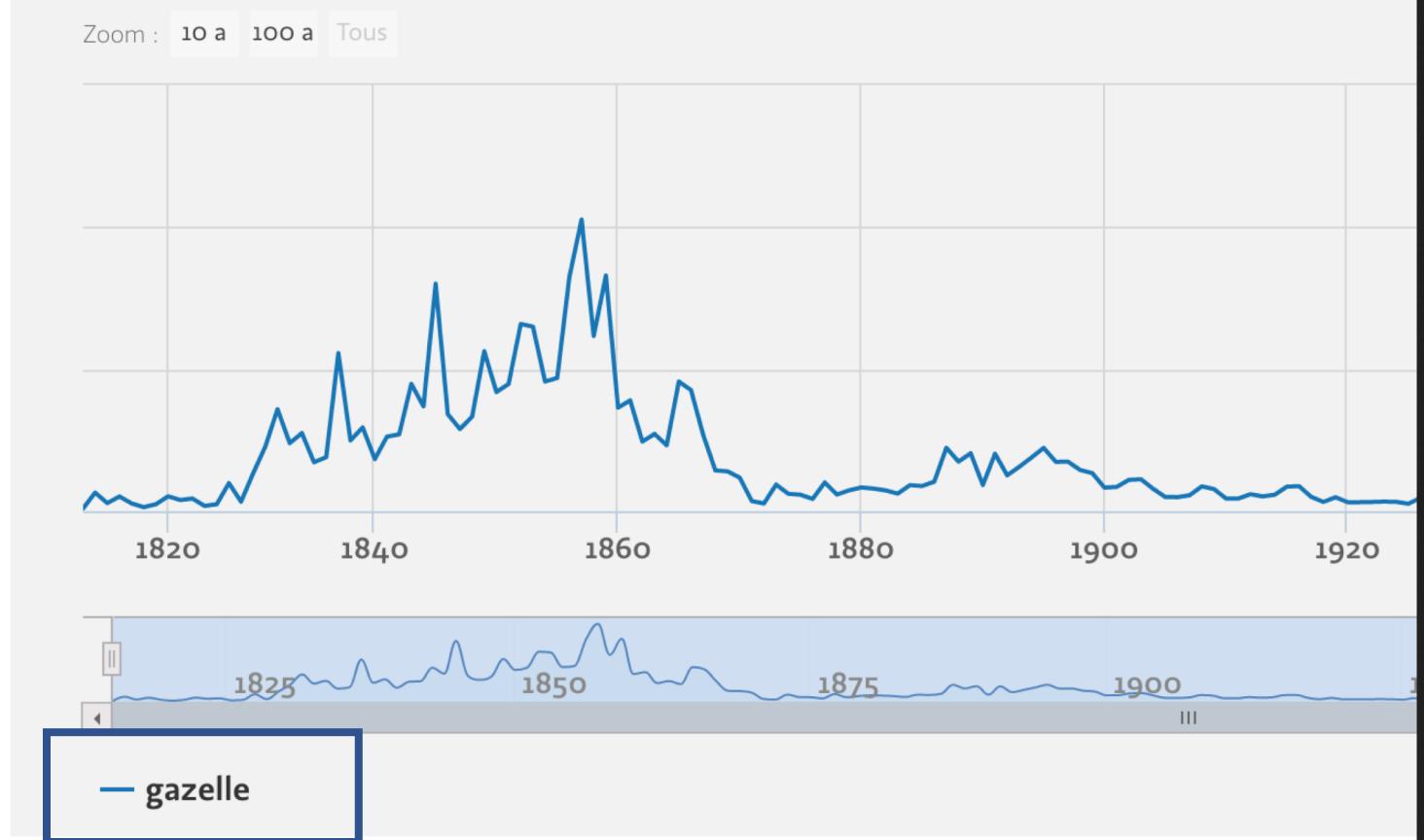


→ We should aim for full tracing of meta-knowledge across actors.

How to enable semantic indexing and exploration of large collections of historic newspapers?



Gazelles in Swiss newspapers



PRIX D'ABONNEMENT / JEURES : Fr. 95.—; FRANCE ET COLONIES : Fr. 95.—; 8 francs, Fr. 15.—; 8 francs, Fr. 2.—; AUTRES PAYS : Fr. 95.—; 8 francs, Fr. 15.—; 8 francs, Fr. 2.—; 8 francs, Fr. 10.—; 8 francs, Fr. 2.— (exempté).

PRIX DES ANNONCES : (C'est à dire déduction) Abonnements, Fr. 60. au min. de tant que leur durée, 80 francs. Membre, 70 francs. Fr. 50. au min. de tant que leur durée, 70 francs.

Compte de chèques IL234

N° 158. — Mardi 10 juin 1930.

REGIE
DES ANNONCES:
SOCIETE AUSTRO-ALPINE DE PUBLICITES
M. PICHAT
et correspondants en Suisse et à l'étranger.

Le Petit Parisien

On a dit dans plusieurs journaux que le mouvement des personnes appartenant aux différentes catégories d'assurances était en augmentation. L'assurance-vieillesse est l'assurance la plus importante.

Ce mouvement a été accéléré par la crise mondiale. Il a été stimulé par le système d'assurance-vieillesse qui existe dans l'Europe, et spécialement en France.

Le système d'assurance-vieillesse a été mis en place par la législation française en 1919. Il a été étendu à toute la population par la loi de 1921. Depuis cette loi, l'assurance-vieillesse a connu une croissance continue.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

L'assurance-vieillesse régulière est une assurance qui couvre toute la vie. L'assurance-vieillesse progressive est une assurance qui couvre une partie de la vie.

La France a mis en place un système d'assurance-vieillesse qui couvre toutes les personnes qui ont atteint l'âge de 50 ans.

Les personnes qui sont assurées contre l'assurance-vieillesse régulière doivent verser une cotisation régulière.

Les personnes qui sont assurées contre l'assurance-vieillesse progressive doivent verser une cotisation progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

Il y a deux types d'assurance-vieillesse : l'assurance-vieillesse régulière et l'assurance-vieillesse progressive.

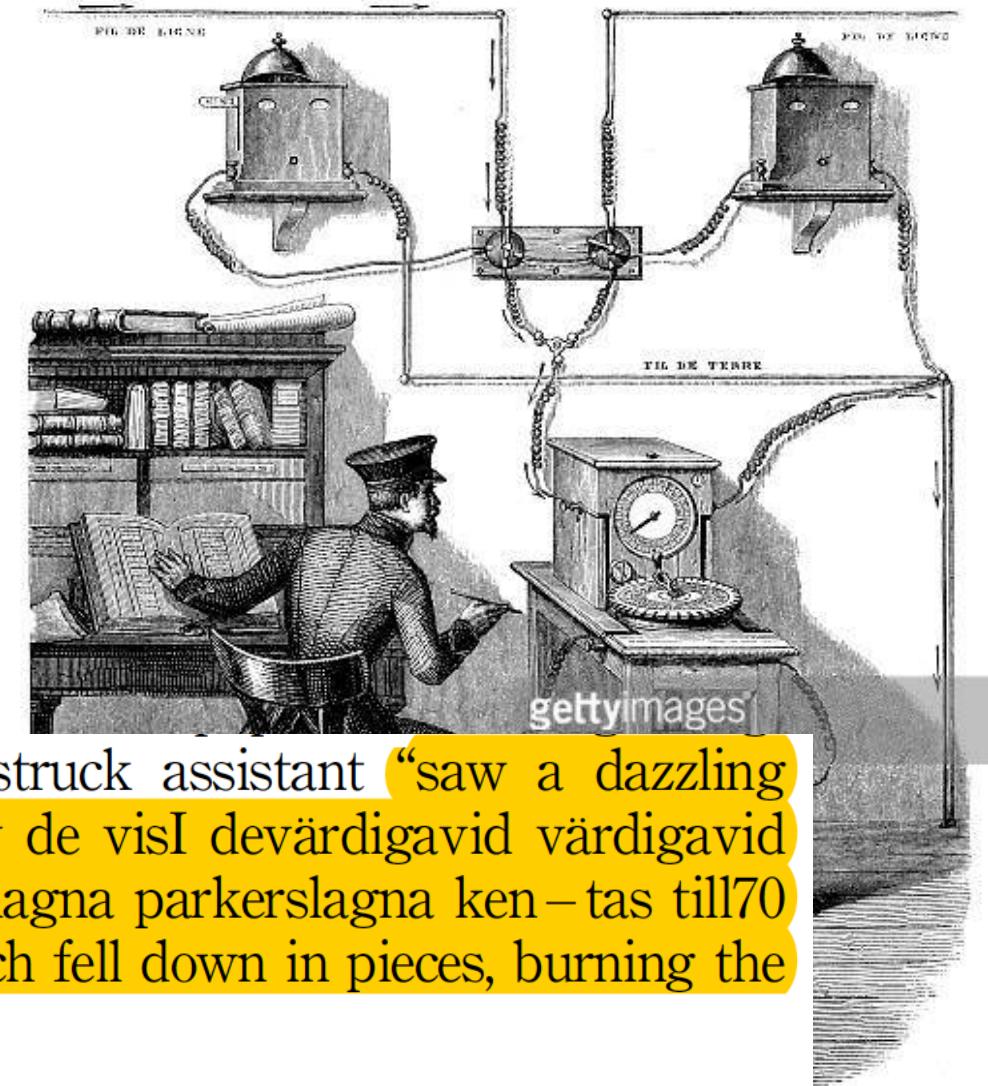
Invented data?

Cultural heritage as digital noise: nineteenth century newspapers in the digital archive

2016

Johan Jarlbrink and Pelle Snickars

telegraph incident, in fact, literally reported that the struck assistant “saw a dazzling light along the wires on the walls conducting electricity de visI devärdigavid värdigavid dejemte fullkommen ihåförvintparkerslagna förvintparkerslagna parkerslagna ken – tas till 70 70 misvärt fruktarsnart tAf eoch sisrans njes ej [...] which fell down in pieces, burning the table and the floor.”



Never printed, but today part of the historical record.

Quality of Images and Original OCR by NZZ

The image displays three rectangular newspaper snippets from the mid-19th century, arranged horizontally. The left snippet is from the 'Zürcher Zeitung' (No. 4), featuring a horse and rider illustration. The middle snippet is from the 'Zürcher Zeitung' (No. 37), dated March 7, 1851. The right snippet is from the 'Neue Zürcher-Zeitung' (No. 89), dated March 29, 1852.

Text from the left snippet:

Nº. 4.
Zürcher
Dienstag.

Text from the middle snippet:

Nº. 37. 1851
Zürcher-Zeitung
Dienstag, den 7. Ma

Text from the right snippet:

Bescheinigung:
Grauds und Co. ganz
Sämtliche
für 10. Mandate
Reu-Banden u.
Bündnungsgeheimnis:
Der Raum einer Reihe
Viele u. Reu-Rappen.
Welt w. Gott fratio.
Montag

Nr. 89.
Neue Zürcher-Zeitung.
Dreiunddreißigster Jahrgang.
Montag den 29. März 1852.

Common text across all snippets:

sagte mir: .Ha. die hauen eben alle ab; ich weide auch bald
vrilchwinden!' ... Leipzig haltr das festlich!!,' «leih an,ir ll>;ss!. daH
es für die ?,>;!?!<;>; z«r Vrnlllnma, hallr, »nd c» war n!ch! schicchi
sscardoilrl. !!!il ütt'l , "fah»rx!»ch in diesel Farbe lax» man in den
Naufhäxlrlu eihalcx u»d viel OrN» ilxii da. Tie (<wststät!e!i
I^rndel?» nur^ nms sie zur U>;ellNs!»!!s>; lmtle», und l>;!r
Lebr,!?i»i!tell<;,!<;onl>;l für die 'lllcssl s!,<;fte l

Bottom captions (approximate translation):

Da die ganze Summe nur ungefähr 500000 zw.
beträgt, so könnte eine solche Kleinigkeit in den
zur Zahlung bestimmten Fonds für diese Zeit keine
sehr auffallende Veränderung verursachen. Die
Stände haben diese Gelegenheit mit Freude ers

höher wie ich. Dehungen wird Sr. Beliebten Macha
schof von
glaubte, e
stet leisten.
und der Um

Die Dinge werden sich, wie man sagt, ein
zößtöre Polnischer und Sachsischer
Offiziere und Gemeine, bis zur Zahl von 300 Kriegsge-
fangenen, auf ihr Ehrenwort bis zu ihrer Aussetzung nicht
zu dienen, zu überbringen. Ein sommerliches Ereignis,

Die Dinge werden sich, wie man sagt, ein
zößtöre Polnischer und Sachsischer
Machta und Waffen, um die Freiheit des Landes zu ver-
tragen. Den Besitztöten, Schlosser und Kompanien, welche die
Befreiung der Abhängigen des Königs im Wege stehen, auf die
Rückführung des Erbfolgers zu unverzüglich, sobald man, nach den
gewölbigen Augenblick abgespielt, wenn man es will, in der
Festen beschäftigt, entstehen kann. Kompanien zu den drei Un-
tergebundenen: Befreiung, die mögl. Kompagnie, welche
Möglichkeit sie gewollt, dass sie, während sie den Dienst, welche
Sache nur zuvinzen oder wenigerden den Erfolg auf Seite des

Politischen Geschehens, ausführlich darstellen. Sie ist gleichzeitig
die Entfernung der Abhängigen des Königs im Wege stehen, auf die
Rückführung des Erbfolgers zu unverzüglich, sobald man, nach den
gewölbigen Augenblick abgespielt, wenn man es will, in der
Festen beschäftigt, entstehen kann. Kompanien zu den drei Un-
tergebundenen: Befreiung, die mögl. Kompagnie, welche
Möglichkeit sie gewollt, dass sie, während sie den Dienst, welche
Sache nur zuvinzen oder wenigerden den Erfolg auf Seite des

Can we do better?

OCR-2005

N °. 86 . Zürcher Samstag , t
793 & deN26 .
Weinmon . MMMM ? urck da
« franiöische Voll Gouverain
geworden , mit dem 3 « nnte
in volitischen Votis » . »
« ...nntn sen ; «
... siecht zufrieden damit , im
Empersts " ; t

F-Measure: 0.147

OCR-2017

Prozeß der Marie Antoinette.
Nachdem dieselbe am 15. Weinm.
alten Styls, oder am
2. des ersten Monat» im 2, en
Jahre der Republik neuen
Kaleuderstpls, in den
Audienz

F-Measure: 0.684

HTR-2018

Prozeß der Marie Antoinette.
Nachdem dieselbe am 15. Weinm.
alten Styls, oder am 23. des ersten
Monats im 2ten Jahre der Republik
neuen Kalenderst...
Audi...
F-Measure: ~0.95
...lassen hatte, fragte sie der
Präsident: Wie sie heis- se? „ Ich
nenne mich, antwortete sie, Marie
Antoinette von Lotharingen-

OCR quality assessment

Goal: allow users to **estimate** the completeness of their keyword search results

Principle: OCR quality \cong ratio of known words compared to all words

Measure:

- **focus** on content words (exit stopwords and punctuation)
- **easy** to grasp by non-technicians

Example – Quality = 10/13 = 0.77

Il en est de même de la constitution d'Argovie de 1841 qui n'a pas encore été garantie

Original	Il en est de mèmede la constitution d'Àrgovie de 1841 qui n'a pas encore été garantie
Normalized	il en est de memede la constitution d argovie de 0000 qui n a pas encore ete garantie
Categorized	.. . *** .. MEMEDE .. constitution . argovie .. **** *** .. *** encore *** garantie

Lucerne, Uri, Schwytz, Unterwald, Zug, Fribourg et Appenzell-intérieur.

Original	Lucerne, Uri, Schwytz, Unterwald, Zug, Fribourg et Appenzell-intérieur.
Normalized	lucerne uri schwytz unterwald zug fribourg et appenzell inlierieur
Categorized	LUCCRNE uri schwytz unterwald zug fribourg .. appenzell INLERIEUR.

. * = masked character of short or frequent word / RED = unknown word / green = known word

Named entity processing

Referential units which underlie the semantics of texts.

(5Ws:Who, What, Where, When, Why)



Proper names, definite descriptions generally of type **Person, Organization, Location**.

Ca. **30%** of content-bearing words in news texts are proper names.

Named entity processing refers to a family of tasks that are part of the larger domain of **Information Extraction**. IE has been defined and formalized in the **90's** with the **Message Understanding Conferences** (MUC) organized by the **US DARPA**. Besides the recognition of proper names of types e.g. **PERS**, **ORG** and **LOC**, named entity processing also implies **disambiguation** and **relation detection**.

PERSON, ORGANISATION, LOCATION, EVENT, DATE, NLP-TASK

Good performances when:

- English
- News, generic domain
- Simple typology

Challenges on historical texts

Suffers from:

- bad OCR

Constap. iipopje, Buch irest, M'' Lucile

- language evolution

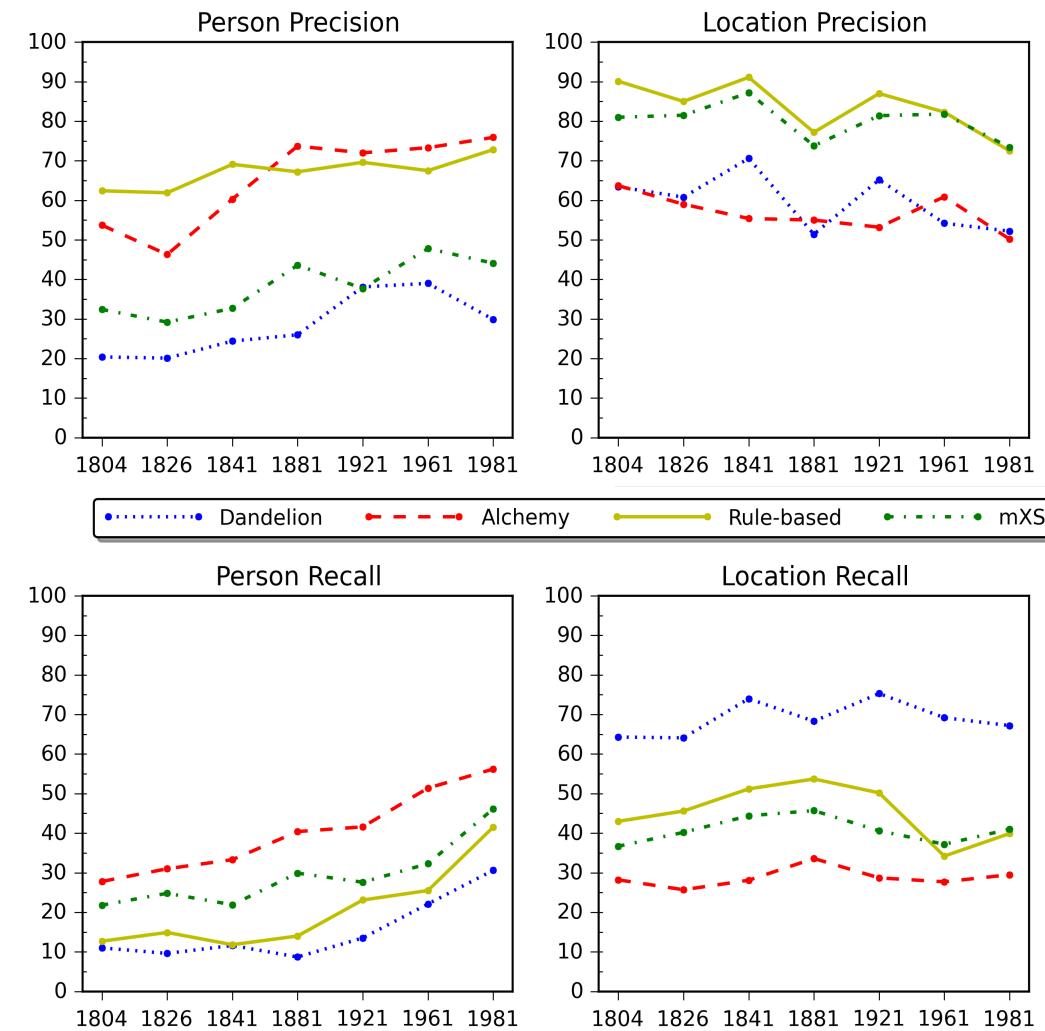
Härnevi → *Arnevi*, *Kallmar* → *Kalmar*

- Poor resource coverage

- non VIPs not recognized

- lack of appropriate trigger words

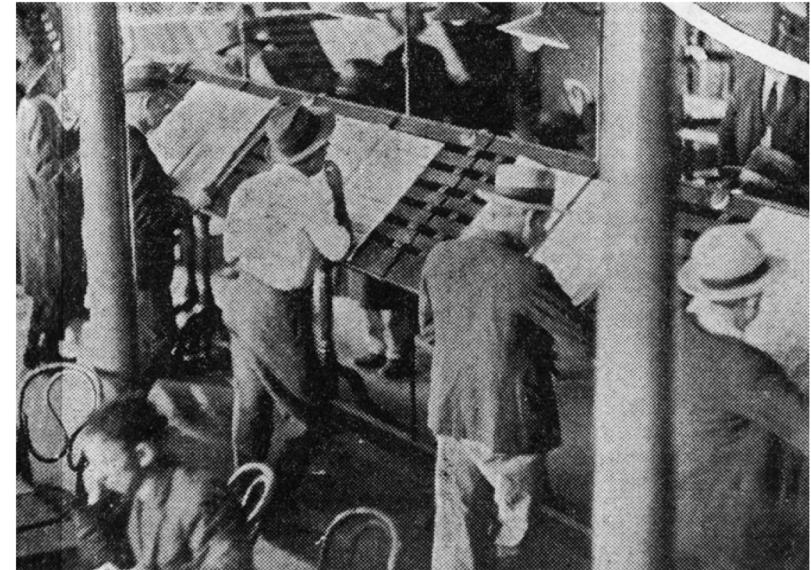
burgomaster, tailor, munition specialist



CLEF-HIPE 2020

- CLEF: Evaluation conference, “Olympic Games of Computational Linguists”
- HIPE (Identifying Historical People, Places and other Entities), “A specific discipline competition”, organized by *impresso*
- Teams try to develop the best system.

Named Entity Processing on Historical Newspapers



Objectives

- strengthen the **robustness** of approaches;
- enable **performance comparison**;
- foster **efficient semantic indexing** of digitized cultural heritage collections.

CLEF-HIPE-2020

Tasks:

- NE recognition and classification
- NE linking



CLEF-HIPE-2020

Tasks:

- NE recognition and classification
- NE linking

Newspapers from CH, LU, US from 1738 to 2019

Participation 13 teams



CLEF-HIPE-2020

Website: <https://impresso.github.io/CLEF-HIPE-2020/>
Overviewpaper: <https://infoscience.epfl.ch/record/281054?ln=en>
Participant papers: <http://ceur-ws.org/Vol-2696/>

Tasks:

- NE recognition and classification
- NE linking

Newspapers from CH, LU, US from 1738 to 2019

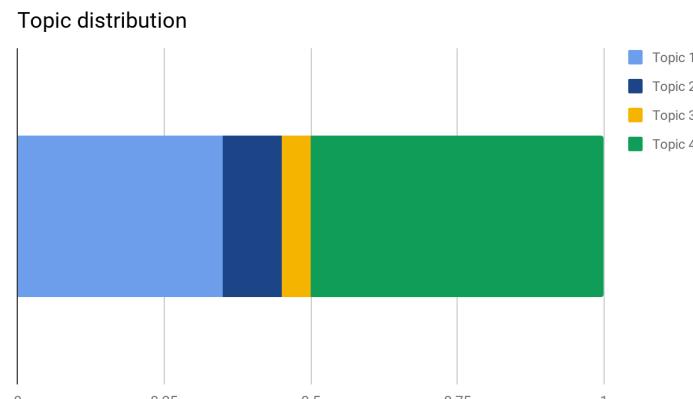
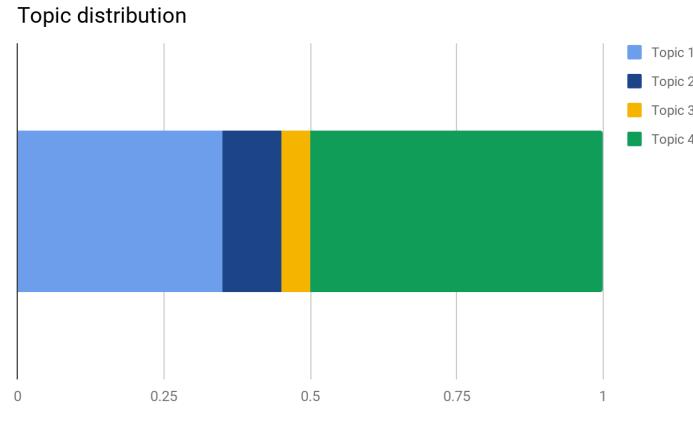
Participation 13 teams

Main outcomes:

- Contextualized neural embeddings (BERT) are strong
- Careful preprocessing helps to achieve a performance not far behind NER/EL on modern newspapers.



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

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

OCR errors dominate topics

13/50

drn verde riner drs nnr
lourde fierer snr znm dri
vrr wnrde tir zürcher sf
hauen bri jür eiue

lil eilen lul nil lin ver
uni llr lullen lun lei van
vol llo vel lol vll ren
lesen lit

nicken bitten vnn icl eit
lill nicht gen liegen clcr
erden essen gen sel st⁺ bereit cke
liegen nde lei chen seln
eilen wild ehren slch
folgen lan schl stellen
suchen eichen

vita dtt chm dje dew dep
ftn stm dtm mem hät ftm
dtr ßch mam
stch

unc clie icl ncl iic ssc unci
icr clen vnn clc cll lcl icli
unss llc nci clic ncni

nicken bitten vnn icl eit
lill nicht gen liegen clcr
die klein cit bereit cke
an immer ckcn

deu leinen tte sep dtt
sev deutsch teu oct
dienstag wey ttt ini ftp
freytag dep beu bep
neapel ste

auk äis äor sen nnä äio
aut virä inä aktie van
veräen obligation versen
it bitten tur vsr oktober
äon

werben imb beten beii
hor finb biefer alt beut
önnfi bieten bafi hub
u bcii bcs bitten jii o

dar lir iiiii iit leihen aii
fiir iri lil tal ifl im
helfen hui iili iti ifi iif
uum lullen

bass werben wild leinen
iahen mehl enn übel wal
fehlen non tann fühlen
zul nui zui übn hatt ihi
wied

Long story short

A LOT of preprocessing:

- OCR error (post-)correction
- old spelling normalization
- lemmatization

Current models in the interface:

- one per language
- only nouns
- only known lemmas
- elimination of rare and very frequent nouns

Text reuse

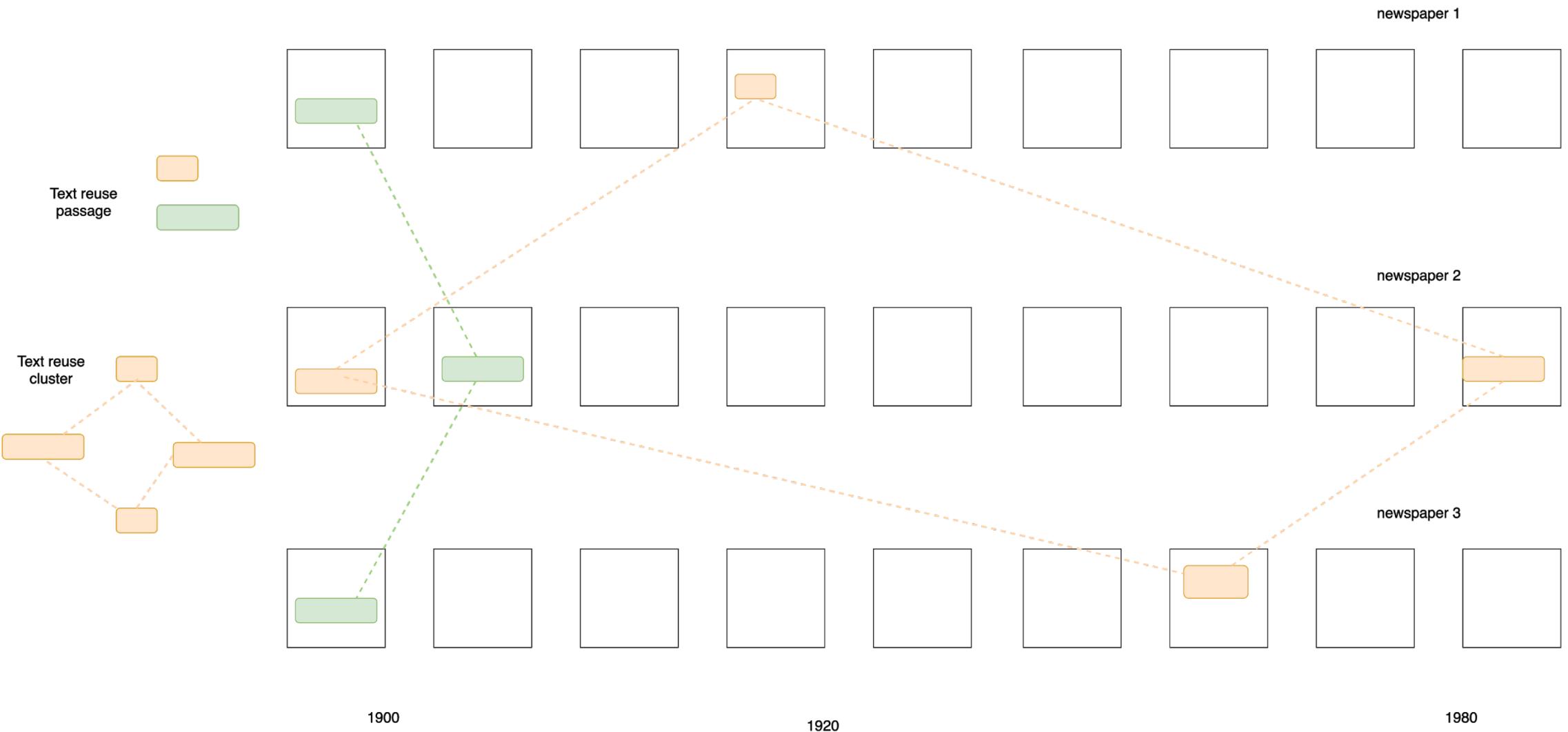
*The meaningful reiteration
of text, usually beyond the
simple repetition of
common language.*

Academic writing → quotation,
plagiarism

Literature → allusion, paraphrase,
quotation (intertextuality)

Newspapers → copy/paste
journalism, republication of news
articles

Text reuse passages and clusters



Passim – output examples

GDL 03/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au bateau à vapeur parti d'Ouchy pour Genève à 2 heures 20 minutes. Le bateau était arrêté pour le débarquement et l'embarquement devant Nyon, lorsque tout à coup on entendit une détonation, un nuage de fumée (ou de vapeur ?) sortit de la machine et le bâtiment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rupture qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soit, le bateau était mis dans l'impossibilité de continuer sa route ; heureusement les voyageurs qui devaient aller plus loin ont pu prendre le train qui passe à Nyon à 5 heures 3 minutes et sont arrivés à bon port, sans autre mal qu'un moment de frayeur. On

JDG 05/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au Guillaume-Tilt, parti d'Ouchy pour Genève à 2 heures 20 minutes. Le bateau était arrêté Pour le débarquement devant Nyou, lorsque tout à coup on entendit une détonation, un nuage de fumée (ou de vapeur ?) sortit de la hiadiine, et le hû liment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rup'" re qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soit, le bateau était mis dans l'impostJuilite de continuer sa route ; h'ureuiipnient les v < y 'geursqui devaient aller plus l <> iri ont pu piendre le train qui passe à Nyon à 5 heures 3 minutes, et sont arrivés à bon poil, sans aulre mal qu'un moment de frayeur. Le



Noisy, historical texts: recap'

Challenging. NLP's legacy is orthogonal to what archives/libraries offer today. Huge time and domain adaptation problem.

Benchmarking. But *to what extent?* Organizing a shared task to benchmark current approaches is a huge effort, but it helps advance research on this topic.

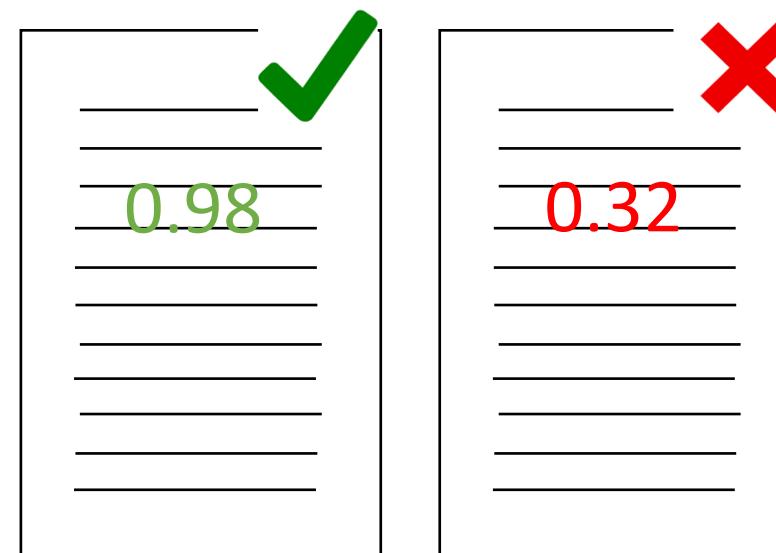
Critical text mining. The application of NLP methods and tools ought to be *functional* to the research questions and needs of historians.

Lexical
processing

language identification
historical spelling
normalization
lemmatization
word embeddings

OCR quality assessment

Lexical
processing



Lexical processing

OCR quality assessment

Named entity processing

Named entity processing refers to a family of tasks that are part of the larger domain of **Information Extraction**. IE has been defined and formalized in the **90's** with the **Message Understanding Conferences** (MUC) organized by the **US DARPA**. Besides the recognition of proper names of types e.g. **PERS**, **ORG** and **LOC**, named entity processing also implies **disambiguation** and **relation detection**.

PERSON, **ORGANISATION**, **LOCATION**, **EVENT**, **DATE**, **NLP-TASK**

Lexical processing

OCR quality assessment



Named entity processing

Topic modeling

kunst konzert bild theater
künstler oper musikalisch
spielen werk künstlerisch

deutschen britisch
deutsch truppe krieg
russisch angriff kampf
london

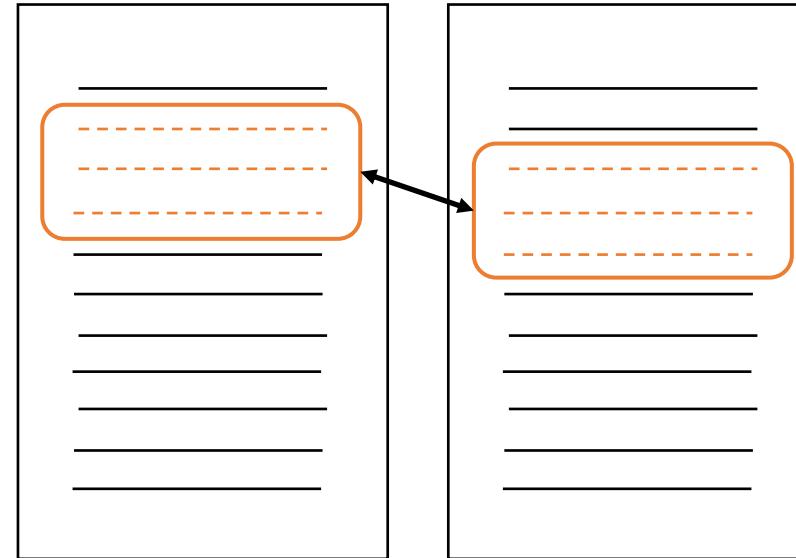
zimmer lage vermieten
verkaufen wohnung
garten groß auskunft

Lexical
processing

OCR quality
assessment

Named entity
processing

Topic
modeling



Text
re-use

Lexical processing

OCR quality assessment

Named entity processing

Topic modeling

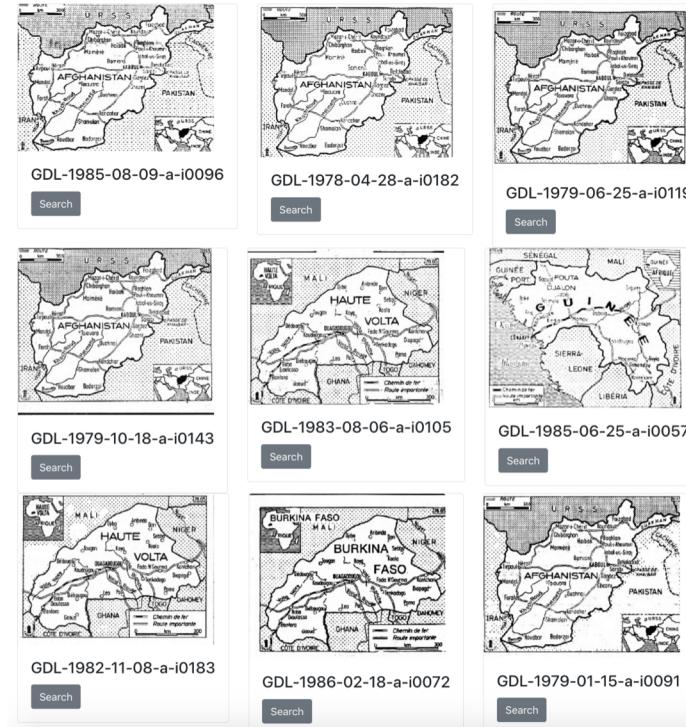


Image
similarity

Text
re-use

Lexical
processing

OCR quality
assessment

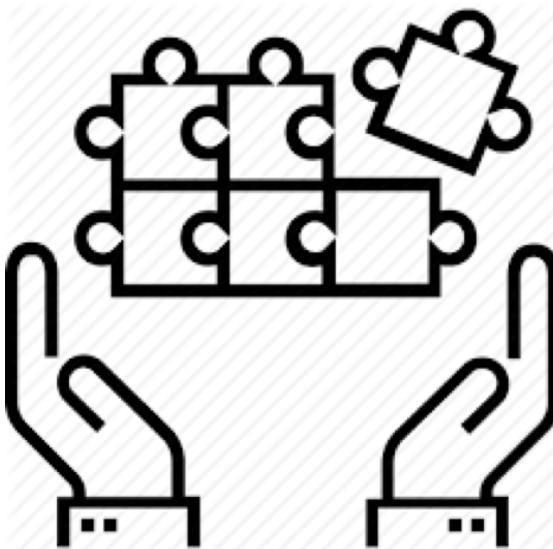
Recommender
system

Named entity
processing

Topic
modeling

Text
re-use

Image
similarity



Visit the impresso YT
channel to learn more
on each component

Named entity
processing

OCR quality
assessment

Topic
modeling

Lexical
processing

78 newspapers
47,876,994
articles
200M entities

Text
re-use

....

Recommender
system

Image
similarity

Named entity
processing

OCR quality
assessment

Topic
modeling

Lexical
processing

search for ...

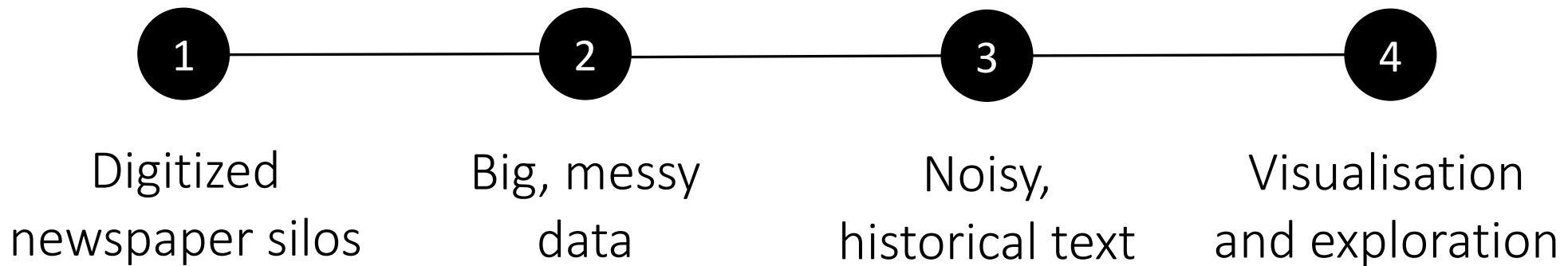


Text
re-use

Recommender
system

Image
similarity

How to enable semantic indexing and exploration of large collections of historic newspapers?



Six assessment criteria for digital scholarship

Source criticism (or *What am I looking at?*)

information on the corpus, on digitization, on OCR acquisition, etc.

Content search (or *How do I engage with the contents?*)

search on various type of (enriched) material

Content filtering (or *How do I select?*)

facets, result sorting, filtering

Generosity (or *How do I discover?*)

corpus presentation, result display modes, recommendation techniques

User interaction (or *How do I work?*)

personal work space functionalities

Connectivity (or *How do I go beyond?*)

interlinking at the level of metadata and contents

Main findings

4 generations of interfaces

1st generation: giving access - publishing keyword-searchable scans

2nd generation: user interactions - user collections, pdf download, store and share content, edit content

3rd generation: semantic enrichment - disambiguated named entities and crowd OCR correction

4th generation: transparency, recommendation

Features are sparsely covered

on average, interfaces cover 38% of the considered features

Best scores:

- keyword search
- viewer
- newspaper metadata

Lowest scores:

- enrichment
- information on collection and on digitization
- query autocomplete
- filter by semantic annotations
- personal account and user interaction
- connectivity

Open questions

Audience - How can we reconcile interfaces made for scholars vs. the general public? Should there be dedicated interfaces for each groups?

Complexity - Should all features and enrichments be visible and accessible?

Tool integration - Within or outside the interface ? Risk of complexifying the uses vs. an externalization via e.g. download functionalities?

What is at stake

How best accommodate text analysis research tools and their usage by humanities scholars?

How did we co-design
the interface



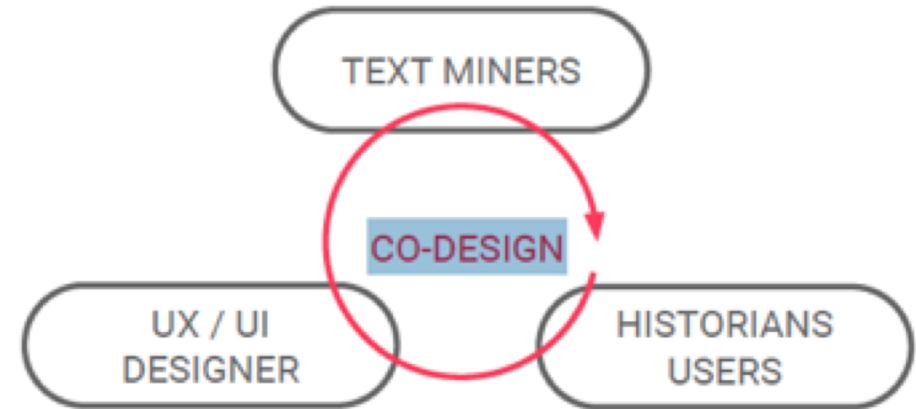
Co-design and fast prototyping

Principle

Continuous exchanges to learn from each other; experiment, evaluate

Implementation

- Workshops
- Associated researchers
- *Impresso* community calls
- One-to-one collaborations



Research scenarios

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

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



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 "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)
 - "alien" often mistaken for "[Woody] alien", "alien", "aliéné" and words with -alien/-alien (312 585 results on Retronews !)
 - "alien": English words with numerous irrelevant meanings

3

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.)
- 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.)

Use-Case for Impresso Project, Tobias von Wachter, 4th and 5th of July 2016

Universität Basel 6

4

Bridging the fields: Research scenarios

How do newspapers reflects the history of computing in Switzerland ?

5 mins

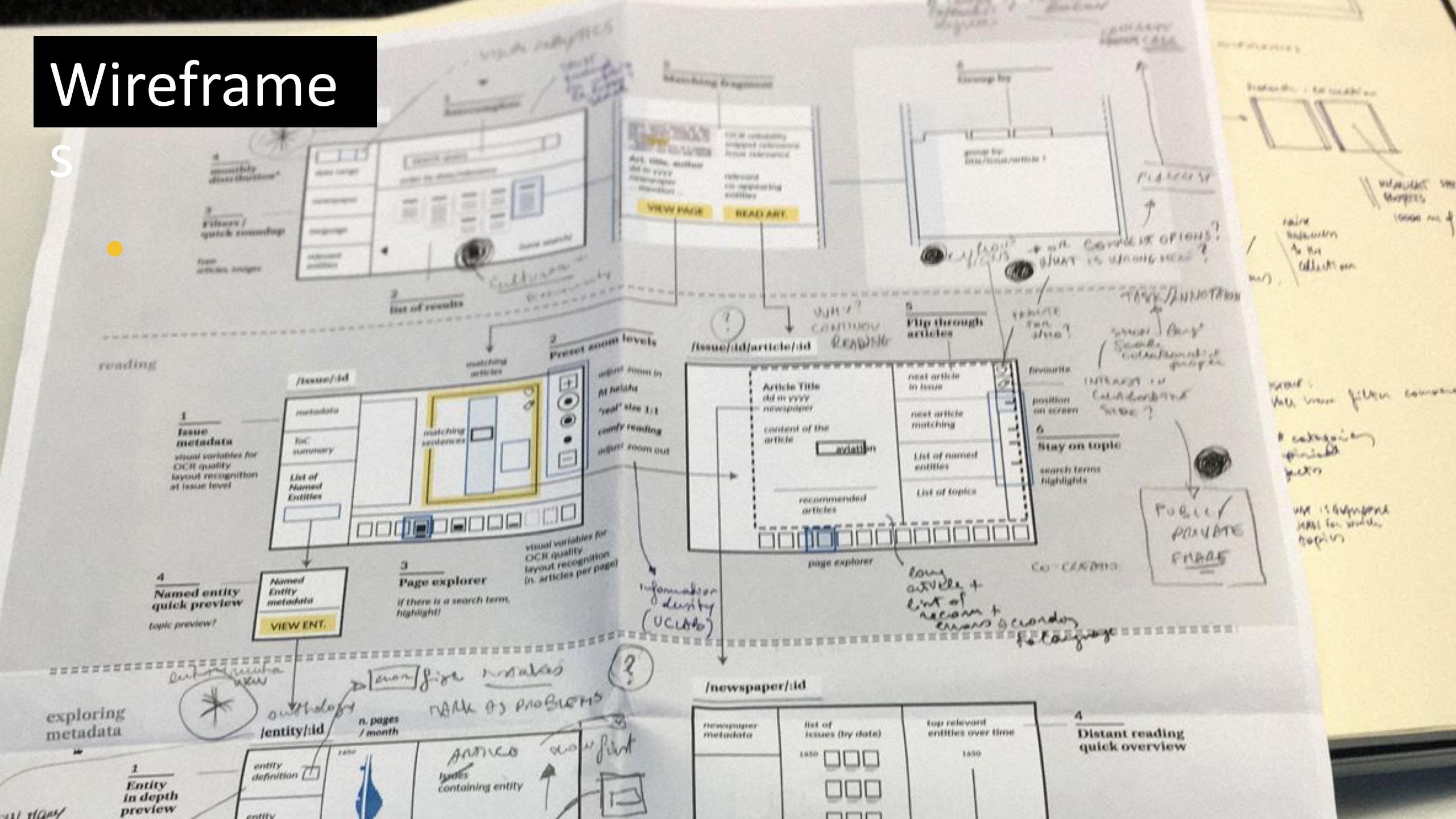
1. Research question
2. Operationalisation/tools
3. Expected results/difficulties

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

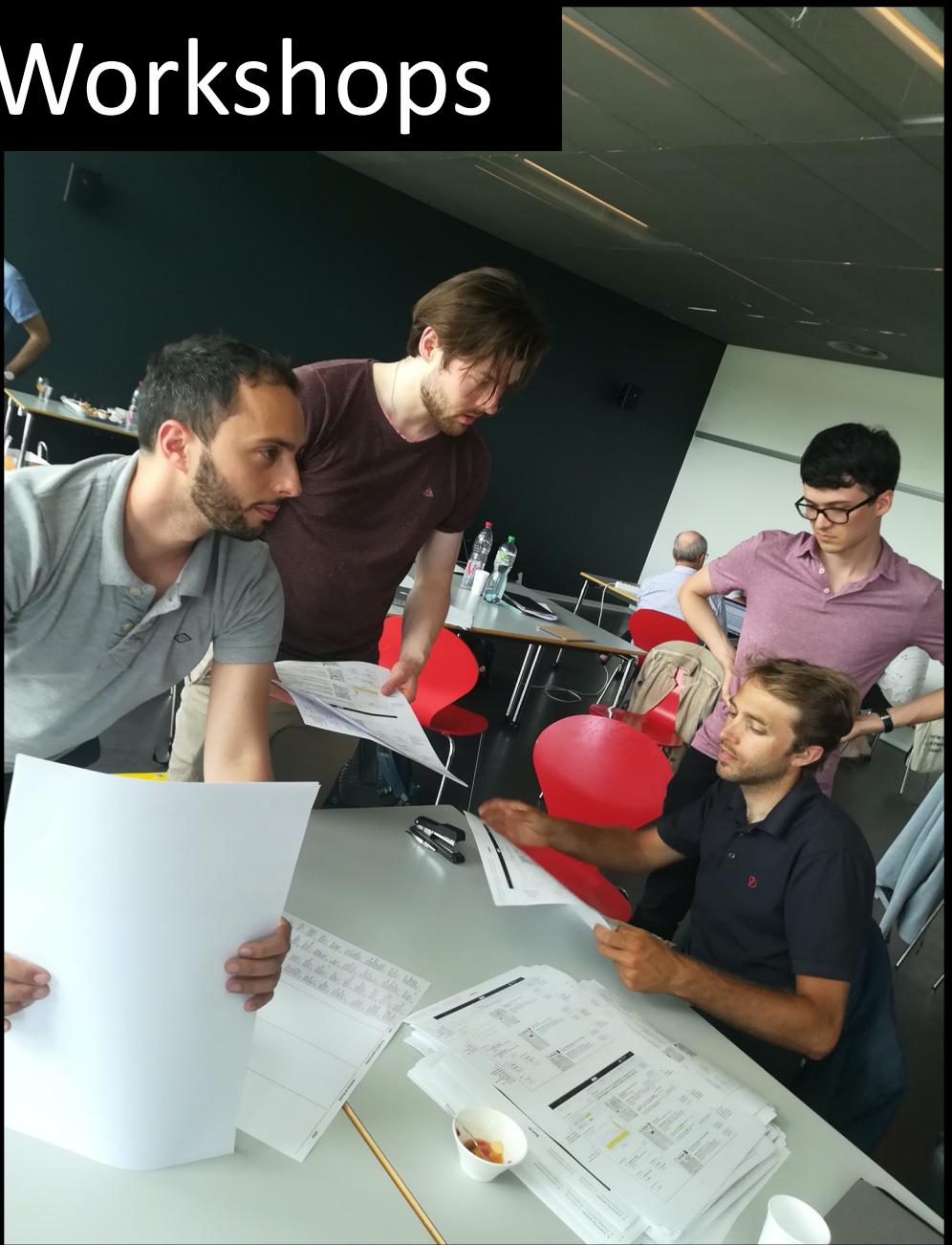
5

6

Wireframe



Workshops



s search for articles **c** collecting
R reading contents **O** organising

M exploring metadata
E visual experiments

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 ▾

YOUR SEARCH PAST SEARCHES

FULL TEXT SEARCH
... type a text or a date or to start ...

PUBLISHED IN (DATE RANGE)
TIMELINE OF
N. ARTICLES N. ISSUES

select a date range to show articles published

NEWSPAPER TITLES

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
MORE ...

112 partisans
103 Napoleon
50 satirique
MORE ...

TOP NAMED ENTITIES
112 French
MORE ...

LANGUAGE
112 French
MORE ...

Search

▼

GROUP RESULTS BY ISSUE PAGE ARTICLES SENTENCES

SEARCH SUMMARY
Find 53625987 articles in our collection.

1 of 112 **Les recherches d'UNI L'INFORMATION**
INTERNATIONAL NEWS

Gazette de Lausanne
Les recherches d'une... possibilités JEAN-PIERRE de multiples...

C.1 Suggest Public Tags (and vocabulary)

C.2 "Format": newspaper or specific page regions

C.3 Contents: related to article content

C.4 See and use other users' public tags

C.5 Explore espresso by tag (coming)

C.6 suggest new tag

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

C.8 suggest new description for the tag

C.9 swtich type of visualisation: list vs graph, vs alluvial diagrams ... Choose within a wide range of visual models

C.10 Visual feedback

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

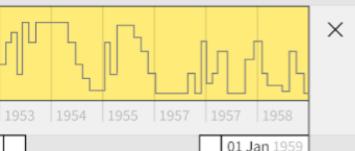
component index

code	Strength & opportunities	Weaknesses
R.4	<p>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)</p> <p>marginalia is great feature</p>	when you select "download as," can you specifically download the marginalia?
R.6	visualization of data can be very effective	alluvial diagrams are difficult to parse, even for data via professionals
S.1	search bar draws on common practice of searching for something, which most people are comfortable with	how else can the front page introduce users to the collection? e.g. featured collection, visual search? current search is designed for subject matter experts only
R.5	this feature is wonderful, saves time by being able to view articles without having to open them	overall, iif server loads somewhat slowly

Feedback

Conception of components

Discussions



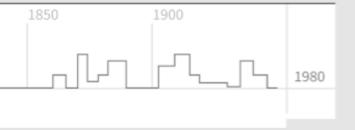
◆ article snippet XL

LAUSANNE

Lausanne — Les récits au caractère de l'actualité sont rares dans les journaux de Lausanne, mais certains d'entre eux sont tout à fait intéressants. C'est le cas du journal *Le Courrier de Lausanne*, qui publie régulièrement des articles sur la vie quotidienne à Lausanne, et l'actualité locale. Ces articles sont écrits par des journalistes locaux et sont très bien écrits. Ils couvrent tous les aspects de la vie quotidienne à Lausanne, de l'économie à la culture, en passant par la politique et l'actualité internationale.

EXPORT CITATIONS

DOWNLOAD AS ...



◆ article snippet XL bis

Gazette de Lausanne

Gazette de Lausanne — La Gazette de Lausanne est un journal suisse fondé en 1850. Il est édité par la Société de la Gazette de Lausanne. Le journal est connu pour ses articles bien écrits et son caractère très pratique. Il couvre tous les domaines de la vie quotidienne à Lausanne, de l'économie à la culture, en passant par la politique et l'actualité internationale.

EXPORT CITATIONS ...

DOWNLOAD AS ...



01 Jan 1959

◆ snippet search timeline

1890 | 310 ISSUES / 259 ARTICLES

SEARCH THIS YEAR

◆ named entity

Napoleon

Napoléon Bonaparte (French: [napoleōn bonapārte]; 15 August 1769 – 5 May 1821) was a French statesman and military leader.

1342 ARTICLES

◆ named entity bis

Pasquale Paoli

Nationalist Corsican leader.

156 ARTICLES

◆ snippet compare

Friday, April 27, 1951

FRANCO PRÉFÈRE LE TÊTE-A-TÈTE AVEC L'AMÉRIQUE**Gazette de Lausanne**

FRANCO PRÉFÈRE LE TÊTE-A-TÈTE AVEC L'AMÉRIQUE SOUS LE PRINTEMPS D'ESPAGNE ATTENDANT UN TARDIF REPENTIR FRANCO-ANGLAIS ;/ Par notre envoyé spécial Michel CLERC ! ! ,,...

VIEW ARTICLE

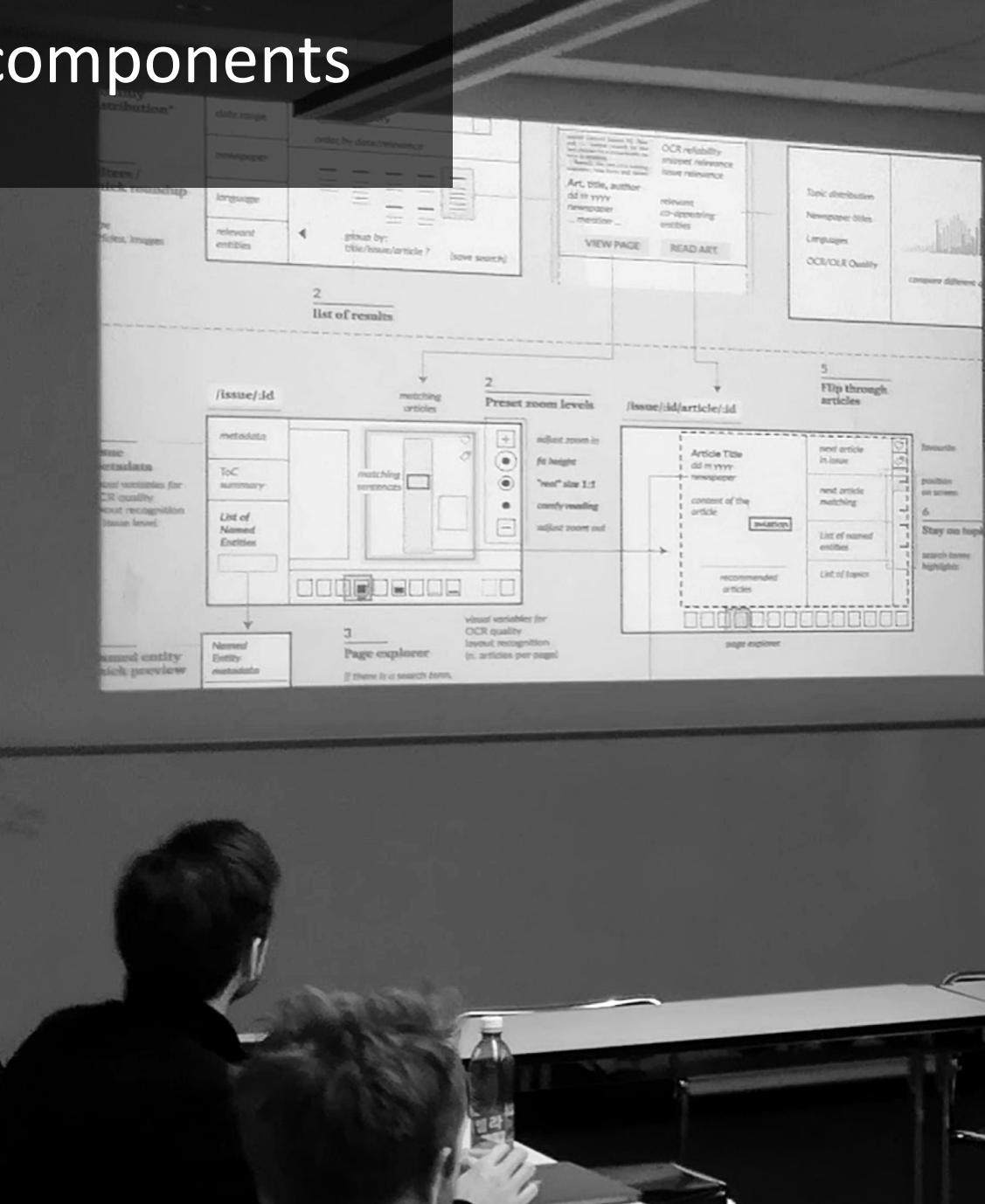
◆ page explorer

Friday, April 27, 1951

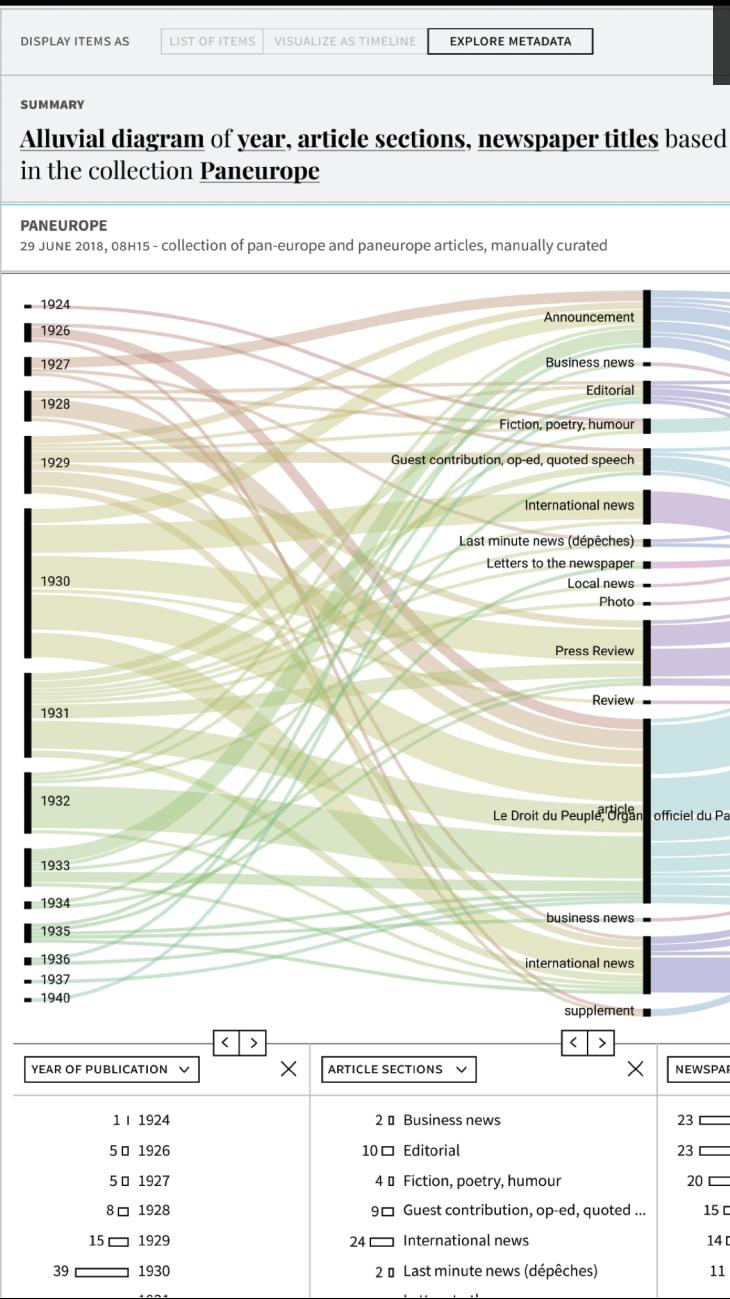
FRANCO PRÉFÈRE LE TÊTE-A-TÈTE AVEC L'AMÉRIQUE**Gazette de Lausanne**

FRANCO PRÉFÈRE LE TÊTE-A-TÈTE AVEC L'AMÉRIQUE SOUS LE PRINTEMPS D'ESPAGNE ATTENDANT UN TARDIF REPENTIR FRANCO-ANGLAIS ;/ Par notre envoyé spécial Michel CLERC ! ! ,,...

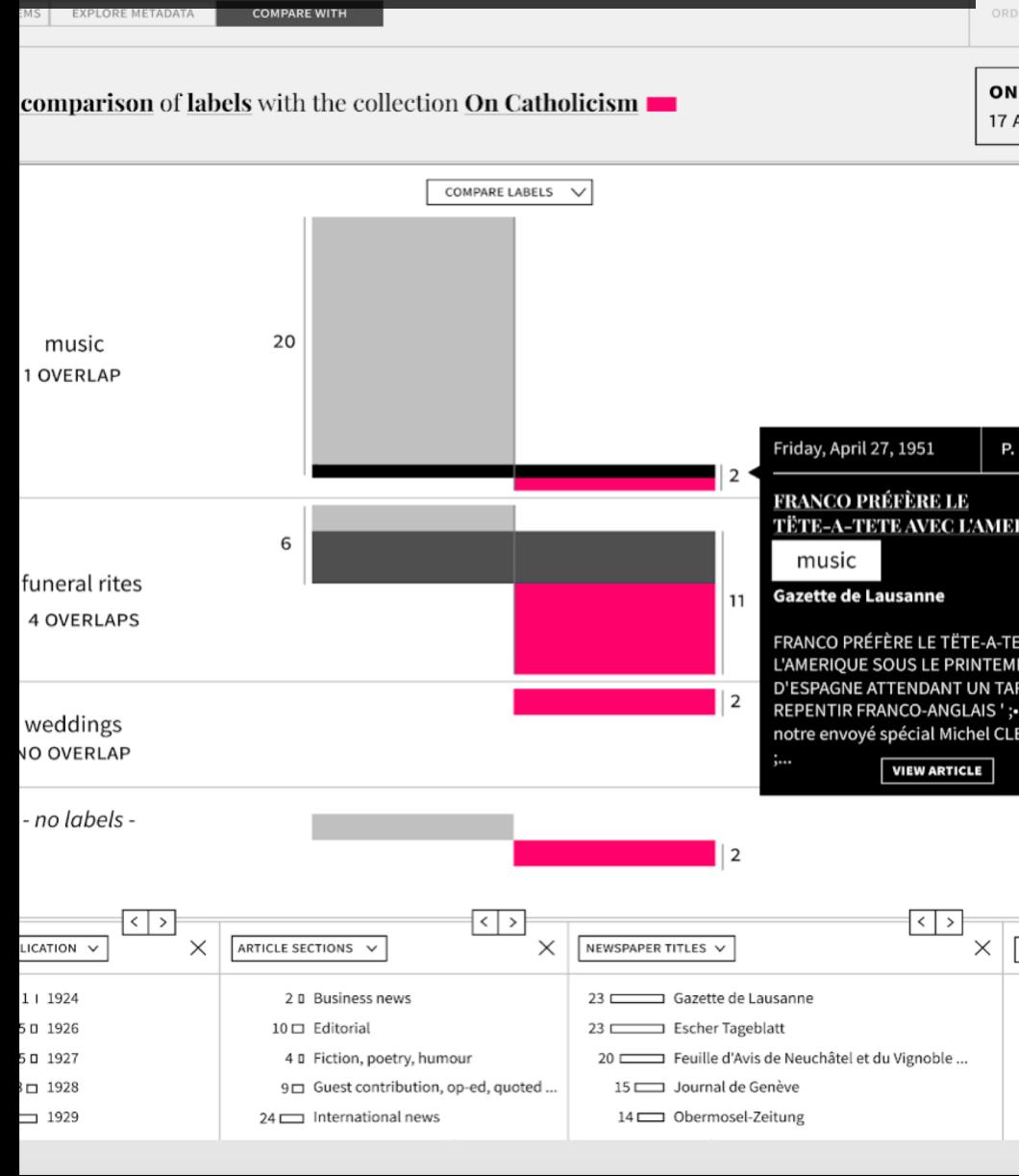
VIEW ARTICLE



personal space / my collections



Exemple: Inspect & Compare



#design ★

12 | 1 | Add a topic

SMH-Zeitung 1971
Le Monde 1999
Le Monde public 4003
St. Galler Zeitung 0
VHT-Zeitung 216
1987 2014

Roman 3:22 PM

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

Screen Shot 2020-04-01 at 15.22.09.png

1143673 Gazette de Lausanne 18413
414645 La Libre 12007
340216 Journal de Genève 180888
302180 L'Express 157392
214785 L'Imperial 118553
63187 Neue Zürcher Zeitung 131240
84511 Freiburger Nachrichten 21477
56295 Confidé 18328
34807 Die Tat 53897
52881 Le Peuple, La Sentinelle 34555
52881 Le Confédéré de Fribourg 20041 9909
11882 Le Chroniqueur 5470
La Libre Syndicale 9683 4770
Gazette du Valais / Nouvelle Gazette du Valais 8703 4672
Solidarité

1 1 1 1

Same with central Y axis

Screen Shot 2020-04-01 at 15.25.09.png

1143673 Gazette de Lausanne 18413
414645 La Libre 12007
340216 Journal de Genève 180888
302180 L'Express 157392
214785 L'Imperial 118553
63187 Neue Zürcher Zeitung 131240
84511 Confidé 21477
56295 Die Tat 12007
34807 Le Peuple, La Sentinelle 34555
52881 Le Confédéré de Fribourg 20041 9909
11882 Le Chroniqueur 5470
La Libre Syndicale 9683 4770
Gazette du Valais / Nouvelle Gazette du Valais 8703 4672
Solidarité

daniele 3:28 PM

fantastic! what if you invert the last two colors? so that the inte

Exemple: Inspect & Compare

A QUERY * COLLECTION

"arnhem OR arnhem(3 more)" match · équipe · ligue · club · saise REFINER ...

search for ... ADD FILTER...

OPEN IN SEARCH PAGE... (5,586 RESULTS)

181 results in common ⓘ

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

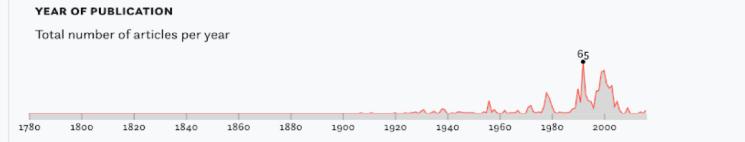
OPEN IN SEARCH PAGE... (181 RESULTS)

B QUERY * COLLECTION

"arnhem OR arnhem(3 more)" "match" REFINER ...

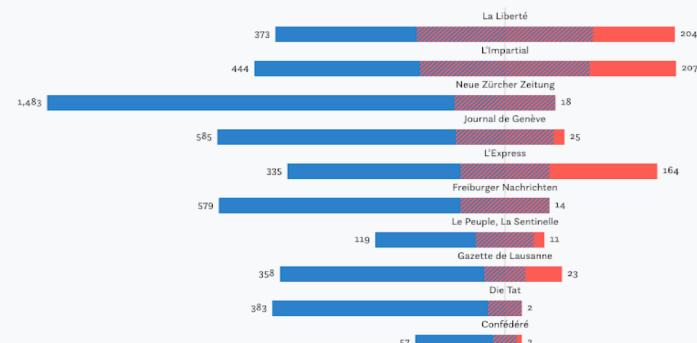
search for ... ADD FILTER...

OPEN IN SEARCH PAGE... (670 RESULTS)



SCALE: SQUARE ROOT ▾ SORT BY ABSOLUTE INTERSECTION ▾

NEWSPAPER (10 OPTIONS)



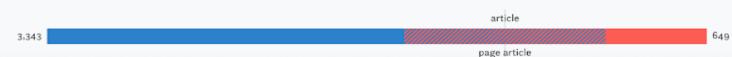
COUNTRY (1 OPTION)



LANGUAGE (2 OPTIONS)

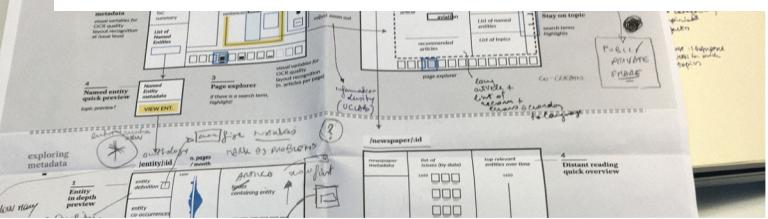


TYPE (3 OPTIONS)





Lessons learned: User requirements & tool adoption



With **time, prototypes and practice** original user requirements emerge (chicken-egg problem).

Users need **time and exposure** appropriate new tools.

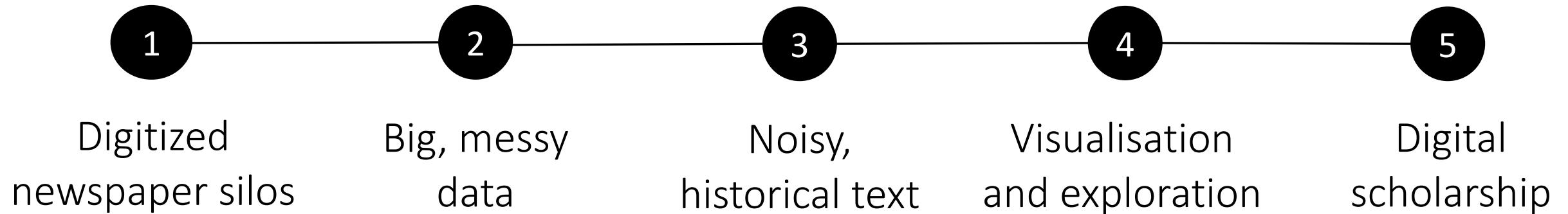
Users are **pragmatic**.

Combinations of **NLP enrichment + IR offer** powerful means to meet the information needs of historians.

Skill development is often needed, users susceptible to jumping to conclusions.



How to enable semantic indexing and of large collections of historic newspapers?



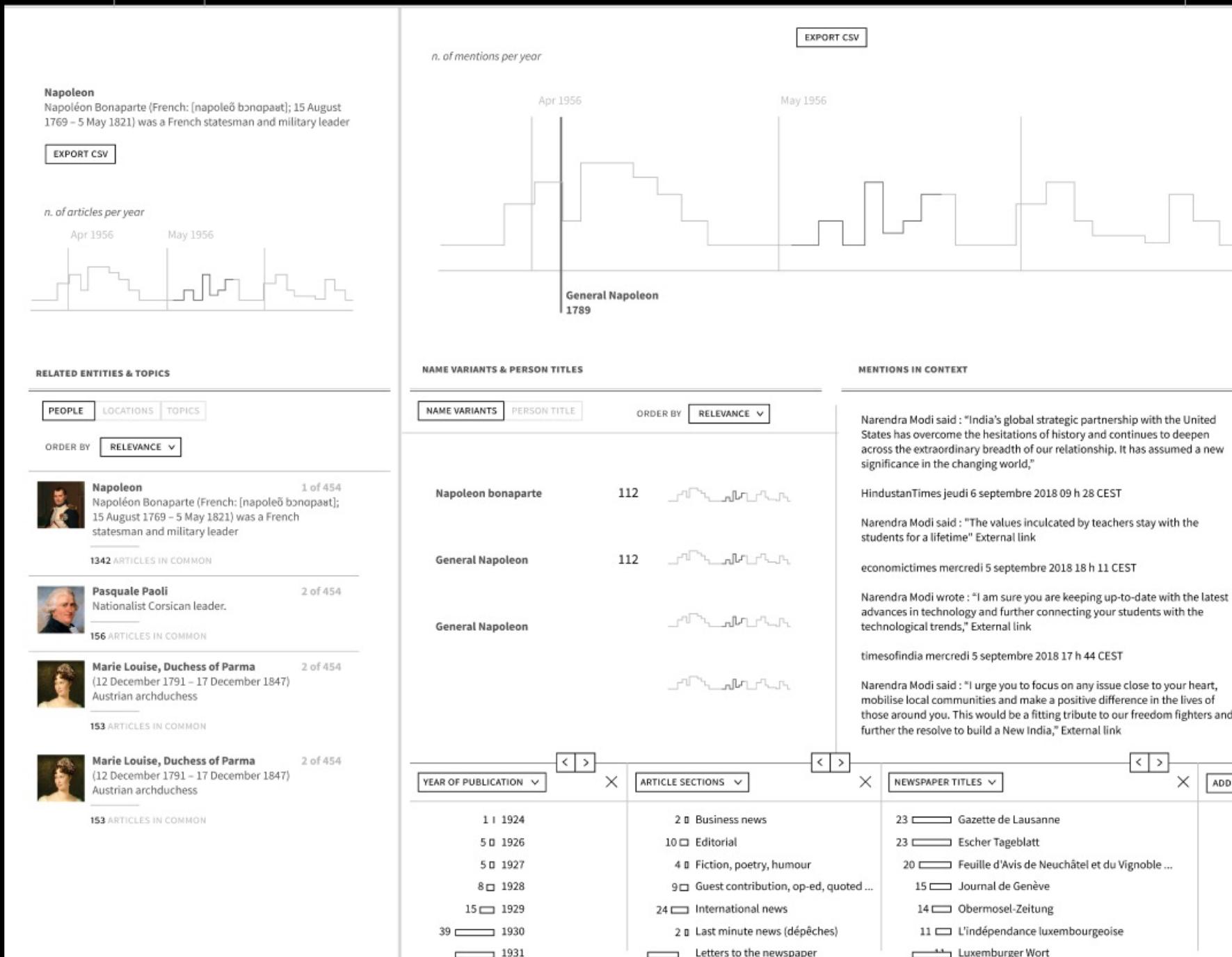
About digital scholarship

New ways to engage with content also means new ways of using it.

How can historians work with

- automatically extracted, imperfect data?
- algorithmically generated, changing data?





MY COLLECTIONS

MY SAVED QUERY

LIST OF MY COLLECTIONS

ORDER BY LATEST MODIFIED ▾

PANEUROPE 1 of 4
YESTERDAY, 12H15 - collection of pan-europe and paneurope articles, manually curated
135 ARTICLES

UNITED STATES OF EUROPE 3 of 10
collection of pan-europe and paneurope articles, manually curated
8 PAGES 7 ARTICLES

GUERRE FROIDE & EUROPE 3 of 10
29 JUNE 2018, 08H15 - collection of pan-europe
1 PAGE 2 ARTICLES

LES CHIMERES 4 of 4
collection of pan-europe and paneurope articles, manually curated
2 ISSUES 8 PAGES 700 ARTICLES

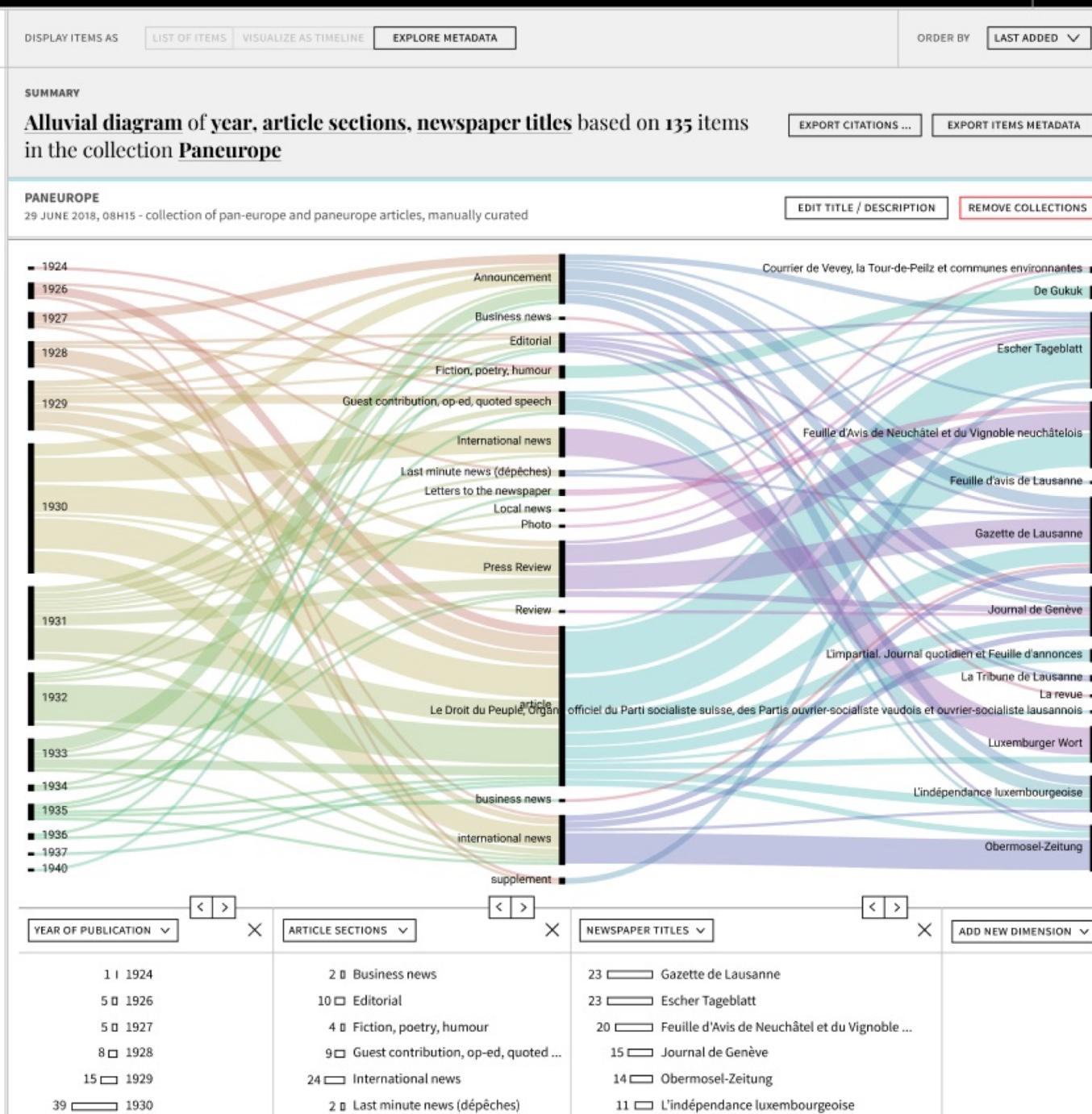
CONFERENCE REPORTS
-
2 ISSUES 8 PAGES 700 ARTICLES

ADD NEW EMPTY COLLECTION

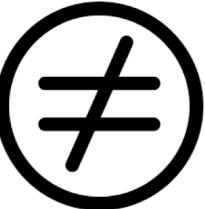
ALWAYS AVAILABLE COLLECTIONS

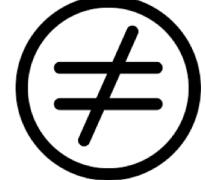
All items in all personal collection contains every article, issue, page you have ever included in one of your collections
2 ISSUES 8 PAGES 700 ARTICLES

FAVOURITE ITEMS
This list contains
2 ISSUES 8 PAGES 700 ARTICLES



Transparency matters

beauty  perfection

imperfection  useless

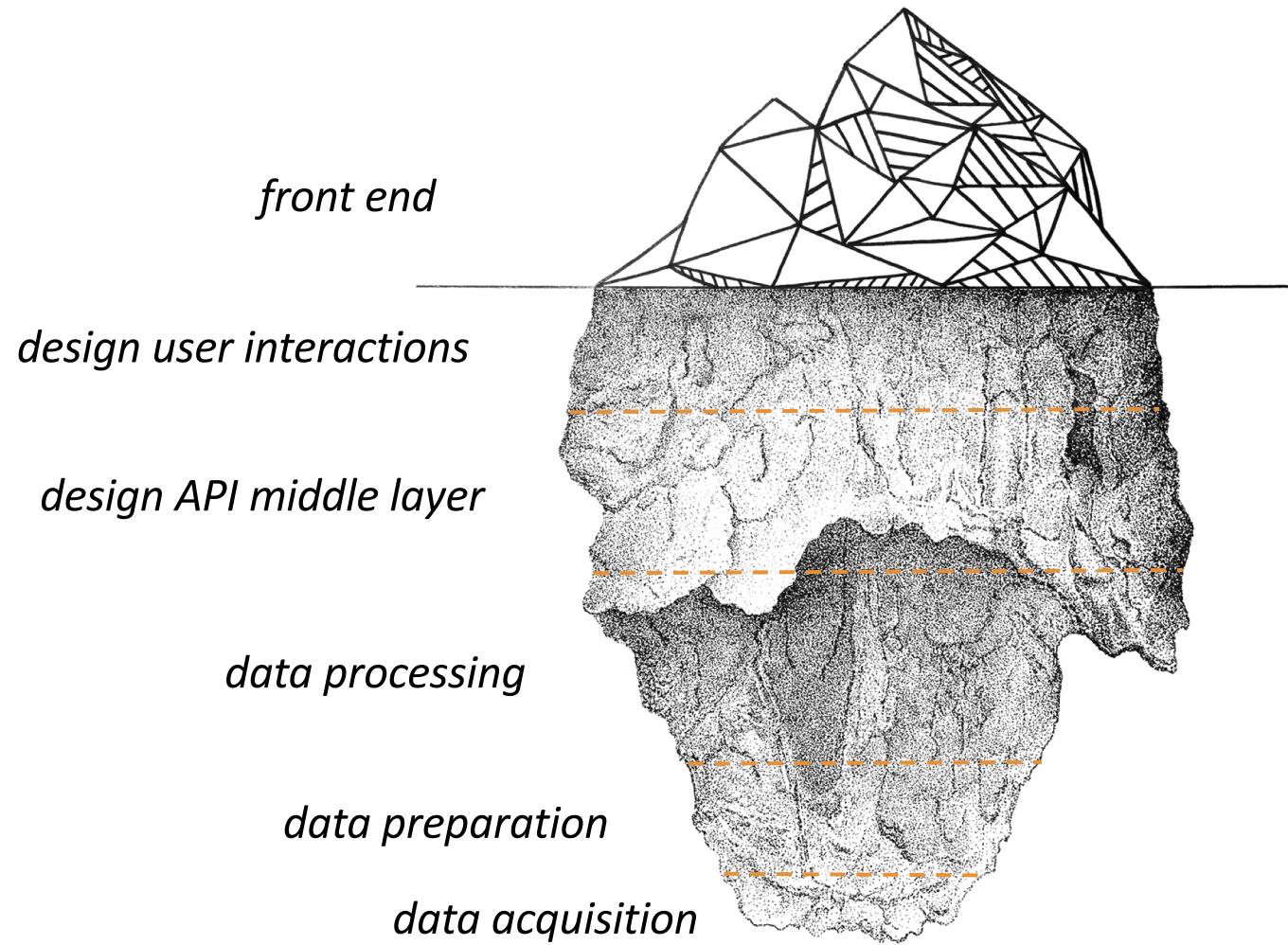
Transparency matters

We inherit, introduce and propagate biases and variances in our corpus.

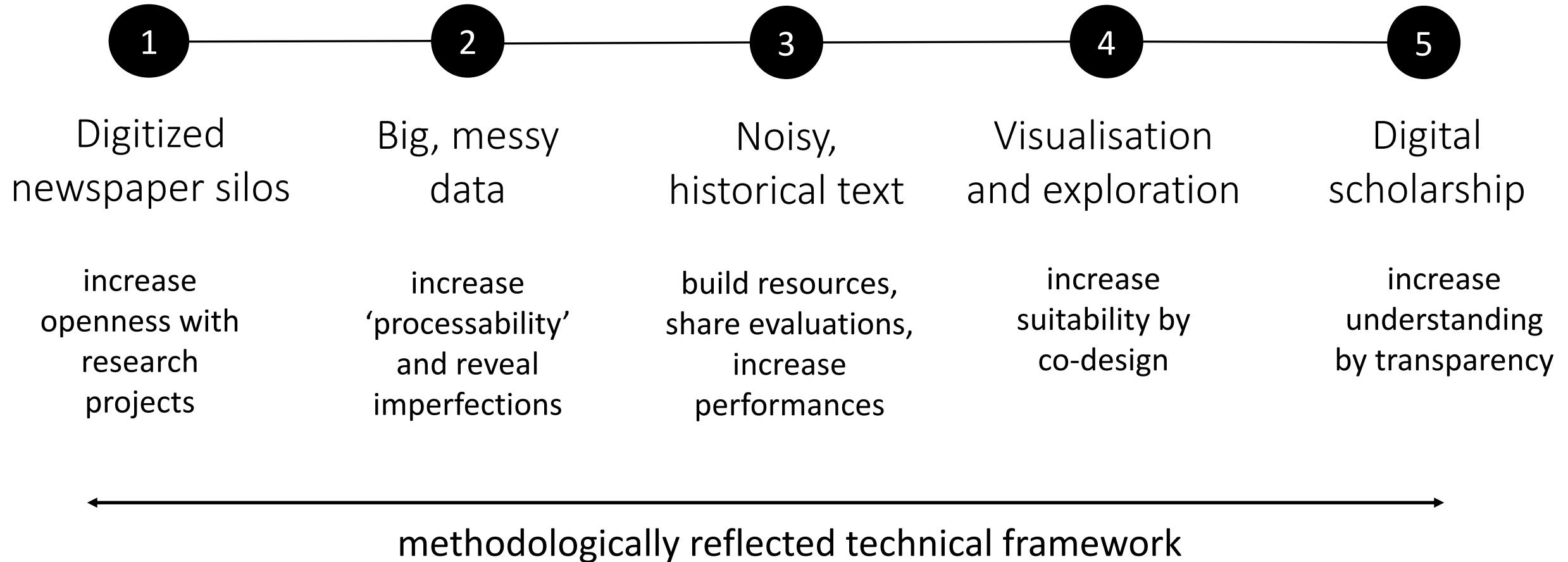
How do they affect content exploration? What exactly do historians need to understand to be able to deal with them?

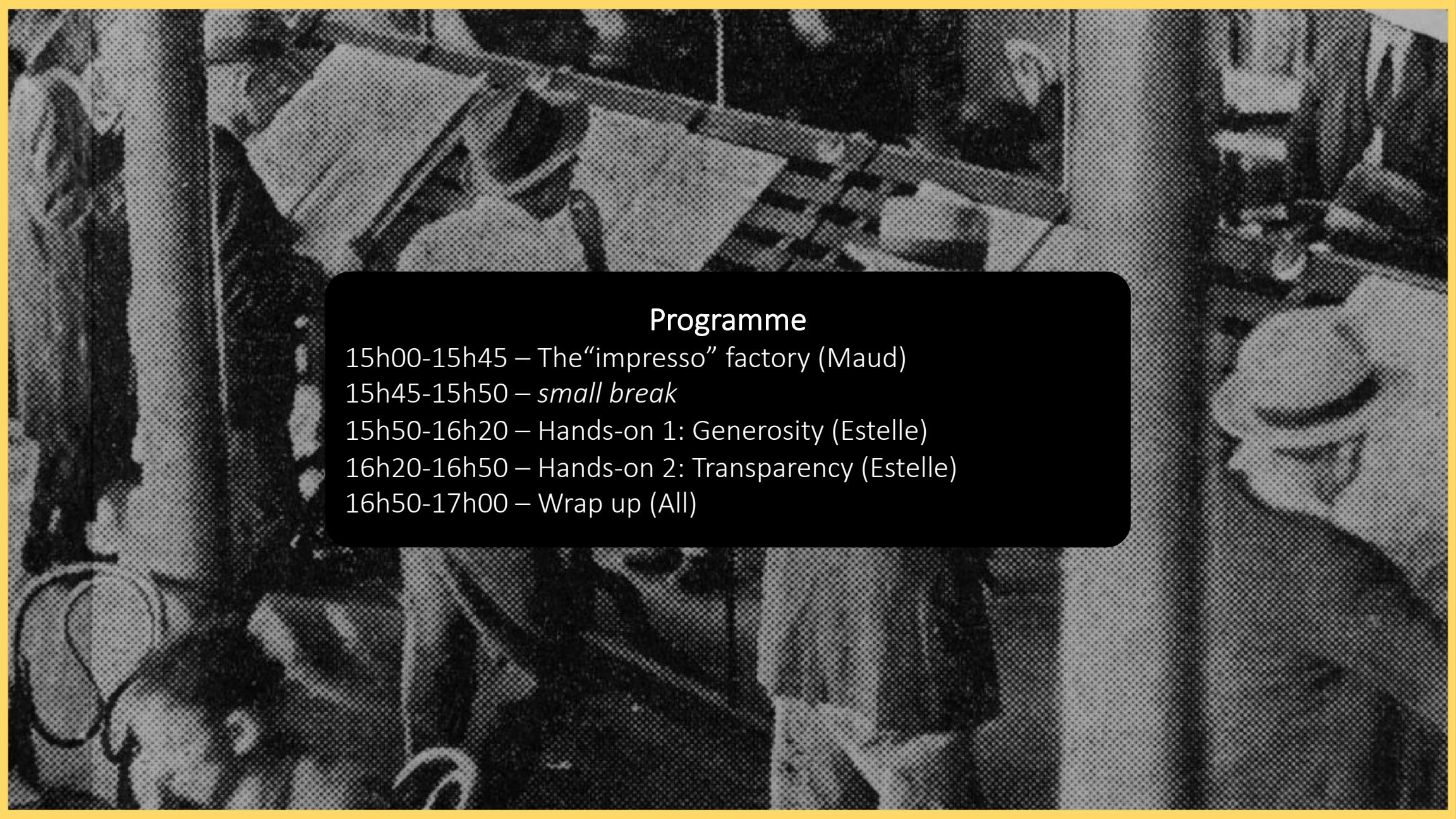
What is the most effective way to communicate these biases and variances?

What's in the *impresso* app?



Newspapers are important, promising but complex digital assets.





Programme

- 15h00-15h45 – The “impresso” factory (Maud)
- 15h45-15h50 – *small break*
- 15h50-16h20 – Hands-on 1: Generosity (Estelle)
- 16h20-16h50 – Hands-on 2: Transparency (Estelle)
- 16h50-17h00 – Wrap up (All)



Part 2
Hands-on session



Media Monitoring
impresso of the Past

Two guiding principles for the app's co-design



Generosity

- **exploration**
- **switching scales (drill down, overview)**
- **keep awareness of the context of the query, of an article collection**

Transparency

- discover novel opportunities in & make **informed** decisions about digitized sources
- provenance and quality to assess the **value** of **imperfect** data

SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE ↴

ORDER BY

RELEVANCE ↴

DISPLAY AS

LIST

TILES

 "robert schuman" X

add keyword to search



7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT ↴

Lycée
Robert-Schuman

Lycée Robert-Schuman Luxembourg Poste vacant

1. How to discover the newspapers' contents via the impresso app (generosity)

- Guided tour of the app
- Hands-on guided by questions on the use of the app's features
- Collection of your answers

2. Taking a step back: understanding the findings (transparency)

- Guided tour of the app
- Hands-on guided by questions on the use of the app's context information
- Collection of your answers

 Frontpage

PUBLICATION DATE

Number of articles per year

 % SUM i

1740 1760 1780 1800 1820 1840 1860

ADD NEW DATE FILTER ...

FILTER BY CONTENT LENGTH i

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results

 French (5,491 results) ↗ German (2,052 results) ↗

Après l'échec de M. Antoine Pinay, le ministre de l'Éducation nationale, M. Robert Schuman est chargé d'une mission de conciliation.

Le conflit modérés-SFIO résolu par le MRP ?

PARIS, 19 (A.F.P.). — Le président René Coty a chargé M. Robert Schuman, leader M.R.P., d'une mission de conciliation.

Après avoir été reçu par M. René Coty, M. Robert Schuman a annoncé que le président de la République venait de lui confier une mission spéciale. Il s'agit d'assurer la conciliation entre les deux groupes qui viennent d'avoir le président de la République, a-t-il dit, ont fait apparaître une commune volonté de conciliation. Il se résout toutefois à ce que la conclusion de la crise soit dominée

M. Robert Schuman est chargé d'une mission de conciliation

L'Express ↗ MONDAY, OCTOBER 21, 1957 – p.1

Personal use

M. Robert Schuman est chargé d'une mission de conciliation Après l'échec de M. Antoine Pinay Le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A.

LOCATIONS France ↗

PEOPLE Antoine Pinay ↗



SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE

"robert schuman"

Search box



add keyword to search

 Frontpage

FIND SIMILAR WORDS

PUBLICATION DATE

Number of articles per year

 % SUM 

ADD NEW DATE FILTER ...

Filters

FILTER BY CONTENT LENGTH

645

0

9,900

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results

 French (5,491 results) German (2,053 results) 7,587 articles found containing **robert schuman**

COMPARE ...

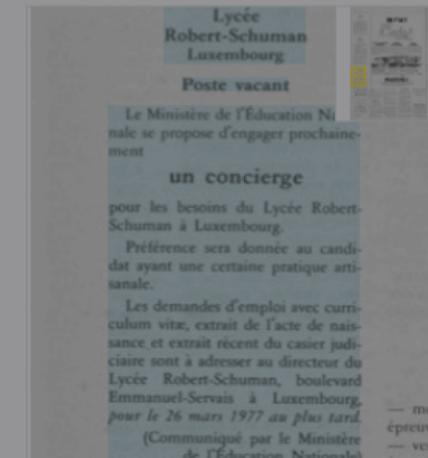
SAVE / EXPORT

Human readable summary

DISPLAY AS

LIST

TILES



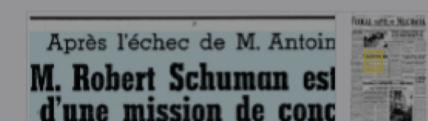
Result list

M. Robert Schuman est chargé d'une mission de conciliation

L'Express MONDAY, OCTOBER 21, 1957 – p.1

Personal use

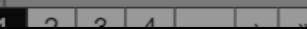
M. Robert Schuman est chargé d'une mission de conciliation Après l'échec de M. Antoine Pinay Le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A.)



Après avoir été reçu par M. René Coty, M. Robert Schuman a annoncé que le président de la République tenait de cette façon une mission de conciliation. Les dernières entrevues que vient d'avoir le président de la République, a-t-il dit, ont fait apparaître que les deux partis étaient toutefois d'accord sur la nécessité d'une conciliation. Il en ressort toutefois que la conclusion de la crise est domi-

LOCATIONS France

PEOPLE Antoine Pinay



SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE ▾

ORDER BY

RELEVANCE ▾

DISPLAY AS

LIST

TILES

"robert schuman" ▾



add keyword to search

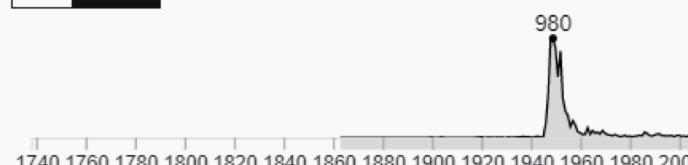


Frontpage

FIND SIMILAR WORDS 

PUBLICATION DATE

Number of articles per year

% **SUM** 

ADD NEW DATE FILTER ...

FILTER BY CONTENT LENGTH 

645

0

9,900

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results

 French (5,491 results)  German (2,053 results) 

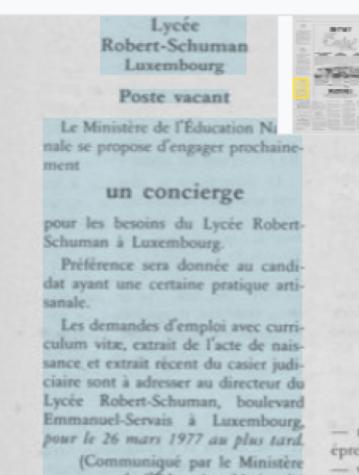
GROUP BY ARTICLE ▾

ORDER BY RELEVANCE ▾

DISPLAY AS LIST TILES

7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT Lycée Robert-Schuman Luxembourg Poste vacantd'Letzeburger Land  FRIDAY, MARCH 18, 1977 – p.14

Personal use (no export)

Lycée Robert-Schuman Luxembourg Poste vacant Le Ministère de l'Éducation Nationale se propose d'engager prochainement un concierge pour les besoins du

LOCATIONS Luxembourg Lycée **Robert - Schuman** Luxembourg Poste vacant Le Ministère de l'Éducation Nationale se proposed'engager prochainement un concierge pour les besoins du Lycée **Robert - Schuman** à Luxembourg. Préférencedu Lycée **Robert - Schuman**, boulevard Emmanuel-Servais à Luxembourg, pour le 26 mars 1977 au plus tard

VIEW

ADD TO COLLECTION ... M. Robert Schuman est chargé d'une mission de conciliationL'Express  MONDAY, OCTOBER 21, 1957 – p.1

Personal use

M. Robert Schuman est chargé d'une mission de conciliation Après l'échec de M. Antoine Pinay Le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A).

LOCATIONS France PEOPLE Antoine Pinay 

SEARCH ARTICLES **SEARCH IMAGES** **NGRAMS**

GROUP BY **ARTICLE** **ORDER BY** **RELEVANCE** **DISPLAY AS** **LIST** **TILES**

"robert schuman" **X** **COMPARE ...** **SAVE / EXPORT**

add keyword to search **🔍** **FILTER**

Frontpage **FIND SIMILAR WORDS** **i**

PUBLICATION DATE
Number of articles per year

% **SUM** **i**

1740 1760 1780 1800 1820 1840 1860 1880 1900 1920

ADD NEW DATE FILTER ...

FILTER BY CONTENT LENGTH **i**

645

0 9,900

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)
check one or more language to filter results

French (5,491 results) **➡**

German (2,053 results) **➡**

GROUP BY **ARTICLE** **ORDER BY** **RELEVANCE** **DISPLAY AS** **LIST** **TILES**

7,587 articles found containing **robert schuman** **COMPARE ...** **SAVE / EXPORT**

Lycée Robert-Schuman Luxembourg Poste vacant
d'Letzeburger Land **FRIDAY, MARCH 18, 1977 – p.14**
Personal use (no export)

Lycée Robert-Schuman Luxembourg Poste vacant Le Ministère de l'Éducation Nationale se propose

d'Letzeburger Land — Friday, March 18, 1977 **p.14**

article **FR | p.14** **FACSIMILE** **TRANSCRIPT** **i** **FULLSCREEN: OFF**

Lycée Robert-Schuman Luxembourg Poste vacant

LOCATIONS Luxembourg **i**

TOPICS

40.8% **FR service · langue · poste · travail · personnel** 5.5% **FR avril · mai · mars · lieu · juin** 4.9% **FR scène · théâtre · spectacle · pièce · public**

d'une mission de conc **Le conflit modérés - SFIO résolu par le MRP**

PARIS, 19 (A.F.P.). — Le président René Coty a chargé M. Robert Schuman, leader M.R.P., d'une mission de conciliation.

Après avoir été reçu par M. René Coty, M. Robert Schuman a annoncé que le président de la République lui demandait de lui confier une mission précise : « Les derniers entretiens que vient d'ecrire le président du Conseil, M. Giscard d'Estaing, ont permis de constater que la situation économique et sociale nécessite une réaction immédiate. Il en résulte toutefois que la conclusion de la crise est dominée par des problèmes économiques et financiers dont il est nécessaire de mettre au point les données techniques et économiques de la situation. »

Il a été décidé de faire appel à M. Robert Schuman pour qu'il mène une étude approfondie de la situation économique et financière.

Ainsi, M. Bernard Clappier, directeur des relations économiques extérieures, qui fait partie de l'équipe de M. Giscard d'Estaing, a été nommé à la tête de la commission de conciliation. Il sera assisté par M. André Devaux, directeur du budget, M. Schwartzenbach, directeur des finances extérieures, M. Bisch-Laloue, directeur de la planification, M. Dreyfus, directeur des relations industrielles, M. Rosenstock-Frank, directeur des prix, M. Robert Schuman a aussi en charge les difficultés

Personal use

M. Robert Schuman est chargé d'une mission de conciliation. Après l'échec de M. Antoine Pinay, le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A).

LOCATIONS France **i**

PEOPLE Antoine Pinay **i**

« < 1 2 3 4 ... > »

SEARCH ARTICLES SEARCH IMAGES NGRAMS

GROUP BY ARTICLE ▾

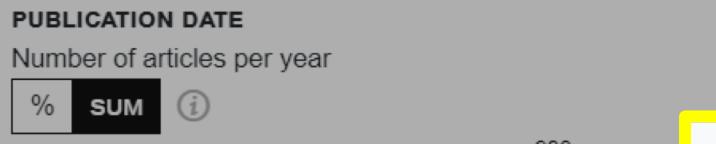
ORDER BY RELEVANCE ▾

DISPLAY AS LIST TILES

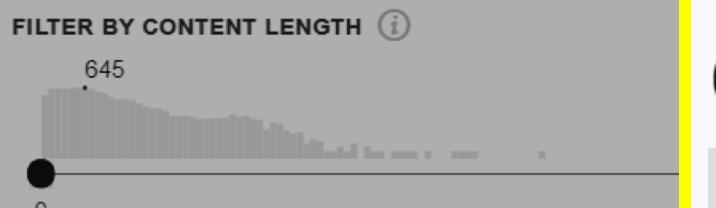
"robert schuman" X

add keyword to search 🔍 FILTER

Frontpage FIND SIMILAR WORDS i



ADD NEW DATE FILTER ...



FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

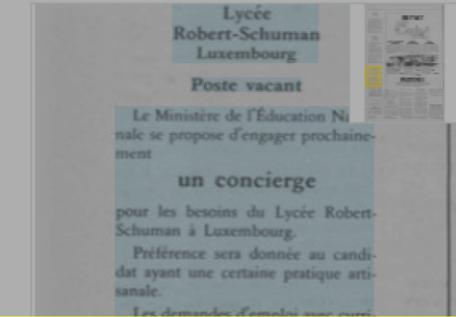
check one or more language to filter results

French (5,491 results) i

German (2,053 results) i

7,587 articles found containing robert schuman

COMPARE ... SAVE / EXPORT i



Lycée Robert-Schuman Luxembourg Poste vacant

d'Letzeburger Land i FRIDAY, MARCH 18, 1977 – p.14

Personal use (no export)

Lycée Robert-Schuman Luxembourg Poste vacant Le Ministère de l'Éducation Nationale se propose d'engager prochainement un concierge pour les besoins du

LOCATIONS Luxembourg i

L'Express — Monday, October 21, 1957 ◀ p.1 ▶

article FR | p.1

FACSIMILE

TRANSCRIPT i

X FULLSCREEN: OFF

M. Robert Schuman est chargé d'une mission de conciliation

LOCATIONS France i

PEOPLE Antoine Pinay i

TOPICS

35.3% FR président · ministre · gouvernement · général · chef i

7.9% FR travail · grève · personnel · chômage · syndicat i

17.7% FR conseil · président · conseiller · directeur · chef i

4.3% FR conseil · commune · construction · crédit · ville i

10.6% FR gouvernement · conférence · accord · question · traité i

4.3% FR canton · projet · développement · recherche · région i



Filters : aggregation of library and NLP metadata

Creation of filters based on the metadata:
help to calibrate the search + information about the collection

FILTER BY NEWSPAPER TITLES

check one or more newspaper to filter results

- Journal de Genève (421,080 results) ↗
- Gazette de Lausanne (314,103 results) ↗
- L'Express (226,843 results) ↗
- L'Impartial (199,024 results) ↗
- L'indépendance luxembourgeoise (111,928 results) ↗
- Le Peuple, La Sentinelle (74,481 results) ↗
- Confédéré (39,085 results) ↗
- Le Confédéré de Fribourg (20,277 results) ↗
- Courrier du Grand-Duché de Luxembourg (17,436 results) ↗
- Luxembourg (1935) (17,004 results) ↗
- Le Chroniqueur (13,478 results) ↗
- L'Union (13,179 results) ↗
- Luxemburger Wort (8,473 results) ↗
- d'Letzeburger Land (8,226 results) ↗
- L'avenir (5,039 results) ↗
- Courrier du Valais (4,448 results) ↗
- Le Bien public (4,370 results) ↗
- Neue Zürcher Zeitung (3,800 results) ↗
- Le Narrateur fribourgeois (2,665 results) ↗
- L'Essor (2,644 results) ↗

[EXPLORE ALL ...](#)

Metadata

FILTER BY PERSON

check one or more topics to filter results

- Francis Matthey (5,247 results) ↗
- Louis-Leopold Robert (4,728 results) ↗
- François Mitterrand (4,652 results) ↗
- Yasser Arafat (4,632 results) ↗
- Jacques Chirac (3,998 results) ↗

[EXPLORE ALL ...](#)

FILTER BY LOCATION

check one or more topics to filter results

- France (355,496 results) ↗
- Paris (281,633 results) ↗
- Lausanne (280,586 results) ↗
- Suisse, Moselle (273,425 results) ↗
- Gare de Cornavin (228,002 results) ↗

[EXPLORE ALL ...](#)

FILTER BY TOPIC

check one or more topics to filter results

- fr ministre · président · gouvernement · affaire · chef (430,702 results) ↗
- fr samedi · dimanche · vendredi · lundi · soir (326,547 results) ↗
- fr président · force · gouvernement · pays · mort (325,304 results) ↗
- fr chambre · gouvernement · ministre · député · ministère (320,971 results) ↗
- fr commission · projet · loi · proposition · conseil (311,543 results) ↗
- fr comité · membre · président · assemblée · association (311,467 results) ↗
- fr loi · droit · initiative · canton · projet (306,266 results) ↗
- fr gouvernement · accord · conférence · question · traité (304,159 results) ↗
- fr parti · gouvernement · socialiste · politique · pouvoir (303,816 results) ↗
- fr guerre · paix · pays · peuple · politique (302,260 results) ↗

[EXPLORE ALL ...](#)

Topics

Named entities

SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE

ORDER BY

RELEVANCE

DISPLAY AS

LIST TILES

 "robert schuman" 

add keyword to search

 Frontpage

PUBLICATION DATE

Number of articles per year

 %

SUM



980

1740 1760 1780

ADD NEW D...

FILTER BY CONTENT LENGTH 

645

9,900

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results  French (5,491 results)  German (2,052 results) 

7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT 

1. How to discover the newspapers' contents via the impresso app (generosity)

2. Taking a step back: understanding the findings (transparency)

Use case: help me find build a collection on the media coverage of Robert Schuman (1886-1963) in Luxembourg and Switzerland via the [impresso app](#).

Après l'échec de M. Antoine M. Robert Schuman est chargé d'une mission de conciliation

Le conflit modérés-SFIO résolu par le MRP ?

PARIS, 19 (A.F.P.). — Le président René Coty a chargé M. Robert Schuman, leader M.R.P., d'une mission de conciliation.

Après avoir été reçu par M. René Coty, M. Robert Schuman a annoncé que le président de la République venait de créer une commission spéciale. « Les derniers entretiens que vient d'avoir le président de la République, a-t-il dit, ont fait apparaître la nécessité d'une mission de conciliation. Il en résulte toutefois que la conclusion de la crise est domi-

M. Robert Schuman est chargé d'une mission de conciliation

L'Express  MONDAY, OCTOBER 21, 1957 – p.1

Personal use

M. Robert Schuman est chargé d'une mission de conciliation Après l'échec de M. Antoine Pinay Le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A.

LOCATIONS France PEOPLE Antoine Pinay 

SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE

ORDER BY

RELEVANCE

DISPLAY AS

LIST

TILES

 "robert schuman" 

7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT 

1. Search for “[Robert Schuman](#)”: autofill indicates the named entity “Robert Schuman” and “Robert Schumann”

2. Add similar keywords, in FR/[DE](#) to expand the result list

3. Frequencies and filters indicate the distribution of the result list

4. Use of NE “Robert Schumann” to [reduce](#) the result list

5. Use of the topics to narrow down the inspection of the result list

6. Store results in a collection

Use case: help me find build a collection on the media coverage of Robert Schuman (1886-1963) in Luxembourg and Switzerland via the [impresso app](#).

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

check one or more language to filter results

 French (5,491 results)  German (2,052 results) 

PARIS, 19 (A.F.P.). — Le président René Coty a chargé M. Robert Schuman, leader M.R.P., d'une mission de conciliation.

Après avoir été reçu par M. René Coty, M. Robert Schuman a annoncé que le président de la République venait de lui confier une mission de conciliation. « Les derniers entretiens que vient d'avoir le président de la République, a-t-il dit, ont fait apparaître l'importance de la question de la conciliation. Il en résulte toutefois que la conclusion de la crise est domi-

nante dans l'économie et l'industrie. Ainsi, M. Robert Schuman a rencontré le ministre des Relations économiques extérieures, qui fut directeur de son cabinet à l'Assemblée nationale, M. André Devaix, et ensuite le M.R.P. Deux autres directeurs du budget, Schwellers, directeur des finances extérieures, Hirsch, haut-commissaire au plan, Blach-Lalos, directeur de la caisse des débiteurs et mini-

LOCATIONS France PEOPLE Antoine Pinay 

Try for yourself! Hands-on „generosity“

On the search page, using the search filters and with the confirmation of the human readable summary, answer the following questions and create collections based on each query.

- Q1:** How many articles which contain the keyword “titanic” have been published on the front page?
- Q2:** Which year has the highest number of hits with “titanic” on the front page?
- Q3:** How many articles which contain the keyword “titanic” have been tagged with a topic related to film?
- Q4:** Which year has the highest number of hits with “titanic” and a topic “film”?
- Q5:** Which people have been the most often detected in articles containing the keyword “titanic”?
- Q6:** (Bonus task) On the search page, with the help of the filters, find movies featuring the Titanic sinking which are *not* by James Cameron. (hint: use the same topic but exclude/select other persons).

Answers to the hands-on session - part 1

Please find here the answers in forms of links to the impresso app, with the relevant filters activated. These are the answers to the hands-on session of the AI4LAM workshop 'Historical Newspaper Content Mining: findings from the impresso project', held on the 31.03.2021.

Q1: How many articles which contain the keyword “titanic” have been published on the front page? [453](#)

Q2: Which year has the highest number of hits with “titanic” on the front page? [1912](#)

Q3: How many articles which contain the keyword “titanic” have been tagged with a topic related to film? [2852](#)

Q4: Which year has the highest number of hits with “titanic” and a topic “film”? [1998](#)

Q5: Which people have been the most often detected in articles containing the keyword “titanic”? [James Cameron](#)

Q6: (Bonus task) On the search page, with the help of the filters, find movies featuring the Titanic sinking which are *not* by James Cameron. (hint: use the same topic but exclude/select other persons). [Titanic](#) by Jean Negulesco.

SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE ↗

ORDER BY

RELEVANCE ↗

DISPLAY AS

LIST ↗

TILES

 "robert schuman" ↗

add keyword to search

 Frontpage

FIND SIMILAR WORDS

PUBLICATION DATE

Number of articles

 %

SUM

ADD NEW DATE FILTER ...

FILTER BY CONTENT LENGTH

645

0

9,900

FILTER BY LANGUAGE OF ARTICLES (4 OPTIONS)

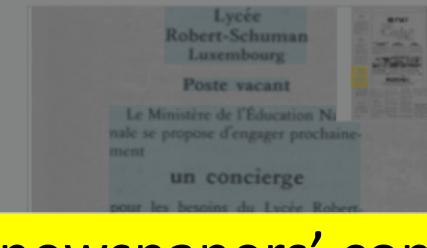
check one or more language to filter results

 French (5,491 results) ↗ German (2,052 results) ↗

7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT ↗

Lycée Robert-Schuman Luxembourg Poste vacant

d'Letzeburger Land ↗ FRIDAY, MARCH 18, 1977 – p.14

Personal use (no export)

Lycée Robert-Schuman Luxembourg Poste vacant Le Ministère de l'Éducation Nationale se propose d'engager prochainement un concierge pour les besoins du

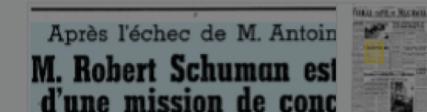
1. How to discover the newspapers' contents via the impresso app (generosity)

2. Taking a step back: understanding the findings (transparency)

d'engager prochainement un concierge pour les besoins du Lycée **Robert-Schuman** à Luxembourg. Préférencedu Lycée **Robert-Schuman**, boulevard Emmanuel-Servais à Luxembourg, pour le 26 mars 1977 au plus tard

VIEW

ADD TO COLLECTION ... ↗

M. Robert Schuman est chargé d'une mission de conciliation

L'Express ↗ MONDAY, OCTOBER 21, 1957 – p.1

Personal use

M. Robert Schuman est chargé d'une mission de conciliation Après l'échec de M. Antoine Pinay Le conflit modérés-SFIO résolu par le MRP ? PARIS, 19 (A.

LOCATIONS France ↗

PEOPLE Antoine Pinay ↗



SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

GROUP BY

ARTICLE

ORDER BY

RELEVANCE

DISPLAY AS

LIST TILES

 "robert schuman" X

add keyword to search



7,587 articles found containing robert schuman

COMPARE ...

SAVE / EXPORT



1. Search for “[Robert Schuman](#)”: autofill indicates the named entity “Robert Schuman” and “Robert Schumann”

2. Add similar keywords, in FR/[DE](#) to expand the result list

3. Frequencies and filters indicate the distribution of the result list

4. Use of NE “Robert Schumann” to [reduce](#) the result list

5. Use of the topics to narrow down the inspection of the result list

6. Store results in a collection

7. Inspect&Compare the result list to expand and curate it

- [Compare Robert Schuman](#) as named entity and as keyword
- Find out when Robert Schuman was [misspelled](#) as Robert Schumann

Use case: help me find build a collection on the media coverage of Robert Schuman (1886-1963) in Luxembourg and Switzerland via the [impresso](#) app.

Try for yourself! Hands-on „transparency“

Q1: From the newspapers page, can you tell which decade has the most issues per year? the most pages?

Q2: Go to the Inspect&Compare component and open in column A the collection "Titanic in movies", and in column B the collection "titanic on front pages". To understand when the overlap peaks and why, please find how many articles appear for the year 1998 in each query (using the mouse over on the frequency graph).

Q3: Open the topic page for the topic "film, cinéma, semaine, jean, john", using the arrow at the end of the topic label and choosing "more". What do you notice about its distribution over time? Who are the most frequently mentioned persons?

Q4: Back to the Inspect&Compare component. Replace the collection in column B with the query for the person "James Cameron". Who are the most associated persons to both queries (i.e. results in common)?

Q5: Switch to the "compare" tab and look at how the collection "titanic in movies" overlaps with the query "James Cameron" as a person. Hover over the graphs and read the overlap proportion. What do you notice when looking at the different facets related to the person James Cameron and the topic "film"?

Answers to the hands-on session - part 2

Q1: From the newspapers page, can you tell which decade has the most issues per year? the most pages?

A1: Have a look [here](#) and change the settings of the total frequency graph from "issues" to "pages" and vice-versa.

Q2: Go to the Inspect&Compare component and open in column A the collection "Titanic in movies", and in column B the collection "titanic on front pages". To understand when the overlap peaks and why, please find how many articles appear for the year 1998 in each query (using the mouse over on the frequency graph).

A2: [in 1998](#), 20 articles containing "titanic" were published on the front page and 1169 containing "titanic" and the topic "film...", resulting in an overlap of 8 articles.

Q3: Open the topic page for the topic "film, cinéma, semaine, jean, john", using the arrow at the end of the topic label and choosing "more". What do you notice about its distribution over time? Who are the most frequently mentioned persons?

A3: [Gerard Depardieu](#) is the most frequently mentioned person in articles containing this topic.

Q4: Back to the Inspect&Compare component. Replace the collection in column B with the query for the person "James Cameron". Who are the most associated persons to both queries (i.e. results in common)?

A4: [Leonardo DiCaprio and Kate Winslet](#).

Q5: Switch to the "[compare](#)" tab and look at how the collection "titanic in movies" overlaps with the query "James Cameron" as a person. Hover over the graphs and read the overlap proportion. What do you notice when looking at the different facets related to the person James Cameron and the topic "film"? Which claim is false?

A5: all articles mentioning James Cameron, that contain the topic "film, cinéma..." are part of "titanic in movies".

Time to sum up

- The *impresso* app was a real *laboratory* where we experimented a lot.
- NLP enrichments can be combined in surprising ways with IR in order to satisfy the information needs of historians.
- Machine understanding at the service of humans: a milestone towards text mining supporting (improving) historical scholarship

What's next?

- Infrastructure maintenance and corpus growth.
- Bringing the interface closer to users: upload your texts, download your models, correct annotations, etc.
- HIPE2 under preparation: contact us in case you are interested!

For more information

- subscribe to the mailing list
- visit the youtube channel
- ask a user account

impresso website:
<https://impresso-project.ch>

impresso youtube channel:
<https://www.youtube.com/channel/UCjRGykH-P9m1aA3amjrQuSQ>

impresso application:
<https://impresso-project.ch/app>

data on Zenodo:
<https://zenodo.org/communities/impresso/search?page=1&size=20>

code on GitHub:
<https://github.com/impresso>

Some publications

More information to come on the new impresso website.

Highlights

- [Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers](#). CLEF 2020
- [Language Resources for Historical Newspapers: the Impresso Collection](#). LREC 2020
- [Historical Newspaper User Interfaces: A Review](#). IFLA 2019
- [Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers](#). JDDH 2020.
- [The impresso system architecture in a nutshell](#). Europeana Tech Insights
- [How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR](#)

Upcoming

- *Digitized newspapers - a new Eldorado for historians ?* Edited volume with De Gruyter
- *Impresso: Historical Newspapers Beyond Keyword Search.* Digital Scholarship in the Humanities.
- *Co-Designing Workflows for the Exploration of Semantically Enriched Historical Newspapers [working title]*



Media Monitoring
of the Past

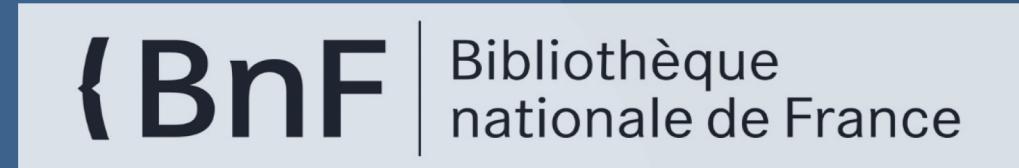
*Estelle Bunout
Simon Clematide
Marten Duering
Maud Ehrmann
Andreas Fickers
Daniele Guido
Frédéric Kaplan
Roman Kalyakin
Peter Makarov
Matteo Romanello
Gerold Schneider
Paul Schroeder
Benoit Seguin
Phillip Stroëbel
Martin Volk
Thijs van Beek
Lars Wieneke*



Thanks!

impresso-project.ch

twitter.com/impressoproject



Slides by Maud Ehrmann, Estelle Bunout and the *impresso* team

License: [CC BY 4.0](#)



Credits: These slides are CC BY. Photographs by LIBER, LILLIAD Learning Centre Innovation, Cantonal and University Library of Lausanne. Template by SlidesCarnival.

