Maud Ehrmann, Estelle Bunout, Frédéric Clavert

Digitised Historical Newspapers: A Changing Research Landscape

Introduction

The application of digital technologies to newspaper archives is transforming the way historians engage with these sources. The digital evolution not only affects how scholars access historical newspapers, but also, increasingly, how they search, explore and study them. Two developments have been driving this transformation: massive digitisation, which facilitates access to remote holdings and, more recently, improved search capabilities, which alleviate the tedious exploration of vast collections, opens up new prospects and transforms research practices.

Since the 2000s, regional and national libraries as well as transnational bodies and commercial operators made considerable investments in historical newspaper digitisation, with the aim of both making them available to larger audiences and ensuring the preservation of sometimes fragile paper originals.[1] This endeavour not only focused on document imaging but also, and importantly, on the transcription of their contents into machine-readable text using optical character and

[1] Natasha Stroeker et al. (2012). *Survey Report on Digitisation in European Cultural Heritage Institutions 2012.* Tech. rep. Brussels. URL: https://www.egmus.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf.

Acknowledgements: The publication of this volume would not have been possible without the support and commitment of many. We express our warmest thanks to the 'Eldorado' workshop programme committee, namely Simon Clematide, Marten Düring, Martin Grandjean, Caroline Muller, Enrico Natale, Matteo Romanello, Raphaëlle Ruppen Coutaz and François Vallotton, for their valuable input and feedback on the scope of the workshop and for their participation in the review process and chairing of sessions. We are grateful to the *impresso* team and consortium for the trusting environment in which, at the onset of a pandemic, the workshop could unfold. We express our greatest appreciation to the people who helped us move this publication forward: Beth Park from C^{2}DH/Luxembourg University for her support with administrative processes, and Rabea Rittgerodt and Jana Fritsche from De Gruyter for their assistance in editorial tasks and their patience and encouragement. This volume is published as part of the research activities of the project "*impresso*– Media Monitoring of the Past", for which we also gratefully acknowledge the financial support of the Swiss National Science Foundation under grant number CR-SII5_173719. Last but not least, we thank all authors for their high-quality contributions and scientific involvement, as well as each one who attended the online workshop, either as panellist or participant.

https://doi.org/10.1515/9783110729214-001

layout recognition technologies (OCR and OLR). These efforts have yielded millions of newspaper facsimiles along with their transcribed text at regional, national and international levels.[2] From manual, on-site exploration of microfilm or paper collections to full-text search over millions of OCRed pages via online portals, digitisation significantly eased the way academic and non-academic users alike can access, visualise and search historical newspapers.[3]

Beyond preservation and accessibility, digitisation also offers the possibility of applying machine-reading techniques to the content of digitised newspapers, with the potential to extend exploration capabilities far beyond keyword searching, browsing, and close reading. In this regard, a diverse research community – including researchers from digital humanities, natural language processing (NLP), computer vision, digital library and computer sciences – started to pool forces and expertise to push forward the processing of digitised newspapers as well as the extraction and linking of the complex information enclosed in their transcriptions. Besides individual works dedicated to the development of tools,[4] evaluation campaigns and hackathons have multiplied[5] and several large

[2] See for example the Impact project (www.impact-project.eu) and the following Centre of Competences (digitisation.eu), as well as the Europeana Newspaper project with Clemens Neudecker and Apostolos Antonacopoulos (2016). "Making Europe's Historical Newspapers Searchable." In: *Proc. of the 12th IAPR Workshop on Document Analysis Systems*. Santorini, Greece: IEEE. DOI: 10.1109/DAS.2016.83.

[3] A. Bingham (2010). "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." In: *Twentieth Century British History* 21.2. DOI: 10.1093/tcbh/hwq007; Bob Nicholson (2013). "The Digital Turn." In: *Media History* 19.1. DOI: 10.1080/13688804.2012.752963.

[4] To cite but a few: Tze-I. Yang et al. (2011). "Topic modeling on historical newspapers." In: *Proc. of the 5th LaTecH workshop*. ACL, pp. 96–104; Jean-Philippe Moreux (2016). "Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment." en. In: *Proc. of IFLA WLIC 2016*, p. 17. URL: http://library.ifla.org/id/eprint/2076; Melvin Wevers (2019). "Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950–1990." In: *Proc. of the 1st International Workshop on Computational Approaches to Historical Language Change*. ACL. URL: https://www.aclweb.org/anthology/W19-4712; Mike Kestemont et al. (2014). "Mining the Twentieth Century's History from the Time Magazine Corpus." In: *Proc. of the 8th LaTeCH workshop*. ACL. URL: https://aclanthology.org/W14-0609; Thomas Lansdall-Welfare et al. (2017). "Content Analysis of 150 Years of British Periodicals." In: *Proceedings of the National Academy of Sciences* 114.4. DOI: 10.1073/pnas.1606380114.

[5] Christophe Rigaud et al. (2019). "ICDAR 2019 Competition on Post-OCR Text Correction." In: *ICDAR Proceedings*. Sydney, Australia. URL: https://hal.archives-ouvertes.fr/hal-02304334; Maud Ehrmann, Matteo Romanello, Alex Flückiger, et al. (2020). "Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Avi Arampatzis et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 288–310. DOI: 10.1007/978-3-030-58219-7_21; Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, et al. (2022). "Extended

consortia projects proposing to apply computational methods to historical newspapers at scale have emerged.[6] Text and image processing techniques are increasingly applied, enriching digitised newspapers with new layers of information in the form of semantic annotations (e.g., n-gram frequencies, named entities, topics, events, text reuse elements, objects in images) and enabling new search possibilities (e.g., keyword suggestion, content recommendation, visual search). Within a few years, these interdisciplinary efforts have produced a set of tools, technical infrastructures, and graphical interfaces that are radically transforming the way digitised newspapers are used. Today, conducting historical research on the basis of automatically enriched newspapers accessible via increasingly sophisticated interfaces is no longer a distant (and debated) prospect, but a tangible reality and a commonplace for many researchers.

In this changing research landscape, historians face a complex mix of opportunities and challenges: while search algorithms, 'datafied' newspapers and visualisation capabilities offer new opportunities, they also confront researchers with new difficulties. Automatically extracted data – most often based on probabilistic approaches–and exploration interfaces are far from neutral and their integration into historical research practices must be accompanied by a critical assessment of their biases and limitations. At the heuristic level, these include the noise introduced by faulty text and document structure recognition processes, with the result that what can be searched is not necessarily what was printed.[7] Inevitably, this noise propagates to downstream processes and affects their performances, in particular with collections digitised long ago.[8]

Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents." In: *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*. Ed. by Guglielmo Faggioli et al. Vol. 3180. CEUR-WS. DOI: 10.5281/zenodo.6979577. URL: http://ceur-ws.org/Vol-3180/paper-83.pdf.

[6] See for example the following projects: Viral Texts (US, 2012–2016); Oceanic Exchanges (US/EU, 2017–2019); *impresso* – Media Monitoring of the past (CH, 2017–2020); Newseye (EU, 2018–2021); and Living with Machines (UK, 2018–2023).

[7] Jarlbrink et al. talk of the 'heritage noise' which creates new perplexities for researchers: Johan Jarlbrink et al. (2017). "Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive." In: *Journal of Documentation* 73.6, pp. 1228–1243.

[8] Daniel van Strien et al. (2020). "Assessing the Impact of OCR Quality on Downstream NLP Tasks." In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. DOI: 10.5220/0009169004840496; Maud Ehrmann, Ahmed Hamdi, et al. (2022). "Named Entity Recognition and Classification in Historical Documents: A Survey." In: *ACM Computing Survey (accepted)*. URL: https://arxiv.org/abs/2109.11406; Estelle Bunout et al. (2019). *Collections of Digitised Newspapers as Historical Sources – Parthenos Training*. URL: https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/collections-of-digital-newspapers-as-historical-sources/ (visited on 09/02/2022).

Although the situation is evolving rapidly thanks to re-OCRisation campaigns, new OCR approaches and text mining models better able to deal with noise,[9] this distortion of the source is, to varying extents, ubiquitous.

Yet, the way to search digitised newspapers and build a research corpus is no longer limited to full-text search alone, and semantic enrichments offer complementary and powerful search and filtering capacities. Besides, source contextualisation is also changing with, on the one hand, a loss of context – keyword search translates into a straight jump to individual articles that conceals the context of surrounding articles and issue – and, on the other, new contextualisations with, for example, information about digitisation processes and links within and across collections. However, this information is not (yet) systematically available in newspaper interfaces, which are also mostly silent on the blind spots of non-digitised (parts of) collections.[10] Overall, the (re) search horizon is subject to various mutations and becomes at the same time broader, less precise, more efficient and in places more fruitful.

The transformation of newspaper sources into complex data objects impacts all phases of historical work and calls for revisited, digitally informed source criticism and interpretation, as well as for the critical analysis of tools and interfaces.[11] In this promising but unsettled context, numerous methodological and epistemological

[9] Clemens Neudecker, Konstantin Baierer, et al. (May 2019). “OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents.” In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. DATeCH2019. New York, NY, USA: Association for Computing Machinery, pp. 53–58. DOI: 10.1145/3322905.3322917; Emanuela Boros et al. (2020). “Robust Named Entity Recognition and Linking on Historical Multilingual Documents.” In: *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*. Ed. by Linda Cappellato et al. Vol. 2696. Thessaloniki, Greece: CEUR-WS, pp. 1–17. URL: http://ceur-ws.org/Vol-2696/paper_171.pdf.

[10] Lara Putnam (Apr. 2016). “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast.” In: *The American Historical Review* 121.2. DOI: 10.1093 / ahr / 121. 2.377; Maud Ehrmann, Estelle Bunout, et al. (2019). “Historical Newspaper User Interfaces: A Review.” In: *Proc. of the 85th IFLA General Conference and Assembly*. Athens, Greece: IFLA Library. DOI: 10.5281/zenodo.3404155. URL: http://infoscience.epfl.ch/record/270246.

[11] Andreas Fickers (2012). “Towards a new digital historicism? Doing history in the age of abundance.” In: *VIEW Journal of European Television History and Culture* 1.1, pp. 19–26; Marijn Koolen et al. (2019). “Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice.” In: *Digital Scholarship in the Humanities* 34.2, pp. 368–385. DOI:10.1093/llc/fqy048; Andreas Fickers (2020). “Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?” In: *Zeithistorische Forschungen* 17.1, pp. 157–168. DOI: https://doi.org/10.14765/zzf .dok-1765; Sarah Oberbichler et al. (2021). “Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians.” In: *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.24565.

questions arise: How, to which extent and under what conditions do automatically generated data support the search and exploration of historical newspapers? What understanding do historians need of digitisation and information extraction techniques to properly interpret enriched historical sources? How can imperfections caused by digitisation and NLP tools best be managed?

Drawing on a growing community of practices, the *impresso* project invited scholars experienced with digitised newspaper collections, tools and interfaces to share their experiences, research practices and findings during a workshop named 'Eldorado' (held online in the early days of the COVID-19 pandemic).[12] This volume brings together the contributions of most of the panellists and offers a snapshot of the research done with digitised newspapers, taking a closer look at the promises and hopes often expressed in the context of digitisation. The primary target audiences are researchers and students in history, as well as researchers and practitioners in the field of digital humanities.

This volume was produced in the context of '*impresso*. Media Monitoring of the Past', an interdisciplinary research project in which a team of computational linguists, designers and historians collaborate on the datafication of a multilingual corpus of digitised historical newspapers.[13] The primary goals of the project are to improve text mining tools for historical text, to enrich historical newspapers with automatically generated data and to integrate such data into historical research workflows by means of a newly developed user interface.[14] Beyond the challenges specific to the different research areas underpinning each of these objectives, the question of how best to adapt text mining tools and their use by humanities scholars is at the heart of the *impresso* enterprise.

Presentation of Contributions

This volume is composed of eighteen articles (13 in English, 4 in German and 3 in French) and is organised in three parts. The first part 'prospects the Eldorado' and focuses on access to digitised newspapers, how it is implemented and with what consequences on research workflows. The second part explores some of the possibilities offered by digital tools to expose or compose new artefacts from these collections, enabling a 'digital reshaping of newspapers'. Finally, the last

[12] https://impresso.github.io/eldorado/.
[13] https://impresso-project.ch.
[14] https://impresso-project.ch/app.

part examines how to effectively mine digital newspapers and create relevant research corpora while incorporating source criticism.

The Allure of Digitised Newspapers: Prospecting the Eldorado

The first part of the volume provides a bird's eye view of the digitised newspaper landscape and initiates a discussion of the impact of digital mediation on the use of newspaper sources for research.[15] Composed of 7 chapters written by scholars from different backgrounds – whether historians, digital librarians or designers – it focuses on the question of which digitised collections we actually have access to, and how. At the intersection of digitisation strategies, curatorial practices, technical implementations, and digitisation quality, this first set of contributions provides an appreciation of both the potential and limitations of the current digitised newspaper landscape, and suggests innovative paths forward.

In the opening of this volume, **Giorgia Tolfo** and her colleagues focus on the very first steps at the origin of any digitised newspaper collection, namely digitisation strategy and selection priorities. Starting from the observation of the recurrent and widespread tension that exists between researchers' expectations of the perfect corpus and the many constraints and decisions that guide library digitisation processes, the authors propose a middle way with the aim of enabling scholars to understand the composition of a digital corpus and make informed decisions about it. The work of Giorgia Tolfo, **Olivia Vane**, **Kaspar Beelen**, **Kasra Hosseini**, **Jon Lawrence**, **David Beavan** and **Katherine McDonough** is part of the open digitisation programme undertaken by the Living with Machine project[16] and the British Library, in which they have attempted to reconcile research-oriented workflows with the diverse requirements of digitisation procedures. They put forward an approach based on a combined selection model, whereby an iterative selection of titles to digitise based on research needs is regularly associated to library digitisation priorities. In the context of heterogeneous and sometimes competing demands, authors emphasise the need to make choices both transparent and pragmatic. To this end, they develop two

[15] The 'Allure of digitised newspapers' echoes Arlette Farge et al. (2013). The Allure of the Archives. The Lewis Walpole Series in Eighteenth-Century Culture and History. New Haven: Yale Univ. Press, as well as its recent digital counterpart: Frédéric Clavert and Caroline Muller, eds. (2017). *Le goût de l'archive à l'ère numérique*. https://goutnumerique.net/.

[16] For URLs of all cited projects, tools or interfaces, we refer the reader to the corresponding chapters.

solutions, the *Press Picker* and the *Environmental Scan*, to support decision making about digitisation and document the selection process and decisions (production of paradata). In a nutshell, the *Press Picker* is a data visualisation tool that allows the selection of newspaper titles to be digitised on the basis of collection metadata, including information on the title's name change, its physical support conservation and possible parts already digitised. The *Environmental Scan* further informs this process by deriving additional newspaper metadata from contemporaneous sources – here the British Newspaper Press Directories, an authoritative list of newspapers that circulated in Britain in the 19th century. This is particularly useful for understanding, beyond those that have already been digitised, which titles have been published, when and where. Overall, drawing on practical experience, Giorgia Tolfo and her colleagues offer both insightful perspectives and practical answers to a hitherto recognised but rarely addressed problem: how to implement research-driven digitisation and support a transparent and informed composition of digital corpora. On the edge of the 'Eldorado', this chapter reminds us of its gaps and boundaries, and show that a close collaboration between researchers and librarians can help tackle them.

While Giorga Tolfo and her colleagues rightly warn of the inevitable incompleteness and partiality of any digitised collection and emphasise the need to understand the contours of what is available, **Andrew Torget** pursues this word of caution and highlights another pitfall of newspaper collections: OCR noise. In his chapter, written as part of the Mapping Texts project, the author highlights the impact of OCR recognition rates on the ability of researchers to not only search and explore but also apply natural language processing to digitised newspaper collections. More specifically, Andrew Torget presents two interactive visualisation modules to support the discovery of high-level trends in collections of digitised newspapers, applied here on a set of titles from Texas. The first offers a quantitative survey of the textual material in these sources and allows to visualise the distribution of words and of OCR noise over time, space, and by newspaper title. Such visualisation reveals, for example, in which geographical locations is the largest quantity of available data, and which time periods have particularly low OCR quality. This allows scholars to better understand what may or may not be accessible or searchable, and to determine the usability of a collection (or a specific title or time period) for their research. The second module exposes high-level thematic trends based on word frequencies, named entity counts and topics and here again OCR noise surfaces as a regular pattern. Whether 'mapping' the quality or content of newspapers, these experiments demonstrate how pervasive OCR noise is and illustrate how communicating this information to researchers. Andrew Torget concludes with a plea for more transparency on OCR quality, both in terms of quality rates and underlying processes – a precondition for enabling the responsible use of these collections.

Moving away from collections of complete newspaper editions, **Irene Amstutz, Martin Reisacher and Elias Kreyenbühl** take us on a journey from an analogue to a digital research infrastructure for a specific type of newspaper archives: press clippings. As part of its economic documentation activities, the Swiss Economic Archives (SWA) has been collecting press clippings from Swiss titles on topics and actors in the Swiss economy since the 1850s, by manually cutting and metadating paper articles until 2012, and then by using specific software on digital news editions. The resulting collection of approximately 2.7 million printed press clippings and about 180,000 digital clippings documents the social and economic history of Switzerland in a unique way, is highly valued by researchers and continues to grow. The shift from manual to software-assisted collection of press clippings and the changes this has brought about are precisely what motivated the SWA to retro-digitise its paper clippings and to work on a new digital infrastructure to improve access to its two collections. To this end, the team (consisting of archivists, digitisation specialists and software engineers) choose the Image Interoperability Framework (IIIF) as their digital compass and defined a set of specifications and guiding principles at the outset of the project, namely: use of existing standards (to benefit from and cooperate with a community of developers), focus on interoperability and infrastructure (to integrate an ecosystem of APIs and tools around digital cultural heritage assets), careful management of authentication, and flexible use and composition of virtual document collections. This undertaking, which is still ongoing, has not been without its difficulties, whether it be the technical implementation, the integration of differently organised collections under a single thesaurus (the Standard Thesaurus for Economics), or the management of copyright. The resulting digital infrastructure is a set of APIs that provides authenticated access to a large collection of news clippings (including images, transcripts and rich metadata), that can be used with existing tools (e.g. viewers) and that offers the possibility to build one's own collection and combine it with other collections available via the same standard. This is a major achievement that bridges the digital and indexing gap between the print and born-digital press clipping collections of the SWA; if some open questions remain (sustainability, persistence of identifiers, transparency), it definitely opens up new opportunities, both for researchers and cultural institutions.

Connecting the preservation and the research perspectives, **Claudia Resch** presents a complete cycle of creating and using a digitised newspaper corpus for research. Focusing on the *Wiener Zeitung*, an Austrian newspaper published since 1703 and recognised as the "oldest newspaper in the world", this chapter examines the issue of OCR quality from two angles, how to improve and adequately communicate it (this element too often remains opaque to users), and what impact it has on research. The author first reports the efforts carried out in the context of the project ‘Das Wien[n]erische Diarium’ to produce gold standard transcriptions for selected issues of the newspaper in order to estimate the efforts needed to produce high-quality transcripts, to train a more efficient OCR model that can be applied to all issues, and to have a corpus of good quality to conduct research. The resulting data is available via an online portal, where the quality level of the transcripts is indicated by a color code. Claudia Resch then takes advantage of this corpus to answer a research question from historical linguistics, that of the gradual adoption of new writing standards emerging during the 18th century. She discusses the evolution of the relative frequencies of two spelling variants of the verb *to be* in German (*seynd* and *sind*), also considering the broader media and political contexts of the definition and implementation of new linguistic norms. Overall, this chapter highlights the diversity of OCR quality within a single collection, illustrates how complex it is to communicate about it, and shows how machine-readable newspaper content can support research questions that remain otherwise difficult to source.

Addressing digitisation, OCR, and image delivery on the Internet, the previous four chapters presents the principles and techniques underlying the digitised newspaper landscape, highlighting its main characteristics, contours, but also its limitations. This new mode of access to historical newspapers also has a significant impact on how researchers interact with this material; the following chapters examine this point more closely.

Claire-Lise Gaillard opens this discussion with a reflection on the impact of digitisation on the search for sources by historians and asks: How to leaf through gigabytes of newspapers? As part of her doctoral research on the dating market from the 19th to the 20th century – a doctoral dissertation largely based on digitised sources, which would have been unimaginable 10 years ago –, Claire-Lise Gaillard shares her experience working with online portals and digitised collections. She examines both the powerful opportunities and the danger of misconceptions or even illusions that can result from them, and shows the ambivalence of this new mode of access. The ergonomics of research is profoundly modified, with the disappearance of the traditional steps towards the archive (exchanges with archivists, attention to the collections and their subdivisions), the loss of the material dimension of the sources and a radically different reading experience. Instead, keyword search becomes central, giving immediate access to masses of ‘immaterial’ documents. Such digital exploration is not without risk, the author reminds us, for example that of a false sense of completeness (not all newspapers are digitised and OCRed), of misconceptions about a corpus whose contours are determined exclusively by a search query, and of neglecting the context. This mode of access shapes historians' corpora, which in turn shape their results. In this regard, the author advocates a reflective posture and recommends considering keyword search results as an indication of where to look rather than as an automatic way to build a corpus. Claire-Lise Gaillard also emphasises the capacities of online portals to support such critical posture, and the added value of distant reading tools (e.g. the *impresso* interface or Numapresse tools) when the eye is no longer sufficient to apprehend it all. Overall, based on her practical experience, Claire-Lise Gaillard draws a sensible picture of this new working environment based on online portals.

Sara Oberbichler and **Eva Pfanzelter** continue the reflection on the impact of mediated access with a chapter devoted to a case study aimed at tracing discourses on return migration in large-scale digital collections of Austrian newspapers. Appraising the unprecedented opportunities introduced by digitisation, in particular the possibility to explore far beyond what manual work could, the authors discuss how to adopt a critical stance when working with digitised newspapers and interfaces and emphasise the importance of digital source, query and interface criticism. On the basis of concrete examples, they examine the possibilities and limitations of keyword search, this ubiquitous functionality provided by all interfaces often used as a first entry point and principal way of building a corpus. The authors show that while keyword search can be a real challenge, due to e.g. word polysemy, synonymy and morphological variation, new functionalities such as keyword suggestions (as implemented in the *impresso* interface and News-Eye demonstrator) and the use of normalised word frequencies can facilitate the exploration and the understanding of a collection. The latter can also help identify which parts of a collection have OCR problems – for instance, if the frequency of a word diminishes during the Great War, it may be due to poor paper quality during that time period. Overall, Sara Oberbichler and Eva Pfanzelter show how data and search engines actually shape what historians can do and the corpus they can build; in response, they demonstrate how heuristics of search can be adapted and call for more documented and transparent tools and interfaces.

In their chapter, **Christoph Hanzig**, **Martin Munke** and **Michael Thoß** address the complex and sometimes controversial question of the National Socialist legacy in memory institutions in Germany. As part of a joint project between the Hannah Arendt Institute for Research on Totalitarianism and the Saxon State and University Library of Dresden, the authors presents the work on the digitisation, semantic indexation and online release of the Saxon Nazi newspaper *Der Freiheitskampf*. In the context of a massive lack of sources for historical studies on national socialism in Saxony (due to war-related destruction), this official organ of the National Socialist German Workers' Party in Saxony published between 1930 and 1945 is of great significance for historical research and education. The authors trace the history of the title as well as the long and meticulous work that was done to assemble the complete collection of the journal's issues. Most importantly, they present the impressive efforts made to manually annotate a selection of about 26,000 articles with entities (persons and locations), thematic categories, as well as articles' summaries where National Socialist expressions and style are explicitly marked. In addition, part of the entities were linked to identifiers of the German National Library authority file (*Gemeinsame Normdatei* GND). The records in this database can be accessed via an online interface which, until recently, did not allow access to facsimiles of the newspaper – only on-site consultation was possible. The issue of access is twofold with, on the one hand, copyright uncertainties and, on the other, the question of how to raise awareness of Nazi propaganda and ideology and prevent the misuse of digitised online content. The authors present the decisions and measures taken in this respect (comments on articles, non-annotation of some persons, educational programs) and discuss further potentialities of the project.

Finally, in the context of contemporary Rwandan history, **François Robinet** and **Rémi Korman** highlight what is still beyond digital reach, with a chapter on the uses of (digital) press collections to write the history of the Tutsi genocide in Rwanda. In the context of their respective research on the memory of the 1994 genocide, the two historians confront their use of the written press – paper and digital, Rwandan and French – published before, during and after the event. Within the general framework of a reflection on the press as a source and as an object of history, the authors report on the difficulties encountered by historians working on the history of Rwanda and discuss the potential benefits of a massive digitisation of the Rwandan written press and what would it take to achieve this. This chapter starts with a review of the history of the Rwandan press and its past and present uses by historians; to varying degrees depending on the period, newspaper archives have always been a central source of historical inquiry into the genocide, whether to consolidate a chronology or describe the evolution of a diplomatic game, or to study the vocabulary, images, and discourses produced in the context of and about the genocide. This historiography also reveals the existence of numerous newspaper collections, which the authors propose to inventory next, focusing in particular on issues of access and digitisation. This mapping of sources shows, in the case of Rwandan press, the incompleteness of the collections, their dispersion in the country and the scarcity of digital access, which makes the collection of sources difficult and time-consuming for the historian. On the basis of these observations, François Robinet and Rémi Korman then question the relevance of a massive collective effort to encourage the use of digital press archives in writing the history of the genocide and, more broadly, of Rwanda.

Considering both the stakes and challenges posed by such enterprise, the authors list the potential benefits of a digitisation effort (in terms of e.g. conservation, accessibility, possibility to combine several sources) but also the political, ethical and technical questions it raises. At the end of this first prospecting of the Eldorado, this contribution reminds us of the incompleteness of the digital newspaper landscape and of the difficulty of the 'step zero' – inventorying, collecting, organising, digitising. In a way, the 'delay' in digitising Rwandan newspaper collections can be seen as an asset, in the sense that it would benefit of lessons learned from previous digitisation campaigns, e.g. around governance and selection of sources to be digitised.

Any historian who has had to go through hundreds or thousands of newspaper issues, leafing through pages or sitting for hours in front of a microfilm reader in order to identify a few relevant articles, knows how difficult it is to work with newspaper archives. Not surprisingly, the contributions of this first part all praise how digitisation considerably pushes back the limits of information retrieval, and the new possibilities it opens. However, they also express cautions about the shape of the digital landscape (incomplete and opaque), point out the far-reaching (and sometimes hidden) consequences of automation, and underline the need to adapt historians' traditional methods to adequately deal with such volumes of primary sources turned into data and unearth new artefacts.

Unearthing New Artefacts: Digital Reshaping of Newspapers

Beyond easier and wider access, digitisation offers the possibility of producing new layers of information for entire collections. This datafication allows for a change in scale in reading this material, as well as automated, data-driven content analyses. This promise is discussed in this second part, where researchers share their proposals for enriching and helping to organise large-scale digitised content.

Taking advantage of the availability of OCR transcriptions for many French daily newspapers of the French National Library, **Pierre-Carl Langlais** proposes to revisit the history of newspaper genre through automatic text classification. Carried out within the framework of the Numapresse project, the large-scale newspaper genre classification approach presented in this chapter is part of the broader context of cultural history and literary analysis. In a first part, Pierre-Carl Langlais presents the design, implementation and evaluation of a supervised text classification approach based on support-vector machines and used to learn three models trained on datasets sampled from different 20-year time periods. The author motivates the choice of a supervised vs. unsupervised approach and details the implementation process, from the iterative definition of genre classes (e.g., reportage, classified ads, international affairs, sports news) to the technical aspects, including the underlying assumptions and choices shaped by historical and cultural analysis needs and considerations. In this regard, the author highlights the specificities of the newspaper as an editorial object where the norms governing content organisation and journalistic writing can be seen as social constructions that develop and consolidate over time. These particularities are taken into consideration in the way the classification models are used with, beyond unary classification, the exploitation of classification probabilities, which allows the study of genre hybridisation based on multi-class labelling, and the exploitation of the probability densities of each class, which allow the study of genre formalism, or how codified or stable a genre is (from its lexical representation). In a second part, the author presents a critical investigation of the results of one model with the aim of uncovering genre patterns. To this end, he proposes a method of 'zoom reading' based on the consideration of various focal distances, based on days and weeks (short-term trends), months and years (seasonal trends), and decades and centuries (long-term trend). The author also experiments with the anachronistic application of a model trained on one dataset (i.e., of a specific time period) to another dataset (i.e., of an earlier time period), allowing for an archaeological investigation of the gradual codification of genres over time.

In the context of public discourse studies, **Melvin Wevers** focuses on a type of content that has so far received little attention, namely advertisements. The author emphasises the importance of historical newspapers for longitudinal studies of public discourse: functioning as 'transceivers', that is, as both producers and messengers of public discourses, newspapers provide access not only to specific views (e.g., those of journalists, editorial staffs, interviewed experts) but also to representations of ideas, values and practices. The same is true for advertisements, which are primarily designed and published for commercial purposes, but which are also vehicles for social and cultural values. Melvin Wevers proposes to study these 'distorted mirrors' – so far mostly studied on the basis of small, manually selected samples – on a macro-scale with the consideration of hundreds of thousands of advertisements in the newspaper collections of the National Library of the Netherlands. His contribution is organised in two parts. First, the author demonstrates the often overlooked value of basic metadata produced during digitisation to better understand the structure and organisation of ads in historical newspapers and unveil advertising trends. Information such as size, number, location on the page and character proportion are extremely valuable hints to categorise and potentially filter out advertisements. Second, on the basis of a case study on cigarette advertisements, Melvin Wevers shows how textual information can be used for the analysis of trends and particularities of ads. A corpus of more than 40,000 cigarette advertisements from 1890 to 1990 is composed and explored in a variety of ways, to study: cigarette advertisements in general, the nationalities of cigarette products, and specific features of cigarettes advertisements in relation to their advertised nationalities. Each time, the author highlights both the opportunities and the limits of the chosen method and illustrates how to switch between perspectives of analysis. Melvin Wevers also discusses the interplay between prior knowledge (original hypothesis), implementation of computational methods and interpretation of results, and emphasises the role of models that, beyond mere prediction, allow to explore and better understand source collections.

After investigating newspaper content from a thematic and journalistic point of view (genre classification with P. C. Langlais) or on the basis of a selected type of content (advertising, with M. Wevers), **Petri Paju, Heli Rantala and Hannu Salmi** illustrate another way of organising and analysing the mass of digitised newspaper content, this time on the basis of large-scale text reuse detection. Yet another purely data-driven process, text reuse corresponds to the meaningful reiteration of text, usually beyond the simple repetition of common language.[17] Text reuse detection can thus help reveal quotations and plagiarism in academic writing, paraphrases and intertextuality phenomena in literary works and, in the case of newspapers, copy-paste journalism, republication of articles and information flows – an area which has received increasing attention in recent years.[18] Drawing on a large corpus of digitised newspapers from the National Library of Finland spanning nearly 150 years (1771–1920), Petri Paju, Heli Rantala and Hannu Salmi investigate what text reuse can reveal about the writing and publishing practices of newspapers across time and space and reflect on the epistemological conditions underlying such examination. Being confronted with masses of text reuse data (i.e., millions of clusters of similar text passages), the authors first examine specific characteristics of text reuse clusters and distinguish between three types of reuse cycles, namely fast, slow, and medium repetition. These cycles, whose definition necessarily impacts how media networks and information flows are conceived, then form the

[17] Matteo Romanello, Aurélien Berra, et al. (2014). *Rethinking Text Reuse as Digital Classicists.* url: https://wiki.digitalclassicist.org/Text_Reuse.

[18] David A. Smith et al. (Sept. 2015). “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.” In: *American Literary History* 27.3. ISSN: 0896–7148. URL: https://doi.org/10.1093/alh/ajv029; Ryan Cordell (Sept. 2015). “Reprinting, Circulation, and the Network Author in Antebellum Newspapers1.” In: *American Literary History* 27.3, pp. 417–445. ISSN: 0896–7148. URL: https://doi.org/10.1093/alh/ajv028; Matteo Romanello and Simon Hengchen (May 2021). *Detecting Text Reuse with Passim.* url: https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim.

basis for the study of historical media phenomena, here in the realm of Nordic press. In a second part, the authors consider the various factors that influence and may bias the results of text reuse detection and analysis, warning against inevitable problems related to digitisation and metadata, but also against the illusion of transnational collections: even digitised, newspapers are still in national silos.

These three contributions demonstrate how the change of time frame and reading scale enables the emergence of new subjects and new lines of research. These studies on the standardisation and circulation of news stories shed new light on the production of information in the 19th and 20th centuries, but also in the present day, by contributing to the historicisation of "new" phenomena such as virality. While these chapters propose ways to partially mitigate the methodological issues associated with OCR and data volume, some shortcomings of computational approaches must be compensated for. To this end, historians can undertake to revisit a long-established tradition: source criticism.

Mining Digitised Newspapers: Source Criticism and the Making of (Digital) History

Against the background of the opportunities and constraints discussed in the first two parts, the contributions of the third part elaborate on research practices in the making of (digital) history. Addressing, among other things, the question of building a research corpus from one or more digitised collections, these chapters integrate the critique of digital sources into their analyses and offer practical research examples linking general issues of using digitised newspapers to specific research topics.

Based on the experience acquired during the *impresso* project and an expert workshop held in January 2020 at the Luxembourg Center for Contemporary and Digital History, **Estelle Bunout** provides an overview of current research practices and the challenges posed by digitised historical newspapers. Three aspects are discussed: the "quietly central" changes brought about by keyword search; understanding the typical content of these collections to make better use of them; and how content-based metadata have helped give more visibility to the structural elements of these collections. Based on a questionnaire of digital source criticism tailored to digitised newspapers and a review of available tools that can inform researchers about newspaper structures, the chapter describes how researchers can, step by step, partially recover the context of such digital sources.

As part of a research on the genesis of financial journalism in Vienna, **Monika Kovarova** presents a study on the popularisation of stock exchange news in

Viennese newspapers (1771–1914) based on the newspaper collections, tools and services of the Austrian National Library. Starting with a traditional keyword search to build a research corpus, the author then experiments with a set of analysis and visualisation tools to investigate various dimensions of the emergence of stock exchange reporting. In the corpus construction phase, Monika Kovarova introduces her workflow with: a description of the Austrian newspaper collection and online portal (ANNO), the presentation of the set of keywords used for the search, the explanation of how the metadata of the corresponding search results were extracted (via web scraping) and, finally, a detailed report and analysis of the search results' profile (e.g., the evolution of hits through time, their distribution per title). In addition to the constitution of a corpus appropriate to her research question, this first phase is also the occasion for the author to inventory some of the limits of digitised newspaper portals such as the "constructed" representativeness of available collections, the influence of the architecture and design of the portal on the research process (interface critique), the impact of linguistic phenomena (polysemy, evolution of meaning and terminology) and the quality of the OCR. In a second phase, the author uses the tools provided by the Austrian library (ONLabs) to perform an exploratory analysis of the previously collected data in a historical communication research perspective. Based on interactive visualisations, this allows the authors to obtain new insights in the popularisation of stock exchange news and the emergence of financial journalism. Overall, by combining research question operationalisation, data collection, digital source criticism, interface critique, visual analytics and historical analysis, this chapter provides a very good overview and illustration of the possibilities offered by digitised newspapers and digital methods.

In her contribution, **Malorie Guilbaud Perez** seeks to trace and analyse the memorial process by which a tragic industrial accident, the fire of a textile manufacturing workshop in New York in March 1911, gradually became part of the American national memory. Trying to reconstruct the diachronic and synchronic circulation of the event over the course of the 20th century, she addresses the challenges of constructing an appropriate research corpus across various collections. The author begins her investigation with the *New York Times*, the famous title published since 1911, digitised by the still existing publisher and available for its subscribers via an online portal (*TimesMachine*). Here Malorie Guilbaud Perez carefully selects, tests and evaluates different set of keywords trying to ensure the best balance between coverage and relevance of results. The obvious obstacles are the polysemy of words and the fact that concepts and proper names can be mentioned in different contexts, but also the lack of segmentation of articles in some parts of the collection, which mixes various unrelated texts together. In order to determine the best set of keywords, the author iteratively tests and carefully documents various combinations of search terms and manually estimates the percentage of relevant result hits. On this basis, she then applies the same search over different newspaper collections, namely *Chronicle America* and *ProQuest*, to expand her initial corpus through time and space. The origins, coverage, benefits and limitations of each portal and collection are presented, as well as dedicated search strategies and how they complement each other. Throughout this process, Malorie Guilbaud Perez shows the complexity of building a relevant research corpus; if the resulting corpus allows her to conduct research with a variety of tools, she emphasises that behind the apparent quantities of documents lie multiple and varied pitfalls and constraints. Overall, this contribution illustrates the central question of how to document and communicate the preparation of research corpora from digitised newspapers; the author provides a neat solution by presenting each step and detailing the results of each decision, for each collection, and comparing them to each other.

Drawing on her experience researching historical newspapers to study changing interpretations of electoral turnouts in Switzerland, France, and Germany after 1945, **Zoe Kergomard** discusses the impact of including digitised newspapers in historical research. In a thorough examination of existing practices and her own, the author examines the challenges and opportunities of using such sources, considering its two facets, that of being a medium, and that of being digitised. The author first reviews the use of the press in general historiography and in the historiography of the media of the 20th century. She notes that historians have always struggled with the place to give to newspapers (digitised or not), with the risk of a "media-centrism" and the need to relate newspapers in their production and reception contexts. She then examines the impact of digitisation with, in addition to the opportunities it opens up (the author's research would not have been possible in an analogue world), the fact that similar issues are reinforced or raised in new ways. Based on her research, the author warns against the danger of viewing digitised newspapers as a proxy for the public sphere and points out the temporal and geographic imbalances in available collections. On the basis of these observations, Zoe Kergomard proposes viewing digitisation as an invitation to multiply the types of sources and perspectives included in a research corpus, and as an incentive for more reflexivity about how historians think about, collect, and analyse their research corpus. Working with political chronicles, press clippings and digitised newspapers accessible via various portals, the author details her journey to build her corpus in an iterative research approach resulting in a heterogeneous and interconnected corpus where digitised newspapers are (almost) "like another source".

In an interesting mise en abyme of reporting practices based on the examination of reports compiled from newspapers and the examination of those same newspapers from which the reports originated, **Suzanna Krivulskaya** addresses the issue of fragmentary evidence from historical sources, digitised or not. The case study underlying this contribution is that of the "runaway reverends" at the turn of the 20th century in the United States, specifically the elopement scandals surrounding Christian ministers that "freethinkers", threatened by a restrictive moral legislation orchestrated by those same clergymen, set out to collect and publicise. The author focuses on a specific publication, *The Crimes of Preachers in the United States and Canada* (1881), which lists alleged crimes of deviance among clergy, all derived from newspapers. Taking advantage of the digital availability of this book, Suzanna Krivulskaya undertook to transform its tabular content into a database. A few technical and methodological challenges later, the author can easily sift through a large set of crime records in digital format, and discovers significant inconsistency and missing information. Is that the fault of the data collectors (the "freethinkers") or the sources themselves, the newspapers? This led the author to confront the information reported by the *Crimes of Preachers* with the content of the contemporary press and examine the reliability of the later in the context of the professionalisation of reporting. Suzanna Krivulskaya details her workflow for tracking elopement scandals in three large U.S. newspaper portals and better understand their media coverage. While problems with OCR quality and spelling variants make it difficult to find information, the author reminds that difficulties also come from the source itself (factual errors, embellished stories) and that newspapers should be treated with critical distance. Paradoxically, the author points out, both the possibilities and shortcomings of digitisation emphasises how troublesome newspapers are as historical sources.

In the context of the emergence of the profession of press correspondents, **Tobias von Waldkirch** analyses the evolution of reporting practices by examining correspondent reports in the *Journal de Genève* during the Crimean War and the Franco-Prussian War (1853–56 and 1870–71 respectively). Switzerland's non-participation in any of these conflicts and the neutral observer status of the Swiss newspaper provide a stable background against which changes in correspondence reporting can be examined as changes in journalistic culture. The author examines several linguistic features and how their use differs between the two time periods, considering, among other things: the first-person singular pronoun (*I/je*), an indicator of personal testimony and interviews when enclosed in quotation marks; the second-person singular pronoun (*you/vous*), an indicator of an address by the reporter to an audience; the grammatical tenses of the verbs of which these pronouns are subjects, an indicator of a perspective related to a current event; and a combination of these features. Tobias von Waldkirch describes the method adopted to build the corpus underlying this investigation, first searching for correspondent reports (often captioned as *correspondances particulières* or variants) via the *impresso* interface, then exporting selected articles with their content and metadata, manually annotating a sample thereof, and using further tools for corpus analysis. Through a fruitful dialogue between quantitative and qualitative analysis, the author is able to identify the central characteristics of the correspondent report genre in the 19th century and to highlight the similarities and differences between two periods. Overall, this chapter illustrates the translation of a research question into measurable indicators, describes the challenges of doing so, and emphasises, once again, the importance of contextualising the results.

Finally, at the intersection of conceptual history, political culture and media history, the study of **Fredrik Norén, Johan Jarlbrink, Alexandra Borg, Erik Edoff and Måns Magnusson** explores the usage of the notion of 'political' in two postwar Swedish newspapers (1945–1989). Based on a combination of text mining techniques applied to a corpus of extracted text blocks containing the term 'political', they attempt to trace how the usage of this notion has evolved over time and in which contexts it has appeared. After a brief overview of the main political and ideological shifts in Sweden from the 1950s to the 1980s and some theoretical perspectives on conceptual history –including the inherent limitation of a study based on what was *explicitly* defined as 'political'– the authors proceed with their investigation using three approaches. The first one is based on the exploration of 'political' bigrams, with an in-depth analysis of their distribution (top and tail elements) and their rank frequency over time. The second one broadens the scope and considers the topics identified in the corpus and their evolution over time. Since topics (generated via topic modelling) are static for the whole corpus, the authors based their diachronic analysis on several topic co-occurrence networks computed for different time windows. They identify three thematic (topic) clusters that emerge and evolve over time, centred around international/foreign affairs, domestic politics, and culture. Finally, the authors carry out a close reading examination of a specific topic over time, the 'women' topic. Overall, this chapter illustrates how text mining techniques, here applied in combination, can help capturing the transformation of a notion as difficult to grasp as the 'political', and also highlights how a better basic processing (OCR, segmentation) could help strengthen finer-grained investigations.

The mythical South American land of Eldorado has never been found. Do digitised newspapers open the way to a vain search for unrealistic promises in historical scholarship? This overview gives a clear answer: the exploration of large collections of machine-readable historical newspapers undoubtedly opens new perspectives and has already led to remarkable results, for example, in understanding the emergence of journalistic genres, in discovering modes of information circulation, or in studying the evolution of certain concepts such as the "political," to name just a few of those presented in this volume. However, if digitised newspapers have something of an Eldorado, it is not an easy one: while all contributors highlight unprecedented opportunities in terms access, search capabilities, temporal and geographic reach, they also point persistent methodological difficulties at several levels, from the nature of newspapers as primary sources to the operation and quality of their digitisation, to interface features and blind spots in the digitised landscape. Coupled with the possibility of applying computational methods to gigabytes of texts and images, this represents a radical transformation of historical newspaper research and of the daily practice of historical inquiry. In response, historians are adapting their methods, drawing on the rich and long-established foundations of source criticism towards the critique of digital sources, tools and interfaces.[19] In the context of different research topics, all authors in this volume emphasise the importance of reflective scholarship when collecting, evaluating and analysing computationally transformed and enriched newspaper archives. In some ways, exploring the Eldorado of digitised newspapers is like walking through the realm of digital history: there are still many unknown areas to explore, and it is an exciting undertaking. Some of the (methodological) paths will be dead ends, others – many? – will lead to new fields of research.

References

Bingham, A. (2010). "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." In: *Twentieth Century British History* 21.2. DOI: 10.1093/tcbh/hwq007.

Boros, Emanuela, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, Jose G. Moreno, Nicolas Sidère, and Antoine Doucet (2020). "Robust Named Entity Recognition and Linking on Historical Multilingual Documents." In: *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. Vol. 2696. Thessaloniki, Greece: CEUR-WS, pp. 1–17. URL: http://ceur-ws.org/Vol-2696/paper_171.pdf.

Bunout, Estelle and Marten Düring (2019). *Collections of Digitised Newspapers as Historical Sources – Parthenos Training*. URL: https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/collections-of-digital-newspapers-as-historical-sources/ (visited on 09/ 02/2022).

Clavert, Frédéric and Andreas Fickers (2021). "On Pyramids, Prisms, and Scalable Reading." In: *Journal of Digital History* 1.1. URL: https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb.

[19] Peter Haber (2011). *Digital Past. Geschichtswissenschaften Im Digitalen Zeitalter*. München: Oldenbourg Wissenschaftsverlag. Frédéric Clavert and Andreas Fickers (2021). "On Pyramids, Prisms, and Scalable Reading." In: *Journal of Digital History* 1.1. URL: https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb.

Clavert, Frédéric and Caroline Muller, eds. (2017). *Le goût de l'archive à l'ère numérique*. https://gout-numerique.net/.

Cordell, Ryan (Sept. 2015). "Reprinting, Circulation, and the Network Author in Antebellum Newspapers1." In: *American Literary History* 27.3, pp. 417–445. ISSN: 0896-7148. URL: https://doi.org/10.1093/alh/ajv028.

Ehrmann, Maud, Estelle Bunout, and Marten Düring (2019). "Historical Newspaper User Interfaces: A Review." In: *Proc. of the 85th IFLA General Conference and Assembly*. Athens, Greece: IFLA Library. DOI: 10.5281/zenodo.3404155. URL: http://infoscience.epfl.ch/record/270246.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet (2022). "Named Entity Recognition and Classification in Historical Documents: A Survey." In: *ACM Computing Survey (accepted)*. URL: https://arxiv.org/abs/2109.11406.

Ehrmann, Maud, Matteo Romanello, Alex Flückiger, and Simon Clematide (2020). "Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Histor- ical Newspapers." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 288–310. DOI: 10.1007/978-3-030-58219-7_21.

Ehrmann, Maud, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide (2022). "Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents." In: *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*. Ed. by Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast. Vol. 3180. CEUR-WS. DOI: 10.5281/zenodo.6979577. URL: http://ceur-ws.org/Vol-3180/paper-83.pdf.

Farge, Arlette, Thomas Scott-Railton, and Natalie Zemon Davis (2013). *The Allure of the Archives*. The Lewis Walpole Series in Eighteenth-Century Culture and History. New Haven: Yale Univ. Press.

Fickers, Andreas (2012). "Towards a new digital historicism? Doing history in the age of abundance." In: *VIEW Journal of European Television History and Culture* 1.1, pp. 19–26.

Fickers, Andreas (2020). "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" In: *Zeithistorische Forschungen* 17.1, pp. 157–168. DOI: https://doi.org/10.14765/zzf.dok-1765.

Haber, Peter (2011). *Digital Past. Geschichtswissenschaften Im Digitalen Zeitalter*. München: Oldenbourg Wissenschaftsverlag.

Jarlbrink, Johan and Pelle Snickars (2017). "Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive." In: *Journal of Documentation* 73.6, pp. 1228–1243.

Kestemont, Mike, Folgert Karsdorp, and Marten Düring (2014). "Mining the Twentieth Century's History from the Time Magazine Corpus." In: *Proc. of the 8th LaTeCH workshop*. ACL. URL: https://aclanthology.org/W14-0609.

Koolen, Marijn, Jasmijn van Gorp, and Jacco van Ossenbruggen (2019). "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice." In: *Digital Scholarship in the Humanities* 34.2, pp. 368–385. DOI: 10.1093/llc/fqy048.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, Find My Past Newspaper Team, and Nello Cristianini (2017). "Content Analysis of 150 Years of British Periodicals." In: *Proceedings of the National Academy of Sciences* 114.4. DOI: 10.1073/pnas.1606380114.

Moreux, Jean-Philippe (2016). “Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment.” en. In: *Proc. of IFLA WLIC 2016*, p. 17. URL: http://library.ifla.org/id/eprint/2076.

Neudecker, Clemens and Apostolos Antonacopoulos (2016). “Making Europe’s Historical Newspapers Searchable.” In: *Proc. of the 12th IAPR Workshop on Document Analysis Systems*. Santorini, Greece: IEEE. DOI: 10.1109/DAS.2016.83.

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann, and Elisa Herrmann (May 2019). “OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents.” In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. DATeCH 2019. New York, NY, USA: Association for Computing Machinery, pp. 53–58. DOI: 10.1145/3322905.3322917.

Nicholson, Bob (2013). “The Digital Turn.” In: *Media History* 19.1. DOI: 10.1080/136888042012.752963.

Oberbichler, Sarah, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen (2021). “Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians.” In: *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.24565.

Putnam, Lara (Apr. 2016). “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast.” In: *The American Historical Review* 121.2. DOI: 10.1093/ahr/121.2.377.

Rigaud, Christophe, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux (2019). “ICDAR 2019 Competition on Post-OCR Text Correction.” In: *ICDAR Proceedings*. Sydney, Australia. URL: https://hal.archives-ouvertes.fr/hal-02304334.

Romanello, Matteo, Aurélien Berra, and Alexandra Trachsel (2014). *Rethinking Text Reuse as Digital Classicists*. URL: https://wiki.digitalclassicist.org/Text_Reuse.

Romanello, Matteo and Simon Hengchen (May 2021). *Detecting Text Reuse with Passim*. URL: https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim.

Smith, David A., Ryan Cordell, and Abby Mullen (Sept. 2015). “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.” In: *American Literary History* 27.3. ISSN: 0896-7148. URL: https://doi.org/10.1093/alh/ ajv029.

Stroeker, Natasha and René Vogels (2012). *Survey Report on Digitisation in European Cultural Heritage Institutions 2012*. Tech. rep. Brussels. URL: https://www.egmus.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf.

van Strien, Daniel, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza (2020). “Assessing the Impact of OCR Quality on Downstream NLP Tasks.” In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. DOI: 10.5220/0009169004840496.

Wevers, Melvin (2019). “Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950–1990.” In: *Proc. of the 1st International Workshop on Computational Approaches to Historical Language Change*. ACL. URL: https://www.aclweb.org/anthology/W19-4712.

Yang, Tze-I., Andrew J. Torget, and Rada Mihalcea (2011). “Topic modeling on historical newspapers.” In: *Proc. of the 5th LaTecH workshop*. ACL, pp. 96–104.