Studies in Communication Sciences 24.3, (2024), pp. 285–301

Digital press archives for media and communication history research: From "reading rooms" to virtual research environments

Hendrik Michael*, University of Würzburg, Human-Computer-Media Institute, Germany
Valentin Werner, University of Bamberg, Institute for English and American Studies, Germany
*Corresponding author: hendrik.michael@uni-wuerzburg.de

Abstract
Digitization facilitates access to large amounts of news data, a fact that also opens new avenues for research in media and communication history. Such resources allow easy access and retrieval as well as means to automatically process relevant materials. The increasing number of digital press archives is particularly valuable for historical research as it offers new insights into the evolution of journalistic language, professional practices, and role performances. This paper surveys current digital sources and reflects on research opportunities and restrictions. To this end, it provides an overview of currently available digital press archives with a focus on German- and English-language archives and offers a typology of archive types along criteria such as availability, accessibility, quality, and usability of the digital source material and its metadata. Simultaneously, it illustrates developments from flat portals to virtual research environments, which arguably offer researchers unprecedented opportunities for establishing personalized workspaces and collaborative research projects. Furthermore, attention is drawn to persistent issues in data compilation, and it is suggested that quantitative computational approaches should be complemented by qualitative analysis and close reading to fully exploit the affordances of current digital press archives.

Keywords
digitization, historical press research, digital humanities, research infrastructure, research methodology

1 Introduction

The present "data-driven times" (Hepp, 2016) present numerous opportunities and challenges for communication and media studies research when assessing the ongoing transformation of communication in a digital world. On the one hand, software-based services assist researchers in collecting, handling, and evaluating increasingly large amounts of data, facilitating sharing and collaboration within the scientific community. Thus, they are shaping new modes of historiography and constructing reality in the process. On the other hand, all aspects of social life that involve the use of digital media leave seemingly fleeting traces that generate data (Hepp, 2016, p. 229). The analysis of such "born-digital materials" (Nanni, 2018, p. 115) has gained attention in the field of media and communication history (Schwarzenegger, Koenen, Pentzold, Birkner, & Katzenbach, 2022) and has been recognized more broadly in the domain of digital history since the early 2000s (Rosenzweig, 2003; Cohen, 2004; see also publications in the *Journal of Digital History*). This early recognition has already touched on questions of how historical sources are "stored, recorded and preserved [...] and will increasingly alter culture itself" in the future (Balbi, 2011, p. 154). Such discussions are not only important for practical reasons relating to archival work and research but also add to the theoretical development of visions that define how we preserve our cultural heritage and establish modes of self-referentiality for contemporary societies under digital transformations (Esposito, 2002; Dimbath, 2014).

This contribution is centrally concerned with such questions, namely if and how the abundance of digitized sources affects historical research practices or even presents a path forward for the broader field of communication and media studies, given its commonly lamented neglect of diachronic perspectives (Kinnebrock, Schwarzenegger, & Birkner, 2015). These questions are discussed

https://doi.org/10.24434/j.scoms.2024.03.4050

in relation to the increasing number of digital press archives. Despite being recently identified as an "Eldorado" for historians (Bunout, Ehrmann, & Clavert, 2023), they are not yet a prominent research object of the digital humanities (Podewski, 2018; Brügger, 2018). Digital press archives make it easier to retrieve and access historical sources than ever before and, in some cases, provide means to automatically search and process relevant materials (Bingham, 2010; Koenen, 2018, 2022a, 2022b; Nicholson, 2013). In this regard, they can become "indispensable sources for research, for both academic and non-academic users" (Ehrmann, Bunout, & Düring 2019, pp. 1–2) and present a hybrid form of what Balbi (2011, pp. 155–156) defines as "old" and "new" sources. The foundation of digital archives is old sources, that is, material artifacts characterized by their scarcity, stability, and accessibility. However, in the digital setting, old sources are also imbued with characteristics of new sources, namely changeability, abundance, and flexible retrievability. In this sense, researchers must acknowledge the "transformation of newspaper [and other] sources into complex data objects" (Ehrmann, Bunout, & Clavert, 2023, p. 4).

This contribution, which offers an applied perspective for researchers in the field of media and communication studies rather than a single piece of empirical research, is dedicated to presenting, comparing, and evaluating the new digital affordances of historical press research. Section 2 provides a brief (and necessarily selective) overview of digital press archives and Section 3 identifies different types of infrastructures as well as user interfaces. It also re-evaluates established distinctions found in the literature and discusses the usefulness of different types of archives (or portals) in the context of media and communication studies. Section 4 reflects on methodological implications when collecting and handling data related to the abovementioned questions of digital hermeneutics when dealing with digital newspapers as hybrid source material. The concluding Section 5 contextualizes the assessment of forms and affordances of digital press archives in light of the evolving nature of digital communication and media history.

2 An overview of digital press archives

The digitization of historical newspapers and periodicals began approximately 15 years ago and has since been successfully advanced by various public and private organizations worldwide (Blome, 2018, B.6, pp. 1–2; Pfanzelter, Oberbichler, Marjanen, Langlais, & Hechl, 2021, pp. 1–2). Blome (2018, B.6, pp. 13–32) summarizes various regional, national, and transnational projects, some of which are introduced below with updated descriptions. Although Wikipedia (2024) provides a comprehensive and constantly updated list of digital press archives, the following overview aims to demonstrate the range and diversity of digital press archives, focusing on English- and German-language databases.

Although this overview does not cover commercial archives (such as Google News, 2024 or the Reuters News Archive) or archives in non-western countries (e. g., Asia, Russia), it is designed to meet the research needs of most (western) researchers. In this regard, the criteria for the selection are primarily practical and draw on our own as well as our expected readers' language skills and their socio-cultural familiarity with the available sources. Another aim of this overview is to provide examples from different socio-political contexts. The selection first presents Germany as an example of a federal system with decentralized (and lagging) efforts to advance the digitization of old sources. On the other hand, we highlight several innovative efforts that benefit from coordinated national programs to build centralized archives. Lastly, examples refer to transnational projects that may offer insight into the future of historical research and thus help in understanding historical processes. As a starting point, it is worth noting that, unlike other countries, German digitization projects lack a centralized initiative and dedicated government funding (Birkner, Koenen, & Schwarzenegger, 2018, p. 1125). Commonly, institutions such as libraries, museums, and archives have made individual efforts to provide resources. However, these are often highly specific and thus possibly too narrow for broader research purposes. As a result, the Deutsche Digitale Bibliothek (DDB),

where such collections are indexed, contains a patchwork of many specialized databases. From 2013 to 2016, the German Research Foundation (DFG), the most important national research funding body in Germany, supported pilot projects to explore means of improving the digitization of historical newspapers and developing strategies for the future (Zeitschriftendatenbank, 2022).

As a result of this initiative, several time- and region-specific digital press archives were created. One of the most recent projects is Zeitpunkt.NRW (2024), funded by the federal state of North Rhine-Westphalia and including digitized local newspapers from North Rhine-Westphalia dating from 1801 to 1945. Other examples include a collection of 17th century-newspapers available through the Staats- und Universitätsbibliothek Bremen (2024), Bavarian newspapers accessible through Bayerische Staatsbibliothek (2024), and ZEFYS (2024), which provides access to over 190 newspapers published in the 19th and 20th century. ZEFYS includes press titles from Prussia and the German Empire, extending to the Weimar Republic and Nazi Germany. In addition, it offers regionally and thematically focused projects, for instance, on official publications of Prussia (Amtspresse Preußens) and on press publications from the GDR (DDR-Presse). Another pertinent example is the digital archive of the party newspaper Vorwärts, which is maintained by the social-democratic Friedrich-Ebert-Stiftung (2024). This project expands its digital library, which contains other publications by progressive social movements, particularly those from the Weimar Republic.

In October 2021, the DDB launched Deutsches Zeitungsportal (2024). Intended as a nationally coordinated initiative, it is the most comprehensive and sophisticated digital press archive in Germany to date. It covers the period 1671 to 1950 and includes 247 newspapers, 591 837 newspaper issues, and a total of 4 464 846 newspaper pages from nine libraries (Bundesregierung, 2021). While the archive is continually expanded, the current process of migrating sources from other libraries (e. g., Staatsbibliothek zu Berlin and ZEFYS) is slow. Thus, unfortunately, as yet, there is no one-stop solution for researchers, meaning, for instance, that *Berliner Morgenpost*, one of the leading popular dailies in the early 20th century, is not yet included in Deutsches Zeitungsportal, while it is accessible through ZEFYS.

Despite these commendable efforts (supported by the fact that access to these resources is commonly free for researchers), the mediocre state of digital press archives in Germany becomes obvious when viewed from an international perspective. For instance, ANNO is Austria's central press archive, maintained by the Austrian National Library (2024). It currently includes over 1 500 titles (approximately 26 million pages) from 1568 to 1951. France, by contrast, lacks a central database but provides digitized media sources through the French National Library (2024a, 2024b) and their projects Gallica and Retronews. It contains close to six million documents (apart from newspapers and revues also books, maps, images, speeches, videos, and audio material) from the 17th to the 20th century. While Gallica (2024) provides access to all digital assets of the national library, Retronews (2024) is exclusively dedicated to the historical press and offers more than nine million digitized pages dating from between 1631 and 1945.

The British Newspaper Archive (2024) contains 60 million newspaper pages dating back to 1704 and is regularly updated with the immense collections of the British Library. The Australian portal, named Trove (2024), is a digital collection of roughly 1 700 newspapers and government gazettes from 1803 to 1954, with selected publications digitized up to the present day. Combined, the archive contains about 25 million pages. Partly, Trove is the result of international crowdsourcing campaigns involving over 900 organizations and individuals who received additional funding and coordination from the National Library of Australia and the State Library of New South Wales.

Blome (2018, B.6, p. 29) lists Chronicling America (2024) as another laudable example. Although the portal is not the largest archive in terms of digitized pages (approximately 21 million by November 2023), it is among the most diverse, with over 3 800 individual titles spanning almost 200 years of Amer-

ican history (1777 to 1963) across 47 States, Washington, D.C., and Puerto Rico. It also includes multilingual sources. The National Digital Newspaper Program, a partnership between the Library of Congress (2024b) and the National Endowment for the Humanities, maintains the database. State partners of the National Digital Newspaper Program contribute content to Chronicling America (Library of Congress, 2024b). In addition to a standardized description of existing cataloging records, newspapers found in this resource include a supplementary essay. It contains ample metainformation about the paper, including place of publication, dates, and schedule of publication (e.g., weekly, daily, morning, or evening), geographic area covered and circulation statistics, political or religious affiliation, specialized audiences, physical attributes, and changes in name, format, and ownership.

Several major transnational efforts have been made to establish and maintain digital press archives. One particularly remarkable one is the Europeana Newspapers Project (2024), which originated from a collaboration of European national libraries. It includes close to 900 000 periodicals in eight languages (German, French, Dutch, Latvian, Finnish, Estonian, Serbian, and Polish), making it one of the largest databases for historical press research in a digital environment. Another recent large-scale European project is NewsEye. As presented in Koenen (2022a), it was funded by the European Union's Horizon 2020 research and innovation program and represents a corpus of 1.35 million newspaper pages from Austria, Finland, and France. The project aimed to introduce "new concepts, methods, and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users," which will be discussed in the following. Similarly, the multi-nation project Impresso – Media Monitoring of the Past (2024) is an interdisciplinary research project and includes a multilingual corpus of 76 digitized historical newspapers in French, German, and Luxembourgish. It is supported by 15 partners, among others, the Swiss National Library, the National Library of Luxembourg, the Media Center and States Archives of Valais, the journals Le Temps, and Neue Zürcher Zeitung. Impresso reviewed positive reviews from members of the research community and has even been acclaimed as a resource collection that represents "the most complete historical newspaper data set [sic] series to date" (Ehrmann, Romanello, Clematide, Ströber, & Barman, 2020, p. 959).

3 A typology of infrastructures and interfaces

As is even evident from the selective listing presented in Section 2, there is a vast landscape of digital press archives that contain source material of varying provenance. While such an overview may be a helpful point of departure for researchers seeking information sources, it is important to acknowledge that digital news archives vary considerably with respect to their interfaces that are designed to "steer what users can learn from digitized newspapers, and […] influence workflows by offering functions and tools" (Pfanzelter et al., 2021, p. 1). In an attempt to establish a general typology, Birkner et al. (2018, p. 1126; see also Koenen, 2022b, pp. 110–112) distinguish three types of interfaces:

(i) flat portals,
(ii) deep but data-restrictive portals,
(iii) virtual research environments.

Although this tripartite classification scheme provides a useful heuristic for reflecting on the affordances and limitations of digital press archives, it implies that each archive neatly fits into one category. However, *in situ,* the distinctions are fuzzier. This is due to a lack of differentiation in defining the above types with respect to the modes or *practices* in different phases of the research process. This conceptualization relates to early reflections by Mintz (Cohen et al., 2008, p. 456), who discusses four stages of future digital history. While stages 1.0 and 2.0 merely provide technological means to make sources available, stage 3.0 puts more emphasis on participatory structures that allow collaboration and interaction, and stage 4.0 offers a "constructivist" perspective to knowledge production that allows us to "navigate and annotate long-lost historical settings."

In this respect, a more elaborated typology could recognize the progression through these four stages by initially focusing on criteria for collecting data and by acknowledging research practices of reading, finding, and selecting sources. Second, criteria can be attributed to practices of analyzing data, such as options for exploration, annotation, and computerization. Distinguishing between these two modes helps to assess the infrastructures and interfaces of different digital archives and thus allows the development of adequate strategies for conducting research.

Ehrmann et al. (2019) propose possible parameters for more fine-grained evaluations: data collection relies on functions such as (1) content search, (2) filtering (e.g., by keyword searches), (3) generosity (i.e., corpus presentation and result display modes); and data analysis benefits from options for (4) user context management and exploration (i.e., allowing researchers to collect, organize, and compare material), (5) connectivity (i.e., interlinking collections by content and metadata), and (6) the option for source criticism (i.e., providing context through metadata) (Ehrmann et al., 2019, pp. 4–5).

The following sections will exemplify how practices of collecting and analyzing data are interconnected but yield different functionalities for conducting research projects. For this purpose, the next section categorizes and discusses digital archives in reference to ZEFYS, Chronicling America, Trove, NewsEye, and Impresso in more detail.

3.1 Flat portals

Flat portals are, in a sense, mere virtual "reading rooms" (Birkner et al., 2018, p. 1126) for a specific digital collection of newspapers. While Google News qualifies as such a portal with an international scope, in Germany, ZEFYS is a pertinent example. Google News provides researchers with a broad database but is not very useful for research purposes due to its limited search options and lack of opportunities to extract or download sources. Relating to Ehrmann et al.'s assessment (2019, pp. 10–14) and the suggested differentiation between data collection and data analysis, ZEFYS also has limited, albeit more expansive, functions. It allows users to access available newspapers through a calendar for browsing by date and title only, lacking functions like sorting, filtering, and displaying results. Findings can only be viewed in the DFG-Viewer, which is based on CMS TYPO3, and the open-source digitization software Goobi (DFG-Viewer, 2021). It provides users with an onscreen facsimile in varying resolutions and allows for the download of single pages or full newspaper issues as PDF files. However, user interaction regarding tagging, recommending, or exporting sources (apart from PDF downloads) is not implemented. Further, ZEFYS provides basic metadata only, including date range, publisher, place of publication, a calendar view of issues, and an indication of archive holder.

From a research perspective, this type of portal seems to offer little. Manual data collection is the only option, which places limits on the size of the material to be analyzed. Additionally, the lack of information-richness of digitized sources limits the options for computerized data analysis. As reading rooms, flat portals allow for a strategy of "top-down research" (Nicholson, 2013, p. 66) that supports qualitative projects. Examples are issue- or event-centered research questions that can work with smaller data sets, for example, historical coverage of catastrophes or conflicts, or projects that focus on case studies of professional practices, such as the diffusion of the inverted pyramid style in news reporting (Birkner et al., 2018) or local reporting in a specific newspaper (Michael, 2017). However, flat portals offer insufficient means for quantitative projects that aim to trace general trends over different media and periods. The infrastructure available through this type of portal merely establishes what could be labeled a "contact function" (in the sense of offering a first yet superficial engagement with the materials), providing orientation for research interests that need historical contextualization.

3.2 Deep portals

The functionality of *deep portals* is less restrictive. They allow for full-text searches and have fewer limitations on user content management (e.g., saving and exporting findings), as well as overall greater generosity in terms of navigation and presentation options. These portals are referred to as "deep" as they allow "bottom-up research" (Nicholson, 2013,

p.66) and provide flexibility that exceeds traditional newspaper research modes.

Cordell (2017) explains that this is mainly possible because the source material is transformed into OCR-derived text that underlies searches and offers opportunities for more complex forms of data collection and, consequentially also, digital text analysis. This means portal users can view a digital facsimile of the original newspaper page but can also search an underlying text file, often hidden from the interface and encoded in a markup language such as XML (eXtensible Markup Language). In most digital newspaper archives, these text files are created with the support of artificial intelligence approaches. In detail, Holley (2009) describes how OCR software "attempts to replicate the combined functions of the human eye and brain" to analyze the structure of a page, recognize alphabetic characters and their irregularities (e.g., old fonts or distorted material) on a page image, and, through continuous training of the algorithm, creates a computer-readable text file. OCR data underlies all deep portals as well as large-scale generic textual databases, such as Google Books (2024) or the Internet Archive (Ben-David & Amram, 2018).

Drawing on Ramsay (2011), Koenen (2022a) classifies such deep portals as "reading machines" that introduce new practices and perspectives for discovering, searching, and acquiring historical sources. However, it is important to note that many deep portals are also "data-restrictive" as they restrict access through institutional memberships or subscription fees (e.g., the British Newspaper Archive). There are several examples of portals that function as reading machines, permitting users to conduct full-text or keyword searches that can be refined by operators. In German-speaking regions, Deutsches Zeitungsportal and Austrian ANNO are among these types. Both offer users the opportunity to conduct advanced searches with multiple keywords within a specific word distance and modify search categories, such as place of publication, newspaper title, language, timescale, and even thematic clusters. Bookmarking results for better management is also possible. Bavarian DigiPress does not have this wealth of search and management options but permits separate headline and article searches. In many ways, these platforms offer an expanded infrastructure and increasingly sophisticated user interfaces.

Australian Trove has set a benchmark in this regard since it provides openly accessible resources, exploiting many options for the reuse of full-text data. Its multi-layered interface displays a digital facsimile directly alongside OCR-text, identifies and allows users to skip to other articles on the image page, to directly cite findings, and to extract results in various formats (image, PDF, and plain text). In addition, users can filter their search results by word count, different illustration types (e.g., photo, cartoon, map, etc.), and article types (e.g., article, advertisement, editorial, obituary, letter, etc.). Trove also allows for user interaction by tagging, listing, and creating categories. This enhances the data analysis process as a collaborative practice and gives Trove some features of a virtual research environment (see Section 3.3), highlighting the issue of (un-)ambiguous classification of individual portals into the tripartite scheme presented above (see Section 3).

By contrast, Chronicling America lacks such a multilayered interface, which limits casual users to full-text searches for keywords or manual searches of front pages, among other options like searching by state, publication, or a specific timescale. However, this portal offers an extensive application programming interface (API) to explore data in various ways (Koenen, 2018, pp. 544–547), relying on common web protocols. Further, Chronicling America enables large-scale data collection through bulk downloads of batches and text files that result from OCR of all digitized newspapers in the collection. Each batch may contain one or more issues from one or more newspapers. The index site linked above provides metadata on the grant awardees, who delivered the content, the number of pages contained in one batch, and the release date. Each subpage of the index gives details on the name of the publication, the number of newspaper issues in the batch, and an overview of each issue. The overview includes a link to the regular portal interface, the supplementary essay, and the catalog records with the Library of Congress Control Number (LCCN). Users

can extract OCR files from the compressed data (tar.bz2) with a debatcher provided through GitHub by the user lmullen (2024). The program takes paths to the tar.bz2 batches of OCR files from the Chronicling America bulk data downloads. It converts each batch into a CSV file, which can be used in many other databases (Mullen, 2019). This enables researchers to semi-automatically retrieve and handle large corpora and apply state-of-the-art instruments and methods from text mining (Viola & Verheul, 2020) and corpus linguistics (Marchi, 2022) as a form of "distant reading" (Moretti, 2016). Distant reading can identify general structures of texts, which contrasts with traditional hermeneutic approaches.

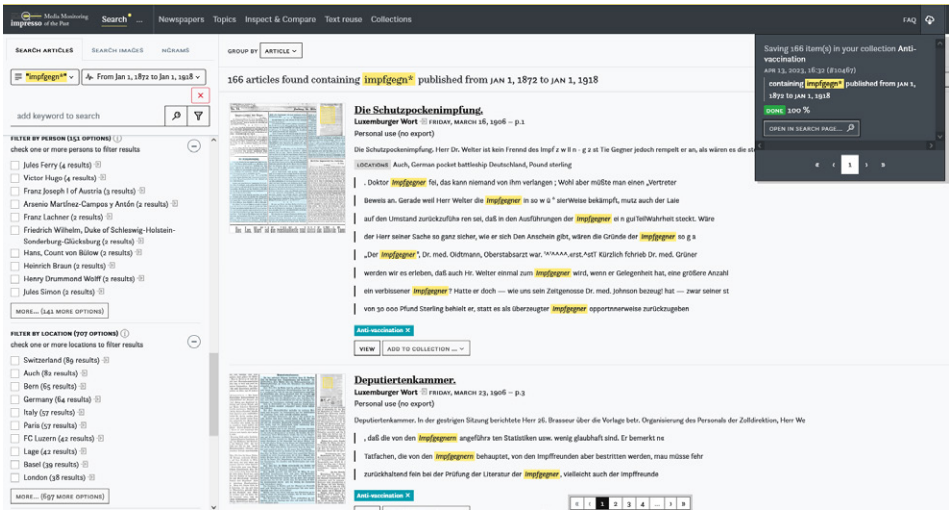
On the surface, the interfaces of deep portals remain data-restrictive. For instance, they lack the means to intuitively collect larger datasets or implement sophisticated options for data analysis. However, as already suggested above concerning Chronicling America, the back end of many portals extends their infrastructure. Two adjacent projects are Digital Borderlands (2024) and Newspapers as Data (2024). Both are based on a dataset of 15 newspapers published in the southwest U.S.-Mexican border region between 1897 and 1963. User Jcoliver (2024) created a GitHub repository that hosts Jupyter notebooks (Dombrowski, Gniady, & Kloster, 2019) to explore text mining analyses on the full dataset in English and Spanish. Similarly, Sherratt (2021) used Trove's API to develop the GLAM Workbench, a collection of Jupyter notebooks that work with data from galleries, libraries, archives, and museums (= GLAM). The notebooks utilize the underlying data from the collections to facilitate data retrieval through various harvester applications that can be run on Binder or Voila. One such application is the Trove Newspaper and Gazette Harvester, which is user-friendly and reliable, and does not require any in-depth coding skills. After requesting an API key for Trove, data can be harvested by simply copying links of search queries conducted through Trove's website. The harvester then downloads all search results in a compressed file and in different formats (plain text, PDF, or image).

Furthermore, there are implementations to extract visual content. Lee et al. (2020) illustrate how to obtain such data from Chronicling America using emerging machine learning techniques that facilitate the automated extraction of headlines, photographs, illustrations, maps, comics, cartoons, and advertisements, for instance. In addition to visual content, these datasets include captions and other relevant text derived from the technical metadata related to layout and text objects as well as image embedding for fast similarity querying. The result of this work is the Newspaper Navigator Dataset, available in the public domain for unrestricted re-use. The goal of Newspaper Navigator is to reimagine searching over the visual content in Chronicling America and present it through innovative modes of display. Similarly, Cordell & Smith (2022) use data from Chronicling America to develop exploratory modes of data analysis. The related website Viral Texts shows examples "to demonstrate how information circulated geographically through the U.S. in the 19th century" or to classify genres and show "the probabilities of newspaper texts belonging to four genres" (Cordell & Smith, 2022). These examples demonstrate how deep portals enable the handling of larger datasets, contextualizing qualitative research and providing statistical evidence of general trends and developments in journalism, media, and society over time.

3.3 Virtual research environments

While Chronicling America or Trove offer many pathways to access, retrieve, and process data, Koenen (2022a) champions the idea of digital portals becoming *virtual research environments* that even transcend advanced aspects of data collection, such as fine-grained filtering, multiple options for saving search results, interactive search options (e. g., tagging and listing), and accessibility of digital newspapers (including metadata). Second, their functionality should approach that of an all-in-one solution that also provides opportunities for data analysis with sophisticated ways for processing metadata, options for managing collaborative workflows, and tools for text mining, all within a single, integrated infrastructure and user interface. Koenen (2022a) argues that the evolution of digital archives into virtual research

Figure 1: Impresso interface / results display for the search query "Impfgegn*"



environments will provide new perspectives and establish concepts and tools for dealing with the challenges of big data in historical research. As Graham, Milligan, & Weingart (2016, p.XVI) noted, this may equip researchers with the "historian's macroscope, [...] a scientist's workbench, where the investigator moves between different tools for exploring different scales, keeping notes in a lab notebook". Overall, by transforming into virtual research environments, digital press archives can merge "data services, digital materials, tools, and an environment for research practices and collaboration" (Pawlicka-Deger, 2020, Sec. 6).

As two cases in point, NewsEye and Impresso (Koenen, 2022a) provide such virtual labs. Both allow users to variably access digitized newspapers and their metadata and channel the research workflow by implementing a set of powerful experimental tools for textual and visual analysis, while also making the technological and methodological process of handling and evaluating the data transparent (Koenen, 2022b, p. 112). During the data collection process, the platforms offer options to identify topics and persons recurring in a specific search query. In a practical sense, Impresso can be useful for a cursory historical contextualization of contemporary issues, for instance when searching for the term "Impfgegn*" to retrieve anti-vaccinationist-related discourse. The screenshot shown in Figure 1 displays results in a way that highlights (very broadly) where and when relevant discourse was prominent and who were vocal agents of this discourse. Registered users can utilize Impresso's tools for topic modeling or data comparison regarding topics, publication frequencies, people, locations, and article types.

NewsEye, on the other hand, allows users to select from ten datasets curated by various project members, covering a range of topics like refugees or the Spanish Flu pandemic. The platform features exemplary case studies on "Women in Pants" (Omari, 2018) or "Interaction between newspapers and readers" (Kapferer, 2018), among others, which demonstrate how large-scale data collection and computerized analysis can help to answer research questions in communication and media history. Additionally, the datasets can be reused for related queries, and the platform interface enables researchers to select secondary information on publication frequencies, organizations, and persons, which can serve as variables in various analyses. Researchers can thus create their own datasets from the corpus of digital newspapers included in NewsEye (or Impresso) and then share them with others. This facilitates the creation of customized workspaces and collaborative research projects and is further aligned with the (emerging) "FAIR" principles for scientific data management. Consequently, the datasets can be utilized to address timely questions

Table 1: Overview of different types of digital press archives

	Flat portal	Deep portal	Virtual research environment
Data collection	› limited searchability › restricted display › download possible in single files (PDF or image) but sometimes only mere reading on screen allowed › limited metadata › manual corpus-building	› searchability improved with various parameters (date, location, page, article type) › varying modes and qualities of result display › metadata provided to varying extent › options for collecting large-scale corpora	› multitude of searching, saving, sharing functions › inherent visualization implemented, various display modes › group results by defined variables (names, places, topics) › different forms of metadata accessible › semi-automatic corpus-building (within the portal)
Data analysis	› typically qualitative research (case studies) with small datasets	› mixed methods: close reading and work with larger corpora possible › limited means for collaboration	› quantitative computerized analysis/distant reading (corpus linguistics, network analysis, topic modelling, etc.) › collaborative workflows
Examples	ZEFYS, Zeitungen des 17. Jahrhunderts, Google News	Chronicling America, Trove, RetroNews, ANNO	NewsEye, Impresso

about public health, migration, or gender not only from a historical perspective but also in comparison to contemporary discourses and research projects (e.g., DFG-funded projects like "Journalismus und sein Publikum" or analyses of media practice in performative publics; Lünenborg, Raetzsch, Reißmann, & Siemon, 2020). Although there are numerous analyses of discourses across media types and genres regarding contemporary public issues, the potential of the material available in such virtual research environments has not been fully exploited yet.

3.4 Overview: Affordances of different types

Table 1 summarizes the potential benefits and drawbacks of each type of digital archive in relation to the examples discussed above. The overview distinguishes two distinct modes/phases that determine research practices: data collection and data analysis.

4 Methodological reflections

Whether the platforms described above simply manifest the archival, library-oriented logic of digital collections or whether such portals create virtual research environments, they require methodological reflection. Several issues occurring with each type of digital archive to a certain degree are salient. They mainly relate to aspects of data generation and data collection caused by constraints in portal infrastructures (Ehrmann, 2020, p. 959; Koenen, 2018, p. 543).

A set of issues arises from a lack of comprehensive standards that lead to uncoordinated, spotty, or selective digitization efforts, missing or insufficient documentation of the data compilation process, and a quality gap with respect to OCR and metadata. This is a generic problem. Many flat and deep portals are persistently growing, which results in an internal quality gap as newer materials benefit from better OCR processing. Older efforts may have been impacted by imperfect OCR, which could have led to errors in image and text recognition, affecting the robustness of mining algorithms, for instance. In this regard, the affordances of digital archives proliferate the production of messy data, that is, digital collections that are characterized by incompleteness and inconsistencies due to technical constraints. Digital historians thus face challenges in developing a "nose" (Balbi, 2011, p. 155) for finding sources and establish-

ing new parameters for selecting, criticizing, interpreting, and comparing historical data (Stöber, 2016, p. 315) by creatively using keyword queries or sampling by limiting regional scope or time-settings (see also Oberbichler & Pfanzelter, 2021).

Further, problems arise with the sheer number of portals and their varying, often inflexible interfaces, "which force all users into a narrow range of interactions" (Gooding, 2017, p. 182) and hamper assessing the scope of material as well as searching specific digital corpora. Additionally, each portal has its own legal and economic restrictions for providing and using source material, which creates issues as regards access permissions and rights of use. Such digitization policy constraints force data providers (libraries, archives, and publishers) to provide non-representative collections. Koenen (2022a) refers to Kitchin's (2014, pp. 22–26) concept of digital archives as "data assemblages" through which informational and technical conditions, as well as institutional and corporate structures, become manifest. This relates to aspects of institutions (not) committing to fund long-term projects, the mitigation of political interests by (not, or only selectively) archiving specific sources (and thus limiting public access to them), the limitations of federal policies to manage digital collections, and the consolidation of different institutional interest groups, including private companies that may archive to make profits. The involvement of a variety of these factors necessarily informs selection priorities and affordances of digital press archives. Digital historians and, more so, media and journalism scholars need to gain a better understanding that an ideal and stable corpus is rare and reflect how underlying constraints shape the heuristics of their research and the materiality of the archival content. In this way, research will share the often diffuse responsibilities assigned to the preservation of cultural heritage (Balbi, 2011, p. 171).

These underlying factors of political economy and institutional bias (Koenen, 2022b, pp. 98–100) affect data generation and the means of data collection. In this sense, Putnam (2016, p. 382) argues that digital archives encourage researchers to focus on the "lamppost" and work with sources that are easily available, while the "shadows" are ignored. This creates digital bias as there is a lack of transparency regarding which sources are missing. For instance, the 33 million pages of the digitized press via the online British Newspaper Archive represent only about 6% of newspapers in the British Library collections (Tolfo et al., 2023, p. 28). As this information is usually not made explicit, researchers are well advised to reflect on such issues before using a specific archive. For instance, Lee et al. (2020, p. 3) describe the unequal distribution of digital sources in Chronicling America over time, across different states, and in terms of regional gaps. These limitations affect the functionality of the archive, which is to return accurate search results that are eventually compiled in larger corpora and analyzed through different methods of semi-automated textual analysis.

While this critique applies to virtual research environments like Impresso or NewsEye as well, such platforms suffer from transparency issues that are more directly related to data analysis. As discussed, researchers have access to a range of different tools to explore the data. However, it is necessary to critically assess inherent biases in these tools and how annotations are extracted from them to make informed use of data in a digital scholarship context. In a similar vein, working with Trove on the popularity of the London Illustrated News, Smits (2017) discusses the quality gap of newspaper digitization and expresses skepticism towards distant reading practices (see Section 3.2) because "the data of Trove is too unstable to draw any more far-reaching conclusions about the nature of this popularity" (Smits, 2017, p. 85). These deficits also affect the efficient maintenance of such a virtual research environment over time because the underlying constellations of agents stem from diverse scientific backgrounds, such as computational linguistics, design, and digital history, all collaborating on the datafication of a multilingual corpus of digitized historical newspapers, but with their own purposes, concepts, and institutional interests.

In addition to institutional and organizational aspects, there are also methodological issues, as argued by Nicholson (2013, p. 64). Digital press archives do not merely *reproduce* historical sources but *remediate* (and in the process change them) as "digitized text." As Ehrmann et al. (2020, p. 960) elaborate, data

is compiled in compressed archives (e. g., bzip2) containing a digital facsimile, with logical and physical representation of newspaper contents, including images and a version that can be automatically processed that is, the content as full texts or continuous strings of OCR token units. However, Cordell (2017, p. 188) emphasizes the relevance of determining where and from what "old" source material a text originated and how "the relationship between the OCR underlying a digital archive and images through which we typically experience those archives" can be understood. The former point addresses the issue that digitized text is often *not* remediated from the original printed edition of a publication but from microfilm copies. This is a first-order remediation process, where information about the original source (e. g., format, paper quality, layout, and coloring) is already becoming opaque (Koenen, 2022b, p. 106). The latter of Cordell's points pertains to the quality of OCR text, as the accuracy of an OCR engine is not always transparent and may change over time, as discussed above. Strange et al. (2014, Sec. 31) conclude that while "OCR achieves relatively accurate results (around 80%) on historical newspaper collections, […] manual correction is required to achieve high accuracy" and "[d]epending on the corpus size and the resources at hand, this two-step process may be no more efficient than directly inputting the original texts from scratch."

In this sense, working in and with digital press archives requires flexibility to switch between modes of distant reading and quantitative analysis to the practice of close reading and qualitative research (Bunout, 2023). While the latter is particularly useful for topic-specific explorations (e. g., Carter Olson, 2021), it is limited by what Nicholson has described as the conventional "top-down" – approach of (historical) research, which is challenged by necessary selectivity and lack of representative results. Modes of distant reading, on the other hand, are a commendable step towards implementing interdisciplinary research perspectives in media and communication history. They encourage a keyword-driven "bottom-up" approach, enabling topic- or discourse-centered analyses. This means approaching the source as a "bag-of-words" (Bubenhofer & Scharloth, 2015, p. 13), which allows for investigations of lexical frequencies, semantic collocates, and concordance plots. Purschwitz & Hinneburg (2019, pp. 55–56) discuss the potentials and drawbacks of this new paradigm as it moves research towards formulating hypotheses that only become apparent through the data-driven macroscope but lack a theoretical grounding.

The challenge of working with lists of words, topic models, and different forms of data visualization is to (re-)contextualize such findings in established theories and is further complicated by the fact that language is a phenomenon characterized by complexity and ambiguity, necessitating analyses that consciously apply mixed-method designs to arrive at valid results (Oberbichler & Pfanzelter, 2021). In this respect, easily accessible and usable corpus-linguistic tools that facilitate such mixed-method approaches can be adopted by media and journalism scholars who want to process material gathered from digital press archives. This suggests the necessity of continuing collaborative efforts between digital historians, media and journalism scholars, as well as corpus linguists, which may certainly bear fruit in the sense of aiming for a holistic picture of news discourse and conducting contextualized "reflected algorithmic textual analysis" (Pichler & Reiter, 2020, p. 58; see also Marchi, 2010, p. 165, 2022, p. 585) as a genuinely interdisciplinary endeavor.

5 Final remarks

This contribution outlined basic considerations when approaching different kinds of digital press archives and discussed research opportunities and restrictions within them. Because historical newspapers are mirrors of past societies and thus reflect the respective political, moral, and economic environments, they hold dense, continuous, and multi-level information that allows exploring how "contemporaries experienced their present" (Ehrmann et al., 2019, pp. 1–2). As such, digital press archives are not merely gateways to past mentalities and languages, public discourses, and symbolic meaning-making (Bingham, 2010, p. 228). Rather, even on a basic level,

portals function as reading rooms that bring researchers (and the broader public) in contact with historical phenomena. Deep portals can add a new dimension to the analysis of the evolution of journalistic language, professional practices, and role performances (Birkner et al., 2018). Koenen (2022b, p. 108) suggests that the availability and retrievability of sources allow for replicating and even expanding on previous research, for instance, on the layout and design of journalistic publications (Barnhurst & Nerone, 2001), the invention of journalistic genres such as the news report (Pöttker, 2003), or journalistic selection routines and news values in their historical development (Wilke, 1984).

In this sense, the continuing effort of newspaper digitization opens new avenues for research in media and communication history in data-driven times that may appear like a route to the land of milk and honey for relevant disciplines in the field of Digital Humanities. However, as was shown, limitations still exist due to the restricted availability of data formats, narrow portal interfaces, and buried or unavailable APIs, which all determine the logic and routines of research. This implicitly suggests that computational approaches should be complemented by qualitative analysis and close reading, which are genuinely possible in deep but data-restrictive portals like Chronicling America. In this spirit, Kergomard (2023, p. 374) suggests a strategy of blended reading, combining large- and small scale-analysis, and asks "that we clarify how we think of, construct and analyse our corpus".

On a different note, the overview presented was also suggestive of strongly diverging individual practices as regards (1) free or limited access to historical news archives and (2) compilation principles of these resources. These aspects may depend heavily on broader issues such as the political organization, and research policies of individual countries and states, which in turn may have a bearing on the allocation of research funding and consequently on the potential aims and scope individual digitization projects can pursue on a regional, national, or transnational scale.

Researchers should consider issues of quality, availability, and usability as part of historical source criticism that helps "to recontextualize the information" (Smits, 2017, p. 84; Blome, 2018, B.6. pp. 9–13), which represents a significant methodological advancement in digital communication (history) research. The evolution of digital press archives makes it paramount to reflect on the "infrastructural turn" that "brings new features to the field: innovation, experimentation, hands-on practices, and collaboration" (Pawlicka-Deger, 2020, Sec. 23). As a result, digital press archives are not mere virtual representations of newspapers and their content as material texts. The multi-layered data derived through OCR and the enrichment of metadata "constitute […] a new edition of that text" (Cordell, 2017, p. 196). As such, digital press archives not only establish the technological and methodological conditions in which these dynamic and flexible digital objects are constructed, shaped, and allocated but – like other digital sources – also change the methods and mindsets of communication and media scholars (Waldherr, 2019). At the same time, it is probable that more digital archives will be added to the current ones (or will be integrated into larger databases). Therefore, research-practical perspectives such as the present one will need to be regularly updated to account for the dynamicity of the field and its evolving opportunities.

With the progress made in developing these archives over the past decade, it is hoped that future efforts in digital humanities and social sciences will enhance access to and improve the quality of such materials (Marchi, 2022, p. 579). This will not only facilitate historical and diachronic research on news discourse through an empirical macroscope but also potentially transform the practices of collaboration, immersion, and critical reflection in the scientific community. This transformative process can assist scholars in achieving what William G. Thomas III already hoped for two decades ago, namely "a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem" (Cohen et al., 2008, p. 454). This can raise new awareness in contemporary societies regarding their cultural heritage.

Conflict of interests

The authors declare no conflict of interests.

References

Balbi, G. (2011). Doing media history in 2050. *Westminster Papers in Communication and Culture*, *8(2)*, 154–177.

Barnhurst, K., & Nerone, J. (2001). *The form of news. A history*. New York: The Guilford Press.

Bayrische Staatsbibliothek. (2024). *digiPress: Das Zeitungsportal der Bayrischen Staatsbibliothek [digiPress: The newpaper portal of the Bavarian state library]*. Retrieved from https://digipress.digitale-sammlungen.de/. Accessed: 25. 03. 2024.

Ben-David, A., & Amram, A. (2018). The *Internet Archive* and the socio-technical construction of historical facts. *Internet Histories*, *2*(1–2), 179–201. https://doi.org/10.1080/24701475.2018.1455412

Bingham, A. (2010). The digitization of newspapers archives: Opportunities and challenges for historians. *Twentieth Century British History*, *21*(2), 225–231. https://doi.org/10.1093/tcbh/hwq007

Birkner, T., Koenen, E., & Schwarzenegger, C. (2018). A century of journalism history as challenge: Digital archives, sources, and methods. *Digital Journalism*, *6*(9), 1121–1135. https://doi.org/10.1080/21670811.2018.1514271

Blome, A. (2018). Zeitungen [Newspapers]. In L. Busse, W. Enderle, R. Hohls, T. Meyer, J. Prellwitz, & A. Schuhmann (Eds.), *Clio Guide – Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften [Clio Guide – A handbook for digital resources in historical studies] (pp. B.6-1–B.6-36). Berlin: Humboldt-Universität.*

Bubenhofer, N., & Scharloth, J. (2015). Maschinelle Textanalyse im Zeichen von Big Data und Datadriven Turn: Überblick und Desiderate [Automated textual analysis in light of Big Data and the Datadriven Turn: Overview and desirables]. *Zeitschrift für germanistische Linguistik (ZGL) [Journal for german linguistics]*, *43*(1), 1–26. https://doi.org/10.1515/zgl-2015-0001

Bundesregierung (2021). Zeitungen aus drei Jahrhunderten im Netz [Newspapers from three centuries online]. Retrieved from https://www.bundesregierung.de/breg-de/suche/zeitungen-aus-drei-jahrhunderten-im-netz-1977752. Accessed: 17. 04. 2023.

Bunout, E. (2023). Contextualising queries: Guidance for research using current collections of digitised newspapers. In E. Bunout, M. Ehrmann, & F. Clavert (Eds.), *Digitised newspapers: A new Eldorado for historians* (pp. 277–300). Berlin: De Gruyter.

Brügger, N. (2018). *The archived web: Doing history in the digital age*. Boston: MIT Press. https://doi.org/10.7551/mitpress/10726.001.0001

Carter Olson, C. S. (2021). "To ask freedom for women": The night of terror and public memory. *Journalism & Mass Communication Quarterly*, *98*(1), 179–199. https://doi.org/10.1177/1077699020927118

Cohen, D. J. (2004). History and the second decade of the Web. *Rethinking History,8*(2), 293–301.

Cohen, D. J., Frisch, M., Gallagher, P. Mintz, S., Sword, K., Taylor, A. M., Thomas III, W. G., & Turkel, W. J. (2008). Interchange: The promise of digital history. *The Journal of American History, 96(2)*, 452–491.

Cordell, R. (2017). "Q i-jtb the Raven": Taking dirty OCR seriously. *Book History*, *20*, 188–225. https://doi.org/10.1353/bh.2017.0006

Cordell, R., & Smith, D. (2022). *Viral texts: Mapping networks of reprinting in 19th-century newspapers and magazines*. http://viraltexts.org. Accessed: 12. 07. 2023.

Deutsche Digitale Bibliothek. (2024). *Deutsches Zeitungsportal [German newspaper portal]*. Retrieved from https://www.deutsche-digitale-bibliothek.de/newspaper. Accessed: 25. 03. 2024.

DFGViewer (2012). Das Projekt [The project]. Retrieved from https://dfg-viewer.de/das-projekt. Accessed: 12. 07. 2023.

Dimbath, O. (2014). *Oblivionismus: Vergessen und Vergesslichkeit in der modernen Wissenschaft [Oblivionism: Forgetting and Forgetfulness in modern science]*. Konstanz: UVK.

Dombrowski, Q., Gniady, T., & Kloster, D. (2019). Introduction to Jupyter Notebooks. *Programming Historian, 8*. https://doi.org/10.46430/phen0087.

Ehrmann, M., Bunout, E., & Düring, M. (2019). Historical newspaper user interfaces: A review. *IFLA WLIC*. https://library.ifla.org/id/eprint/2578/1/085-ehrmann-en.pdf. Accessed: 14. 12. 2022.

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P., & Barman, R. (2020). Language resources for historical newspapers: The Impresso collection. European Language Resources Association (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 958–968). Marseille:

ELRA. Retriever from https://www.aclweb.org/anthology/2020.lrec-1.121. Accessed: 17. 04. 2023.

Esposito, E. (2002). *Soziales Vergessen [Social Forgetting]*. Frankfurt a. M.: Suhrkamp.

Europeana (2024). *Collections*. Retrieved from https://www.europeana.eu/en/collections/topic/18-newspaper. Accessed: 25. 03. 2024.

Friedrich Ebert Stiftung (2024). *Vorwärts [Forward]*. Retrieved from https://www.fes.de/adsd50/vorwaerts. Accessed: 25. 03. 2024.

Google (2024). *Google News*. Retrieved from https://news.google.com/newspapers?hl=en. Accessed: 25. 03. 2024.

Google (2024). *Google Books*. Retrieved from https://books.google.de/. Accessed: 25. 03. 2024.

Gooding, P. (2017). *Historic newspapers in the digital age: Search all about it!* London: Routledge. https://doi.org/10.4324/9781315586830

Graham, S., Milligan, I., & Weingart, S. (2016). *Exploring big historical data: The historian's macroscope*. London: ICP.

Hepp, A. (2016). Kommunikations- und Medienwissenschaft in datengetriebenen Zeiten [Communication and Media Studies in datadriven times]. *Publizistik, 61*, 225–246. https://doi.org/10.1007/s11616-016-0263-y

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine, 15(3–4)*. Retreived from http://www.dlib.org/dlib/march09/holley/03holley.html. Accessed: 12. 07. 2023.

Impresso Project. (2024). *Impresso: Media Monitoring of the Past*. Retrieved from https://impresso-project.ch/app/?sq=. Accessed: 25. 03. 2024.

Jcoliver. (2024). *Dig-coll-borderlands*. Retrieved from https://github.com/jcoliver/dig-coll-borderlands. Accessed: 25. 03. 2024.

Journal of Digital History. (2024). *Write (digital) history*. Retrieved from https://journalofdigitalhistory.org/en. Accessed: 25. 03. 2024.

Kapferer, B. (2018). Interaction between newspapers and readers. *Newseye*. Retrieved from https://www.newseye.eu/case-studies/case-study-4-media-and-journalism/interaction-between-newspapers-and-readers/. Accessed: 12. 07. 2023.

Kergomard, Z. (2023). A source like any other? In E. Bunout, M. Ehrmann, & F. Clavert (Eds.), *Digitised newspapers: A new Eldorado for historians?* (pp. 359–378). Berlin: De Gruyter. https://doi.org/10.1515/9783110729214-016

Kinnebrock, S., Schwarzenegger, C., & Birkner, T. (2015). Theorien des Medienwandels – Konturen eines emergierenden Forschungsfeldes? [Theories of media change – contours of an emerging field of research] In S. Kinnebrock, C. Schwarzenegger & T. Birkner (Eds.), *Theorien des Medienwandels* [Theories of media change] (pp. 11–28). Köln: Herbert von Halem.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. New York: Sage.

Koenen, E. (2018). Digitale Perspektiven in der Kommunikations- und Mediengeschichte [Digital perspectives in communication and media studies]. *Publizistik, 63, 535–556*. https://doi.org/10.1007/s11616-018-0459-4

Koenen, E. (2022a). Epistemologie digitaler Experimentalsysteme am Beispiel von Zeitungsportalen: Methodologische und praktische Herausforderungen, Probleme und Perspektiven. [Epistemology of digital experimental systems – the example of newspaper portals: Methodological and practical challenges, problems, and perspectives] *Zeitschrift* für *digitale Geisteswissenschaften*. [Journal for digital historical studies] https://doi.org/10.17175/sb005_013

Koenen, E. (2022b). Digitalisierte Zeitungen des 19. und 20. Jahrhunderts in der historischen Presseforschung: Dimensionen und Probleme einer digitalen Quellenkritik [Digitized newspapers of the 19. and 20. century in historical press research: Dimensions and problems of digital source critique]. *Jahrbuch für Kommunikationsgeschichte [Yearbook for Communication History]*, *24*, 95–114.

Lee, B., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & Weld, D. (2020). The Newspaper Navigator Dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in *Chronicling America*. *arXiv*. https://doi.org/10.48550/arXiv.2005.01583

Library of Congress. (2024a). *Chronicling America*. Retrieved from https://chroniclingamerica.loc.gov/. Accessed: 25. 03. 2024.

Library of Congress. (2024b). *National Digital Newspaper Program*. Retrieved from https://www.loc.gov/ndnp/. Accessed: 25. 03. 2024.

Lmullen (2024). Chronam OCR debatcher. Retrieved from https://github.com/lmullen/chronam-ocr-debatcher/releases. Accessed: 25. 03. 2024.

Lünenborg, M., Raetzsch, C., Reißmann, W., & Siemon, M. (2020). Media Practice in performativen Öffentlichkeiten. Für eine praxistheoretische Positionierung der Journalismusforschung. [Media Practice in performative publics. For a practical-theoretical positioning in journalism research] In J. Schützeneder, K. Meier & N. Springer (Eds.), *Neujustierung der Journalistik / Journalismusforschung in der digitalen Gesellschaft. Jahrestagung der Fachgruppe Journalistik / Journalismusforschung der DGPUK 2019* [Readjustment of journalism research in the digital society. Annual conference of the DGPUK-section journalism research 2019] (pp. 34–51). Eichstätt. https://doi.org/10.21241/ssoar.70817

Marchi, A. (2010). "The moral in the story": A diachronic investigation of lexicalised morality in the UK press. *Corpora*, *5(2)*, 161–189. https://doi.org/10.3366/E1749503210000432

Marchi, A. (2022). Corpus linguistics in the study of news media. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 576–588). New York: Routledge.

Michael, H. (2017). Konstitution und Synthese der Reportage in der Lokalberichterstattung der New Yorker Massenpresse vor dem Bürgerkrieg: ein Beitrag zur historischen Genreforschung. [Constitution and synthesis of reportage in local reporting of New York mass periodicals before the Civil War: A contribution to historical genre analysis] *Medien & Zeit*, *32*(2), 4–16.

Moretti, F. (2016). *Distant reading*. Konstanz: Konstanz UP.

Mullen, L. (2019). *Chronam-ocr-debatcher*. Retrieved from https://github.com/lmullen/chronam-ocr-debatcher/releases. Accessed: 12. 07. 2023.

Nanni, F. (2018). Collecting primary sources from web archives: A tale of scarcity and abundance. In N. Brügger, & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 112–124). New York: Sage.

National Library of Australia. (2024). *Trove*. Retrieved from https://trove.nla.gov.au/. Accessed: 25. 03. 2024.

NewsEye (2022). *A Digital investigator for historical newspapers*. Retrieved from https://doi.org/10.3030/770299

Nicholson, B. (2013). The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History*, *19(1),* 59–73. https://doi.org/10.1080/13688804.2012.752963

Oberbichler, S., & Pfanzelter, E. (2021). Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods. *Journal of Digital History*, *1*(1). https://doi.org/10.1515/JDH-2021-1003

Österreichische Nationalbibliothek. (2024). ANNO Historische Zeitungen und Zeitschriften. [ANNO historical newspapers and journals] Retrieved from https://anno.onb.ac.at/. Accessed: 25. 03. 2024.

Omari, N. (2018). Women in pants. Case study 2: Gender (tansl. Ummel, A.). *Newseye*. Retrieved from https://www.newseye.eu/case-studies/case-study-2-gender/women-in-pants/. Accessed: 12. 07. 2023

Pawlicka-Deger, U. (2020). The laboratory turn: Exploring discourses, landscapes, and models of humanities labs. *Digital Humanities Quarterly, 14*(3). Retrieved from http://www.digitalhumanities.org/dhq/vol/14/3/000466/000466.html. Accessed: 12. 07. 2023.

Pfanzelter, E., Oberbichler, S., Marjanen, J., Langlais, P.-C., & Hechl, S. (2021). Digital interfaces of historical newspapers: Opportunities, restrictions and recommendations. *Journal of Data Mining and Digital Humanities*. https://doi.org/10.46298/jdmdh.6121

Pichler, A., & Reiter, N. (2020). Reflektierte Textanalyse. [Reflected textual analysis] In N. Reiter, A. Pichler, & J. Kuhn (Eds.), Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt [Reflected algorithmic textual analysis: Interdisciplinary works in the CRETA-workshop] (pp. 43–60). Berlin: de Gruyter. https://doi.org/10.1515/9783110693973-003

Podewski, M. (2018). 'Kleine Archive' in den Digital Humanities: Überlegungen zum Forschungsobjekt 'Zeitschrift.' ['Small' archives in the digital humanities: Reflections on the research object 'journal'] *Zeitschrift für digitale Geisteswissenschaften [Journal for digital*

historical studies]. https://doi.org/10.17175/sb003_010

Pöttker, H. (2003). News and its communicative quality. The inverted pyramid: when and why did it appear? *Journalism Studies, 4*(4), 501–511. https:/doi.org/10.1080/1461670032000136596

Purschwitz, A., & Hinneburg, A. (2019). Funktionsmechanismen gesellschaftlicher Wissensproduktion – Chancen und Grenzen des Topic-Modeling in den Geisteswissenschaften. Die halleschen Zeitungen und Zeitschriften 1688–1815. [Functional mechanisms of social knowledge production – chances and limitations of topic-modeling in the humanities. Newspapers and journals from Halle 1688–1815] *Medien & Zeit*, *34*(2), 50–64.

Putnam, L. (2016). The transnational and the text-searchable. Digitized sources and the shadows they cast. *The American Historical Review*, *121*(2), 377–402.

Ramsay, S. (2011). *Reading machines: Toward an algorithmic criticism*. Champaign: UI Press.

Reuters. (2024). *News archives*. Retrieved from https://www.reutersagency.com/en/media-solutions/archive/. Accessed: 25. 03. 2024.

Rosenzweig, R. (2003). Scarcity or abundance? Preserving the past in a digital era. *The American Historical Review, 108(3), 735–762.*

Schwarzenegger, C., Koenen, E., Pentzold, C., Birkner, T., & Katzenbach, C. (2022). Der digitalen Kommunikation eine Vergangenheit geben: Gegenstände und Perspektiven eines überfälligen Unterfangens. [Giving digital communication a past: Objects and Perspectives of an Outstanding Endeavour] In C. Schwarzenegger, E. Koenen, C. Pentzold, T. Birkner, & C. Katzenbach (Eds.), *Digitale Kommunikation und Kommunikationsgeschichte: Perspektiven, Potentiale, Problemfelder* [Digital Communication and Communication History: Perspectives, Potentials, and Problems] (pp. 9–27). Berlin: SSOAR. https://doi.org/10.48541/dcr.v10.1

Sherratt, T. (2021). GLAM Workbench (version v1.0.0). *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.5603060. Accessed: 12. 07. 2023.

Smits, T. (2017). Looking for *The Illustrated London News* in Australian digital newspapers. *Media History*, *23*(1), 80–99. https://doi.org/10.1080/13688804.2016.1196585

Staatsbibliothek zu Berlin. (2024). ZEFYS: Zeitungsinformationssystem [ZEFYS: Newspaper information system]. Retrieved from https://zefys.staatsbibliothek-berlin.de/. Accessed: 25. 03. 2024.

Staats- und Universitätsbibliothek Bremen. (2024). *Historische Zeitungen [Historical newspapers]*. Retrieved from https://brema.suub.uni-bremen.de/zeitungen17. Accessed: 25. 03. 2024.

Stöber, R. (2016). Historische Methoden in der Kommunikationswissenschaft: Die Standards einer Triangulation. [Historical methods in communication studies: Standards of triangulation] In S. Averbeck-Lietz, & M. Meyen (Eds.). *Qualitative Methoden der Kommunikationswissenschaft* [Qualitative methods in communication studies] (pp. 303–318). Wiesbaden: Springer.

Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, *8*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html. Accessed: 14. 12. 2022.

The Bibliothèque nationale des France. (2024a). *Gallica*. Retrieved from https://gallica.bnf.fr/accueil/de/content/accueil-de?mode=desktop. Accessed: 25. 03. 2024.

The Bibliothèque nationale des France. (2024b). *Retronews*. Retrieved from https://www.retronews.fr/. Accessed: 25. 03. 2024.

The British Newspaper Archive. (2024). Retrieved from https://www.britishnewspaperarchive.co.uk/. Accessed: 25. 03. 2024.

The University of Arizona. (2024). *Digital Borderlands.* Retrieved from http://borderlands.digitalscholarship.library.arizona.edu/. Accessed: 25. 03. 2024.

The University of Arizona. (2024). *Newspapers as data: A collections as data project.* Retrieved from https://libguides.library.arizona.edu/newspapers-as-data. Accessed: 25. 03. 2024.

Tolfo, G., Vane, O., Beelen, K., Hosseini, K., Lawrence, J., Beavan, D., & McDonough, K. (2023). Hunting for treasure: Living with machines and the British Library Newspaper Collection. In E. Bunout, M. Ehrmann, & F. Clavert (Eds.), *Digitised newspapers: A new Eldorado for historians?* (pp. 25–46). Berlin: De Gruyter.

Viola, L., & Verheul, J. (2020). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. *Digital Scholarship in the Humanities*, *35*(4), 921–943. https://doi.org/10.1093/llc/fqz068

Waldherr, A. (2019). Messinstrumente und Sinnkonstruktionen: Methoden als Antreiber und Taktgeber der Kommunikationswissenschaft. [Measuring instruments and meaning making: methods as drivers and clock of communication studies] *Medien & Zeit*, *34*(1), 40–47.

Wikipedia. (2024). *Wikipedia: List of online newspaper archives.* Retrieved from https://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives. Accessed: 25. 03. 2024.

Wilke, J. (1984). *Nachrichtenauswahl und Medienrealität in vier Jahrhunderten.* [News selection and media reality in four centuries] Berlin: De Gruyter.

Zeit.punkt NRW. (2024). *Zeitungsportal NRW.* [Newspaper Portal Northrhine Westfalia] Retrieved from https://zeitpunkt.nrw/?lang=en. Accessed: 25. 03. 2024.

Zeitschriftendatenbank (2022). *Zeitungsdigitalisierung.* [Newspaper digitization] Retrieved from https://zeitschriftendatenbank.de/zeitungsdigitalisierung. Accessed: 12. 07. 2023.