4.3 Fairness and Transparency throughout a Digital Humanities Workflow: Challenges and Recommendations

Kaspar Beelen (The Alan Turing Institute – London, GB)
Sally Chambers (Ghent University, BE & KBR, Royal Library of Belgium, Brussels, BE)
Marten Düring (Luxembourg Centre for Contemporary and Digital History, LU)
Laura Hollink (CWI – Amsterdam, NL)
Stefan Jänicke (University of Southern Denmark – Odense, DK)
Axel Jean-Caurant (University of La Rochelle, FR)
Julia Noordegraaf (University of Amsterdam, NL)
Eva Pfanzelter (Universität Innsbruck, AT)

4.3.1 Main challenges and aim

How can we achieve sufficient levels of transparency and fairness for (humanities) research based on historical newspapers? Which concrete measures should be taken by data providers such as libraries, research projects and individual researchers? We approach these questions from the vantage point that digitised newspapers are complex sources with a high degree of heterogeneity caused by a long chain of processing steps, ranging, e.g., from digitisation policies, copyright restrictions to the evolving performance of tools for their enrichment such as OCR or article segmentation. Overall, we emphasise the need for careful documentation of data processing, research practices and the acknowledgement of support from institutions and collaborators.

Increasingly, historical newspaper data undergoes automatic processing using probabilistic methods. For example, topic modelling may inspire the identification of semantic facets within a set of articles, and word embeddings can suggest new keywords and as such different contexts or semantic shifts over time. The acknowledgement of such input matters inasmuch as it holds novel analytical potential and constitutes opportunities to broaden researchers' views on their sources. At the same time, it can mislead researchers due to the underlying principles which govern their creation and make them neither neutral nor objective. We therefore emphasise that researchers benefit from accessible information regarding the processing of data and its fairness. Still, at some point they will nevertheless have to trust systems' output and accept that their findings also depend on factors beyond their understanding, e.g., the impact of different constellations of search engine settings or the outcome provided by topic modelling tools.

Our goal is to compile recommendations for different aspects of transparency and fairness required for the analysis of digitised and enriched historical newspaper collections. We focus on aspects with a potentially high impact on the outcome of research. We distinguish between the need of researchers to obtain information for processes which lie beyond their control, such as institutional digitisation policies and OCR, and their obligation to provide information on aspects they can control, such as the documentation of their modus operandi and sharing research data to allow the traceability of their research. In this report we focus on the former.

The authors of this report have backgrounds in computer science (AI, visualisation, engineering), history (media history, contemporary history, digital history) and library science. This report is the result of one week of exchange and discussion on the topic of data transparency and fairness.

4.3.2 Approach

In a first exploration phase we started with a round-table discussion about fairness and transparency in the context of humanities research based on digitised historical newspapers. For a more formal and systematic review of interface features for historical newspapers see [4, 15].

Second, we performed an initial exploration of seven portals that provide access to historical newspapers. Several issues related to fairness and transparency surfaced in the round-table discussion and in the platform exploration which was centred on the needs of researchers in the historical disciplines.

Third, we used the output of the exploration phase to identify six focus areas which play a key role for historical newspaper research and formulated accompanying recommendations for measures to improve transparency and fairness. The focus areas and measures are organised along the lines of a typical digital humanities workflow.

In a final application phase we used the identified focus areas and recommendations to evaluate the *impresso* interface[32] which was developed with particular attention to transparency. We tested the portal and discussed to what extent each issue plays a role, and to what extent *impresso* implements or enables the recommended strategies. This resulted in insights regarding how far one of the state-of-the-art portals is when it comes to facilitating fair and transparent research on digitised historic newspapers.

4.3.3 Definitions

User Various types of persons work with digitised historical newspaper data, for example humanities scholars, interested lay people, collection owners, and portal developers, as well as scientists from other fields, such as natural language processing (NLP) researchers, who use newspapers as training sets. In this report, our point of reference are foremost the needs of historians, but we nevertheless expect that our recommendations are also relevant for other user groups.

Collection A comprehensive body of materials, in our case digitised historical newspapers, that is curated by a library, museum, or archive.

Corpus or Research Dataset The dataset that a researcher has compiled and on which they will do their analysis. The research dataset may be a subset of one or more collections. The researcher may have used one or more portals to compile the research dataset. The research dataset has often undergone multiple (iterative) processing and enrichment steps.

Portal An access point to one or more collections of digitised historic newspapers, providing functionality such as keyword search, faceted search, browsing or the inspection of raw scans.[33]

Workflow A sequence of actions executed by a researcher to find, collect, transform, enrich, and/or analyse documents.

Fairness We define fairness as the absence of bias. Fairness and equity can relate to a collection, a corpus, or the input and output of a tool. Fairness can be improved by raising awareness of biases as well as unwanted over- and under-representations. Lacking fairness can either be the result of culturally ingrained biases or technical processing on any stage of newspaper digitisation and enrichment.

[32] `https://impresso-project.ch/app/`

[33] In this report the terms *platform*, *interface*, *web application* describe the same thing.

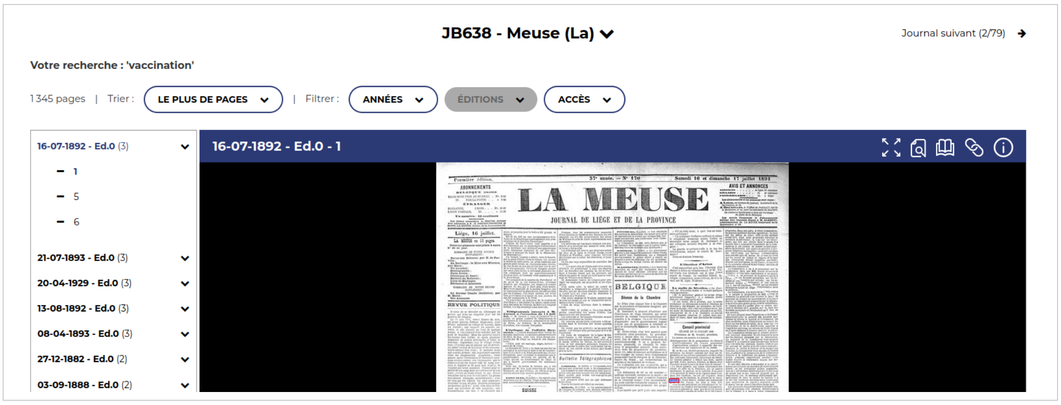


Figure 8 Screenshot of the KBR Belgica Press search interface (© KBR, Royal Library of Belgium).

Transparency Explicit, accessible information regarding the content of a collection or corpus regarding the workflow that was followed to create, process, and enrich it and/or regarding what is known about its fairness.

4.3.4 Exploration: Initial use case-based exploration of platforms

Here we present the findings of our initial exploration of platforms that provide access to historical newspapers. The findings were used as input for the workflow requirements regarding fairness and transparency described in the next section. Seven platforms were investigated. This list is not complete: not covered are, for example, Delpher[34], the CLARIN Newspapers Resource Family[35], and *impresso*.

The initial exploration was guided by a use case on the topic of "vaccination". We have documented this exploration in the form of short reviews which are structured as follows:

- Overview of the portal and its collections including a characterisation of the titles and main features for search, exploration and opportunities to interact with the data.
- Vaccination case study with a focus on the following questions: When was the first article which mentions vaccination published? Which bursts/peaks can be observed in the coverage?
- Summary and assessment of the level of transparency and fairness.

In the following sections we provide reports on the results of these experiments for different portals.

BelgicaPress

The landing page of the Belgica Press portal[36](Figure 2) gives information about the content of the available collection: 121 titles published between 1814 and 1970. Some details are given about the selection of this collection, as well as the information that only one title has been digitised until 1970. However, there is no further information concerning the availability of other titles.

[34] https://www.delpher.nl/
[35] https://www.clarin.eu/news/clarin-resource-families-newspaper-corpora
[36] https://www.belgicapress.be

The interface itself is simple. A search bar can be used to query for keywords and an advanced search allows for date filtering as well as Boolean conditions on the presence or absence of keywords in the results. A first query for "vaccin" yields 12.721 pages in 93 newspapers. The results are grouped by newspaper title which makes the search for the first occurrence and the overall distribution over time within the entire corpus rather laborious. Copyright-protected content is accessible for registered researchers with a MyKBR account. The results are presented as a list of newspaper titles sorted by the number of pages containing mentions of the keywords. When clicking on a result, a new page opens with a viewer allowing the user to see mentions of the keywords. The user can navigate through a list of other pages of this title. It is also possible from this page to switch to another title. There is apparently no relevance ranking for search results but there is the possibility to sort results by newspaper title, by date or by number of pages containing a keyword.

German Newspaper Portal

The German Newspaper Portal[37] has a very simple, "clean" interface that is available in German and English. It is not immediately clear if the search is also bilingual; testing reveals that this is not the case. The caption on the search page has a very minimal indication of the scope of the collection: one can search newspapers from 1671 to 1950. The first thing users see is a search box which invites for a direct keyword search. If users scroll down, three different browsing options are provided. Underneath those is a graph visualising the total amount of newspapers. At the bottom there is a display of a historical newspaper issue of the same date 100 years ago. The interface is clearly designed for a general audience, that is: users focused on encyclopedic use and browsing.

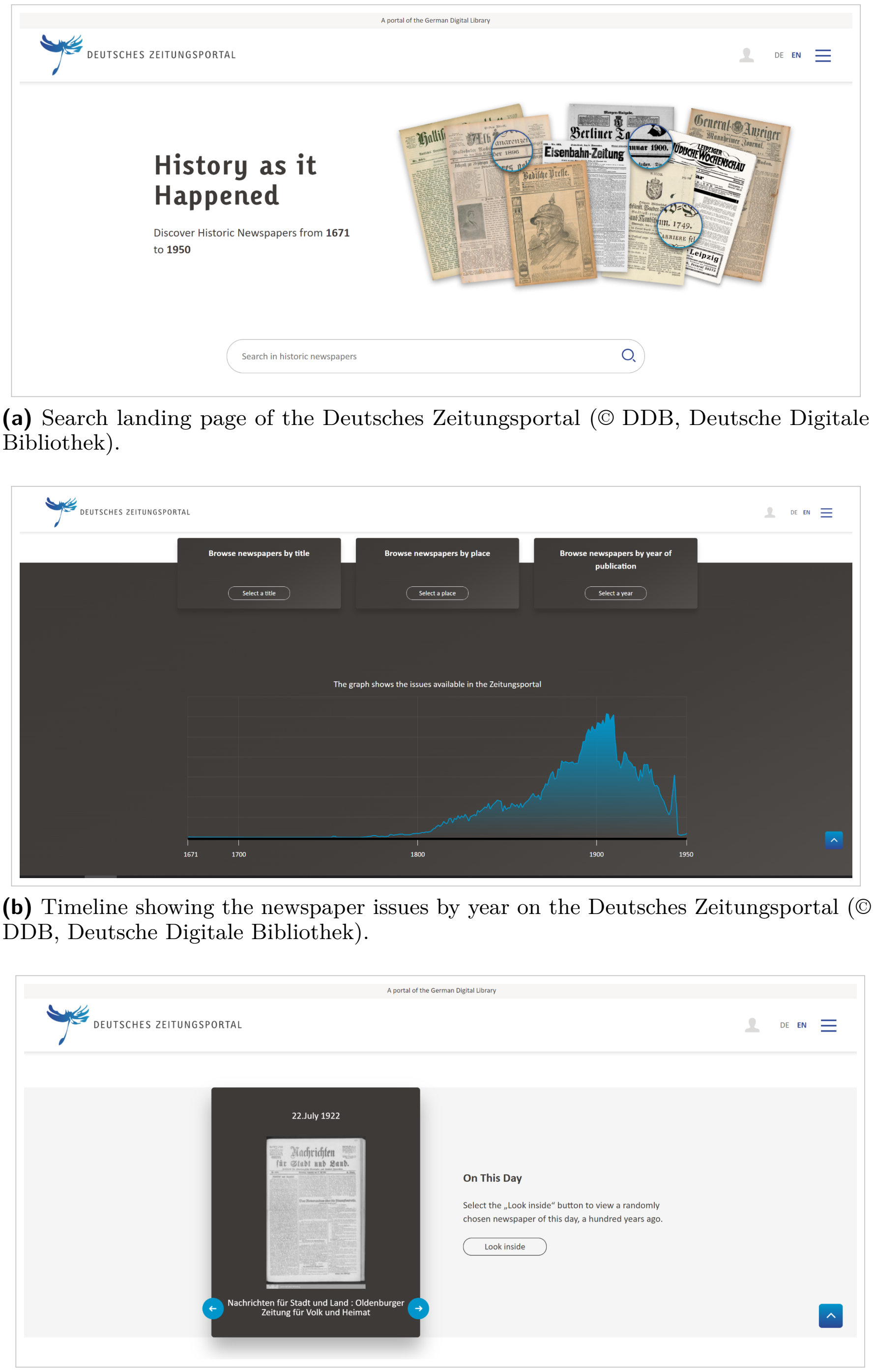
The "About" page indicates that it is a federated site that provides access to newspapers held at different German institutions. It provides data on the total number of newspapers: "The Deutsches Zeitungsportal was launched in October 2021 with 247 newspapers, 591,837 newspaper issues and a total of 4,464,846 newspaper pages from nine libraries. The offerings are being continually expanded and, in the long run, should comprise all digitised historical newspapers which are stored in German cultural and scientific institutions." They also indicate that it is not a representative selection[38], but how "not representative" it is, is not indicated. There is an alphabetical list of all the newspapers with information on their publication history, frequency of publication, and area of distribution.[39] Only 82% of the articles are full-text indexed, but it is unclear which parts of the collection it concerns. This makes it very hard to do source criticism on this collection.

A keyword query for "vaccination" in the search box generates a graph and result-list with snippets organised by titles: apparently 279 results from 26 June, 1802, until 5 June, 1950, were found. Results can be sorted by relevance, but it is unclear how that is defined. Alternatives are sorting functions by publication date (oldest first, newest first) or A-Z or Z-A, where results are apparently ranked by newspaper title.

The earliest mention of "vaccination" is in the *Hallesches Tageblatt* of 26 June, 1802, where it is mentioned in a section on "Kuhpocken" ("cow pocks").

However, a wildcard search of "vaccin*" gives 759 results with the oldest in the *Gülich und bergische wöchentliche Nachrichten* of 20 May 1783, but there it mentions the Latin

[37] https://www.deutsche-digitale-bibliothek.de/newspaper
[38] https://www.deutsche-digitale-bibliothek.de/content/newspaper/fragen-antworten
[39] https://www.deutsche-digitale-bibliothek.de/newspaper/select/title



A portal of the German Digital Library

DEUTSCHES ZEITUNGSPORTAL

DE EN

History as it Happened

Discover Historic Newspapers from **1671** to **1950**

Search in historic newspapers

(a) Search landing page of the Deutsches Zeitungsportal (© DDB, Deutsche Digitale Bibliothek).

DEUTSCHES ZEITUNGSPORTAL

DE EN

Browse newspapers by title

Select a title

Browse newspapers by place

Select a place

Browse newspapers by year of publication

Select a year

The graph shows the issues available in the Zeitungsportal

(b) Timeline showing the newspaper issues by year on the Deutsches Zeitungsportal (© DDB, Deutsche Digitale Bibliothek).

A portal of the German Digital Library

DEUTSCHES ZEITUNGSPORTAL

DE EN

22.July 1922

Nachrichten für Stadt und Land : Oldenburger Zeitung für Volk und Heimat

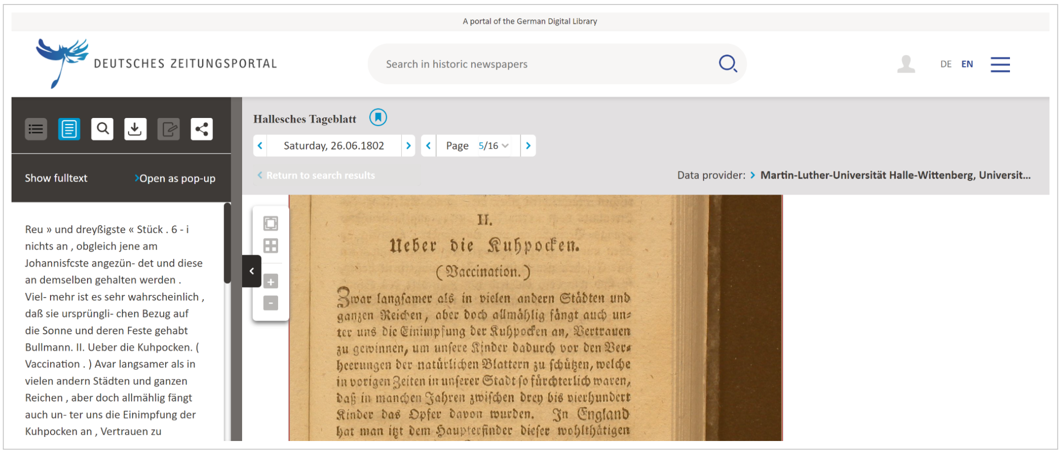
On This Day

Select the „Look inside“ button to view a randomly chosen newspaper of this day, a hundred years ago.

Look inside

(c) Example newspaper page on the Deutsches Zeitungsportal (© DDB, Deutsche Digitale Bibliothek).

Figure 9 User interface of the Deutsches Zeitungsportal.


A portal of the German Digital Library
DEUTSCHES ZEITUNGSPORTAL
Search in historic newspapers
DE EN
Show fulltext
Open as pop-up
Hallesches Tageblatt
Saturday, 26.06.1802
Page 5/16
Return to search results
Data provider: Martin-Luther-Universität Halle-Wittenberg, Universit...

Reu » und dreyßigste « Stück . 6 - i nichts an , obgleich jene am Johannisfcste angezün- det und diese an demselben gehalten werden . Viel- mehr ist es sehr wahrscheinlich , daß sie ursprüngli- chen Bezug auf die Sonne und deren Feste gehabt Bullmann. II. Ueber die Kuhpocken. (Vaccination .) Avar langsamer als in vielen andern Städten und ganzen Reichen , aber doch allmählig fängt auch un- ter uns die Einimpfung der Kuhpocken an , Vertrauen zu

II.

Ueber die Kuhpocken.

(Vaccination.)

Zwar langſamer als in vielen andern Städten und ganzen Reichen, aber doch allmählig fängt auch unter uns die Einimpfung der Kuhpocken an, Vertrauen zu gewinnen, um unſere Kinder dadurch vor den Verheerungen der natürlichen Blattern zu ſchützen, welche in vorigen Zeiten in unſerer Stadt ſo fürchterlich waren, daß in manchen Jahren zwiſchen drey bis vierhundert Kinder das Opfer davon wurden. In England hat man itzt dem Haupterfinder dieſer wohlthätigen

Figure 10 Earliest example of “Vaccination” as shown on the Deutsches Zeitungsportal (© DDB, Deutsche Digitale Bibliothek).[40]

“Vaccinium” which refers to a blueberry[41] – considering that the word vaccination was invented by Jenner in 1796 this result clearly is off topic. The earliest mention from 22 February 1802, is in the *Karlsruhe Zeitung*, the newspaper that most often contains the term (107 articles, 14% of the total). The results page contains a result hit timeline that reveals peaks in 1871-1874 (coinciding with the smallpox pandemic of 1870-1874), 1884 (perhaps a late response to Pasteur’s publication on vaccination of 1880?), one around 1890 (perhaps new vaccinations found) and a final one in 1913 (with reference to the use of vaccinations at war time), after which the references decline.

The portal allows users to filter by newspaper title or distribution area (or period), but the functionalities are too limited for putting together a research corpus for our question. The FAQ section points to the well-documented API[42] where the portal allows digitally literate and registered users to extract data.

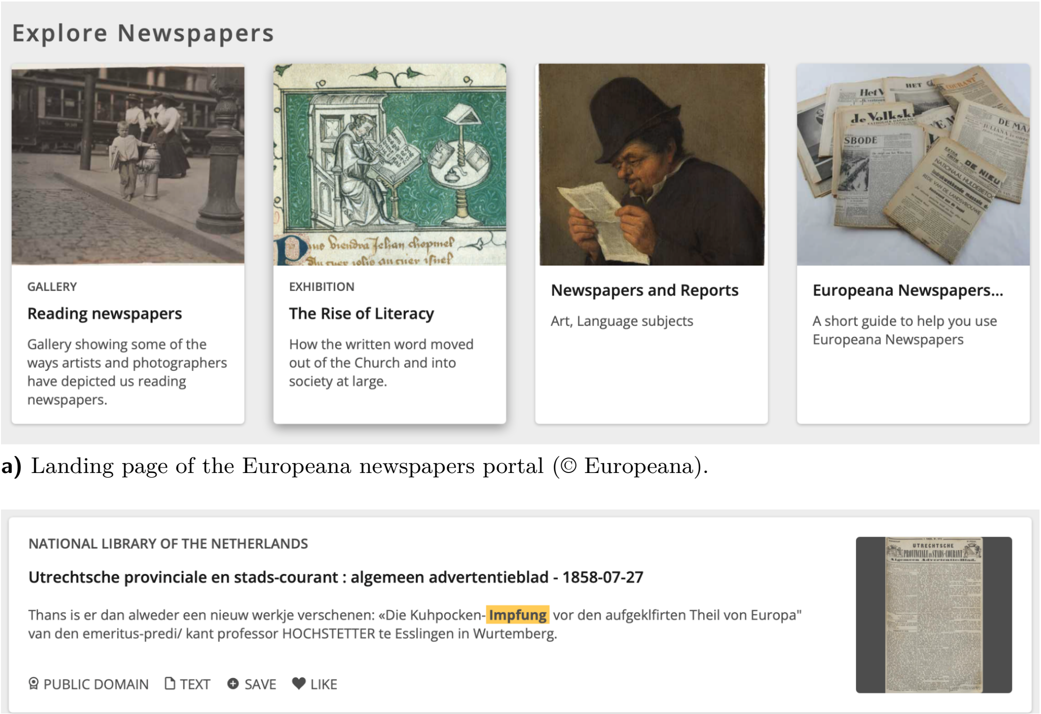
To conclude: the portal allows for exploratory search but is not suited for building a research corpus due to a lack of transparency on the scope and quality of the underlying collections and their processing. The API should be used to extract a corpus and for quality assessments, but this requires technical expertise most historians do not have.

Europeana Newspapers

The Europeana portal includes a “Newspapers Theme”[43]. The title of the theme is “Explore the headlines, articles, advertisements, and opinion pieces from European newspapers from 20 countries, dating from 1618 to the 1980s.” It includes 887,607 items from ten European countries (Austria, Estonia, Finland, Germany, Italy, Latvia, Luxembourg, The Netherlands, Poland and Serbia). However, it is not possible to see a listing of which titles are included.

[40] https://www.deutsche-digitale-bibliothek.de/newspaper/item/N5UFNO5HCR36P7BJK3TYEI5XM4IR6UUK?issuepage=5
[41] https://www.deutsche-digitale-bibliothek.de/newspaperitem/R2LPDFW7YX27WTEOLNLEBIE4PCDKF66Q?issuepage=4
[42] https://labs.deutsche-digitale-bibliothek.de/app/ddbapi/
[43] https://www.europeana.eu/en/collections/topic/18-newspapers

Additional content is provided at the end of the page, including a gallery on Reading Newspapers, Exhibition on the Rise of Literacy, teaching information on Newspapers, Reports, and a short guide to the use of Europeana Newspapers.



Explore Newspapers

GALLERY
Reading newspapers
Gallery showing some of the ways artists and photographers have depicted us reading newspapers.

EXHIBITION
The Rise of Literacy
How the written word moved out of the Church and into society at large.

Newspapers and Reports
Art, Language subjects

Europeana Newspapers...
A short guide to help you use Europeana Newspapers

(a) Landing page of the Europeana newspapers portal (© Europeana).

NATIONAL LIBRARY OF THE NETHERLANDS

Utrechtsche provinciale en stads-courant : algemeen advertentieblad - 1858-07-27

Thans is er dan alweder een nieuw werkje verschenen: «Die Kuhpocken-Impfung vor den aufgeklfirten Theil von Europa" van den emeritus-predi/ kant professor HOCHSTETTER te Esslingen in Wurtemberg.

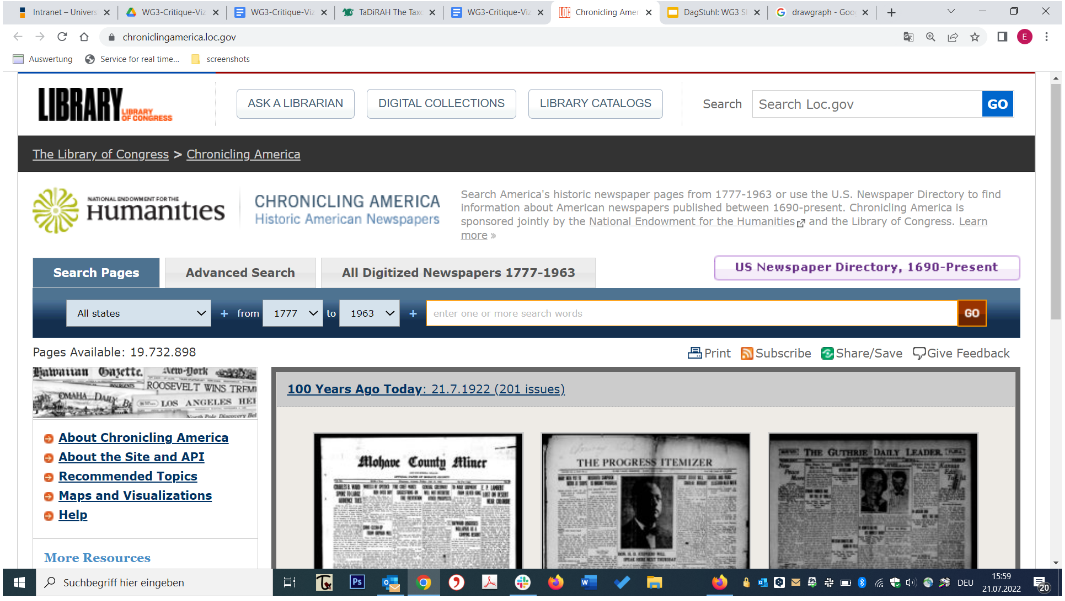
(b) Example of search results for the query "vaccination" as shows on the Europeana newspaper portal (© Europeana).

Figure 11 User interface of the Europeana newspaper archive.

For the "vaccination" research questions, since it is a multilingual portal it is important first to assess which search terms could be used to find relevant newspaper articles on vaccines. Based on this author's language skills a search was undertaken on "vaccine" (EN: 4,626 results), "Impfung" (DE: 12,647 results) and "vaccin" (FR: 4,607 results). Using a wildcard, e.g. "vaccin*" (10,734 results), articles in other languages (e.g., Italian) were included. It was not possible to sort the results. However, a range of result filters are available, e.g., by language or providing country. Additionally there was a "date issued" filter which, however, was not easy to use. Each of the search results provided an overview of the providing institution, the title of the newspaper, as well as a text snippet including the search term, and a thumbnail of the newspaper in question (see below).

Relevant articles can be saved by the user in a personal "gallery" (following the creation of a Europeana account). This gallery can be kept "private" or made "public". It is possible to share a public gallery on social media, e.g. on Twitter[44]. It is possible to download individual search results as page images (jpeg). There does not seem to be an advanced search function. There is some filtering of search results, however, they are not sufficient enough to answer our research questions.

[44] https://www.europeana.eu/en/set/7143



chroniclingamerica.loc.gov

Search Loc.gov

The Library of Congress > Chronicling America

Search America's historic newspaper pages from 1777-1963 or use the U.S. Newspaper Directory to find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress. Learn more »

All states + from 1777 to 1963 + enter one or more search words

Pages Available: 19.732.898

100 Years Ago Today: 21.7.1922 (201 issues)

Suchbegriff hier eingeben

Figure 12 Screenshot of the landing page of Chronicling America (© Library of Congress).

Chronicling America

The landing page of the Library of Congress collection of historical US-newspapers[45] offers several facets and has tabs to give access to the collection and offers links to information pages, APIs, as well as help files, thematic corpora, maps, and visualisations. A search bar and tabs in the background indicate that there are more search options available for advanced search and more complex investigations of the collection. The attention of users is drawn to the centre of the page where a selection of newspaper front pages are displayed under the heading “100 Years Ago Today”, today’s date and the number of newspaper issues collected.

The collection is a composition of “historic US-newspapers from 1690 to the present”. The interface is the result of a collaboration between the Library of Congress and the National Endowment for the Humanities. It includes 3,758 newspapers with 19,7 million digitised pages. The APIs[46] enable expert users to perform the following tasks: search, auto-suggest from newspaper titles, link to stable URLs, linked data views of the collection, JSON view of data, bulk data to use with external services, and CORS- and JSONP-support for JavaScript applications. For all APIs explanations on use and examples are given. Under the heading “Recommended Topics” thematic features in Chronicling America are collected. These corpora are arranged alphabetically, by category, and by date range. They cover a growing number of different themes, time-spans, and genres. The section heading “Maps and Visualizations” leads to a number of graphs and data visualisations of the collection. These pages are updated on a regular basis. So, while the landing page and the simple search bar may give the impression that this collection is meant for a general audience, both the sub-sites and the accessible design of the “Advanced Search” function are clearly intended for expert users.

[45] `https://chroniclingamerica.loc.gov/`
[46] `https://chroniclingamerica.loc.gov/about/api/`

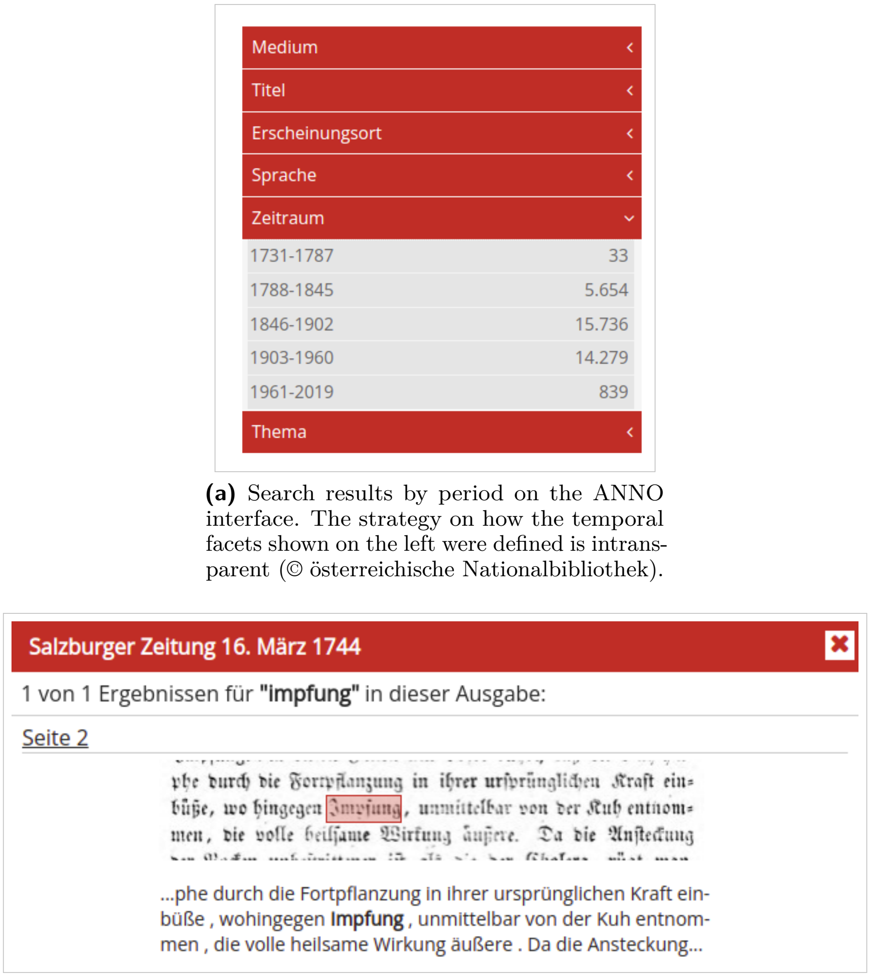
The collection is composed of newspapers in 19 languages: English is the dominant language (with 18,7 mio pages), followed by German (500,000), and Spanish (330,000). At the end of the scale Hebrew (830) and Arabic (2,000) can be found.

With regard to transparency and fairness we wish to highlight dedicated visualisations on the distribution of ethnic press coverage within the corpus. A keyword query for "vaccine" using the basic search bar leads to 208,360 results. The wildcard search for "vaccin*" to capture also results for "vaccination" or "vaccinated" led to slightly over 207,000 results. This apparently wrong output was quickly resolved by reading the help files which indicated that wildcards, as well as upper-/lower-case search, and simple Boolean operators are not implemented. However, the search engine utilises language specific dictionaries which use stemming to include word variants. In order to limit (or increase) the search results, combinations of words or the features offered in the "Advanced Search" should be used (here filters on states, titles, years, front pages, language, combination of words, phrase search, and distance search are implemented). The search for "vaccine fear" produces 69,762 results. A quick scan of the results showed, however, that the two terms often do not occur in the same news item so that this keyword search does not produce usable results. A distance search of the two terms (with a distance of 10 words) produced 1,344 results which did not prove more appropriate (corresponding to sentences similar to "I fear that . . . "). Finally, the combination of the terms "vaccine" and "effect" in a distance search of 10 words led to 5,634 results that could be used to study newspaper reporting of this topic. However, it remains uncertain if the word "effect" really covers what a user was looking for in the context of discourses on vaccination. Bulk downloads are not possible at this level. Another point of "granular access" (as opposed to bulk download) is the Chronicling America API. Programmatic access is often preferable for computational analysis, as retrieving and processing data can be easily integrated into one workflow. However, the API functionality is in many ways similar to keyword search. The main functionality is search defined by a query term and refined by a few additional parameters. As can be gathered from the online documentation, the API is especially useful when a researcher wants to retrieve documents related to a specific topic in bulk. Of course, additional filtering can happen downstream in custom-made scripts, but it does not seem to be part of the API functionality (or at least is not very well publicised on the main page).[47]. Having said that, the API is undoubtedly easy to use and the examples are easily adaptable. We successfully used the search endpoint to retrieve articles that mention "vaccination" as a starting point for further processing.

ANNO

AustriaN Newspapers Online (ANNO)[48] is a digitisation project of the Austrian National Library for Austrian historical newspapers and magazines. The project was launched with 15 newspapers in August 2003, and now, more than 25 million pages of more than 1,500 newspapers and magazines can be read and downloaded free of charge and in full text from the portal. The oldest editions date back to 1568. Like other newspaper portals, ANNO offers users to browse and read digital newspapers, and to search for articles based on keyword or an advanced search with additional filters (publication place, date, language, and topic). First hits for "vaccin" are found in the Italian paper *Il Corriere ordinario* from 1679, however referring to cows, a common false positive result we have also observed in other portals.

[47] https://chroniclingamerica.loc.gov/about/api/
[48] https://anno.onb.ac.at/



1731-1787	33
1788-1845	5.654
1846-1902	15.736
1903-1960	14.279
1961-2019	839

(a) Search results by period on the ANNO interface. The strategy on how the temporal facets shown on the left were defined is intransparent (© österreichische Nationalbibliothek).

Salzburger Zeitung 16. März 1744

1 von 1 Ergebnissen für **"impfung"** in dieser Ausgabe:

Seite 2

phe durch die Fortpflanzung in ihrer urſprünglichen Kraft einbüße, wo hingegen Impfung, unmittelbar von der Kuh entnommen, die volle heilſame Wirkung äußere. Da die Anſteckung

...phe durch die Fortpflanzung in ihrer ursprünglichen Kraft einbüße , wohingegen **Impfung** , unmittelbar von der Kuh entnommen , die volle heilsame Wirkung äußere . Da die Ansteckung...

(b) Screenshot of an individual result (© österreichische Nationalbibliothek).

Figure 13 User interface of the ANNO portal.

ANNO does not provide visual cues that summarise metadata of the retrieved results such as, e.g., *impresso* does. The results are displayed in a faceted browser environment, and facets, for which numerical information are provided, can be selected and deselected. The default ranking of results is by relevance, however, it is not traceable how relevance is defined. Ordering of results by other metadata like date is also possible. Clicking a result opens a popup that juxtaposes scan and OCR transcript, and highlights the search term(s) in both views. ANNO also includes the option to use Boolean operators. Alongside this common search and filter features, ANNO supports filtering by language and themes such as “science” or “agriculture” but it remains unclear, how these filters and the underlying data were generated. A Help and FAQ sections explains available functionalities but do not include information about the technical processing.

NewsEye

The NewsEye portal[49] includes newspaper data from various countries. The data are from different time periods, which makes a comparative analysis difficult. Keyword search in combination with filters can be used to to search through the data. Results are presented as snippets, allowing to quickly assess the relevance of the results, and thus the appropriateness of the keywords. The portal allows users to create and store a custom research dataset by selecting articles or newspaper issues from the search results page. This increases transparency for peers with regard to which data a researcher used for their analysis. The portal does not support transparency with respect to the methods used to select data. A systematic method could be, for example, to fix a set of keywords/facets, and include the resulting articles.

The portal contains an interface to create and store experiments, i.e., sequences of data processing steps (Figure 14a). This functionality increases transparency of the analysis step: not only can the pipeline be stored for future use, it also makes it easy to compare output of different pipelines.

The NewsEye platform used for the current exploration was the experimental platform of the NewsEye project. So, some functionalities were not implemented in this interface yet: the help-button did not work yet; some of the facets still produced unexpected results; only a small number of simple data processing tools were included in the experiment interface (e.g., stopword removal).

The search for the truncated word "vaccin*" produced 25,255 search results in Swedish, French, English, German, and Finnish (Figure 14b). A random check of documents in the different languages confirmed that these were indeed related to vaccinatation. This result is not surprising as in many languages the word vaccination was derived from the Latin "vaccinus" (from the cow). A graph gives an overview of the distribution of the search hits over time. Several facets allow researchers to dig deeper into the search results. Results can additionally be sorted by date or relevance score and the function "random sample" gives a quick overview of what the reader can expect to find in the results. The first mention of "vaccin*" in the newspapers aggregated in NewsEye is in Swedish from the Finnish title *Abo Underrattelser* from 13 March, 1824, where the distribution of vaccines in Finland is discussed.

Trove

The Trove portal[50] aggregates a wide range of textual and visual digitised and born-digital resources (books, newspapers, websites, images), hosted by Australian cultural heritage institutions. Trove Newspapers offers a clean interface with common (advanced) search and filtering options alongside the notebook-based GLAM-workbench[51]. An informative About[52] section gives a concise overview of the whole "Trove ecosystem", its construction and guiding principles and is accompanied by a Research Guide[53] which offers basic insights into the availability and legal status of the collection. User expectations are managed effectively through additional documentation, e.g. on a variety of errors and instructions for correction

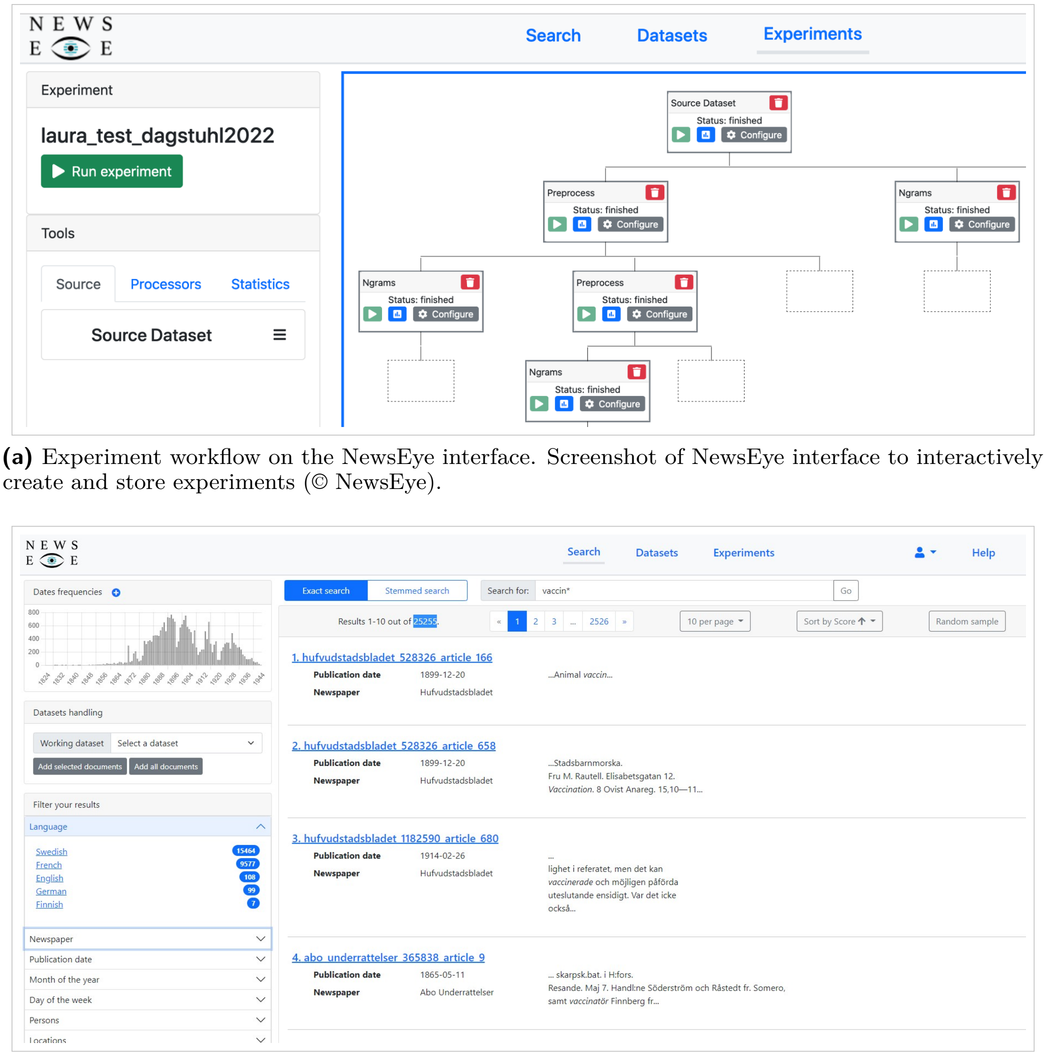
[49] `https://platform2.newseye.eu/`

[50] `https://trove.nla.gov.au/newspaper/`

[51] `https://mybinder.org/v2/gh/GLAM-Workbench/trove-newspapers/master?urlpath=lab/tree/index.ipynb`

[52] `https://trove.nla.gov.au/about`

[53] `https://www.nla.gov.au/research-guides/australian-newspapers`



(a) Experiment workflow on the NewsEye interface. Screenshot of NewsEye interface to interactively create and store experiments (© NewsEye).

Results 1-10 out of 25255.

1. hufvudstadsbladet_528326_article_166
Publication date 1899-12-20
Newspaper Hufvudstadsbladet
...Animal *vaccin*...

2. hufvudstadsbladet_528326_article_658
Publication date 1899-12-20
Newspaper Hufvudstadsbladet
...Stadsbarnmorska.
Fru M. Rautell. Elisabetsgatan 12.
Vaccination. 8 Ovist Anareg. 15,10—11...

3. hufvudstadsbladet_1182590_article_680
Publication date 1914-02-26
Newspaper Hufvudstadsbladet
...
lighet i referatet, men det kan
vaccinerade och möjligen påförda
uteslutande ensidigt. Var det icke
också...

4. abo_underrattelser_365838_article_9
Publication date 1865-05-11
Newspaper Abo Underrattelser
... skarpsk.bat. i H:fors.
Resande. Maj 7. Handl:ne Söderström och Råstedt fr. Somero,
samt *vaccinatör* Finnberg fr...

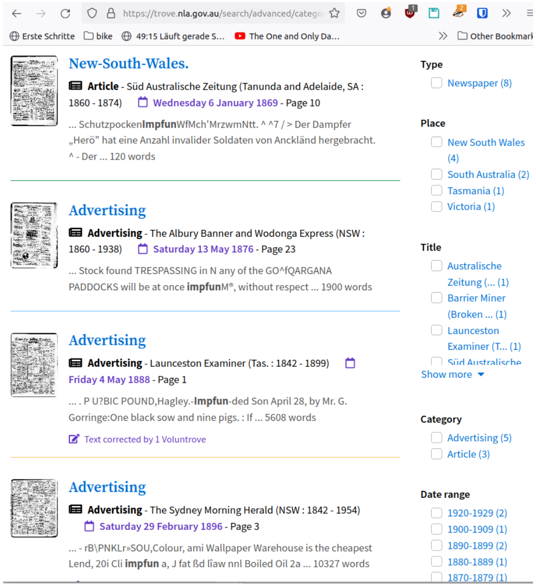
(b) Search results for "vaccin*" in the NewsEye platform (© NewsEye).

Figure 14 User interface of the NewsEye portal.

for volunteers.[54] Noteworthy is also optional information concerning cultural sensitivity which users are free to en- or disable throughout their interaction with the portal. During this limited testing we were however not able to see it in action but learned that users are encouraged to ammend DublinCore and MARC metadata of affected articles with the reference "Culturally sensitive". Users are furthermore able to filter content by region, content type, media, and content length with the notable absence of language.

Our case study on vaccination reveals the tremendous added value of crowd sourcing and its effective implementation in Trove. The earliest reference can be found with a query for "vaccin*" and retrieves an article published in the *Sydney Gazette and New South Wales Advertiser* in 1803. The article covers an experimental treatment of orphans with early vaccines against cow pocks including the assertion that "It is believed, that it never has been

[54] `https://trove.nla.gov.au/help/become-voluntrove/text-correction`



https://trove.nla.gov.au/search/advanced/catego

Erste Schritte · bike · 49:15 Läuft gerade S… · The One and Only Da… · Other Bookmark

New-South-Wales.

Article - Süd Australische Zeitung (Tanunda and Adelaide, SA : 1860 - 1874) Wednesday 6 January 1869 - Page 10

… Schutzpocken**Impfun**WfMch'MrzwmNtt. ^ ^7 / > Der Dampfer „Herö" hat eine Anzahl invalider Soldaten von Anckländ hergebracht. ^ - Der … 120 words

Advertising

Advertising - The Albury Banner and Wodonga Express (NSW : 1860 - 1938) Saturday 13 May 1876 - Page 23

… Stock found TRESPASSING in N any of the GO^fQARGANA PADDOCKS will be at once **impfun**M®, without respect … 1900 words

Advertising

Advertising - Launceston Examiner (Tas. : 1842 - 1899) Friday 4 May 1888 - Page 1

… . P U?BIC POUND,Hagley.-**Impfun**-ded Son April 28, by Mr. G. Gorringe:One black sow and nine pigs. : If … 5608 words

Text corrected by 1 Voluntrove

Advertising

Advertising - The Sydney Morning Herald (NSW : 1842 - 1954) Saturday 29 February 1896 - Page 3

… - rB\PNKLr»SOU,Colour, ami Wallpaper Warehouse is the cheapest Lend, 20i Cli **impfun** a, J fat ßd lîaw nnl Boiled Oil 2a … 10327 words

Type

☐ Newspaper (8)

Place

☐ New South Wales (4)
☐ South Australia (2)
☐ Tasmania (1)
☐ Victoria (1)

Title

☐ Australische Zeitung (… (1)
☐ Barrier Miner (Broken … (1)
☐ Launceston Examiner (T… (1)
☐ Süd Australische

Show more

Category

☐ Advertising (5)
☐ Article (3)

Date range

☐ 1920-1929 (2)
☐ 1900-1909 (1)
☐ 1890-1899 (2)
☐ 1880-1889 (1)
☐ 1870-1879 (1)

Figure 15 Ranking of search results for “Vaccination” as shown in the Trove interface (© National Library of Australia and Partner Institutions).

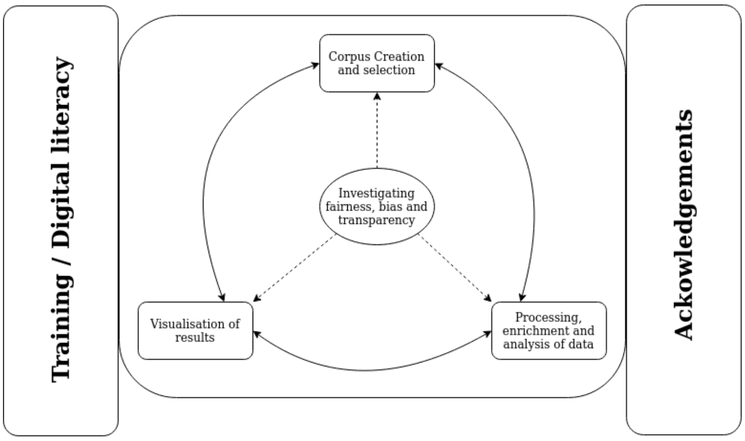
fatal, and never will be”. The query term “vaccin*” does not occur in the text which has been manually transcribed by volunteers. Instead, the article has been tagged manually with the term “Vaccination” alongside other helpful yet anachronistic tags such as “Bioethics” or “Clinical trial”. Trove search includes such tags as well and thereby helped retrieve this article. A distribution over time of search results is possible via hits per year counts but this feature is basic. Information e.g. regarding the breadth of content type detection across the corpus is missing as is information concerning the overall representativity of the corpus for the Australian press. The interface supports crowdsourced OCR correction and informs about the number of “Voluntroves” who worked on a given article. Overall, Trove stands out regarding audience-integration: Crowdsourcing and -annotation features, notebook-infrastructure and cultural sensitivity are well integrated and cater to the needs of different user groups.

4.3.5 Consolidation: Issues and recommendations with respect to fairness and transparency in each stage of the workflow

We identified distinct stages in a typical digital humanities workflow for the analysis of digitised historical newspapers:

1. Research corpus creation and selection,
2. Processing and enrichment,
3. Data analysis,
4. Visualisation of results,
5. Training,
6. Acknowledgements.

Figure 16 illustrates how the stages interrelate. The stages are often performed in an iterative fashion. In this section, we discuss issues with respect to fairness and transparency in each of the stages, and present recommendations for how to deal with them.


Training / Digital literacy
Corpus Creation and selection
Investigating fairness, bias and transparency
Visualisation of results
Processing, enrichment and analysis of data
Ackowledgements

Created with https://app.diagrams.net/

Figure 16 Typical workflow of a researcher. It is essential to investigate fairness, bias and transparency at various stages and in a continuous fashion. It is to be noted that if the training is a prerequisite to the creation of a research dataset, it never really stops (© Axel Jean-Caurant).

4.3.6 Research corpus creation, selection and sharing

▶ Focus areas for transparency and fairness

Awareness of the digitisation and preservation policies. The "digital sample" constitutes a subset of the totality or population of newspapers which existed at some point in time. This population is unknown, but can be approximated using contextual resources such as newspaper press directories or library catalogues [2]. Both are simultaneously useful and problematic. Catalogues record mainly what sources are preserved and are close to providing an overview of the complete newspaper record especially for countries with legal deposit. Newspaper press directories are a useful historical source, but also come with their own issues: classification of what "is" a newspaper changes over time and varies by directory. Ultimately, the population or "newspaper landscape" remains unknown, but we can describe those that have been recorded and/or preserved using metadata derived from catalogue or contextual resources [2].

A rich description of the newspaper landscape helps to contextualise and situate the "sample" of digitised newspapers. In other words, it enables us to at least approximate the "representativeness" of the collection, bearing in mind that the latter concept is more complex than simple proportionality and is always defined in relation to the research questions and ethical values or priorities of the researcher. When it comes to the composition of a corpus, we as researchers and content providers will never be able to get rid of biases and unwanted over- and under-representations. Our goal must rather be to identify, understand, acknowledge them, to infer how they may influence the research outcomes and to make them clearly visible within portals.

When using newspapers at scale and-to repeat the metaphor-as a "mirror" of the past, assessing diversity of collections emerges as a critical issue alongside the processes of media production which heavily influences how the presence was reflected. Researchers need to acknowledge whose perspectives or voices are absent in the data and which social categories dominate a collection.

The question of representativeness is closely intertwined with the issue of diversity and inclusion: which (social) perspectives are present in our data, which are missing? Coming back to the metaphor of newspapers as a "mirror" of the past, we can not simply trust the

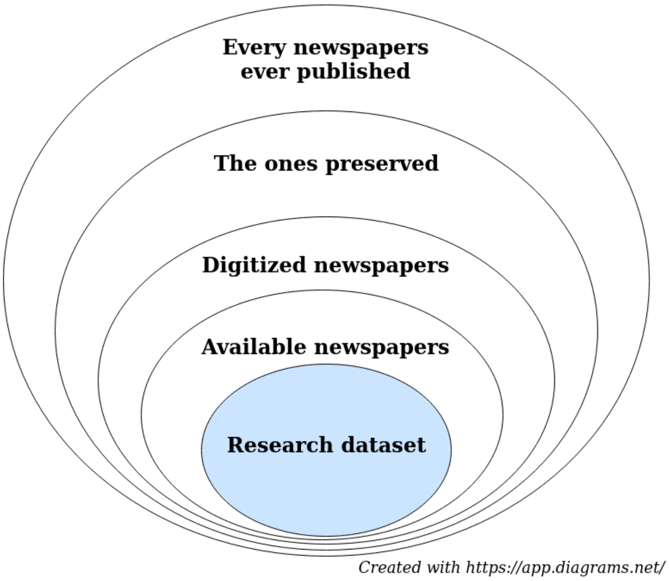


Figure 17 A research dataset is inherently biased, as it is impossible to create a complete dataset because of missing or unavailable sources (© Axel Jean-Caurant)

reflection but need to assess how it potentially distorts our image of the past. With rich descriptions of the sample (and population) we can situate data historically and socially. Of course these will always be rough and approximate descriptions, but nonetheless a crucial part of contextualising (the results derived from) big data. Finally, digitisation does not automatically mean accessibility which depends on the institutional policies (e.g. paywalls) and legal restrictions.

Diverse user needs. Keyword search may satisfy a large group of researchers (and laymen), but others may want to go beyond simply retrieving and reading newspaper content. Interfaces simultaneously provide and restrict access, i.e., the inbuilt functionalities set the limits to how users can navigate and analyse historical materials. They provide the tools and heuristics via which content becomes visible and users can create their research corpus or data set (see below).

However, while such type of access generally works for humanities' scholars like historians, it does not necessarily meet the needs of those who follow more data-driven approaches, such as computational humanities researchers, computational historians, or NLP researchers. The latter often wish to process larger datasets for automatic enrichment and filtering, among other tasks. While most libraries or platforms provide access via search, accessing "newspaper collections as data" (i.e., at scale) is becoming more prevalent, but contemporary portals remain limited in their support for such interactions with the data.

Toxicity and cultural bias. As newspapers are embedded in specific spatial and temporal contexts, their content also contains traces of historical biases, both in text and image. In the most extreme cases, historical newspapers contain "toxic" content, to use a term common in today's research on language models and ethical AI. The textual (or visual depiction) of people, especially the more marginalised and underprivileged, articulate attitudes which are considered offensive within contemporary norms.

But not all biased language is “toxic”. A more neutral term would be “overrepresentation” of specific textual patterns among certain subsets of the data, for example conservative newspapers may mention words such as “agriculture” more frequently than newspapers of other political leaning.

▶ Measures to help achieve transparency and fairness

Related to the activity of “research corpus creation, selection and sharing”, there are a number of measures that could help or improve transparency and fairness. These measures are both at the level of the cultural heritage institutions providing the digitised newspaper collections as well as at the level of the researchers who create their research corpora.

Collection documentation. Providers of digitised historical newspapers, such as cultural heritage institutions and specific newspaper portals (e.g., ANNO, BelgicaPress, Chronicling America, Delpher, impresso, NewsEye, etc.) can provide detailed information regarding the collection. For example: list of newspaper titles, dates of publication, how much of a newspaper title has been digitised. It could also be useful to provide whatever contextual information about a newspaper and the entire collection is available, e.g., concerning selection criteria, geographical scope, number of editions, print runs, publishers and editors. Information such as political orientation of the newspaper titles, even when imperfect and tied to specific time periods, will be useful here. Ideally, such contextual information is accompanied by sources such as bibliographic references. The question of who is responsible for providing this information was raised, e.g., the cultural heritage institution or the researcher undertaking the research. Perhaps a partnership between these two actors would be most valuable.

Figures 18 illustrates how contextual information could be displayed: firstly, the Newspaper Timelines[55] from the *impresso* project, and secondly, the Press Picker[56] from the Living with Machines project.

The provision of explicit information regarding digitisation quality (e.g., Optical Character Recognition, OCR) would also be useful, ideally provided at a number of levels: for the whole newspaper title, for an issue, or per article.

Terms of use. Digitised newspaper providers should provide explicit guidelines regarding terms of use, particularly in terms of legal consideration. For example, the *impresso* platform requires users to sign a Non-Disclosure Agreement (NDA)[57] before access to full collection is granted. Furthermore, it would be useful for cultural heritage providers to provide information about what percentage of the total collection has been digitised. This helps to provide transparency on the “missingness” in a collection, ideally at the level of each of the newspaper titles.

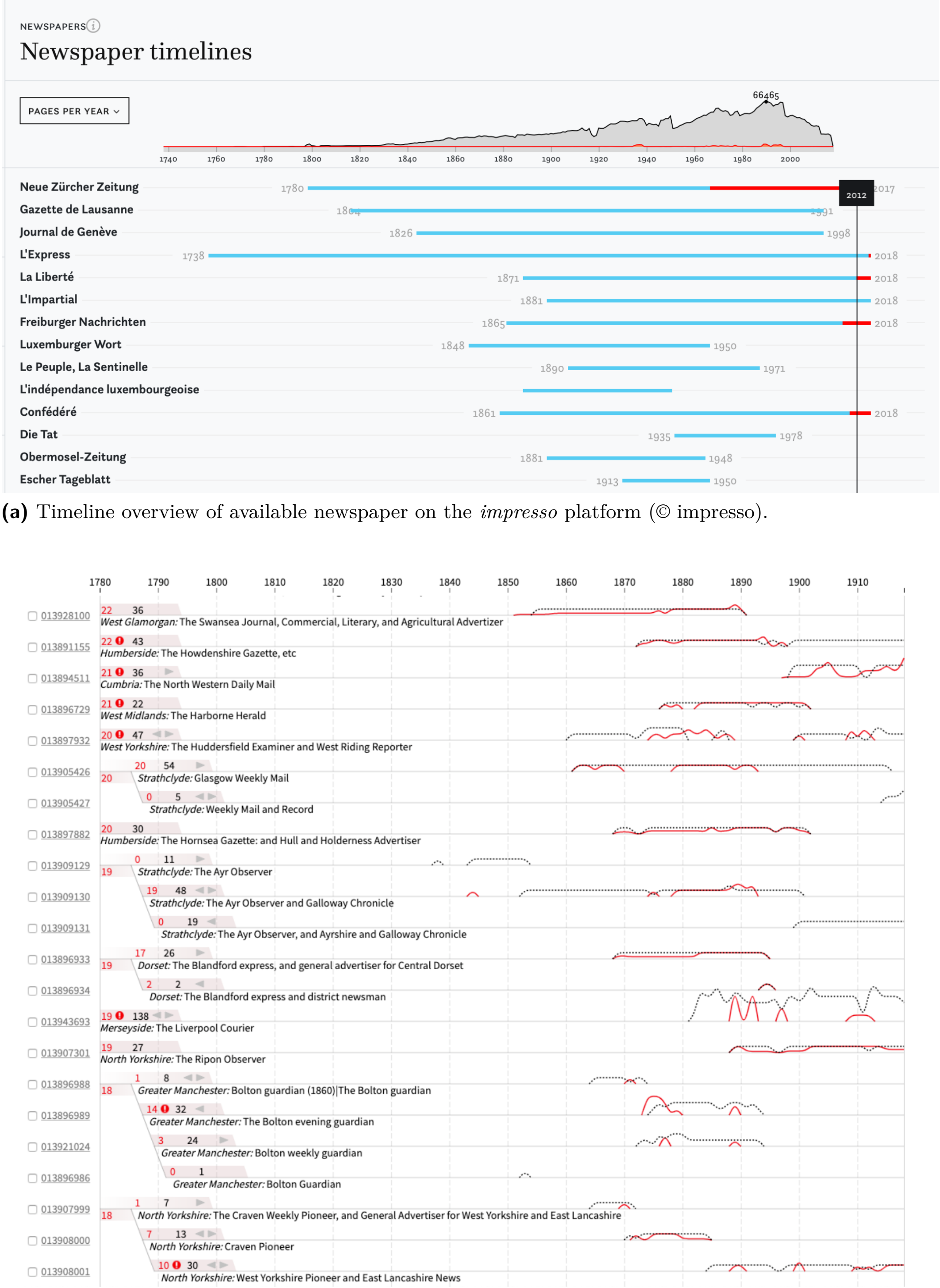
Contested terms. When considering measures to ensure transparency and fairness of research corpus creation, selection and sharing, it is important to consider diversity, equality, equity. To assist both researchers and cultural heritage institutions with this, an equity monitor could be developed. A number of aspects could be considered; for example, the identification of contested terms in a corpus (see [11, 6], as well as [3] and Conconcor[58]). If collection holders are aware that their corpora include contested terms, a disclaimer

[55] https://impresso-project.ch/app/newspapers/

[56] https://livingwithmachines.ac.uk/press-picker-visualising-formats-and-title-name-changes-in-the-british-librarys-newspaper-holdings/

[57] https://impresso-project.ch/assets/documents/impresso_NDA.pdf

[58] https://www.cultural-ai.nl/conconcor


NEWSPAPERS
Newspaper timelines
PAGES PER YEAR
66465
1740
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Neue Zürcher Zeitung 1780 2012 2017
Gazette de Lausanne 1804 1991
Journal de Genève 1826 1998
L'Express 1738 2018
La Liberté 1871 2018
L'Impartial 1881 2018
Freiburger Nachrichten 1865 2018
Luxemburger Wort 1848 1950
Le Peuple, La Sentinelle 1890 1971
L'indépendance luxembourgeoise
Confédéré 1861 2018
Die Tat 1935 1978
Obermosel-Zeitung 1881 1948
Escher Tageblatt 1913 1950
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
☐ 013928100 22 36 West Glamorgan: The Swansea Journal, Commercial, Literary, and Agricultural Advertizer
☐ 013891155 22 43 Humberside: The Howdenshire Gazette, etc
☐ 013894511 21 36 Cumbria: The North Western Daily Mail
☐ 013896729 21 22 West Midlands: The Harborne Herald
☐ 013897932 20 47 West Yorkshire: The Huddersfield Examiner and West Riding Reporter
☐ 013905426 20 20 54 Strathclyde: Glasgow Weekly Mail
☐ 013905427 0 5 Strathclyde: Weekly Mail and Record
☐ 013897882 20 30 Humberside: The Hornsea Gazette: and Hull and Holderness Advertiser
☐ 013909129 19 0 11 Strathclyde: The Ayr Observer
☐ 013909130 19 48 Strathclyde: The Ayr Observer and Galloway Chronicle
☐ 013909131 0 19 Strathclyde: The Ayr Observer, and Ayrshire and Galloway Chronicle
☐ 013896933 19 17 26 Dorset: The Blandford express, and general advertiser for Central Dorset
☐ 013896934 2 2 Dorset: The Blandford express and district newsman
☐ 013943693 19 138 Merseyside: The Liverpool Courier
☐ 013907301 19 27 North Yorkshire: The Ripon Observer
☐ 013896988 18 1 8 Greater Manchester: Bolton guardian (1860)|The Bolton guardian
☐ 013896989 14 32 Greater Manchester: The Bolton evening guardian
☐ 013921024 3 24 Greater Manchester: Bolton weekly guardian
☐ 013896986 0 1 Greater Manchester: Bolton Guardian
☐ 013907999 18 1 7 North Yorkshire: The Craven Weekly Pioneer, and General Advertiser for West Yorkshire and East Lancashire
☐ 013908000 7 13 North Yorkshire: Craven Pioneer
☐ 013908001 10 30 North Yorkshire: West Yorkshire Pioneer and East Lancashire News

(a) Timeline overview of available newspaper on the *impresso* platform (© impresso).

(b) Timeline overview of British newspapers on the PressPicker tool (© Living With Machines).

Figure 18 Top: Timeline overview of available newspaper on the *impresso* platform (© impresso); Bottom: Timeline overview of British newspapers on the PressPicker tool (© Living With Machines).

alerting users to this could be added to the website. This could be particularly relevant when contentious terms are used for query expansion or are used in a visualisation. For researchers whose corpus includes contested terms, it is advisable to explicitly acknowledge this in their publications.

4.3.7 Processing and enrichment

▶ Focus areas for transparency and fairness

In almost all cases, a raw, digitised newspaper corpus is processed in several ways before it is suitable for analysis. This could include, for example, Optical Character Recognition (OCR), Optical Layout Recognition (OLR), lexical processing such as part-of-speech tagging, named entity recognition (NER), linking to knowledge graphs, topic detection, or sentiment analysis. It could also include manual annotation of, for example, topics, people, or viewpoints. Each processing and enrichment step introduces bias or unwanted over- and under-representations into the data. When data is retrieved via a search engine or recommendation system, the ranking algorithm of that system also plays a role.

Tool performance. Regarding "low" level processing tasks (OCR/OLR), bias is mainly related to the quality of the tools. Do they work equally well on each part of the collection? How does OCR/OLR quality impact the retrievability and accessibility of each (type of) document? The quality of the tools on each part of the collection will depend on the data that they were originally trained on or developed for. They will likely work best on data that resembles the training/development set.

Similarly, for higher level, automated enrichment tasks (part-of-speech tagging, NER, linking to knowledge graphs, topic detection, sentiment analysis), we can ask: how well do they work for each part of the collection? What data were the tools trained on or developed for, and to what extent is this different from the data currently under investigation? NER tools may show a higher performance on some entities than others. Knowledge graphs may not cover all relevant entities.

Ingrained bias. In some cases, bias is ingrained in the collection [3]. As a product of their times, historical newspapers will also reflect the norms, values and language of e.g. colonising nations and their perspectives on their colonies. The use of automated enrichment tools may lead to unwanted side effects with respect to colonial or otherwise outdated terminology. Words may be taken out of context. Consider, for example, that a topic detection algorithm may define a geographically-focused topic as a list of terms including racist references to people.

Posterior annotation. Manual annotation is prone to bias that relates to the viewpoints, background and knowledge of the annotator. In some cases, these highly personal characteristics will be unknown, such as when making use of crowdsourcing. In some cases, we might not want to expose anonymous crowd workers to bias that is ingrained in the collection, such as when offensive, colonial terminology is present.

Ranking. Search engines typically rank documents based on a combination of the following factors: a matching score between a query and the content of the document, usage data in the form of previous queries and clicks, and an importance score of the document, e.g., using a PageRank-like algorithm. Whether a document will appear high in the ranking will therefore be influenced by its popularity (if usage data is taken into account), connectedness (if PageRank is used), and document properties such as length, which impact the matching score [20]. This may introduce distorted perceptions of search results. Therefore ranking-algorithms should be made transparent to users who rely on them.

▶ Measures to help achieve transparency and fairness

Fine-grained OCR performance metrics. Detailed information about OCR and segmentation quality helps a user to decide not only whether the quality is good enough, but also whether bias towards certain parts of the research corpus is to be expected. This requires fine-grained performance metrics, for example at the level or articles, newspaper titles or time periods.

Access to "raw" data. Another solution to mitigate or at least understand bias due to OCR errors is to provide access to the original "raw" scans, i.e., the images of pages.

Systematic documentation of tools and training sets. For automated enrichment tools, documentation of how, for what purpose, and on which training set they were created, helps a user to assess whether bias is to be expected when these tools are applied to their data. Several documentation approaches have been proposed in recent years, the most notable being Datasheets for Datasets [6] for documentation of training sets, and Model Cards [10] for documentation of trained models. Also, the research on provenance is relevant here, i.e. a formal representation of the consecutive processes involved in the creation of the enrichments. PROV[59] is an approach to formally capture provenance information on the semantic web.

Scanning for contentious terms. The content of historic newspaper collections will often be "biased" in the sense that the articles display the perspectives of the time in which they were created. Removing this type of bias will mostly not be feasible or desirable in the context of historical research. We recommend to include an "equity monitor" as part of a research design, where a user critically assesses whether contentious terminology is present in the corpus, and whether this is problematic. As noted above, contentious terminology could be problematic when used as input to automated enrichment tools, or when presented to crowd workers. In these cases, a user could decide to not include certain articles in their research. Note that detection of contentious terminology is not trivial and automation of this task is still in its infancy [3].

Disclaimer about contentious language. A user may include a disclaimer as part of the dataset, to warn (other) users and/or annotators that there may be offensive content. This is especially recommended when sharing the corpus for reproducibility or future research.

Representative annotators. We consider human annotators to be always biased. A diverse or representative group of annotators helps to avoid annotations that are skewed towards one background or viewpoint.

Transparency about annotators. Explicit information about who created the annotations helps users to assess whether an (unwanted) bias in the annotations is to be expected. This could consist of age, gender, country of citizenship, (native) language, level of expertise, and way of recruitment of the annotator(s). Note that this information is often not available when using crowdsourcing.

Multiple relevance rankings. Bias introduced by the ranking algorithm of a search engine may be explicated and mitigated by providing multiple rankings. Many search engines already include additional rankings next to relevance ranking, such as a chronological order. However, specifically the inclusion of multiple relevance rankings would allow a user to understand to what extent the ranking algorithm impacts their goals.

[59] https://www.w3.org/2001/sw/wiki/PROV

4.3.8 Data analysis

Once a research corpus has been selected and processed it is ready for analysis. Analysis already is an integral part of data selection, processing and enrichment. At first, we discussed this stage as part of the processing and enrichment stage. There are, however, tasks that clearly come after data processing and enrichment; for instance, the identification of named entities in the corpus has to be undertaken with NER software. Therefore, we have decided to identify it as a separate step in the research workflow.

Depending on the research question and the skill set of researchers, the analysis may be performed on the entire collection (e.g., downloading all the data and analysing it with a Jupyter notebook) or of a subset of the collection generated via the search and filter options in a portal. It also may be performed qualitatively, with a scholar browsing and reading specific articles, or quantitatively, applying computational approaches and tools.

▶ Focus areas for transparency and fairness

Traceability. In order to make the research traceable, researchers have to be explicit about the methods and tools they use including, for the latter, the version, the used settings, and why they were appropriate for the task in question.

Often, the analysis of a newspaper corpus involves a set of tools organised in a pipeline. In order to obtain transparency, users should have insight into the composition and performance of the various components of the pipeline. In the case of machine learning tools, it should be clear which versions are used and on which dataset they have been trained.

In order to be fully transparent, ideally all these things are documented, and the tools or queries stored alongside the data. This raises the question what level of documentation is required to make the research traceable or repeatable for others. Some researchers provide tools to document the settings, such as the Gephi Fieldnotes plugin developed by [22]. Others have proposed strategies for tracing all the data handling steps [7]. We see, however, the risk that the effort to produce such documentation may take a disproportionate amount of time and effort. Tool standardisation may make this need less urgent (e.g., the role of SPSS[60] in Social Science research) and therefore reduce this burden.

▶ Measures to help achieve transparency and fairness

Access to facsimiles. To facilitate qualitative research, where users explore the corpus at object level, an interface should present research results in the form of scans next to the OCR and metadata (which most portals currently afford). For quantitative analysis, tools will be used both inside and outside the portals. In order to improve the traceability of the research, researchers should have the ability to store tools and their settings alongside the datasets, perhaps on publicly accessible platforms such as GitHub and Zenodo, or using tools specifically designed for this purpose, such as the Gephi Fieldnotes plugin. "How to cite" text blocks with detailed and multimodal information on the tools and their settings could be helpful here.

Comparative perspectives. The transparency of analysis pipelines can be supported by interfaces that allow researchers to compare the performances of different algorithms on a specific task. An example is the NEWSGAC platform[61] that allows users to compare different

[60] https://www.ibm.com/de-de/analytics/spss-statistics-software

[61] https://github.com/newsgac/platform

algorithms for automatic genre detection in newspapers. In order to increase the transparency of machine learning tools, references to publications on models used and documentation on training datasets is provided (e.g., in the form of “datasheets for datasets”[6]).

Replication. Ideally, users should have the possibility to save, export and reuse their own analysis pipeline and the results. This output should ideally connect to the changing publication and presentation modes for the research results, that allow researchers to include data, code and narratives alongside each other (e.g., the Journal of Digital History[62] that publishes the data alongside the narrative and a description of the methodological issues, or the ESWC conference[63] where linking to data, code and other resources is a review requirement); this is further discussed in the Acknowledgements section below.

4.3.9 Visualisation of data, results and bias thereof

Visualisations have the added value of showing complex matter e.g. in graphs and images that help users to get a better overview of collections, corpora, data sets and also the content of these. Visualisations are important because they support the exchange between data and users since they can help to contextualise the collection on the one hand and research on the other. Using graphs, timelines, charts, maps, word-clouds, bubbles, and similar transparency concerning the collection and the research method is offered. The possibilities of how to support transparency by visualising collections and data span a wide range: e.g., research questions and methods are made explicit, topics can be contextualised, a classification of genres and faults or missing/biased data in the collection can be made visible, OCR-/layout-quality and research approaches can be identified, and comparisons of topics (and many other similar things) are possible. As a consequence interfaces can be designed to offer possibilities for visualisation.

▶ Focus areas for transparency and fairness

General Guidelines for Visual Design. Visual interfaces are in many contexts suitable, necessary means to make patterns inherent in the data set in question salient to the observer. However, visualisations are abstract representations of (typically) numerical data, and the visual mapping of the underlying information always imposes a level of distortion because numbers are rather easier to compare when they are served in textual form than when our brain has to approximate them when they appear in the form of visuals such as bars in a bar chart or dots in a scatter plot. The complexity of comprehending data in textual form increases with the size of the data set, but visualisations help to arrange the data in a way that users get a quick, understandable overview even for vast data sets. In order to limit the level of distortion, visual representations of data have to be carefully designed.

Accurate representations of data . Following Edward Tufte’s guidelines for graphical excellence, first of all, visualisations should “show the data”, make it coherent and avoid distorting what the data has to say [19]. An appropriate indicator for good visual design is when viewers are induced to think about the substance rather than about methodology, graphic design or the technology of graphic production. Moreover, visualisations should not “lie,” i.e., the size of effect shown in the visual display needs to correspond to the size of effect in the data.

[62] https://journalofdigitalhistory.org/en/about

[63] https://2022.eswc-conferences.org/call-for-papers-research-track/

Choice of colour. Of particular importance for visual design is the selection of appropriate colour maps. Qualitative colour maps (a set of different hues) should be used to display categorical data, and continuous colour maps (sequential or diverging) to communicate quantitative data (colour gradients). Although powerful, a general advice is not to encode the most important feature with colour ("get it right in black and white"). Visualisations should also be colorblind-safe, i.e., one should not mix diverse shades of green and red. Several online tools support defining accurate colour maps, e.g. ColorBrewer[64].

Clarity. Next to choosing inappropriate colour maps, visualisation designers should avoid visual clutter that reduces the readability of the displayed data and conceals occurring patterns. Clutter can also occur when choosing 3D over 2D representations, which are the means of choice when visualising data that does not inhere 3D structures. Textures that cause visual stress (moiré vibrations) should furthermore be avoided.

Visual Exploration. In order to support Information seeking, visualisation tools should implement Shneiderman's mantra "Overview first, zoom and filter, then details on demand" [17]. Whereas the overview corresponds in digital humanities terminology to distant reading, details on demand refers to close reading. Thus, visual interfaces should support gradual zooming and filtering of the data to be analysed.

Visualising uncertainty. Especially, data in the context of humanities applications often embody uncertainty of different kinds (imprecision, inhomogeneity, incompleteness). Visualisations are suitable means to communicate these uncertainties, for example through transparency or grey glyphs, indispensable for increased reliability of visual display of information.

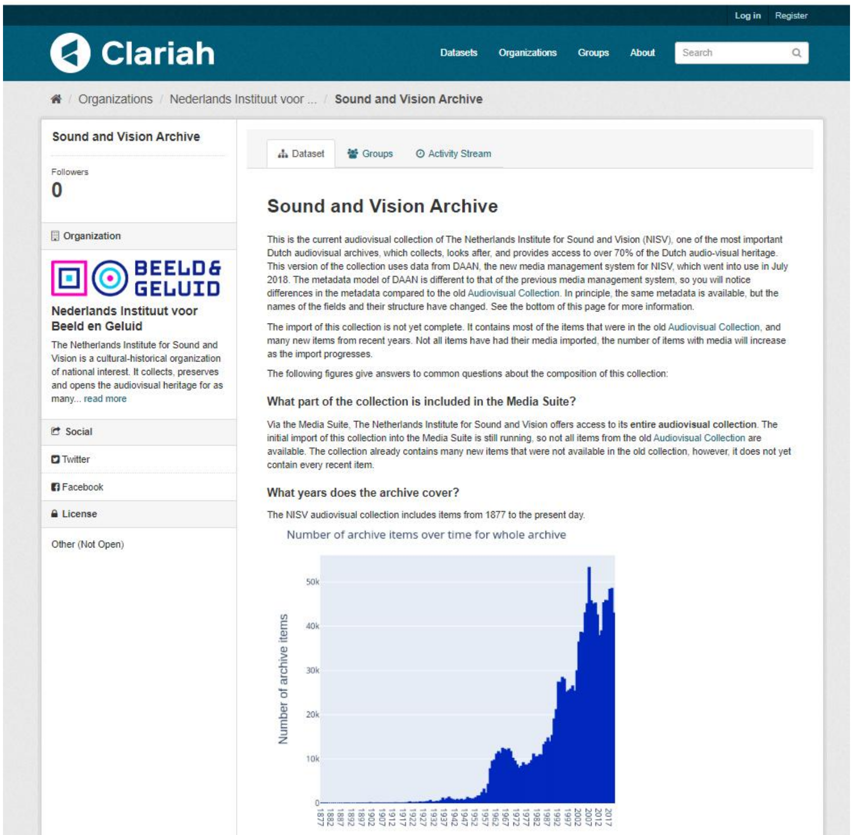
▶ Measures to help achieve transparency and fairness

Participatory design. To ensure a transparent visual interface with a minimised level of data distortion, we suggest conducting a participatory visual design process that involves visualisation experts on the one hand, and domain experts that ensure the suitability of the visual design for its intended purpose on the other. An exhaustive overview of visual design principles can be found in [12].

Collection visualisation. There are some good examples on how interfaces can offer transparency on the collections. The CLARIAH-NL Media Suite[65] , and the Sound and Vision Archive[66], offer explanations and graphs to contextualise the audiovisual collection of the Netherlands Institute for Sound and Vision (NISV), explaining both the role of the archive in the archival field in the Netherlands as well as the time the collection spans, what the digital archives cover, what kind of media is included, when updates happen, what part or the collection is digitised, how the collection is enriched, where additional information and help can be found and how it can be searched (as exemplified in Figure 19). It also indicates what the differences between the metadata of two media management systems are and offers links as well as downloads to the description of the metadata fields.

Digital newspaper collections also contextualise their datasets to a certain extent although additional information like the one offered for the Sound and Vision archive might be added. The *impresso* interface indicates the provenance of its collection in the detail view of

[64] `https://colorbrewer2.org/`
[65] `https://mediasuitedata.clariah.nl/dataset/nisv-catalogue`
[66] `https://mediasuitedata.clariah.nl/dataset/audiovisual-collection-daan`



Sound and Vision Archive

Followers
0

Organization

Nederlands Instituut voor Beeld en Geluid

The Netherlands Institute for Sound and Vision is a cultural-historical organization of national interest. It collects, preserves and opens the audiovisual heritage for as many... read more

Social

Twitter

Facebook

License

Other (Not Open)

Sound and Vision Archive

This is the current audiovisual collection of The Netherlands Institute for Sound and Vision (NISV), one of the most important Dutch audiovisual archives, which collects, looks after, and provides access to over 70% of the Dutch audio-visual heritage. This version of the collection uses data from DAAN, the new media management system for NISV, which went into use in July 2018. The metadata model of DAAN is different to that of the previous media management system, so you will notice differences in the metadata compared to the old Audiovisual Collection. In principle, the same metadata is available, but the names of the fields and their structure have changed. See the bottom of this page for more information.

The import of this collection is not yet complete. It contains most of the items that were in the old Audiovisual Collection, and many new items from recent years. Not all items have had their media imported, the number of items with media will increase as the import progresses.

The following figures give answers to common questions about the composition of this collection:

What part of the collection is included in the Media Suite?

Via the Media Suite, The Netherlands Institute for Sound and Vision offers access to its **entire audiovisual collection**. The initial import of this collection into the Media Suite is still running, so not all items from the old Audiovisual Collection are available. The collection already contains many new items that were not available in the old collection, however, it does not yet contain every recent item.

What years does the archive cover?

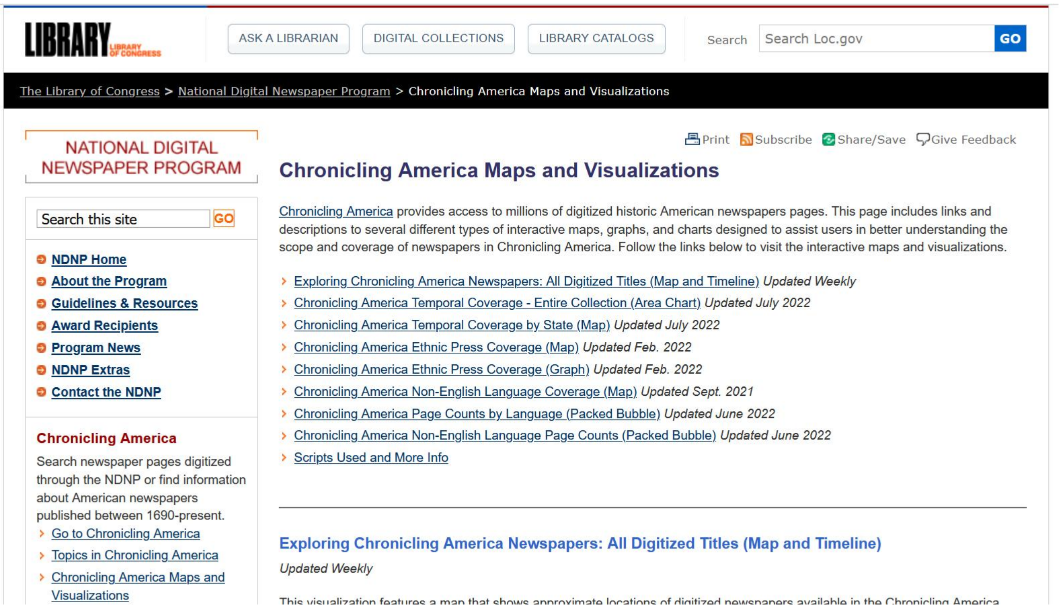
The NISV audiovisual collection includes items from 1877 to the present day.

Figure 19 Landing Page of the Sound and Vision Archive which is one of the datasets of the Nederlands Instituut voor Beeld en Geluid within CLARIAH (© Nederlands Instituut voor Beeld en Geluid).

individual titles. A simple graph visualises the overall temporal distribution of the newspaper collection and a bar chart enables users to get a quick overview of the time span and amount of data each newspaper contributes to the collection (see Figure 17). Due to the efforts of the National Digital Newspaper Program[67] which "is a long-term effort to develop an Internet-based, searchable database of U.S. newspapers with descriptive information and select digitization of historic pages" Chronicling America is currently also adding information on its dataset. Figure 20 shows that contextualisation regarding the collection is made in the "Maps and Visualisation" section of the portal and it can be seen that most of this information is of very recent date (mostly updated February and June 2022) and that the "Map and Timeline" of the collection are updated on a weekly basis using the ArcGIS Instant App (see example in Figure 21). The interactive map visualisation has the added value that it supports a scalable reading of the collection, which is often required by humanities researchers.

These examples show that awareness for the necessity of transparency and biases is growing constantly. In this context visualisations can be helpful to support the communication of complex issues at a glance. The visualisation in Figure 16 frames the issue at hand very

[67] https://www.loc.gov/ndnp/



The Library of Congress > National Digital Newspaper Program > Chronicling America Maps and Visualizations

NATIONAL DIGITAL NEWSPAPER PROGRAM

- NDNP Home
- About the Program
- Guidelines & Resources
- Award Recipients
- Program News
- NDNP Extras
- Contact the NDNP

Chronicling America

Search newspaper pages digitized through the NDNP or find information about American newspapers published between 1690-present.

- Go to Chronicling America
- Topics in Chronicling America
- Chronicling America Maps and Visualizations

Chronicling America Maps and Visualizations

Chronicling America provides access to millions of digitized historic American newspapers pages. This page includes links and descriptions to several different types of interactive maps, graphs, and charts designed to assist users in better understanding the scope and coverage of newspapers in Chronicling America. Follow the links below to visit the interactive maps and visualizations.

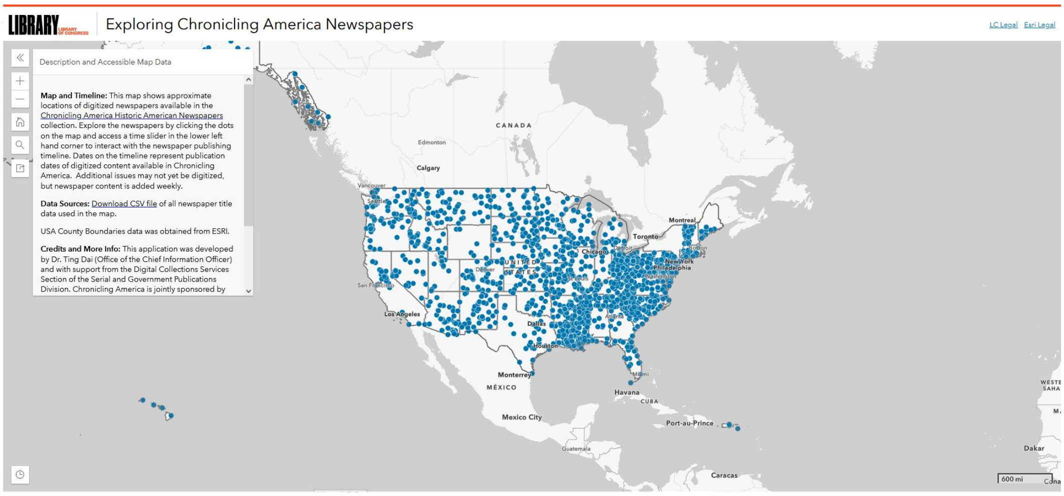
- Exploring Chronicling America Newspapers: All Digitized Titles (Map and Timeline) *Updated Weekly*
- Chronicling America Temporal Coverage - Entire Collection (Area Chart) *Updated July 2022*
- Chronicling America Temporal Coverage by State (Map) *Updated July 2022*
- Chronicling America Ethnic Press Coverage (Map) *Updated Feb. 2022*
- Chronicling America Ethnic Press Coverage (Graph) *Updated Feb. 2022*
- Chronicling America Non-English Language Coverage (Map) *Updated Sept. 2021*
- Chronicling America Page Counts by Language (Packed Bubble) *Updated June 2022*
- Chronicling America Non-English Language Page Counts (Packed Bubble) *Updated June 2022*
- Scripts Used and More Info

Exploring Chronicling America Newspapers: All Digitized Titles (Map and Timeline)

Updated Weekly

This visualization features a map that shows approximate locations of digitized newspapers available in the Chronicling America

Figure 20 Screenshot of "Chronicling America Maps and Visualizations" where information about the collection can be found (© Library of Congress).



Exploring Chronicling America Newspapers

Description and Accessible Map Data

Map and Timeline: This map shows approximate locations of digitized newspapers available in the Chronicling America Historic American Newspapers collection. Explore the newspapers by clicking the dots on the map and access a time slider in the lower left hand corner to interact with the newspaper publishing timeline. Dates on the timeline represent publication dates of digitized content available in Chronicling America. Additional issues may not yet be digitized, but newspaper content is added weekly.

Data Sources: Download CSV file of all newspaper title data used in the map.

USA County Boundaries data was obtained from ESRI.

Credits and More Info: This application was developed by Dr. Ting Dai (Office of the Chief Information Officer) and with support from the Digital Collections Services Section of the Serial and Government Publications Division. Chronicling America is jointly sponsored by

Figure 21 Screenshot of the visualisation of Chronicling America Newspapers, an interactive map created using the ArcGIS Instant App which enables an in depth exploration of the newspapers available in the collection ranked by the primary place of publication (© Library of Congress).

clearly. It is a good example of how well visualisations are able to communicate difficult and complex information. This is also true for analyses that are being done by researchers (see Section Data analysis).

4.3.10 Acknowledgement

▶ Focus areas for transparency and fairness

Reveal hidden labour in digitisation. There is a considerable amount of hidden labour in the process of generating digital newspaper collections. It would be ideal if this hidden labour could be made visible and all the actors in the process could be acknowledged for their work. This would help increase the level of transparency and fairness of the whole digitised historical newspaper ecosystem.

In order for historical newspaper collections to exist, they have to have been acquired by libraries, archives or museums, often in physical form. They then need to be digitised, processed (e.g. metadata generation, creation of lower resolution images for public display), and then enriched (e.g., OCR and article segmentation). It is not feasible to undertake the whole digitisation workflow at once, and therefore selection or prioritisation is needed. All these steps involve the expertise of cultural heritage professionals, computer and data scientists and software engineers, as well as (digital) humanities researchers who would like to use the digitised newspapers as historical sources for their research.

Reveal hidden labour in data curation and enrichment. In addition, the creation of research data sets or corpora, which requires sustained intellectual effort, is often not recognised or acknowledged as formal research output. This both discourages researchers to spend time and properly document this crucial step in the research workflow, but also devalues this work as a necessary evil before the "real" research work can begin. Both cultural heritage professionals and computer scientists contribute significantly to this phase, e.g., by providing historical context information about the historical newspaper collections or by working on information extraction methods to computationally facilitate the corpus building process. It is therefore important that this valuable work is uncovered and made visible.

Finally, such work is often made possible by the financial contribution of (public) funding agencies, the acknowledgement of which is also important.

▶ Measures to help achieve transparency and fairness

Acknowledgement in portals. We recommend the formal and visible acknowledgement of all contributing partners at each step of the digitisation process. This includes the contributions by cultural heritage professionals, software engineers, researchers, and (public) funding agencies. For example, when building a platform for the exploration and analysis of digitised historical newspapers, both the funding agency can be acknowledged, such as is the case with the impresso project ("impresso. Media Monitoring of the Past. Supported by the Swiss National Science Foundation under grant CR- SII5 173719, 2019." [68]) or the NewsEye project ("This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299." [69]). Additionally, when building a research dataset or corpus, it is important to acknowledge all contributors, such as "Biltereyst, Daniël, Philippe Meers, Dries Moreels, Julia Noordegraaf and Christophe Verbruggen. Cinema Belgica: Database for Belgian Film History." [70]. Not only does this publicly acknowledge the work of people involved in the development of the platform or

[68] https://impresso-project.ch

[69] https://www.newseye.eu/about/

[70] https://www.cinemabelgica.be, all consulted on July 27, 2022

dataset in question, but it also enables it to be cited in articles or other research outputs that have made use of it. Being able to demonstrate the impact of such platforms becomes increasingly important for their sustainability.

Acknowledgement in publications. Acknowledgement in publications is another key method to make all parties who contributed to the development of data and the design of platforms visible. For instance, the domains in which each actor has contributed to such a dataset or platform could be stated explicitly, as is already required with regard to authorship by some journals. The Journal of Open Humanities Data Author Guidelines[71] for example provides a number of recommendations based on the ICMJE (Internal Committee of Medical Journal Editors)[72], outlining criteria for authorship. An area where there is still room for further development is the (academic) recognition of more innovative digital research outputs. Innovative Journals such as the Journal of Open Humanities Data (JOHD)[73], which focuses on the publication of "peer reviewed publications describing humanities data or techniques with high potential for reuse"[74]; the Journal of Digital History which intends "serve as a forum for critical debate and discussion in the field of digital history by offering an innovative publication platform and promoting a new form of data-driven scholarship and of transmedia storytelling in the historical sciences"[75] and the Journal of Data Mining & Digital Humanities (JDMDH)[76] which is situated at "the intersection of computing and the disciplines of the humanities, with tools provided by computing such as data visualisation, information retrieval, statistics, text mining by publishing scholarly work beyond the traditional humanities."[77]

4.3.11 Training

Training is essential to raise awareness of the complexity of enriched historical sources and variability in the quality of available data. In addition, and as we have outlined above, historical newspaper data may reproduce biases present in past societies. Training can provide the necessary understanding and tools to help deal with this.

Digital literacy. Training already exists on various levels. More and more, digital literacy is a part of the curriculum of (digital) humanities students. Some newspaper portals also offer domain-specific training. For example, *impresso* offers 1) General training on research using digitised historical newspapers, 2) Training on how to use the *impresso* processing tools and 3) Platform specific training about the functionality of the interface, linked to the FAQ, which contains references to literature.[78] *NewsEye* provides access to various training materials for schools and universities as well as material targeted at a general audience.[79] The *Ranke 2* platform, for example, offers a series of lessons on Digital Source criticism that can be helpful not only for students.[80]

[71] https://openhumanitiesdata.metajnl.com/about/submissions/

[72] https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

[73] https://openhumanitiesdata.metajnl.com

[74] https://openhumanitiesdata.metajnl.com/about/

[75] https://www.degruyter.com/journal/key/jdh/html

[76] https://jdmdh.episciences.org

[77] https://jdmdh.episciences.org/page/editorial-policies

[78] https://impresso-project.ch/theapp/usage/

[79] https://www.newseye.eu/

[80] https://ranke2.uni.lu/

Table 5 Stages of digital humanities workflows for historical newspapers including focus areas and measures to implement.

	Focus area	**Measures**
Research corpus	Awareness of dig. & pres. policies	Collection documentation
	Diverse user needs	Multiple data access points
	Toxicity and cultural bias	Terms and conditions
		Contested terms
Processing & Enrichment	Tool performance	Performance metrics
		Access to “raw” data
		Doc. of tools and training sets
	Ingrained bias	Scanning for contentious terms
		Disclaimers
	Posterior annotation	Representative annotators
		Transparency about annotators
	Ranking	Offer multiple relevance rankings
Data analysis	Traceability	Access to facsimiles
		Replication
		Comparative perspectives
Visualisation	Visual design guidelines	
	Accurate data representations	
	Colour choices	Collection visualisation
	Clarity	Participatory design
	Visual exploration	
	Uncertainty	
Acknowledgement	Reveal hidden labour in dig.,	Ackn. in portals
	data curation and enrichment	and publications
Training	Digital literacy	Publications with best practices
		Code examples
		Example workflows and use cases
		Platform-specific training
		API training

We identified the following types of training as contributing to the skills and knowledge of researchers with respect to transparency and fairness when studying digitised historical newspapers:

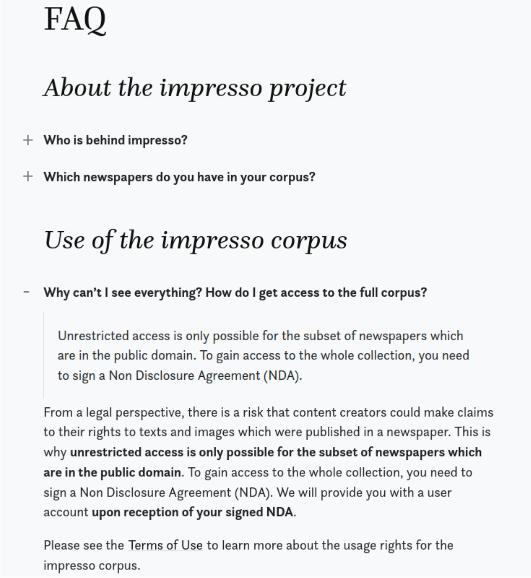
1. Publications with best practices;
2. Code examples, for example in the form of Jupyter notebooks;
3. Example workflows and example use cases;
4. Example lesson plans and course material;
5. Platform specific training, in the form of interface walk-throughs;
6. API documentation.

Table 5 offers a high-level overview of the proposed focus areas and measures to achieve transparency and fairness.

In the following last section we undertake a review of the *impresso* portal and assess to which extent it fulfils the above-mentioned criteria for transparency and fairness.

4.3.12 Application: Analysis of *impresso* portal

We revisited the vaccination case study and its underlying questions to apply it to the *impresso* interface for historical newspapers which was developed in close cooperation between historians, computer scientists and designers and with special emphasis on transparency. We revisited the focus areas and accompanying measures we identified above and concentrated on opportunities for improvement of the current interface.



FAQ

About the impresso project

+ **Who is behind impresso?**

+ **Which newspapers do you have in your corpus?**

Use of the impresso corpus

- **Why can't I see everything? How do I get access to the full corpus?**

> Unrestricted access is only possible for the subset of newspapers which are in the public domain. To gain access to the whole collection, you need to sign a Non Disclosure Agreement (NDA).

From a legal perspective, there is a risk that content creators could make claims to their rights to texts and images which were published in a newspaper. This is why **unrestricted access is only possible for the subset of newspapers which are in the public domain.** To gain access to the whole collection, you need to sign a Non Disclosure Agreement (NDA). We will provide you with a user account **upon reception of your signed NDA.**

Please see the Terms of Use to learn more about the usage rights for the impresso corpus.

Figure 22 FAQ page on the impresso website (© impresso).

Corpus creation and selection. The Newspapers overview page offers a good overview of the print runs of the newspapers in the collection. Information about missing pages and issues (mismatches between available data and expected data based on library metadata) is available but could be explained more clearly. Information on the origins of the collections from different partners is available albeit scattered across the interface and could be further improved by links to the respective digitisation policies of the partnering institutions. *impresso* offers rich metadata for individual titles (e.g. name, print run, number of pages, orientation, regional focus) and links to their respective pages on the websites of the institutions from which they originate. The value of this information could be increased by allowing users to use them as filters in the search component and by offering a download of the data for individual processing and analysis outside the interface. More detailed information about the alignment between *impresso*'s collection and the collection of the partnering institutions should be added. The hitherto unsolved problem to relate the "tip of the iceberg" of the available digital content to the rest of it, i.e., the total record of newspapers in circulation in the past persists in the *impresso* interface as well. Search results are sorted by "relevance" by default, the underlying settings are not explained. Access is granted via a browser interface and following approval after signing an Non-Disclosure-Agreement. Users are able to export metadata for articles and, depending on legal agreements, also full text. The corresponding FAQ entry should be expanded to explain the content of the export file.

Processing, enrichment, and analysis of data. The FAQ section of the interface collects important information about the project, corpus and interface. Semantic enrichments such as topic modelling, named entity recognition or text reuse are well explained as is the entry on OCR quality. Filtering by OCR quality should be enabled. The FAQ entries for data export and legal restrictions are valuable but could more concisely explain the legal status of exported data and link directly to project publications on system architecture and the processing pipeline.

Visualisation of results. The application makes use of data visualisations in multiple forms which are accompanied by corresponding FAQ entries. This includes frequently distributions over time e.g. for search results, graphs to represent overlaps between topics. The Inspect&Compare[81] component uses small multiples of bar charts to reveal overlaps and dissimilarities between two queries or article collections and can e.g. be used to evaluate search strategies [1]. The interface is clearly designed for research purposes. Basic knowledge about data and interactive visualisations is a prerequisite.

Training and digital literacy. From a user perspective, the portal supports both scholars with low digital literacy (as tools for analysis are built in) and for more advanced skilled researchers (because data and metadata can be exported as csv). The project has compiled and integrated educational materials for researchers ranging from beginner to advanced level which cover digitised newspapers per se as well as the functionalities of the interface.[82] Tutorials could be placed more prominently in the interface and specific support for visually impaired users is missing.

Acknowledgements. As stated above, the project indicates the source of funding by the Swiss National Science Foundation together with an overview of the full project consortium[83] and the contributions of individuals.

4.3.13 Conclusions

It was our goal to compose a set of recommendations for content providers such as libraries and archives as well as developers of research interfaces, in order to help individual researchers in the field to gain as much transparency and fairness as is required for the analysis of digitised and enriched historical newspaper collections. We did so by focusing on aspects with a potentially high impact on the outcome of research. We found that efforts to increase transparency and fairness have to be made on all stages of the workflow. The stages we identified were 1) corpus creation and selection; 2) processing, enrichment, and analysis of data; 3) visualisation of results; 4) training and digital literacy; and 5) acknowledgements. These stages interrelate and are dependent on each other. It was therefore not always possible to make clear distinctions. We discussed issues of transparency and fairness in each of the stages and also reflected on some measures that can help mitigate biases, lacks of transparency and fairness.

Overall it can be summarised that we found a lot of variability in the current landscape of digital newspapers. This applies within and across newspaper collections and includes differences in metadata standards such as METS/ALTO, the quality of OCR with older processing tending towards lower quality, but also in regard to the scope of enrichment: whereas some content providers invested in the correct identification of even small news items (e.g., obituaries), others only offer PDFs of scanned images. These variations have a high impact on research but are still poorly communicated in contemporary interfaces.

For researchers, therefore, some challenges remain and some of these challenges can be countered by content and interface providers. For example, it remains crucial to understand the “digital sample”, i.e., researchers have to be given the ability to assess the (non-)

[81] https://impresso-project.ch/app/inspect

[82] https://impresso-project.ch/theapp/usage/

[83] https://impresso-project.ch/consortium/people/

representativity of the small number of digitised and available newspapers against the background of all past and potentially not archived newspapers. Also, providing detailed information – such as metadata, information on the digitisation process, distribution over time, political orientation of newspapers, links to historical contextual information, etc. – of the nature of the collections that can be found via the interfaces, can be of great help to researchers. Interfaces should therefore provide intuitive guidelines for and explanation of the collection. Also, information regarding diversity, equality, and equity (such as a notification of contested terminology) should be found.

Since the quality of OCR and layout recognition as well as classification issues remain a challenge, it also remains crucial for researchers to be able to always have access to the "raw" data, i.e., the images. If automated tools or training sets are available, a proper documentation has to be provided. The experimental nature of the analysis tools still poses challenges regarding replication and sustainability (e.g., the query storage facility, notebooks for re-training on the same corpora). For example, the CLARIAH Media Suite allows users to rerun queries but updates to the underlying collections cause this feature to break. Again, the progress in automation creates a need for extensive documentation (e.g., the Gephi Fieldnotes Plugin) which could be diminished once there is some standardisation of the tools (e.g., as in the case of SPSS that is more integrated in a methodological framework and also generally accepted as reliable). Another option could be to allow researchers to compare the results of different algorithms. For many research questions, however, it also remains important to work with external tools and adaptable methods. The possibility to download data sets and analyse on (or publish about) them outside the interfaces is one more important feature we identified. Overall, a higher degree of transparency in visual interfaces can be a real asset for humanities research.

References

1 Düring, M., Kalyakin, R., Bunout, E. & Guido, D. Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles. *Information*. **12**, 348 (2021,9), `https://www.mdpi.com/2078-2489/12/9/348`, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute

2 Beelen, K., Lawrence, J., Wilson, D. & Beavan, D. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship In The Humanities*. (2022)

3 Brate, R., Nesterov, A., Vogelmann, V., Van Ossenbruggen, J., Hollink, L. & Van Erp, M. Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus. *Proceedings Of The 11th On Knowledge Capture Conference*. pp. 17-24 (2021)

4 Ehrmann, M., Bunout, E. & Düring, M. Historical Newspaper User Interfaces: A Review. *Proceedings Of The 85th International Federation Of Library Associations And Institutions (IFLA) General Conference And Assembly*. pp. 24 (2019), `https://infoscience.epfl.ch/record/270246?ln=en`

5 Fry, B. Visualizing data. (" O'Reilly Media, Inc.",2008)

6 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Iii, H. & Crawford, K. Datasheets for datasets. *Communications Of The ACM*. **64**, 86-92 (2021)

7 Hoekstra, R. & Koolen, M. Data scopes for digital history research. *Historical Methods: A Journal Of Quantitative And Interdisciplinary History*. **52**, 79-94 (2019)

8 Linhares Pontes, E., Cabrera-Diego, L., Moreno, J., Boros, E., Hamdi, A., Doucet, A., Sidere, N. & Coustaty, M. MELHISSA: a multilingual entity linking architecture for historical press articles. *International Journal On Digital Libraries*. **23**, 133-160 (2022)

9 McGillivray, B., Poibeau, T. & Ruiz Fabo, P. Digital humanities and natural language processing "Je t'aime . . . Moi non plus". (Alliance of Digital Humanities,2020)

10 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. & Gebru, T. Model cards for model reporting. *Proceedings Of The Conference On Fairness, Accountability, And Transparency.* pp. 220-229 (2019)

11 Modest, W. & Lelijveld, R. Words Matter, Works in Progress I. National Museum of World Cultures. (2018)

12 Munzner, T. Visualization analysis and design. (CRC press,2014)

13 Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A. & Pletschacher, S. A survey of OCR evaluation tools and metrics. *The 6th International Workshop On Historical Document Imaging And Processing.* pp. 13-18 (2021)

14 Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. & Tolonen, M. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal Of The Association For Information Science And Technology.* **73**, 225-239 (2022)

15 Hechl, S., Langlais, P., Marjanen, J., Oberbichler, S. and Pfanzelter, E., Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal Of Data Mining & Digital Humanities.* (2021)

16 Schneider, P. Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction. *ArXiv Preprint ArXiv:2110.01661.* (2021)

17 Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. *The Craft Of Information Visualization.* pp. 364-371 (2003)

18 Van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B. & Colavizza, G. Assessing the impact of OCR quality on downstream NLP tasks. (SCITEPRESS-Science,2020)

19 Tufte Edward, R. The visual display of quantitative information. (Cheshire, Connecticut: Graphic Press,2001)

20 Traub, M., Samar, T., Van Ossenbruggen, J., He, J., Vries, A. and Hardman, L. Querylog-based assessment of retrievability bias in a large newspaper corpus. *2016 IEEE/ACM Joint Conference On Digital Libraries (JCDL).* pp. 7-16 (2016)

21 Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D. and Helberger, N. Recommenders with a mission: assessing diversity in news recommendations. *Proceedings Of The 2021 Conference On Human Information Interaction And Retrieval.* pp. 173-183 (2021)

22 Wieringa, M., Geenen, D., Es, K. and Nuss, J. The Fieldnotes Plugin: Making Network Visualization in Gephi accountable. *Good Data.* **14** pp. 277 (1988)

4.4 Towards an International Historical Newspaper Infrastructure

Clemens Neudecker (Staatsbibliothek zu Berlin, DE)
Maud Ehrmann (EPFL – Lausanne, CH)
Matteo Romanello (EPFL – Lausanne, CH)
Martin Volk (Universität Zürich, CH)
Lars Wieneke (C2DH – Esch-sur-Alzette, LU)
Dario Kampkaspar (TU Darmstadt, DE)

Portals and platforms that aggregate digitised newspapers from multiple sources and institutions have added great value for researchers, as they e.g. allow the comparative study of newspaper data from different countries and in multiple languages from within a uniform