DE GRUYTER
OLDENBOURG

DIGITISED NEWSPAPERS – A NEW ELDORADO FOR HISTORIANS?

REFLECTIONS ON TOOLS, METHODS AND EPISTEMOLOGY

Edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert

STUDIES IN DIGITAL HISTORY AND HERMENEUTICS

DE G

Digitised Newspapers – A New Eldorado for Historians?

Studies in Digital History and Hermeneutics

Edited by
Andreas Fickers, Valérie Schafer, Sean Takats,
and Gerben Zaagsma

Volume 3

Digitised Newspapers – A New Eldorado for Historians?

Reflections on Tools, Methods and Epistemology

Edited by
Estelle Bunout, Maud Ehrmann, and Frédéric Clavert

DE GRUYTER
OLDENBOURG

With the support of the Swiss National Science Foundation for the "*impresso* – Media Monitoring of the Past" project under grant CR-SII5_173719.

Swiss National Science Foundation

ISBN 978-3-11-072971-9
e-ISBN (PDF) 978-3-11-072921-4
e-ISBN (EPUB) 978-3-11-072926-9
ISSN 2629-4540
DOI https://doi.org/10.1515/9783110729214

Library of Congress Control Number: 2022944039

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Cover image: Deauville, baigneuses sur la plage [i.e. vendeurs de journaux sur la promenade], 19-08-1926, Bibliothèque nationale de France.
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Unearthing New Artefacts: Digital Reshaping of Newspapers

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

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 NewsEye 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.

The Allure of Digitised Newspapers: Prospecting the Eldorado

Giorgia Tolfo, Olivia Vane, Kaspar Beelen, Kasra Hosseini,
Jon Lawrence, David Beavan, Katherine McDonough

Hunting for Treasure: Living with Machines and the British Library Newspaper Collection

Abstract: This chapter discusses the open access digitisation programme undertaken by Living with Machines, exploring the range of constraints that inform digitisation strategies and selection priorities. Because the landscape of digitised newspaper collections is so complex, and research and digitisation processes operate on different timelines, we have focused on opportunities to make digitisation choices both transparent and pragmatic. Working towards solutions that reflect collaborations between library staff and scholars, we introduce: a) *Press Picker*, our custom visualisation tool designed to support decision making about digitisation; and b) the *Environmental Scan*, a process of automatic metadata generation from the *Newspaper Press Directories*, a contemporaneous record of British newspapers.

Keywords: digitised newspaper collections, digitisation strategy, research workflows, digital corpus, interdisciplinarity

1 Introduction

In his "Archaeology" of British Library newspapers, Paul Fyfe sets out some material causes that have shaped the composition of the digitised version of

Acknowledgments: Work for this chapter was produced as part of Living with Machines. This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London. This work was also supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. We thank Yann Ryan and the entire British Library Newspapers Collection curatorial team for their guidance, Heritage Made Digital and the Imaging Studio for digitisation guidelines, and all the other members of the LwM team that have contributed to the work presented here.

https://doi.org/10.1515/9783110729214-002

one of the largest national newspaper collections in the world.[1,2] We take that story about the British Library (BL) collection forward and relate it directly to the open access digitisation programme undertaken by Living with Machines (LwM).[3] Rather than working from a question-driven proposal with pre-defined source materials, LwM was designed to be an iterative research project with developing and shifting topics of focus. The project is innovative because it combines digitisation, infrastructure and research, activities which are usually supported by separate funding instruments.

If Fyfe looks backwards at digitisation from the perspective of a historian who could not alter his corpus, we discuss the advantages and challenges of combining digitisation and research workflows. We highlight the tensions between researcher expectations of access to a stable, "ideal" corpus and library digitisation processes and timelines, which necessarily move at their own rhythm. We discuss the range of constraints informing selection priorities, and reflect on how both the heuristics of historical research and the materiality of the content continue to impact one another.

Because the digitised newspaper collection landscape is complex, and research and digitisation processes operate on different timelines, we have focused on opportunities to make choices both transparent and pragmatic. Working towards solutions that reflect collaborations between library staff and project researchers, we introduce: a) our own custom visualisation tool – *Press Picker* – designed to support decision making around digitisation; and b) an overview of the project's digitisation of Mitchell's *Newspaper Press Directories*, a contemporaneous record of British newspapers, which will help shape future BL digitisation priorities by affording a better understanding of existing bias in the digitised corpus.

1 G.T., O.V., and K.B. are the principal contributors to conceptualisation and writing and J.L., and K.M. also contributed to conceptualisation and writing. O.V., K.H. and G.T. developed the *Press Picker* tool. K.B. developed the *Environmental Scan* method. D.B supervised the *Environmental Scan* work.

2 Paul Fyfe (2016). "An Archaeology of Victorian Newspapers." In: *Victorian Periodicals Review* 49.4, pp. 546–577. DOI: 10.1353/vpr.2016.0039. URL: https://muse.jhu.edu/article/644183.

3 LwM is a five-year collaborative project using digital collections and methods to explore lived experiences of industrialisation in nineteenth-century Britain. For more information and to see all the project team members: livingwithmachines.ac.uk and livingwithmachines.ac.uk/team-2/.

2 Related Work

Increasing use of digital newspapers has spurred a small but growing scholarship on the digitisation process and the creation of digital newspaper collections. Such scholarship builds on important early histories of reproduction in libraries and archives,[4] but pays special attention to the specific conditions of digitisation. In a recent article, Hauswedell et al. explore the implicit and explicit selection criteria that shape digital collections of historical newspapers.[5] The authors point out that so far there have not been many in-depth analyses of the processes and motivations influencing inclusions and exclusions in digitised newspaper archives. Similarly Gabriele argues that "the residual layers of policy, practices and politics are utterly invisible in the digital record."[6] Jarlbrink and Snickars show the difficulty in reconstructing the provenance of digital copies created of physical newspapers.[7] Following in the footsteps of Mak,[8] Fyfe exposes the invisibility of corporate histories of digital scholarly resources in the public record. He suggests that the legacy and functionality of such resources depends deeply on "paradata": the "procedural contexts, workflows, and intellectual capital generated by groups throughout a project's life cycle." Paradata, Fyfe argues, might include "commentary, rationale, process notes, and records of decisions about projects recording what its participants chose to include or exclude".[9] Researcher access to paradata depends on early and thorough documentation of digitisation processes and institutional capacity to share this information publicly. They provide information that complement metadata and help acknowledge possible unavoidable bias inherent any selection of digitised material. These can include for

4 Monika Dommann and Sarah Pybus (2019). *Authors and apparatus: a media history of copyright*. In collab. with Ebook Central. Ithaca; London: Cornell University Press. 1 p.

5 Tessa Hauswedell et al. (2020). "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers." In: *Archival Science* 20.2, pp. 139–165. DOI: 10.1007/s10502-020-09332-1.

6 Sandra Gabriele (2003). "Transfiguring the newspaper: from paper to microfilm to database." In: *A modern 2: Network Archaeology*. URL: https://amodern.net/article/transfiguring-the-newspaper/.

7 Johan Jarlbrink and Pelle Snickars (2017). "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive." In: *Journal of Documentation* 73.6.

8 Bonnie Mak (2014). "Archaeology of a digitization." In: *Journal of the Association for Information Science and Technology* 65.8, pp. 1515–1526. DOI: 10.1002/asi.23061. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23061.

9 Paul Fyfe (2016). "An Archaeology of Victorian Newspapers." In: *Victorian Periodicals Review* 49.4, pp. 546–577. DOI: 10.1353/vpr.2016.0039. URL: https://muse.jhu.edu/article/644183, p. 550.

example the software used to produce the Optical Character Recognition (OCR), the timestamp or the assessment of its quality, or they could be related to selection criteria, like – in the case of newspapers – the preference for a complete run to a one full of gaps, a geographic area of coverage or the format of the source (microfilm or hard copy).

This literature argues that the pragmatic decisions shaping the digital corpora available to scholars have in turn profoundly influenced not just research practices (who can resist a digital word search?), but also what is actually consulted in the first place (although often the use of the digitised version of a source goes unacknowledged).[10] When scholars can secure access to more than 33 million pages of the digitised press via the online *British Newspaper Archive* it is all too easy to forget that this represents only about 6% of newspapers in the British Library collections. While the British Library has been a legal deposit library since its inception,[11] its holdings represent an unknown percentage of what was originally printed across Great Britain. Inclusion and exclusion from the digital corpus is not random. It is also easy to forget that uneven transcription quality resulting from processes such as optical character recognition means that an unknown sub-sample of those 33 million pages will not in fact be reliably searchable at the word[12]. Mussell and Smits carefully navigate both the pitfalls of using two versions of the digitised British Library newspaper collection and the challenge of working with commercial online resources (ones

10 Tim Hitchcock (2011). "Academic History Writing and its Disconnects." In: *Journal of Digital Humanities* 1; Ian Milligan (2013). "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." In: *The Canadian Historical Review* 94.4, pp. 540–569; Jon Giullian (2013). "« "Seans Chernoi Magii Na Taganke": The Hunt for Master and Margarita in the Pravda Digital Archive »." In: *Slavic East European Information Resources* 14.2–3, pp. 102–26. DOI: 10.1080/15228886.2013.813374; Richard Abel (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1–2, pp. 1–10. DOI: 10.2979/filmhistory.25.1–2.1.

11 There is a large literature on copyright history including the provision for legal deposit in guild, royal, and university libraries. See especially Will Slauter (2019). *Who Owns the News?: A History of Copyright*. Google-Books-ID: FZSFDwAAQBAJ. Stanford University Press. 488 pp. and Edmund King (2005). "Digitisation of Newspapers at the British Library." In: The *Serials Librarian* 49.1, pp. 165–181. DOI: 10.1300/J123v49n01_07. URL: https://doi.org/10.1300/J123v49n01_07.

12 Guillaume Chiron et al. (2017). "Impact of OCR errors on the use of digital libraries: towards a better access to information." In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–4; Mark J Hill and Simon Hengchen (2019). "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study." In: *Digital Scholarship in the Humanities* 34.4, pp. 825–843; Daniel van Strien et al. (2020). "Assessing the Impact of OCR Quality on Downstream NLP Tasks." In: *ICAART (1)*, pp. 484–496.

which are not always, or only, geared towards academic research).[13] Building on this work, one aim of LwM is to enable scholars to understand the composition of a digital corpus, as well as to make informed decisions about mitigating inevitable exclusions.

3 Negotiating Digitisation and Research Workflows

LwM is privileged when it comes to digitisation. The project has a reasonable budget to invest in digitisation and acquisition of digitally available historical sources. In addition, one of the project's partners is the British Library, whose British newspaper collection is one of the largest national newspaper collections in the world. Even from such a favourable starting point, however, there were a number of difficulties with digitisation planning.

Setting up digitisation workflows, protocols, and copyright assessment procedures, as well as defining services, requirements, and quality control procedures is a complicated, costly and lengthy process. LwM benefited from best practices already set in place by previous and in-progress large-scale newspaper digitisation projects at the British Library: the JISC-funded digitisation programme[14] and, most recently, Heritage Made Digital.[15] These projects established precedents for workflow and copyright management and represent a benchmark when setting expectations and scoping potential risks.

There are some key differences, however, between these digitisation projects and LwM. Crucially, they were not driven by research questions, but by curatorial practices, resulting in different selection constraints. These pre-established

13 James Mussell (2014). "Elemental Forms." In: *Media History* 20.1. Publisher: Routledge _eprint: https://doi.org/10.1080/13688804.2014.880264, pp. 4–20. DOI: 10.1080/13688804.2014.880264. URL: https://doi.org/10.1080/13688804.2014.880264; Thomas Smits (2016). "Making the News National: Using Digitized Newspapers to Study the Distribution of the Queen's Speech by W. H. Smith & Son, 1846–1858." In: *Victorian Periodicals Review* 49.4, pp. 598–625. DOI: 10.1353/vpr.2016.0041. URL: https://muse.jhu.edu/article/644185.

14 *10 Billion Words: The British Library British Newspapers 1800–1900 Project. Some Guidelines for Large-Scale Newspaper Digitisation* (2005). URL: http://www.webarchive.org.uk/wayback/archive/20140615090156/http://www.jisc.ac.uk/uploaded_documents/IFLA_2005.pdf.

15 See Jane Shaw (2009). *British Newspapers 1620–1900: Final Report. J.* JISC and the British Library. URL: http://www.webarchive.org.uk/wayback/%20archive/20140614080134/http://www.jisc.ac.uk/media/documents/programmes/digitisation/blfinal.pdf and the blog post: blogs.bl.uk/thenewsroom/2019/01/heritage-made-digitalthe-newspapers.html.

workflows do not leave flexibility for mid-project alterations and are demanding in terms of rigour and precision during the planning stage.

If digitisation is done internally, as in the case of the British Library Imaging Studios, any new work request must fit within an existing, institution-wide delivery plan. This means that timeframes and resources must be planned in advance and allocated carefully. Sources must be selected with enough time to ensure that copyright can be assessed. In some cases, a conservation assessment might need to be performed. This becomes more difficult as the scale of the work requests increases. At the outset, the Imaging Studio needs a list of titles, items, source type, and number of pages.

This, however, is at odds with document selection flexibility preferred by a research-driven project. Scholars ideally want to select titles iteratively throughout the project as questions are refined and new directions emerge or as the factors which contribute to a balanced corpus are refined. LwM did not wish to commit to a complete selection at the start of the process, allowing ongoing research to drive subsequent selections. This flexibility complicated planning the Imaging Studio work because many variables like page or volume format could not be predicted. Such predictions are essential because they determine how much time must be allocated for specific equipment and studio staff.

Normally, scoping collections to digitise would be performed before scholarly research starts, but for LwM this could only start once the project received funding. Our uncommon funding design led to digitisation planning and setup commencing simultaneously with researchers joining the project. On the positive side, instead of expecting researchers to define digitisation needs before funds have been released, LwM enables researchers to devote time to this while they are funded project participants. This allows an alternative approach to preliminary workload expectations – like data acquisition – that often are invisible in grant-funded projects and therefore challenging for any staff who are precariously employed. However, this equitable workload provision complicates project planning. It takes time to process acquired data to make it useful for research applications. In order to make data available well before the grant ends, digitisation must be complete even earlier. Such restrictions mean that we have had to increase the monthly data delivery in order to spend the available budget. This raises more workflow questions and puts yet more pressure on the studios.

Another digitisation aim in LwM was to rebalance bias in the existing digitised corpus, for example in geographical coverage or political leaning. But addressing such issues of representativeness requires analysis of the existing corpus. Digitising “blindly” (without having analysed the existing corpus), on the other hand, risks selecting titles that may actually deepen existing social, cultural or geographical biases.

As we can see, the pace at which digitisation and research move are substantially different and attempting alignment, rather than staggering work, can generate a more complex and extended timeline. In LwM, negotiation between these two sets of priorities has been the way forward. Negotiations between researchers and the studio (who typically do not interact) have been challenging but informative. In this chapter, we highlight key considerations for research-driven digitisation for future projects.

To reach an agreement between research and digitisation priorities we have had to make some choices. We opted for a combined selection model that would bring forward the initial selection of titles to digitise ("frontloading"). We could then progress through an iterative process of selection informed by developing research. Simultaneously, we intentionally chose to document requirements, choices, and tools developed in LwM using paradata. Part of this process includes referring to paradata as thoroughly as possible in our outputs, of which this chapter is an example.

LwM creates paradata in two key areas. First, the *Press Picker* tool consolidates newspaper metadata, explicit selection criteria, and digitisation planning needs. It enables quick selection by subject experts informed by preliminary research interests and collection availability. Second, digitising the *Newspaper Press Directories* improves selection and contributes to the reduction of selection bias.[16] Our work sets the stage for future projects to use both tools in negotiations between librarians and researchers. We demonstrate how enriching collection metadata and making it available to researchers to improve discovery and corpus creation is a key element in the preliminary stages of digital humanities research.

4 Selection Criteria

Our selection criteria for newspapers to digitise in LwM were shaped by a mix of research-led and practical factors. (Criteria were determined with input from multiple parties: academics – mostly historians –, newspaper collection curators and cataloguers, copyright experts, and conservation experts. Digitisation practices, timeframes and costs were provided by imaging technicians). Research-led criteria were both temporal (choosing titles within with the project's time period

16 For more information about the Newspaper Press Directories, see Susan Gliserman (1969). "Mitchell's "Newspaper Press Directory": 1846–1907." In: *Victorian Periodicals Newsletter* 4, pp. 10–29 and section 5 below.

[1780–1918]) and geographic (prioritising industrial areas that appear to be underrepresented in current digital corpora). A newspaper 'title' here means, for example, *The Times* or the *Blackpool Herald*. A single title series has potentially many variants (unlike *The Times*, local newspapers frequently modified their masthead titles). We preferred titles with longer publication runs as possible evidence of sustained readership, and titles published weekly over those published daily so that our digitisation budget could be spread over a wider range of localities currently under-represented in the digital archive. In addition, we prioritised early newspapers not previously digitised, and titles known to have been aimed specifically at a plebeian readership (the aim here was not just to redress the class bias inherent in the nineteenth-century press, but also to create a 'plebeian' sub-corpus large enough to enable separate analysis).

There were also practical criteria. As a time-limited project, it was important that the digitisation was completed quickly enough for us to use. This translated into a preference for digitising microfilms. Digitising microfilm avoids potential delays due to paper conservation and is also cheaper than working from hard copies, allowing us to maximise the number of newspapers digitised with our budget. The British Library has microfilmed a significant number of newspapers.[17] This microfilming, however, proceeded unevenly across the country and sometimes patchily across a title's issues (some but not all may be microfilmed). Additionally, some early microfilming at the BL was done on acetate, a material that has since degraded. This makes it unsuitable for digitisation and so these reels needed to be avoided. Our intention was, on picking a title, to digitise every page of every issue held for completeness. It was necessary, therefore, to be able to select titles where a majority of issues were available as non-acetate microfilm to maximise both speed and economy.

The task of selecting these newspapers was too complex for BL discovery systems. We therefore developed a data-driven approach. To begin, we used specific criteria to narrow down a definitive dataset of BL newspaper titles[18] For the geographic focus, in consultation with the historians on LwM, we chose six counties across Great Britain: Lancashire, Warwickshire, Yorkshire, Glamorgan, Lanarkshire, and, as a non-industrial control, Dorset. With shifting county borders over time and inconsistent granularity in the geographic metadata, this filtering was implemented to some extent manually by compiling a list of places or

17 https://www.bl.uk/collection-guides/newspapers# accessed 09/03/2020.

18 'British and Irish Newspapers', 2019, https://bl.iro.bl.uk/work/7da47fac-a759-49e2-a95a-26d49004eba8 accessed 09/03/2020. Dataset credit: Contemporary British (British Library), Collections Metadata (British Library).

administrative jurisdictions in the metadata to be included or excluded.[19] This allowed us to be more faithful to historic county boundaries (had we not done this important population centres such as Liverpool, Manchester, Middlesbrough, Barrow-in-Furness, and Birmingham would have been excluded from consideration). Narrowing down the list to undigitised (and not due to be digitised by other projects like Heritage Made Digital) titles overlapping 1780–1918 and from the six chosen counties left approximately 2,500 titles. Further conversations with image technicians, cataloguers, curators and archivists helped us decide to make an initial selection of approximately 400 items[20] from these, fitting with our budget and frontloading digitisation within the project. The number of items per title varies; we aimed to select roughly 50 titles from the 2,500 to align with this number. How could we make the selection?

5 *Press Picker*

Our solution is a custom tool – *Press Picker* – for overviewing BL newspaper holdings over time, foregrounding those factors most important for making a selection. *Press Picker* consists of two Jupyter notebooks[21] (a graphical user interface environment for writing and running computer code alongside outputs and narrative). The first notebook performs data filtering and processing. The second notebook produces an interactive visualisation of the processed data, which can be used to select titles. Digitising cultural heritage collections has opened up new opportunities for their analysis, exploration and presentation through data visualisation.[22] Visualisation, for example, can help cultural institutions to administer and

19 Recent work led by Yann Ryan continues to explore advanced geographic metadata analysis, including filtering by historic counties. See Yann Ryan et al. (2020). "Using smart annotations to map the geography of newspapers." In: DH2020. URL: https://www.conftool.pro/dh2020/index.php?page=showAbstract&form_id=532&show_abstract=1.

20 An 'item' is equal to one hardcopy bound volume or 100ft microfilm reel, each of these covering usually one year of publication. The number of issues (frequency of publication) per year and pages per issue can vary and is not usually recorded in the library catalogue. A 'title', on the contrary, refers to the unique title under which a sequence of issues were published. This chronological sequence is known as a 'run'. Sometimes, a hardbound volume contains issues from multiple titles bound in alongside each other. Titles can also change over time. A complete run often includes the various permutations of the title.

21 https://jupyter.org/.

22 F. Windhager et al. (2018). "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges." In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*; Olivia Vane (2019). "Timeline design for visualising cultural heritage data." PhD

manage collections.[23] Bringing this full circle, *Press Picker* is a data visualisation tool for digitisation selection that links researchers to curatorial decisions.

The *Press Picker* notebooks use the BL newspaper title list[24] and additional datasets detailing the hard copy and microfilm holdings for each title (retrieved internally at the BL). Since we were making our selection over a series of rounds, we also designed the tool so that a list of previous selections can be loaded in and these titles excluded in subsequent use.

The visualisation could be described as a branching sparklines design.[25] Each newspaper title is represented as a small line graph with time running horizontally (see Figure 1). The title's 'general area of coverage' (a metadata field containing content such as 'Merseyside') and its name (eg. *Liverpool Telegraph and Shipping Gazette*) is shown at the left side of each line (see Figure 2). The line graph has two lines: red for microfilms and black/dashed for hard copies. For each, the line height represents the number of records per year. (In footnote 4 we expand on what a record means in terms of the items they represent for these different cases). The total hard copy records count (in black) and microfilm records count (in red) for 1780–1918 are shown as numbers above the title name. A numeric code at the far left (hyperlinked to the title's BL catalogue record online) is the title's ID in BL systems. If some of the title's microfilms are likely on acetate (i.e., the degraded material that cannot be easily digitised), a warning icon (a red circle containing an exclamation mark) is shown next to the microfilm count. Hovering over this icon with the cursor reveals the number of likely acetate records in a tooltip.[26] The data does not explicitly state which are acetate. Based on advice from BL staff, we inferred this from the record's

thesis. Royal College of Art, London. URL: http://researchonline.rca.ac.uk/4325/1/TimelineDesignForVisualisingCulturalHeritageData_OliviaVane_redacted. pdf, pp. 11–16.

23 Jefferson Bailey and Lily Pregill (2014). "Speak to the Eyes: The History and Practice of Information Visualization." In: *Art Documentation: Journal of the Art Libraries Society of North America* 33.2, pp. 168–191. DOI: 10.1086/678525. eprint: https://doi.org/10.1086/678525. URL: https://doi.org/10.1086/678525.

24 'British and Irish Newspapers', 2019, https://bl.iro.bl.uk/work/7da47fac-a759-49e2-a95a-26d49004eba8 accessed 09/03/2020. Dataset credit: BL Contemporary British and BL Collections Metadata.

25 Sparklines are very small charts/graphics (often line graphs), typically drawn without axes, presenting some measurement in a simple and highly condensed way. (Edward Tufte [2006]. *Beautiful Evidence*. Graphics Pr) It is a way of achieving high density in data display. In our design, the sparklines are sometimes branched on the left, indicating where newspaper titles are connected through name changes over time.

26 A tooltip is a message that appears when a cursor is positioned over an element in a graphical user interface.

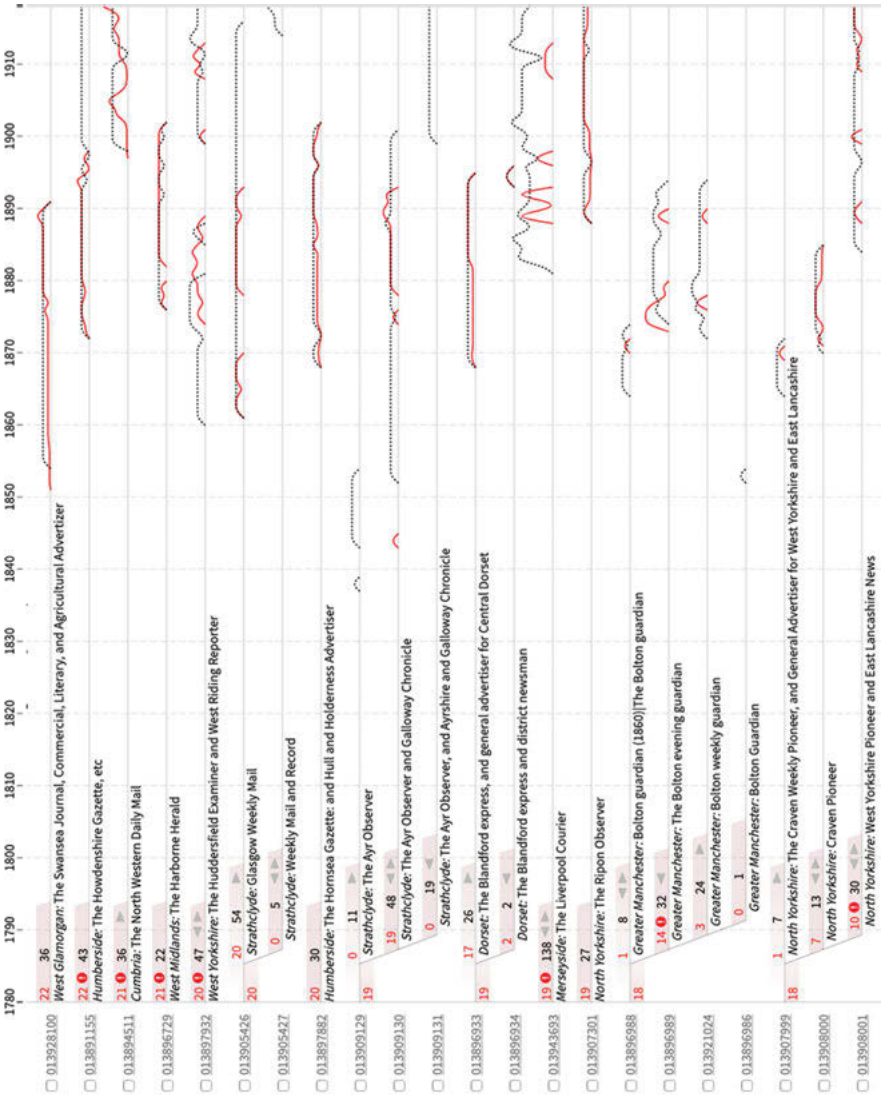


Fig. 1: *Press Picker* visualisation: users can scroll vertically through the titles.

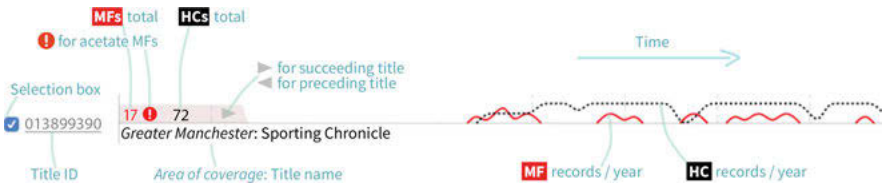


Fig. 2: How to read the *Press Picker* visualisation.

can number.[27] However, we cannot guarantee that this will identify all the acetate films.

A single record for hard copies means something different than it does for microfilms (see Figure 3 and Footnote 9). Hard copy and microfilm records are not directly comparable, therefore, in terms of newspaper 'volume' (they do not directly correspond to measures we might use for capturing the 'amount' of newspaper, e.g. text length, number of pages or number of issues). All we have access to is the number of records in the BL system. This is a challenge for reading the visualisation, but there is no more precise data alternative.

Fig. 3: Hardbound volumes of newspapers in the BL collection. Photos by Luke McKernan. (CC BY-SA 2.0).

The time data attached to each record was either a single year or a date span of varying granularity. We converted these to time series data for each title, normalising it to the number of records per year. Where a single record covers multiple years (eg. 1834–1837), we represent this in the line graph as a line of stable height, where the height is 1(years covered). For a single record covering 1834–1837, the

27 Microfilm cans were numbered sequentially during production, meaning we can assume numbers below a threshold were created earlier and are on acetate.

line graph displays a stable line between 1834 and 1837 of height ¼. Because of these kinds of inconsistencies in the metadata, we were comfortable smoothing the line graph curve to increase readability over the dense data display.

5.1 Branching Design to Represent Title Name Changes

There is further complexity in the newspaper title data because titles can experience incorporations, amalgamations, and name changes through time. For example, *The Athletic Reporter* in 1886 becomes *The Reporter*, which in 1888 becomes *The Midland Counties Reporter and General Advertiser*, which in 1889 becomes *The Reporter and General Advertiser*, which in 1890 becomes *The Coventry Reporter and General Advertiser*. In the BL title dataset, each subsequent publication name is considered a new and separate title. The connections between titles are recorded under two facets: 'preceding title' and 'succeeding title' (the nature of the connection, eg. amalgamation, is not made clear.) Name changes are fairly common in this collection. Our intention was to treat connected titles as a group, digitising them together. These connections, therefore, needed to be apparent in the visualisation.

The metadata fields 'preceding title' and 'succeeding title' contain a title name and, sometimes, issue numbers and dates, all in free text format. In order to identify links across the dataset, we used string cleaning with Regular Expressions to find matching title names. To help distinguish between titles where the name is more generic (eg. *Daily News*), we only connected records with a matching 'general area of coverage' data facet. An informal inspection suggests this approach is successful for the vast majority of titles, so we did not try more sophisticated string matching techniques. Creating these links in the data meant we could visualise the connections.

Where a number of titles are connected via 'preceding' and 'succeeding title', we use a branching design to connect the line graphs (see Figure 4). The total number of microfilm records across connected titles is shown as a number at the stem of the branches. The *Press Picker* tool orders all the line graphs vertically by microfilm count (connected titles are treated as a group). Customising our visualisation design meant that we could communicate these kinds of important complexities in the newspaper data.

While the linking technique was largely successful, we made sure the underlying data could still be inspected for checking. If a title has a 'preceding' and/or 'succeeding title' entry, small grey arrow icons (backwards pointing for preceding,

Fig. 4: How to read the *Press Picker* visualisation: connected titles.

forwards pointing for succeeding) are displayed above the title name. Hovering over an arrow icon reveals the full metadata entry from the original dataset as a tooltip.[28]

5.2 Using *Press Picker* to Select Titles

The visualisation is embedded, as part of the data processing, inspecting, and selecting workflow, within the second Jupyter notebook (see Figure 5). Titles can be 'picked' using selection boxes to the line graphs' left. The associated hard copy and microfilm records can be viewed using print commands in the subsequent notebook cells.

Once the user is happy with their selection, the final cell in the notebook can be used to export the selected titles' records in .csv format.

Press Picker is a powerful tool for surveying the BL's newspaper holdings. But some knowledge about the historical context of newspapers simply isn't recorded in BL metadata: identifying key early publications, for example, required additional period expertise to complement what information was made visible via the tool. Our title selection, therefore, was performed by a historian from the LwM team. Selections have been successfully made and have progressed to digitisation.

While the motivation for creating *Press Picker* was to support making newspaper selections within LwM, we are also exploring its potential to support managing newspaper metadata internally at the BL, and for BL readers interested in exploring the library's newspaper holdings, particularly for queries that would benefit from incorporating digitisation paradata. It may also be of interest to other libraries with significant newspaper holdings.

28 A tooltip is a message which appears when a cursor is positioned over an element in a graphical user interface.

List selected microfilms

In the following cell, some info about the selected microfilms will be shown.

In [20]:
```
if type(selected_IDs_list) == str:
    selected_IDs_list = selected_IDs_list.split(",")

selected_titles_microfilm = records_mf[records_mf['Title.ID'].isin(selected_IDs_list)]
print("#Selected unique titles: %s" % len(selected_IDs_list))
print("#Selected microfilms: {}".format(selected_titles_microfilm.shape[0]))
selected_titles_microfilm
```

```
#Selected unique titles: 2
#Selected microfilms: 5
```

Out[20]:

	Title.ID	Publication title	edition	locale	canNo	startReel	endReel	startDate	endDate	duplicate	LastModifiedOn	NewspaperItemID	TitleItemID	HoldingItemID
8650	013901125	The Wimborne Journal and East Dorset Advertiser	NaN	LondonProvincial	8659.0	03	03	1869 Jul 3	1870 Dec 06	false	2009-11-30 17:17:36.963000	029-000233926	030-000187097	031-000187099
8651	013901125	The Wimborne Journal and East Dorset Advertiser	NaN	LondonProvincial	8659.0	04	04	1871	1872	false	2008-06-26 17:34:03.440000	029-000233926	030-000187097	031-000187100
8652	013901125	The Wimborne Journal and East Dorset Advertiser	NaN	LondonProvincial	8659.0	05	05	1873	NaN	false	2008-06-26 17:34:03.440000	029-000233926	030-000187097	031-000187101
10818	013901126	The Dorset Free Press and Wimborne & South Wes...	NaN	LondonProvincial	8659.0	06	06	1874 Jan 0	1875 Mar 12	false	2009-11-30 17:21:22.047000	029-000233926	030-000129549	031-000187102
10819	013901126	The Dorset Free Press and Wimborne & South Wes...	NaN	LondonProvincial	10838.0	014	014	1875 Jan 0	1875 Mar 12	true	2009-11-30 17:20:33.033000	029-000233926	030-000129549	031-000187103

Fig. 5: Detail from *Press Picker* Jupyter notebook, showing a selection made in the visualisation and the microfilm records from that selection displayed below.

6 Generating Metadata from Contextual Resources

Press Picker provides a bird's eye view on the newspapers as held and catalogued by the BL. While it records a newspaper's genealogy and publication timeline, the metadata remains scarce, especially when it comes to identifying producers and audiences. In this section we describe another strategy for improving collaborative decision-making about digitising historical newspapers, one that draws upon contextual resources to extract newspaper metadata. It takes a sketch of the newspaper 'landscape' – i.e. the contours of all newspapers that (we assume) circulated in Britain during the second half of the nineteenth century – as a starting point, from which it derives future digitisation priorities.

This approach, which we dubbed the *Environmental Scan*, combines research and digitisation workflows, with the goal of rebalancing digital archives by reducing their latent bias. Below, we critically discuss the theoretical and practical issues that arise from such an undertaking.

6.1 The Bias of Big Data

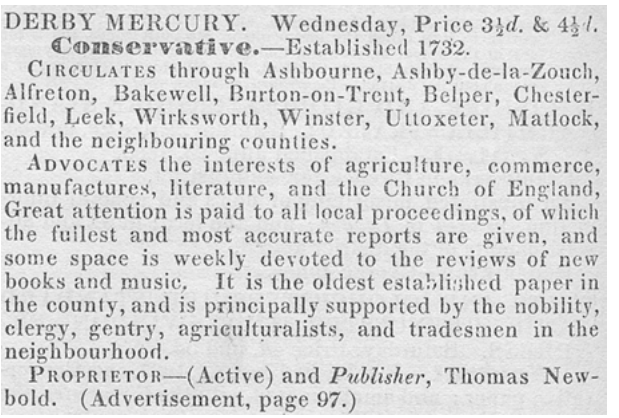
Even though digital newspaper collections have expanded considerably in recent years, they still constitute a small portion of all the newspapers that have existed. In the case of Britain, various initiatives have altogether digitised only around 6% of estimated number of printed newspaper pages. This may be large in absolute counts, but small in relative terms. More problematic is that we simply do not know how representative this digital sample is of the newspaper landscape because we lack (a) relevant contextual information for individual newspapers and (b) an accurate description of the population of newspapers (e.g. a sort of census of the newspaper landscape).

Given the current speed at which digitisation proceeds, this situation will not change anytime soon – it will take decades to have all newspapers digitised. However, instead of waiting and wailing, this section shows how we can mitigate the issue of bias by ensuring that digitisation is informed by contextual knowledge; by an understanding of the boundaries and the gaps of the digital archive. Below, we discuss how we can correct some of these imbalances, especially in relation to the social, political and cultural dimensions of the Victorian press – i.e. the various commercial stakeholders, political and economic interest groups, and local or national audiences, all of whom influenced the shape and content of the newspaper landscape, but are currently absent in the existing metadata.

6.2 Deriving Metadata from Contextual Resources

To repopulate the landscape of the Victorian press we turned to contemporaneous sources of information that offer glimpses into the circulation and consumption of periodicals beyond the metadata currently available in the BL catalogue. More specifically, we systematically interrogate the Newspaper Press Directories, which were published almost annually by Charles Mitchell from 1846 onwards. Initially intended to keep a more "dignified and permanent" record of the press, the directories emerged as an authoritative list of London and "Provincial" newspapers.[29]

In addition to these noble intentions, the directories also served the more narrow commercial interests of proprietors and advertisers. The catalogue gave proprietors space to profile their publication and helped advertisers (who wanted to buy space in those newspapers) find the right audience. Newspapers circulating in London or in the provinces (and beyond, in the later editions) were often described in great detail, with special attention given to the interests and orientations of the producers and their audiences.



DERBY MERCURY. Wednesday, Price 3½*d.* & 4½*d.*
Conservative.—Established 1732.
CIRCULATES through Ashbourne, Ashby-de-la-Zouch, Alfreton, Bakewell, Burton-on-Trent, Belper, Chesterfield, Leek, Wirksworth, Winster, Uttoxeter, Matlock, and the neighbouring counties.
ADVOCATES the interests of agriculture, commerce, manufactures, literature, and the Church of England. Great attention is paid to all local proceedings, of which the fullest and most accurate reports are given, and some space is weekly devoted to the reviews of new books and music. It is the oldest established paper in the county, and is principally supported by the nobility, clergy, gentry, agriculturalists, and tradesmen in the neighbourhood.
PROPRIETOR—(Active) and *Publisher*, Thomas Newbold. (Advertisement, page 97.)

Fig. 6: Description of the *Derby Mercury* (Mitchell's 1857).

29 Susan Gliserman (1969). "Mitchell's "Newspaper Press Directory": 1846–1907." In: *Victorian Periodicals Newsletter* 4, pp. 10–29.

What information do the directories contain, exactly? Figure 6 shows an example from Mitchell's 1857 edition. This entry describes the *Derby Mercury*, listing the day of publication, the price (range) and the political leaning, followed by the date the newspaper was established. These aspects are systematically added for each newspaper and adhere to a regular pattern. An even more detailed profile is recorded in the 'free text' field (to project the language of contemporary databases back to the nineteenth century). Here we encounter an enumeration of the places in which the newspaper reportedly circulated, but also some qualitative pointers to who may be reading or supporting the *Derby Mercury* (in this case the local nobility, clergy, gentry, agriculturists and tradesman), linked to the interests it promoted (agriculture, commerce, etc.). These descriptions mostly end with a mention of the proprietor and/or publisher, with their occupations.

In LwM we use these press directories as a source to automatically generate metadata[30] for the digitised *as well as* the non-digitised newspapers, which will inform future digitisation priorities wherever possible. The figures below show a few concrete examples of opportunities created by linking digital newspapers to the contextual information derived from the directories. In this initial phase we only included newspapers from two counties, Lancashire and Dorset, as these show widely divergent trends in their industrialisation process: while the former industrialised early, the latter remained largely rural, only changing much later in the nineteenth century.

Figure 7 compares places of publication (as recorded in the BL newspaper title list) to those where newspapers circulated (as mentioned by the directories). The latter allows a more complex and regional analysis of the press, away from the traditional centres where the newspapers were printed.

Besides the more fine-grained spatial metadata, the directories' complex categorisation provides other lenses for rediscovering the Victorian press. Figure 8 compares, for example, the distribution of political labels (such as liberal and conservative, as recorded by Mitchell) between the digital sample and the newspaper landscape. It suggests that current digital newspapers (for the two counties) are permeated by a liberal bias, with a substantial under-representation of the conservative press. Insights such as these – derived by incorporating contextual resources – can then serve as guidelines to define future digitisation priorities.

The directories contain elaborate profiles for the newspapers in our collection, and allow us to sketch, roughly, the contours of the newspaper landscape. However,

30 Kaspar Beelen, Jon Lawrence, Daniel C.S. Wilson and David Beavan, 'Bias and Representativeness in Digitized Newspapers: Introducing the Environmental Scan,' Digital Scholarship in the Humanities Advance Access DOI: 10.1093/llc/fqac037.

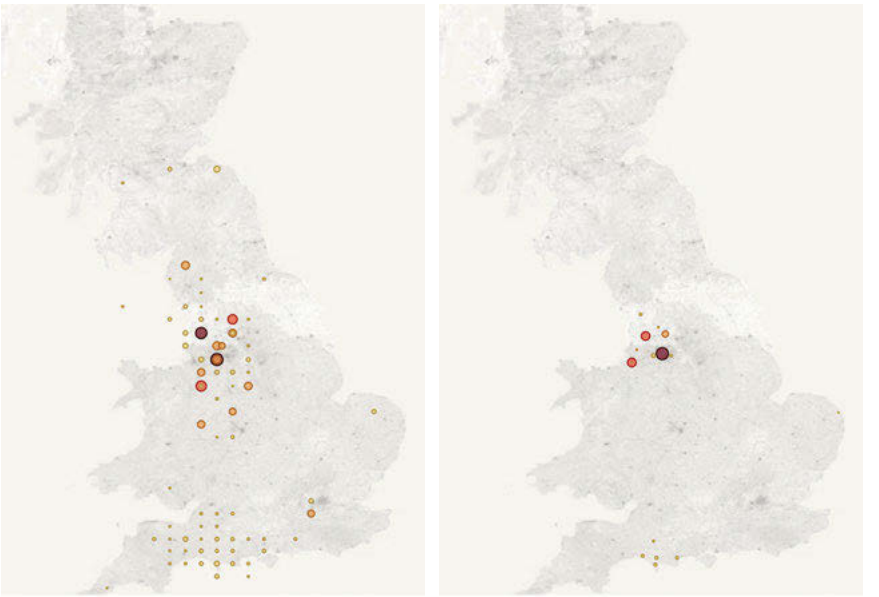


Fig. 7: Newspapers by place of circulation (left) and place of publication (right). The size and the colour of the individual circles indicate the number of newspaper pages published (or circulating) in an area (Lancashire and Dorset titles only).

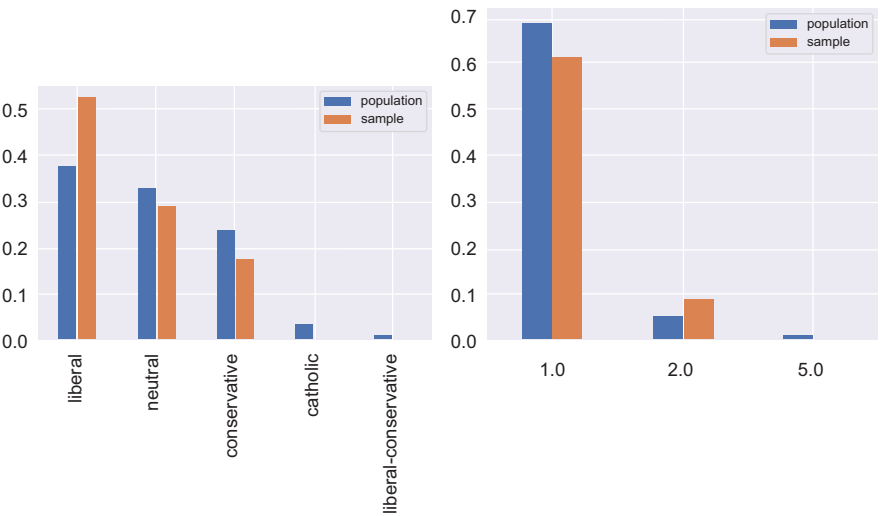


Fig. 8: Comparing the distribution of political labels (left) and prices in pennies (right) between the digital sample and the population of newspaper.

some critical caveats are in place here: while valuable, the directories have to be handled carefully, and can't be treated as an unproblematic source for metadata.

The nineteenth century was an age of classification. Mitchell's attempt to bring order to the often inchoate reality of the periodical press springs from a tradition that produced knowledge by the means of categorisation.[31] This does not mean that Mitchell's classification of the Victorian press was stable, nor that its categories (e.g. liberal and conservative) were unproblematic. O'Malley, for example, gives a detailed account on how the notion of "a national press" was articulated by the directories, pointing out how the directories were a constituent force that reflected and shaped the press landscape. By producing a comprehensive record, Mitchell was "imposing a typology on the industry through its classifications". His attempt to make sense entails some performativity.

7 Conclusions

LwM is still in progress, but we can begin to evaluate work to date. Conducting digitisation as part of a digital humanities project is valuable as, in theory, it provides an opportunity to select sources in parallel with developing research. The reality, however, has proven to be more complicated. Attuning the pace of digitisation practices to that of research inevitably requires negotiation and compromise. The bigger the scale of digitisation, the less flexibility is possible. Planning and flexibility are two (often) opposing project management styles which LwM has had to move between, sometimes leaning towards one, sometimes the other. Digitisation requires early decisions followed by a long setup process. Research, meanwhile, follows a more iterative, interactive course where staggering digitisation decisions enables building on preliminary findings. In this chapter, we introduced innovative solutions developed by the LwM team representing our negotiations: the *Press Picker* and the *Environmental Scan*. Both are helpful tools for reconciling competing demands. They generate paradata that monitors the creation of the digital corpus and records the publishing context at the time when titles were printed. Even these, however, require a certain amount of preparation: access to collection metadata, consensus on selection criteria, and preparation of contextual sources like the press directories. LwM experiences demonstrate that these all take time, impacting planning. Furthermore, any attempt to obtain a

31 Tom O'Malley (2015). "Mitchell's Newspaper Press Directory and the Late Victorianand Early Twentieth-Century Press." In: *Victorian Periodicals Review* 48 (4), pp. 591–606.

balanced corpus, or simply to understand the contours of what is available, still needs to recognise the incompleteness and partiality of any collection.

By traveling together to discover what has been digitised, what can be digitised, and the significance of these choices, LwM researchers and librarians have developed a shared understanding of the origins and futures of digitised newspaper collections.

Bibliography

Abel, Richard (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1–2, pp. 1–10. DOI: 10.2979/filmhistory.25.1-2.1.

Bailey, Jefferson and Lily Pregill (2014). "Speak to the Eyes: The History and Practice of Information Visualization." In: *Art Documentation: Journal of the Art Libraries Society of North America* 33.2, pp. 168–191. DOI: 10. 1086/678525. eprint: https://doi.org/10.1086/678525. URL: https://doi.org/10.1086/678525.

Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux (2017). "Impact of OCR errors on the use of digital libraries: towards a better access to information." In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–4.

Dommann, Monika and Sarah Pybus (2019). *Authors and apparatus: a media history of copyright*. In collab. with Ebook Central. Ithaca; London: Cornell University Press. 1 p.

Fyfe, Paul (2016). " An Archaeology of Victorian Newspapers." In: *Victorian Periodicals Review* 49.4, pp. 546–577. DOI: 10.1353/vpr.2016.0039. URL: https://muse.jhu.edu/article/644183.

Gabriele, Sandra (2003). " Transfiguring the newspaper: from paper to microfilm to database." In: *A modern 2: Network Archaeology*. URL: https://amodern.net/article/transfiguring-the-newspaper/.

Giullian, Jon (2013). "« "Seans Chernoi Magii Na Taganke": The Hunt for Master and Margarita in the Pravda Digital Archive »." In: *Slavic East European Information Resources* 14.2–3, pp. 102–26. DOI: 10.1080/15228886.2013.813374.

Gliserman, Susan (1969). "Mitchell's" Newspaper Press Directory": 1846–1907." In: *Victorian Periodicals Newsletter* 4, pp. 10–29.

Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras, and Emily Bell (2020). "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers." In: *Archival Science* 20.2, pp. 139–165. DOI: 10.1007/s10502-020-09332-1.

Hill, Mark J and Simon Hengchen (2019). "Quantifying the impact of dirty OCR on histor- ical text analysis: Eighteenth Century Collections Online as a case study." In: *Digital Scholarship in the Humanities* 34.4, pp. 825–843

Hitchcock, Tim (2011). "Academic History Writing and its Disconnects." In: *Journal of Digital Humanities* 1, 1.

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. 1228–43.

King, Edmund (2005). "Digitisation of Newspapers at the British Library." In: *The Serials Librarian* 49.1, pp. 165–181. DOI: 10.1300/J123v49n01_07. URL: https://doi.org/10.1300/J123v49n01_07.

Mak, Bonnie (2014). "Archaeology of a digitization." In: *Journal of the Association for Information Science and Technology* 65.8, pp.1515–1526. DOI: 10.1002/asi.23061. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23061.

Milligan, Ian (2013). "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." In: *The Canadian Historical Review* 94.4, pp. 540–569.

Mitchell, C., and Co. (1857). The Newspaper Press Directories.

Mussell, James (2014). "Elemental Forms." In: *Media History* 20.1. Publisher: Routledge_eprint: https://doi.org/10.1080/13688804.2014.880264, pp. 4–20. DOI: 10.1080/13688804.2014.880264. URL: https://doi.org/10.1080/13688804.2014.880264.

O'Malley, Tom (2015). "Mitchell's Newspaper Press Directory and the Late Victorian and Early Twentieth-Century Press." In: *Victorian Periodicals Review* 48 (4), pp. 591–606.

Ryan, Yann, Mariona Coll Ardanuy, Daniel van Strien, Kasra Hosseini, Kaspar Beelen, James Hetherington, Katherine McDonough, Barbara McGillivray, Mia Ridge, Olivia Vane, and Daniel CS Wilson (2020). "Using smart annotations to map the geography of newspapers." In: DH2020. URL: https://www.conftool.pro/dh2020/index.php?page=showAbstract&form_id=532&show_abstract=1.

10 Billion Words: The British Library British Newspapers 1800– 1900 Project. Some Guidelines for Large-Scale Newspaper Digitisation (2005). URL: http://www.webarchive.org.uk/wayback/archive/20140615090156/http://www.jisc.ac.uk/uploaded_documents/IFLA_2005.pdf.

Shaw, Jane (2009). *British Newspapers 1620–1900: Final Report. J.* JISC and the British Library. URL: http://www.webarchive.org.uk/wayback/%20archive/20140614080134/http://www.jisc.ac.uk/media/documents/programmes/digitisation/blfinal.pdf.

Slauter, Will (2019). *Who Owns the News?: A History of Copyright*. Google-Books-ID: FZSFDwAAQBAJ. Stanford University Press, Stanford, CA. 488 pp.

Smits, Thomas (2016). "Making the News National: Using Digitized Newspapers to Study the Distribution of the Queen's Speech by W. H. Smith & Son, 1846–1858." In: *Vic- torian Periodicals Review* 49.4, pp. 598–625. DOI: 10.1353/vpr.2016.0041. URL: https://muse.jhu.edu/article/644185.

van Strien, D., K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray and G. Colavizza. (2020). "Assessing the Impact of OCR Quality on Downstream NLP Tasks." In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence – Volume 1: ARTIDIGH*, pp. 484–496. DOI: 10.5220/0009169004840496.

Tufte, Edward (2006). *Beautiful Evidence*. Graphics Pr, Cheshire, Conn.

Vane, Olivia (2019). "Timeline design for visualising cultural heritage data." PhD thesis. Royal College of Art, London. URL: http://researchonline.rca.ac.uk/4325/1/TimelineDesignForVisualisingCulturalHeritageData_OliviaVane_redacted.pdf.

Windhager, F., P. Federico, G. Schreder, K. Glinka, M. Dörk, S. Miksch, and E. Mayr (2018). "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges." In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Andrew J. Torget

Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections

Abstract: This paper documents the "Mapping Texts" project, an experiment focused on the problem of OCR noise in historical newspapers. The purpose of the project was to combine natural language processing with data visualization to measure both OCR noise rates and their effects on how scholars detect meaningful high-level trends embedded in large-scale digital newspaper collections. The project developed two interactive visualizations measuring OCR quality and its effects on detecting high-level trends in the data, revealing the depth of the challenges facing humanities scholars seeking greater transparency of OCR data in historical newspaper databases.

Keywords: historical newspaper, OCR noise, data visualization

1 Introduction

"Mapping Texts" (http://mappingtexts.org) is a collaborative project between the University of North Texas and Stanford University whose goal has been to develop a series of experimental models for combining the possibilities of natural language processing and data visualization in order to enable researchers to develop better methods for discovering and analyzing meaningful high-level trends embedded within massive collections of historical newspapers. The broader purpose behind this effort has been to help scholars develop new tools for coping effectively with the growing challenge of doing research in the age of information abundance, as historical newspapers are digitized and made available online at a rapidly expanding pace. The *Chronicling America* project (a joint endeavor in the United States of the National Endowment for the Humanities and the Library of Congress), for example, has made more than seventeen million historical newspaper pages freely available online. What can scholars do with such an immense wealth of information? Without research tools and methods capable of sifting out meaningful and novel patterns embedded within such massive datasets, scholars typically find themselves confined to performing only basic word searches across enormous collections, which become increasingly less useful as datasets grow in size. If, for example, a search for a particular term yields 4,000,000 results, even

https://doi.org/10.1515/9783110729214-003

those search results produce a dataset too large for any single scholar to analyze in meaningful ways using traditional methods. The age of information abundance, it turns out, can simply overwhelm researchers, as the sheer volume of available digitized historical newspapers is beginning to do.

To address this, the "Mapping Texts" project sought to create two different interactive, online models that experiment with new methods for discovering high-level trends embedded in collections of digitized historical newspapers. For this work, we experimented on a collection of 232,500 pages of historical newspapers from Texas digitized by the University of North Texas (UNT) Library as part of the U.S. National Digital Newspaper Program (NDNP)'s *Chronicling America* project (http://chroniclingamerica.loc.gov/).(For access to the full dataset, as well as a discussion of why we chose this particular collection, please see the project's website.)The first model we built offers users a quantitative survey of these digital newspapers, providing interactive visualizations of (a) the quantity of words in the collection over time, (b) the quantity of words both by location and newspaper title, and (c) the optical character recognition (OCR) accuracy rates of the digitized newspapers. The second model offers users a high-level survey of the content of these digital newspapers by using natural language processing techniques to expose various high-level trends – such as ranked word frequencies, named entity counts, and topic models – that can be customized by users to focus on particular time periods, locations, and newspaper titles.

This paper documents the creation of the experimental models and the insights we gleaned from the "Mapping Texts" project. Our intention was for the project to provide concrete examples of new ways to uncover high-level trends embedded in large datasets of digitized historical newspapers, which we believe it did. But perhaps more importantly, the two models we built also demonstrated the powerful role that OCR accuracy rates play in determining what researchers can and cannot glean from digitized historical newspapers. Indeed, the central finding of our work was that OCR accuracy rates are the single most important metric in determining the usability of a dataset of digitized historical newspapers for humanities research. (All of the source code and datasets supporting the experiments described below are available at: http://mappingtexts.org/data).

This work has built upon advancing work in both the application of natural language processing to humanities texts and the broadening use of visualizations for sorting out patterns embedded in humanities datasets. With the massive increase in digital materials available to researchers, far more humanities scholars have sought in recent years to employ text analysis or data visualization tools for

sorting overwhelming piles of digitized records.[1] Yet there remain relatively few projects like "Mapping Texts" that attempt to combine text analysis with data visualization, even though the work of Stéfan Sinclair and Geoffrey Rockwell have demonstrated the power of integrating the two. Even fewer projects have combined text-analysis and data visualization in order to explore digitized American newspapers.[2] The most notable exception is Ryan Cordell and David Smith's "Viral Texts," which experiments with how U.S. newspapers reprinted from one another during the nineteenth century and visualizes identified patterns. Much like "Mapping Texts," the "Viral Texts" project encountered significant challenges presented by OCR noise.[3] Indeed, this project's central findings detailed below directly align with a burgeoning body of work that has collectively come to similar conclusions on the impact of OCR recognition rates on the ability of scholars to apply natural language processing and data visualization to collections of digitized historical newspapers.[4]

---

1 For discussions on how humanities scholars have used natural language processing, see Ted Underwood, "Seven Ways Humanists Are Using Computers to Understand Text. The Stone and the Shell," June 4, 2015, https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/; Matthew K. Gold et al., "Forum: Text Analysis at Scale," in *Debates in the Digital Humanities* (University of Minnesota Press, 2016), 525–568, http://dhdebates.gc.cuny.edu/debates/text/93; Tim Hitchcock and William J. Turkel, "The Old Bailey Proceedings, 1674–1913: Text Mining for Evidence of Court Behavior," *Law and History Review* 34:4 (August 2016): 1–27, doi:10.1017/S0738248016000304.For examples of large-scale data visualization projects focused on historical analysis, see Scott Nesbit and Edward Ayers, *Visualizing Emancipation,* https://dsl.richmond.edu/emancipation/, published 2013; Gregory P. Downs and Scott Nesbit, *Mapping Occupation: Force, Freedom, and the Army in Reconstruction*, http://mappingoccupation.org, published 2015.

2 Stéfan Sinclair and Geoffrey Rockwell, "Text Analysis and Visualization: Making Meaning Count," in *A New Companion to the Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth (Wiley Blackwell, 2016), 274–290.

3 Ryan Cordell and David Smith, Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines (2017), http://viraltexts.org.Attempted to analyze and visualize the reprinting of stories in nineteenth-century U.S. newspapers. This project dealt with OCR noise, in part, by comparing clusters of similar passages reprinted in multiple newspapers – see David A. Smith, et al., "Detecting and Modeling Local Text Reuse," published in the Proceedings of IEEE/ACM Joint Conference on Digital Libraries (IEEE Computer Society Press, 2014), doi: https://www.ccs.neu.edu/home/dasmith/infect-dl-2014.pdf. For another example of visualizations of a historical newspaper, see Cameron Blevins, "Space, Nation, and the Triumph of Region: A View of the World from Houston," *Journal of American History*, 101:1 (June 2014), 122–147.

4 See, for example van Strien, Daniel, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. "Assessing the Impact of OCR Quality on Downstream NLP Tasks." In ICAART (1), 484–496. 2020. https://www.staff.universiteitleiden.nl/bina

2 Mapping Newspaper Quality

The project's first model (http://mappingtexts.org/quality/) plotted the quantity and quality of the newspaper content. Such an interface, we hoped, would provide users with transparency into the amount of information available in our collection of digitized newspapers by providing visualizations of quantifiable metrics such as the sheer quantity of data available, the geographic locations of that information, and how much of that information was concentrated into any given span of time. The purpose was to empower users of digitized historical newspapers to make more informed choices about what sort of research questions could be answered by the available sources. If, for example, a scholar was interested in how U.S. newspapers discussed Abraham Lincoln during the American Civil War, it would be highly useful to be able to determine how much information in a given database came from the 1861–1865 era. Without such information, it would be remarkably difficult for a researcher to evaluate whether or not a given collection of digitized historical newspapers would hold potentially useful information. Our first task in that process was simply to count the number of words published by each newspaper in the collection by year so we could plot the quantity of available information in any given time period. In so doing, however, we quickly realized that we would have to confront the problem of OCR noise.

2.1 The Problem of OCR Noise

Optical character recognition (OCR) is a process by which a computer program scans images of words and attempts to identify alpha-numeric symbols (letters and numbers) so they can be translated into electronic text. In doing an OCR scan of an image of the word "the," an effective OCR program should be able to

---

ries/content/assets/governance-and-global-affairs/isga/artidigh_2020_7_cr.pdf; Sunghwan Mac Kim and Steve Cassidy, "Finding Names in Trove. Named Entity Recognition for Australian Historical Newspapers," in *Proceedings of Australasian Language Technology Association Workshop* (2015), 57–65, https://www.aclweb.org/anthology/U15-1007.pdf; Chiron, G., Doucet, A., Coustaty, M., Visani, M., & Moreux, J. P. (2017, June). Impact of OCR errors on the use of digital libraries: towards a better access to information. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (1–4), IEEE, https://ieeexplore.ieee.org/abstract/document/7991582 See also: Hill, Mark J., et Simon Hengchen. "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study," Digital Scholarship in the Humanities, vol. 34, no. 4, Oxford Academic, December 2019, 825–843, doi: 10.1093/llc/fqz024.

recognize the individual “t” “h” and “e” letters and then save those as “the” in text form. Various versions of this process have been around since the late 1920s, although the technology has improved drastically in recent years. Today most OCR systems achieve a high-level of recognition accuracy when used on printed texts and calibrated for specific fonts.

Images of historical newspapers, however, present particular challenges for OCR technology for a variety of reasons. The most prolific challenge is simply the quality of the images of newspaper pages. Most of the OCR done on historical newspapers relies upon scanning microfilmed versions of those newspapers, and the quality of those microfilm images can vary enormously. Microfilm imaging done decades ago, for example, often did not film in grayscale (which meant that areas poorly lit during the imaging process often became largely blacked out in the final image) and so OCR performed on badly imaged newspapers regularly achieved poor results. A related challenge is that older newspapers typically employed very small fonts in very narrow columns, which often compounds the problem of poor imaging. The combination of these limitations often introduce mistakes into scanned texts (such as replacing the letter “l” with the numeral “1” as in “1imitations” for “limitations”) that can, in turn, matter enormously for a researcher attempting to determine how often a certain term was used in a particular newspaper or time period. If poor imaging – and therefore poor OCR results – meant that “Lincoln” was often rendered as “1inco1n” in a data set, that *should* affect how a scholar researching newspaper language patterns surrounding Abraham Lincoln would go about his or her work.

As we sought to measure the number of words in the collection, we realized that it would be important to develop programmatic methods for scrubbing the corpus so as to correct simple and artificial recurring errors introduced by the OCR process. Common misspellings introduced by OCR, for example, could be detected and corrected by systematically comparing the words in our corpus to English-language dictionaries. For this task, we used the GNU Aspell dictionary and then ran a series of processes over the corpus that checked every word in our newspaper corpus against the dictionary. Within Aspell we also used an additional dictionary of place names gathered from Gazetters. This way, Aspell could also recognize historically relevant place names, such as “Denton” or “Cuahtemoc,” and then suggest them as alternatives when there was a slight misspelling. Whenever a word was detected that did not match an entry in the dictionary, we checked if a simple replacement for letters that are commonly mis-rendered by OCR (such as “Linco1n” for “Lincoln”) would then match the dictionary. We made these replacements only with the most commonly identified errors (such as the numeral “1” for the letter “l” and “@” for “a”), and we

experimented with this numerous times in order to refine our scripts based on hand-checking the results, before running the final process over the corpus.

End-of-line hyphenations and dashes could also be programmatically identified and corrected in the OCR text. If a word in the original newspaper image had been hyphenated to compensate for a line-break (a highly common occurrence), that would create in the OCR text two nonsensical words (such as "hist-" and "orical" from "hist-orical" which would not match any text searches for "historical") even though the word did appear in the original text. To correct for this, we ran a script over the corpus that looked for words that ended with a hyphen and was followed by a word that did not match any entries in our dictionary. The two parts ("hist-" and "orical") were then reconnected with the hyphen removed ("historical") and if that reconnected word now matched an entry in the dictionary, we made the correction.(We guarded against these programmatic corrections introducing new words that were not in the original by rigorously spot-checking with human readers, which confirmed that our scripts were overwhelmingly correcting OCR errors.)

To give an idea of the coverage and efficiency of this spelling correction phase, we collected statistics on a random sample of 100 documents. From a total of 209,686 words, Aspell identified as correct 145,718 (70%), suggested acceptable replacements for 12,946 (6%), and could not find a correction for 51,022 (24%).

The objective of this work was to perform basic clean-up of known OCR errors so that we could get a finer and more accurate sense of how much true "noise" (that is, unrecognizable information) was present in the corpus compared to recognizable words and content. A 6% improvement on the accuracy of the text was, indeed, meaningful. Yet the fact that nearly a full quarter of the collection still composed unusable noise rather than explorable content was sobering to us. Researchers need to know the quality of the digitization of the collections they are exploring in order to be able to account for such large error rates, and so we concluded that making OCR accuracy rates both available and transparent for scholars would be indispensable for enabling the responsible use of these collections. As such, we decided to include them in our first model (see figure 1).

2.2 Constructing the Interactive Model

Once we had our corpus scrubbed of easily corrected errors introduced by the OCR process, we then ran the full newspaper data set against the dictionary once more to produce a final word count of recognized words ("good" content, in the sense that the OCR rendered usable text) to unrecognized words ("bad"

content, or "noise" introduced by the OCR process). We also used human readers to spot-check batches of both "good" and "bad" content to ensure we were not producing false positives or negatives, and the batches proved highly accurate. This provided a database we could use to measure the quality of the data, which we then organized by newspaper title and year. So, for every newspaper title, we had counts of the "good" and "bad" words per year – effectively measures of OCR quality by newspaper – which we then organized into a series of interactive visualizations that offer users multiple ways to access and parse the quantity and quality of the digitized newspapers.

The model is organized by time and geography (see figure 1). At the top is a timeline that plots the quantity of words (both in the complete corpus, and the "good words") over time, providing an overall sense of how the quantity of information ebbs and flows with different time periods. Users can also adjust the dates on the timeline to focus on a particular date-range in order to explore in more detail the quantity of information available. Adjusting the timeline also affects the other major index of the content: an interactive map of Texas. For the visualization, all the newspapers in the database were connected by their publication city, so the map reveals the geographic distribution of the newspaper content by city. This can be adjusted to show varying levels of quality in the newspaper corpus (by adjusting the OCR-ratio bar for "good" to "bad" words) in order to find areas that had higher or lower concentrations of quality OCR text. The size of the circles for cities show their proportion of content relative to one another, which the user can switch from a logarithmic view (the default view, which gives a wider sense of the variety of locations) to a linear view (which provides a greater sense of the disparity and proportion of scale between locations).

Viewing the database geographically reveals that two locations dominated the collection. Newspapers from Houston and Fort Worth. Combined, those two locations outstrip the quantity of information available from any other location in Texas, which is interesting in part because neither of those locations became dominant populations centers in Texas until the post-World War II era (and therefore well after the 1883–1911 and 1925–1940 time periods that compose the majority of the newspaper content in our sample).This would suggest that the newspapers of rural communities, where the majority of Texans lived before the 1930s, are underrepresented among the newspapers of this collection, and that urban newspapers – and therefore urban concerns – are likely overrepresented. While scholars of urbanization would be well-served by this dataset, scholars interested in rural developments would be advised to be wary of this imbalance when conducting research with this collection.

Because we had concluded that making OCR accuracy rates transparent is crucial, we focused on creating an OCR visualization. The third major window

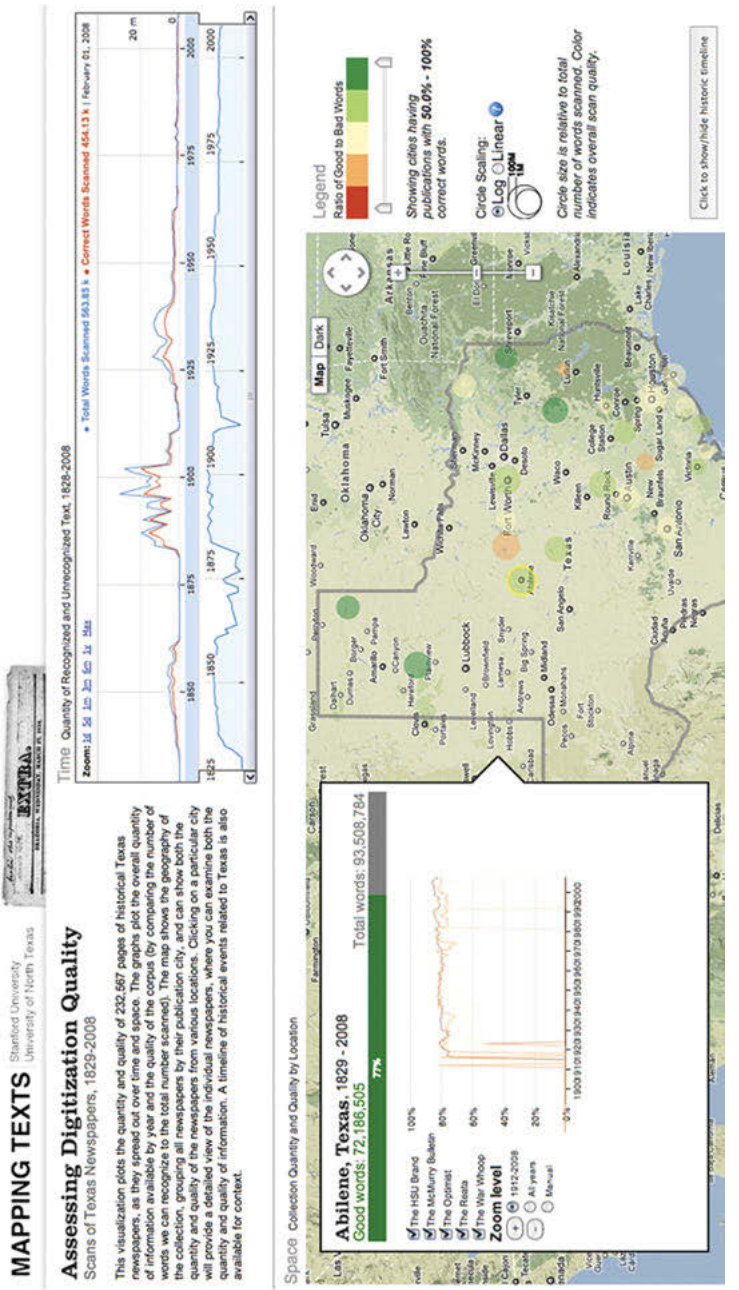


Fig. 1: First Model, offering a quantitative survey of the newspapers. URL: http://mappingtexts.org/ [Accessed 21.6.2022].

into the collection, therefore, is a detail box that (see figure 2), for any given location (such as Abilene, Texas), provides a bar of the good-to-bad word ratio (OCR accuracy), as well as a complete listing of all the newspapers that correspond to that particular location and OCR metrics on the individual newspapers. The detail box also provides access to the original newspapers themselves, as clicking on any given newspaper title will take the user to digitised versions of the newspapers.

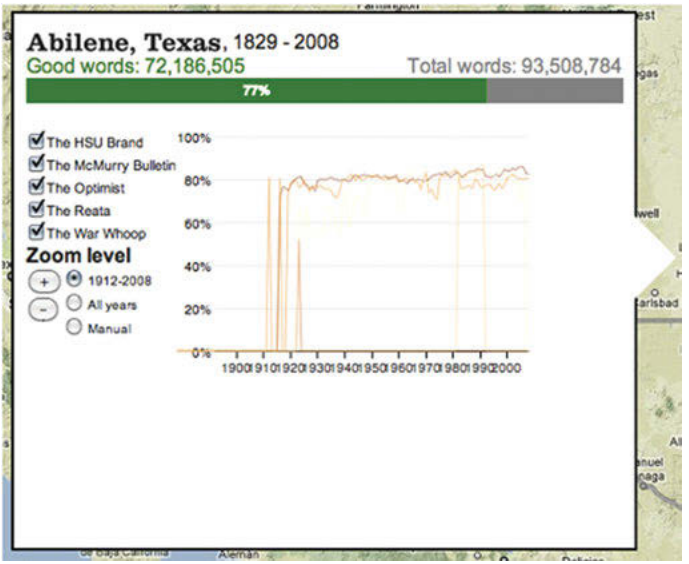


Fig. 2: Detail box, from First Model, for Abilene, Texas, newspapers.

Exploring the various geographic locations with the detail box reveals again the overwhelming power of OCR quality on the usability of these digital newspapers for scholars. Although Houston and Fort Worth represent the locations with the *largest* quantity of available data, they are not the locations with the *highest* quality of available data. The overall accuracy rate for the OCR of Houston newspapers was only 66% (although this varied widely between various newspapers) and for Fort Worth the overall rate was 72%. By contrast, the newspaper in Palestine, Texas, achieved an 86% quality rate, while the two newspapers in Canadian, Texas, achieved an 85% quality rate. At the lowest end of quality was the OCR for newspapers from Breckenridge, Texas, which achieved only a 52% accuracy rate. Scholars interested in researching places like Breckenridge or Houston, therefore, would need to consider that anywhere

between a third to a half of the words OCR'd from those newspapers were rendered unrecognizable by the digitization process. Scholars who decided to focus on newspapers from Palestine or Canadian, on the other hand, could rely on the high quality of the digitization process for their available content. The broad diversity of OCR recognition rates, we determined, was directly related to the quality of the image capture, as poor digital imaging of the original newspapers then led to poor OCR results.

3 Mapping Newspaper Language

Once we had completed our quantitative survey of the collection, we turned our attention to building a second model for assessing high-level trends embedded within our digitized newspaper collection. With this model we wanted to experiment with ways for people to explore thematic trends in the newspaper corpus as they spread out across both time and space, with the driving purpose of experimenting with ways to allow users to discover novel and meaningful patterns scattered across the collection. Toward that end, we chose to focus on three methods widely used by humanities scholars for surveying high-level trends in large bodies of text and use those to visualize patterns embedded in the collection:

(1) Word Counts. One of the most basic and widely used metrics for assessing language use in text has been word counts. The process is simple: after stripping out stop words, we run a script to count all the words in a body of text and then rank the results by absolute frequency in order to reveal which words and phrases appear most frequently.[5]
(2) Named Entity Recognition (NER) Counts. This is a more finely-grained version of basic word counts. In collecting NER counts, a program will attempt to identify and classify various elements in a text (usually nouns, such as people or locations) in a body of text. Once that has been completed, the absolute frequency of those terms can then be tallied and ranked, just like

5 One of the most famous examples of humanities research utilizing word counts is Michel, Jean-Baptiste et al. "Quantitative analysis of culture using millions of digitized books." *Science* (New York, N.Y.) vol. 331,6014 (2011) 176–182, doi:10.1126/science.1199644. For a discussion of limitations of this approach, see Brysbaert, Marc et al. "Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing." *Frontiers in Psychology* 2:27, 2 Mar. 2011, doi:10.3389/fpsyg.2011.00027.

with basic word counts. The result is a more specific and focused ranking of frequency of language use in the text.[6]

(3) Topic Modeling. This method of text-analysis has grown in popularity among humanities scholars in recent years, with the greater adaption of programs like the University of Massachusetts's MALLET (MAchine Learning for LanguagE Toolkit). The basic concept behind topic modeling is to use statistical methods to produce collections of words that appear to be highly correlated to one another (which are called "topics") that appear in a given collection of texts. So, for example, running the statistical models of MALLET over a body of text will produce a series of "topics," which are strings of words (such as "Texas, street, address, good, wanted, Houston, office") that may not necessarily appear next to one another within the text, but, nonetheless, are likely to appear together within the same text and therefore have a statistical relationship to one another. The idea behind topic modeling is to identify collections of statistically related words as a means of exposing larger, wider patterns of language spread across many documents, revealing hidden thematic structures which would be impossible to discover in a large corpus by doing a close-reading of the same texts.[7] Evaluating the topics produced by Latent Dirichlet Allocation (LDA) modeling for humanities research is famously difficult, largely because the string of words produced for each "topic" requires interpretation by human readers of the "meanings" that might connect those words.[8] Topic modeling has nonetheless gained increasingly popularity among humanities scholars, in large measure because the

6 For examples of this work being done on historical newspapers, see Sunghwan Mac Kim and Steve Cassidy, "Finding Names in Trove. Named Entity Recognition for Australian Historical Newspapers," in *Proceedings of Australasian Language Technology Association Workshop* (2015), 57–65, https://www.aclweb.org/anthology/U15-1007.pdf; Kimmo Kettunen, et al., "Old Content and Modern Tools: Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910," *Digital Humanities Quarterly* 11(3) (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html; Teemu Ruokolainen and Kimmo Kettunen, "Name the Name: Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data," *Proceedings of the Digital Humanities in the Nordic Countries 5th* Conference (October 2020), 137–156, http://ceur-ws.org/Vol-2612/paper10.pdf.

7 For excellent discussions of topic modeling in humanities research, see the winter 2012 issue of the *Journal of Digital Humanities* (http://journalofdigitalhumanities.org/2-1/) with articles by David Blei and David Mimno, among others, on the use of topic modeling in humanities research.

8 Jian Tang, et al., "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis," *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. http://proceedings.mlr.press/v32/tang14.pdf.

statistical models appear to produce topics that seem both relevant and meaningful to human readers.[9]

3.1 Collecting Word and NER Counts

All of this work built upon our work on the first model – including the extensive scrubbing of the OCR texts – and generating the word counts was a simple matter of tallying absolute word frequencies, ranking them, and then organizing them by newspaper and location. Generating the Named Entity Recognition dataset was somewhat more complicated. There are a number of available programs for performing NER counts on bodies of text, and we experimented with a variety to see which achieved the best results. To determine the accuracy of the candidate parsers, we manually annotated a random sample of one hundred named entities from the output of each parser considered. To measure the efficiency (because scale, again, was a primary consideration), we also measured the time taken for the parser to label 100 documents. Among those we tried that did not – for a variety of reasons – achieve high levels of accuracy for our collection were LingPipe, MorphAdorner, Open Calais, and Open NLP. We had a great deal more success with the Illinois Named Entity Tagger (http://cogcomp.cs.illinois.edu/page/publication_view/199). It was, however, the Stanford Named Entity Recognizer (http://www-nlp.stanford.edu/software/CRF-NER.shtml) that achieved the best parser accuracy while also maintaining a processing speed comparable with the other taggers considered. We, therefore, used the Stanford NER to parse our collection and then ranked the NER counts by frequency and organized them by newspaper and year.

9 See, for example, D.J. Newman and Sharon Block, "Probabilistic topic decomposition of an eighteenth-century American newspaper," *Journal of the American Society for Information Science and Technology* 57(6) (2006), 753–767, doi:10.1002/asi.20342.; Sharon Block, "Doing More with Digitization," 6.2 (2006) *Commonplace: The Journal of early American Life*, http://commonplace.online/article/doing-more-with-digitization/; Robert K. Nelson, "Mining the Dispatch," (revised November 2020), https://dsl.richmond.edu/dispatch/; Cameron Blevins, "Topic Modeling Martha Ballard's Diary," (posted April 2010) http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/. For examples of topic modeling work on nineteenth-century American literature, see Matthew Jockers *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013) and Matthew Jockers and David Mimno. "Significant Themes in 19th-Century Literature." *Poetics* 41.6 (2013), 750–769.

3.2 Topic Modeling

For our topic modeling work, we decided to use the University of Massachusetts's MALLET package (http://mallet.cs.umass.edu/) for a number of reasons, the most prominent of which were that (a) it is well documented, and (b) other humanities scholars who have used the package have reported high quality esults.[10] MALLET also uses the probabilistic latent semantic analysis (pLSA) model that has become one of the most popular within the field of natural language processing, and so we decided to use the package for our experiments in testing the effectiveness of topic modeling on our dataset.[11]

We spent far more time working on and refining the topic modeling data collection than any other aspect of the data collection for this project. Much of that work concentrated on attempting to assess the quality and relevance of the topics produced by MALLET, as we ran repeated tests on the topics produced by MALLET that were then evaluated by hand to see if they appeared to identify relevant and meaningful trends within our newspaper collection. Our close examinations of the topics produced by MALLET convinced us that the statistical program did appear to identify meaningful high-level trends in our newspaper collection. We therefore processed our entire newspaper corpus using the program.[12]

Because we needed to preset specific timeframes for our topic models, we selected commonly recognized historical eras among historians who study Texas and the U.S.-Mexico borderlands: 1829–1835 (Mexican National Era), 1836–1845 (Republic of Texas), 1846–1860 (Antebellum Era), 1861–1865 (Civil War), 1866–1877 (Reconstruction), 1878–1899 (Gilded Age), 1900–1929 (Progressive Era), 1930–1941 (Depression), 1942–1945 (World War II), 1946–2008 (Modern Texas).For each of these eras, we used MALLET to generate a list of topics drawn from all newspapers in the dataset from those particular years.

10 See Robert K. Nelson, "Mining the Dispatch," (revised November 2020), https://dsl.richmond.edu/dispatch/; andCameron Blevins, "Topic Modeling Martha Ballard's Diary" (posted April 2010), http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

11 David Blei, "Probablistic Topic Models," *Communications of the ACM*, 55:4 (April 2012) doi: 10.1145/2133806.2133826; Sanjeev Arora, et al., "A Practical Algorithm for Topic Modeling with Provable Guarantees" (December 2012), https://arxiv.org/abs/1212.4777.

12 For a detailed discussion of how we implemented our topic modeling work, see Tze-I Yang, Andrew J. Torget, and Rada Mihalcea, "Topic Modeling on Historical Newspapers," conference proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011), June 2011, 96–104.

3.3 Building the Second Model

Constructing the online interface for the second model (see figure 3) presented challenges not posed by the first model. The temporal and spatial dimensions were roughly the same, but this visualization needed to show the results from three separate types of natural language processing, not all of which could be sliced into the same temporal chunks. After much experimentation, we decided the best approach would be to repeat the time slider and map interface, while adding a three-part display to present the individual results of the three natural language processing techniques. The result is an interactive visualization (http://mappingtexts.org/language) that maps broad trends embedded within our newspaper collection over any particular time period and geography selected by the user.

Once a user has selected a time frame and geography, they can examine the three associated categories which are listed below the map in their own "widgets" (see figure 4). In the word counts and named entity counts widgets, there are two ways to look at the language data. As a ranked list with the most frequently appearing (in absolute numbers) words at the top followed by a descending list, or as a word cloud that presents the data as a constellation of words with their size representing their relative frequencies. In the topic model widget, the user is offered ten "topics" generated by MALLET from the newspapers associated with a particular date range. Within each topic is a list of 100 words that have a statistical relationship to one another in the collection, in that these words have a high likelihood of occurring within the same document.

3.4 The Power of OCR

This model again highlighted the problem of OCR noise as a central challenge facing scholars who seek to use databases of digitized historical newspapers. While we hoped that each of the three metrics in this model would expose novel trends embedded in the collection, the most prominent pattern they identified was OCR noise. In the word counts, for example, the third most frequently occurring term for the entire collection was "nnd" (likely a corruption of "and"), which was second only to "Texas" and "city" for prominence in the dataset. Exploring word counts for particular time periods, locations, or newspapers confirmed that OCR noise was so woven throughout the collection as to produce some of the most frequently occurring terms. Within the twenty-five most frequently occurring words for the "Antebellum Era," more than a quarter of them (seven out of twenty-five) were nonsensical noise. The topic modeling work exposed these "noise" threads in even greater detail. For each preset historical era,

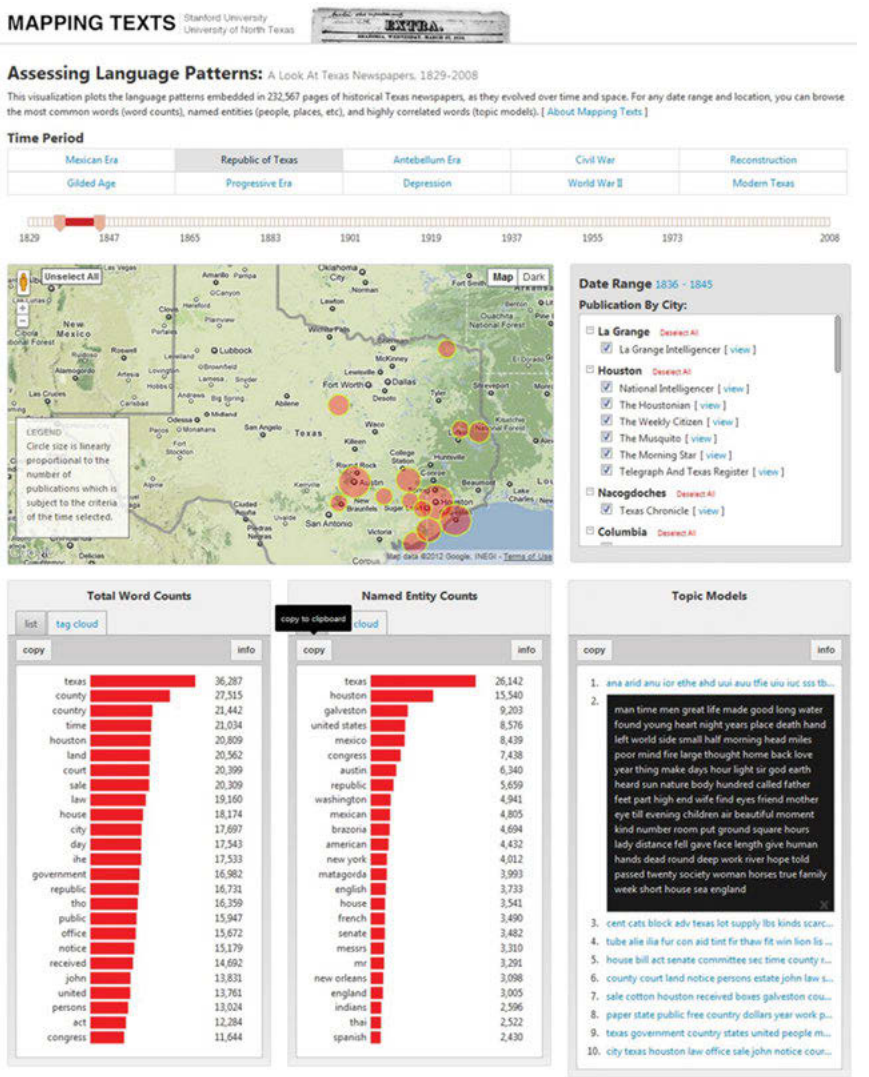


Fig. 3: Second Model, offering a survey of high-level trends within the newspapers.

the online model produced ten "topics" and in every single era at least one (and sometimes several) of those topics were composed entirely of OCR noise. In the "Republic of Texas" era, some of the topics offered by MALLET appear to expose discussions in the newspapers about the local economy ("sale, cotton, Houston, received, boxes, Galveston"), local government ("county, court, land, notice,

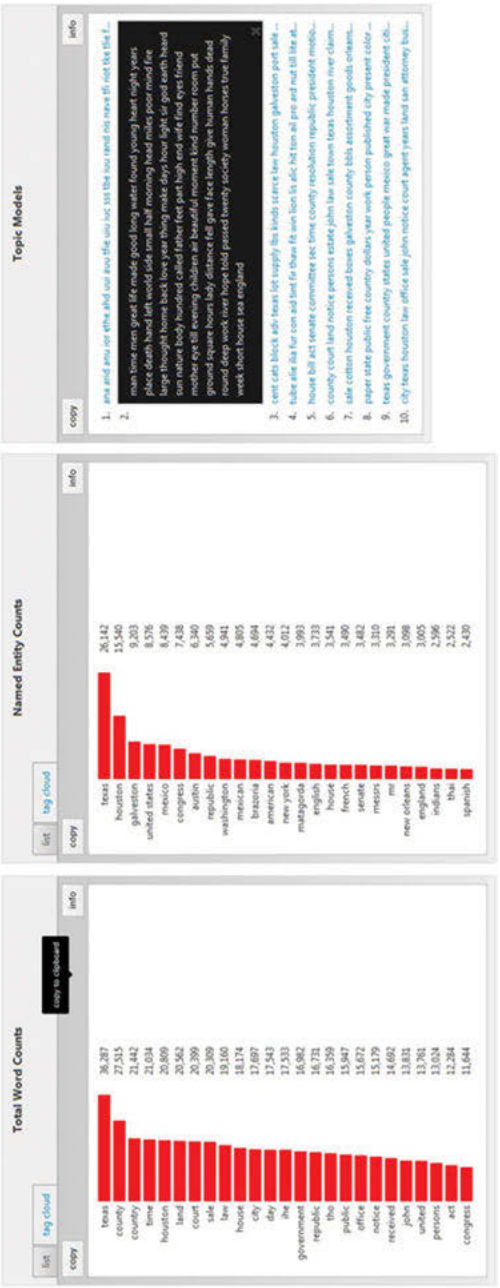


Fig. 4: Detail of Language Widgets in the Second Model.

persons, estate"), and social concerns ("man, time, men, great, life"), which we found intriguing patterns. Yet two of the topics were made up solely of OCR noise, including one topic which collected "words" such as "ana, arid, anu, ior, ethe, ahd, uui, auu, tfie, uiu, iuc, sss, tbe, iuu" as one of the most prominent threads of the collection. Scrambled content, in other words, was so statistically prominent as to be highlighted by MALLET as some of the most consistent topics in our digitized newspapers.

4 Conclusion

OCR accuracy rates, we determined, are one of the most important metrics to identify in assessing the usability of a particular collection of digitized historical newspapers for humanities research. Although the level of noise present in our collection of 232,500 pages of historical newspapers from Texas varied over time and by newspaper title, in all our experiments it nonetheless emerged as one of the most persistent patterns woven throughout the collection. We attempted to mitigate this problem by performing a programmatic "scrubbing" of the corpus and found that we could, indeed, improve the accuracy of the OCR'd texts. Yet that improvement could not make up for the 24% noise level that remained, which had a distorting effect on all our other efforts to collect and visualize meaningful high-level trends within the newspapers.

As such, the project highlighted the need for data transparency about OCR quality in online databases of digitized historical newspapers. Scholars can only determine how best to explore a digitized newspaper if they know how clean or noisy the text of that newspaper is. A researcher working with a digital newspaper with a 90% OCR accuracy rate, for example, could rely on keyword searches as a reasonable method for finding terms in the text relevant to their project. If that same scholar, however, found themselves working with a newspaper with only a 52% OCR accuracy rate, such high noise levels mean that keyword searches would be wholly unreliable, and the researcher would likely be better served in doing a close reading of the scanned pages. Yet such a determination by the scholar can only happen if they are first provided ready access to OCR accuracy rates. This underscored, in turn, the need to develop a standardized vocabulary for measuring and expressing OCR quality (which we attempted to experiment with in our visual scales for "good" to "bad" word ratios in our first model) that will make this information readily apparent to users of these databases.

We found that visualizing high-level trends in the newspapers could yield interesting insights into the collection. It was highly useful, for example, to be

able to plot the overall quantity of newspaper data in the first model because it demonstrated where the collection was particularly content-rich in terms of time periods (1883–1911 and 1925–1940) and locations (Houston and Fort Worth), both of which are important considerations for scholars when determining if a given newspaper database would be useful to their particular research topic. Topic modeling, as another example, showed significant promise for helping humanities scholars identify meaningful trends in large collections of text. Yet we found the overall distorting effect of OCR noise in the collection made it almost impossible to discern whether many of those patterns reflected the content of the newspapers or the warping of that content by the OCR process. This, we believe, is one of the central challenges that scholars will have to address in order to unlock the full research potential of digitized historical newspapers.

Bibliography

Arora, Sanjeev, et al., "A Practical Algorithm for Topic Modeling with Provable Guarantees" (December 2012), https://arxiv.org/abs/1212.4777.

Blei, David, "Probablistic Topic Models," *Communications of the ACM*, 55:4 (April 2012), doi: 10.1145/2133806.2133826.

Blevins, Cameron, "Topic Modeling Martha Ballard's Diary," (posted April 2010) http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

Blevins, Cameron, "Space, Nation, and the Triumph of Region: A View of the World from Houston," *Journal of American History*, 101:1 (June 2014), 122–147.

Block, Sharon, "Doing More with Digitization," *Commonplace: The Journal of early American Life*, 6:2 (2006), http://commonplace.online/article/doing-more-with-digitization/.

Brysbaert, Marc et al. "Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing." *Frontiers in Psychology* vol. 2:27. 2 Mar. 2011, doi:10.3389/fpsyg.2011.00027.

Chiron, G., Doucet, A., Coustaty, M., Visani, M., & Moreux, J. P., Impact of OCR errors on the use of digital libraries: towards a better access to information. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (1–4). IEEE.https://ieeexplore.ieee.org/abstract/document/7991582.

Cordell, Ryan and David Smith, *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (2017), http://viraltexts.org.

Downs, Gregory and Scott Nesbit, *Mapping Occupation: Force, Freedom, and the Army in Reconstruction*, published 2015, http://mappingoccupation.org.

Gold, Matthew, et al., "Forum: Text Analysis at Scale," in *Debates in the Digital Humanities* 2016 (University of Minnesota Press, 2016), 525–568, http://dhdebates.gc.cuny.edu/debates/text/93.

Hill, Mark J., et Simon Hengchen, "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study," *Digital Scholarship in the Humanities*, vol. 34, no. 4, Oxford Academic, December 2019, 825–843, doi: 10.1093/llc/fqz024.

Hitchcock, Tim and William J. Turkel, "The Old Bailey Proceedings, 1674–1913: Text Mining for Evidence of Court Behavior," *Law and History Review* 34:4 (August 2016): 1–27, doi:10.1017/S0738248016000304.

Jockers, Matthew, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013).

Jockers, Matthew and David Mimno, "Significant Themes in 19th-Century Literature." *Poetics* 41.6 (2013): 750–769.

Journal of Digital Humanities (Winter 2012), http://journalofdigitalhumanities.org/2-1/.

Kettunen, Kimmo, et al., "Old Content and Modern Tools: Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910," *Digital Humanities Quarterly* 11(3) (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html.

Michel, Jean-Baptiste et al., "Quantitative analy-sis of culture using millions of digitized books," *Science* (New York, N.Y.) vol. 331,6014 (2011): 176–182, doi:10.1126/science.1199644.

Nelson, Robert. "Mining the Dispatch," (revised November 2020), https://dsl.richmond.edu/dispatch/.

Nesbit, Scott and Edward Ayers, *Visualizing Emancipation*, published 2013, https://dsl.richmond.edu/emancipation/.

Newman, D.J., and Sharon Block, "Probabilistic topic decomposition of an eighteenth-century American newspaper," *Journal of the American Society for Information Science and Technology* 57(6) (2006): 753–767, doi:10.1002/asi.20342.

Ruokolainen, Teemu and Kimmo Kettunen, "Name the Name: Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data," Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (October 2020): 137–156, http://ceur-ws.org/Vol-2612/paper10.pdf.

Sinclair, Stéfan and Geoffrey Rockwell, "Text Analysis and Visualization: Making Meaning Count," in *A New Companion to the Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth (Wiley Blackwell, 2016), 274–290.

Smith, David, et al., "Detecting and Modeling Local Text Reuse," published in the Proceedings of IEEE/ACM Joint Conference on Digital Libraries (IEEE Computer Society Press, 2014), https://www.ccs.neu.edu/home/dasmith/infect-dl-2014.pdf.

Sunghwan Mac Kim and Steve Cassidy, "Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers," in Proceedings of Australasian Language Technology Association Workshop (2015), 57–65, https://www.aclweb.org/anthology/U15-1007.pdf.

Tang, Jian, et al., "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis," Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. (http://proceedings.mlr.press/v32/tang14.pdf).

Underwood, Ted, "Seven Ways Humanists Are Using Computers to Understand Text. The Stone and the Shell," June 4, 2015, https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/.

van Strien, Daniel, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza, "Assessing the Impact of OCR Quality on Downstream NLP Tasks," in ICAART (1), 484–496, 2020. https://www.staff.universiteitleiden.nl/binaries/content/assets/governance-and-global-affairs/isga/artidigh_2020_7_cr.pdf.

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea, "Topic Modeling on Historical Newspapers," conference proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011), June 2011, 96–104.

Irene Amstutz, Martin Reisacher, Elias Kreyenbühl

Von der analogen Sammlung zur digitalen Forschungsinfrastruktur

Die IIIF APIs als digitaler Kompass für die Überführung der historischen Zeitungsausschnittsammlung des SWA

Abstract: The Swiss Economic Archives (SWA) owns a collection of newspaper clippings on companies, persons, and factual topics related to the Swiss economy, dating back to the 1850s. The 2.7 million clippings represent a well-used, central source collection for economic and social history. Already in 2012, a shift from analogue to digital clipping took place. Since then over 180.000 digital-born articles have been added and made available online as PDFs. Starting 2018, as part of a still ongoing project, about half of the analogue newspaper clippings have been digitized and OCRised. This combined digital collection is accessible to Swiss researchers (upon authentication due to copyright restrictions).

The retrodigitisation project marked the move from digital publication to digital infrastructure based on the International Image Interoperability Framework (IIIF) standard and its Application Programming Interface (APIs) for images, full text, authentication and, last but not least, presentation. Thanks to IIIF and compliant applications it has been possible to combine the benefits of digital and analogue collections and provide innovative and flexible access to over 800.000 clipped articles.

However, building and maintaining a digital infrastructure brings new challenges. When using interoperable standards, persistence, reliability, sustainability and transparency must be ensured. We recognise this responsibility and look forward to explore these challenges in depth with the researchers who use this infrastructure.

Keywords: digitised newspapers clippings, IIIF, interface design

1 Einleitung

Am Anfang steht die etwas altbacken wirkende Sammlung von 2,7 Millionen Zeitungsausschnitten, die, in grauen Schachteln liegend, vor Ort genutzt werden kann. Am Schluss eine Infrastruktur, die mittels IIIF APIs (Application Programming Interface) auch für Laien nutzbar ist, weitgehende Interoperabilität der Daten schafft und die Möglichkeit, sich seine eigenen virtuellen themati-

https://doi.org/10.1515/9783110729214-004

schen Sammlungen nachhaltig nutzbar herzustellen. Von den Zwischenschritten dieses langen Wegs und den zukünftigen Herausforderungen, die damit verbunden sind, handelt der vorliegende Artikel.

Das Schweizerische Wirtschaftsarchiv SWA verfügt über eine Zeitungsausschnittsammlung zu Firmen, Personen und Sachthemen der Schweizer Wirtschaft, die bis in die 1850er-Jahre zurückreicht. Die 2,7 Millionen Ausschnitte stellen eine gut genutzte, zentrale Quellensammlung dar. Die Artikel sind thematisch geordnet und erschlossen: Im Rahmen eines noch laufenden Projekts wurde die Hälfte der Zeitungsausschnitte digitalisiert und im Volltext auf einer innovativen IIIF-basierten Oberfläche zur Verfügung gestellt, die neue Einstiege in die Sammlung bietet.[1]

2 Die analoge Sammlung: Entstehung und Nutzung

Seit seiner Gründung 1910 dokumentiert das Schweizerische Wirtschaftsarchiv SWA die Wirtschaft. Dokumentieren bedeutet im Wesentlichen das gezielte Sammeln von meist publizierten Informationen über einen Gegenstand. Eine andere Art der Quellensicherung stellt das Archivieren dar. Beim Archivieren werden von einer Person oder Organisation selbst hergestellte Unterlagen von einer Archivinstitution übernommen. Die Unterlagen erfahren anlässlich der Übergabe ans Archiv eine Zweckänderung, vom aktiven Gebrauch zur Erfüllung einer Aufgabe hin zum Nutzen für die historische Forschung. Für Forschende ist die Unterscheidung wichtig, denn es spricht eine andere Perspektive aus dem Quellenmaterial. Der Unterschied fließt in die Quellenkritik ein. Die Sicherung von Archiven der privaten Wirtschaft ist in der Schweiz nicht gesetzlich geregelt und deshalb freiwillig. Firmen archivieren in unterschiedlichem Ausmaß: Dies reicht von der Nullarchivierung bis zu einigen professionell geführten und umfangreichen Firmenarchiven. Weil schon zur Gründung des SWA klar war, dass deshalb keine systematische und nur eine lückenhafte Überlieferung im Wirtschaftsbereich entstehen würde, legte das SWA eine Wirtschaftsdokumentation an. Diese dokumentiert bis heute die Wirtschaft in systematischer Weise in Form von einzelnen Dokumentensammlungen: Im Blick steht die gesamte Schweizer Wirtschaft, gesammelt wird Material

1 https://ub-zas.ub.unibas.ch (zugänglich momentan nur aus dem IP Range der Uni Basel bzw. zukünftig aus dem Schweizer Hochschulnetz).

zu wirtschaftlichen Sachthemen, über Firmen- und Verbände sowie wirtschaftsrelevante Personen.

Eine Säule der Wirtschaftsdokumentation besteht darin, die wesentlichen Tages- und Wochenzeitungen der Schweiz hinsichtlich wirtschafts- und wirtschaftspolitisch relevanter Artikel auszuwerten.[2] Die Auswahl erfolgt intellektuell, selektiv und auf der fundierten Kenntnis der gesamten Sammlung. Vor dem digitalen Zeitalter wurden die Artikel ausgeschnitten, aufgeklebt und mit Metadaten versehen (Zeitungsname, Ausgabe, Signatur). Jeder Artikel wurde inhaltlich erschlossen, indem er in einer thematischen Dokumentensammlung, einer Dokumentensammlung über Firmen und Organisationen oder über eine Person abgelegt wurde. Die Titel der Dokumentensammlungen können als Schlagworte betrachtet und verwendet werden, welche die Sammlungen in einer qualitativ hochwertigen Weise erschließen. In der mehrsprachigen Medienlandschaft der Schweiz wird hiermit auch eine sprachübergreifende Erschließung nutzbar gemacht. Entstanden ist in den vergangenen mehr als hundert Jahren eine herausragende Sammlung von ca. 35.000 Dokumentensammlungen, in denen ca. 2,7 Millionen Print-Zeitungsausschnitte des Zeitraums von 1850 bis 2012 liegen und die digital weitergeführt wird (siehe unten).

Darüber hinaus enthalten die Dokumentensammlungen „Graue Literatur", also nicht von Verlagen publizierte Schriften wie Expertenberichte, Studien, Statuten, Jahresberichte etc. Zudem liegen darin auch in Verlagen publizierte Veröffentlichungen, wie etwa Festschriften über Firmen. Heute werden die Dokumentensammlungen zu nahezu 100% mit digitalen Publikationen ergänzt, welche online zur Verfügung gestellt werden, lokal abgespeichert, langfristig archiviert und vom SWA dauerhaft online vermittelt werden. Die historischen Dokumentensammlungen bestehen aus Schachteln, in denen in Mappen die einzelnen Teile der Dokumentensammlung strukturiert sind. Eine Dokumentensammlung kann einen Umfang von einer Mappe bis hin zu unzähligen Schachteln aufweisen.[3] Die Schachteln werden vor Ort im Sonderlesesaal des SWA zur Einsicht zur Verfügung gestellt (siehe Abb. 1). Die Mappen mit den historischen Zeitungsausschnitten können ausschließlich vor Ort konsultiert werden. Katalogisierte Einzelschriften können ausgeliehen werden.

2 Für weitergehende Informationen zu den ausgewerteten Zeitungstiteln, der Struktur und Erschließung der Sammlung etc. siehe Rechercheportal Zeitungsausschnitte – Informationen | Universitätsbibliothek (https://ub.unibas.ch/de/historische-bestaende/wirtschaftsdokumentation/rechercheportal-zeitungsausschnitte-informationen, Zugriff am 29.06.2022).

3 So besteht etwa die Zeitungsausschnitte-Sammlung zum „Münz- und Währungswesen im Allgemeinen" aus über 8.000 Artikeln.

Abb. 1: Die analoge Form der Zeitungsausschnittsammlung und die Schachteln, die deren Struktur abbilden.
Copyright: Universitätsbibliothek Basel, SWA.

Die Dokumentensammlungen werden stark genutzt und erfreuen sich großer Beliebtheit. Forschende unterschiedlicher Fächer wie Geschichte und Wirtschaft, aber auch Medienwissenschaften, Politologie, Soziologie, Geografie, Volkskunde, Gender Studies u. a. sind daran interessiert. Neben Wirtschaftsfragen im engeren Sinn lassen sich auch Fragen zu Konsum, Sozialpolitik, Verkehr, Bildung, Raumplanung, Technik, Architektur, Gleichstellung etc. bearbeiten. Die Zugänglichkeit ist niederschwellig, da es sich um gedrucktes Material ohne schwer entzifferbare Handschriften handelt. Zudem liegt die Information häufig in gut aufbereiteter Form vor (Überblicke, Zusammenfassungen, Berichte, Statistiken). Man kann sich anhand der Dokumentensammlungen in kurzer Zeit einen guten Überblick über die Entwicklung einer Firma oder eines Themas aneignen. Gerade die Zeitungsausschnitte bilden den öffentlichen Diskurs zum jeweiligen Thema ideal ab. Der teilweise lange Zeitraum der Berichterstattung pro Thema ist ein weiterer Pluspunkt. Eine umfangreiche Dokumentensammlung stellt aber auch große Anforderungen daran, den Überblick zu halten. Das analoge Material kann nur zu den Öffnungszeiten eingesehen werden. Reproduktionen müssen extra hergestellt werden.

In der analogen Arbeitsweise musste jedes Dokument zwingend in eine der Dokumentensammlungen abgelegt werden, auch wenn vom Thema her vielleicht zwei oder mehr Dokumentensammlungen dafür in Frage gekommen wären. Zum Beispiel musste der Zeitungsausschnitt, in dem Roche-CEO Severin Schwan über die Entwicklung eines Medikaments von Roche berichtete und gleichzeitig zu seiner Person befragt wurde, entweder in die Dokumentensammlung über die Firma Roche oder in diejenige zu seiner Person gelegt werden. Arbeitet man an einem Thema, muss man sich überlegen, welche Dokumentensammlungen dazu Material bieten könnten. Die Titel der Dokumentensammlungen werden zudem mit der

Zeit historisch und das Thema kann seine Benennung über die Jahre verändern. So wurde beispielsweise eine Dokumentensammlung über „Reklamewesen" angelegt, das heute „Werbung" oder sogar „PR" heißt oder „Fremdenverkehr" wurde zu „Tourismus".[4] Übrigens war die Veränderung der Begriffe auch immer ein Argument dafür, bei der intellektuellen Auswahl der Artikel zu bleiben. So landen auch Zeitungsartikel in einer Dokumentensammlung, welche den Begriff, der für den Titel der Dokumentensammlung verwendet wird, nicht im Inhalt tragen.

3 Erste Schritte in Richtung digitale Transformation

2012 begann mit dem Projekt E-ZAS (Elektronische Zeitungsausschnitt-Sammlung) eine neue Ära für die Zeitungsausschnitt-Sammlung. Die Artikel werden mittels einer Clipping-Software, die für die kommerzielle Medienbeobachtung entwickelt wurde, aus den als E-Paper zur Verfügung gestellten Zeitungen nach einer halbautomatischen Layout-Erkennung digital ausgeschnitten und mit Metadaten versehen. Im Gegensatz zur Print-Sammlung werden zusätzliche Metadaten erfasst (neu auch Autor, Titel des Artikels, Seitenzahl). Die Artikel stehen in einem Portal online zur Verfügung (siehe Abb. 2). Jeder Artikel wird mit einem Schlagwort des Standard Thesaurus Wirtschaft versehen und damit einer digitalen

4 Dieser Veränderung wurde in der Erschließung in den vergangenen 15 Jahren Rechnung getragen, indem Ontologien und Linked Open Data eingesetzt werden. Die Dokumentensammlungen zu Personen sowie über Firmen und Organisationen werden mit dem Namen betitelt. Namenswechsel von Firmen und weiteren Organisationen werden nachgeführt. Firmen und Organisationen sind zudem mit den Normdaten der GND (Gemeinsame Normdatei) verknüpft. Die Dokumentensammlungen zu Wirtschaftsthemen trugen bis 2004 hausintern gewählte, historische Titel, z. B. „Gewerbe- und Fabrikinspektionswesen". 2005 fand eine Revision der Titel statt: Die Anzahl der weiter zu pflegenden Dokumentensammlungen wurde von 3.500 auf 1.200 reduziert; vor allem auf eine starke geografische Gliederung wurde verzichtet. Die neu verwendeten Titel bzw. Begriffe wurden in Anlehnung an den Standard Thesaurus Wirtschaft STW vergeben. Der STW stellt zudem eine hierarchische Gliederung zur Verfügung (Volkswirtschaft, Betriebswirtschaft etc.), welche für die Implementierung einer systematischen Suche genutzt wird. Die Dokumentensammlungen mit den historischen Benennungen wurden mit den neuen STW-basierten Titeln der Dokumentensammlungen verknüpft. So gelangt man in der Suche mit einem STW-Schlagwort zur historischen Dokumentensammlung. Oft führt ein STW-Schlagwort zu mehreren historischen Dokumentensammlungstiteln. Seit 2020 werden statt der adaptierten STW-Begriffe die Begriffe des originalen STW verwendet (als linked open data-Verknüpfung). Alle früher benutzten Begriffe bzw. Titel von Dokumentensammlungen werden bei einer Suche als Synonyme mitberücksichtigt.

Dokumentensammlung zugeordnet (es wird dasselbe Schlagwort wie in der analogen Sammlung verwendet, siehe dazu auch Fussnote 4). Nun können pro Artikel mehrere Schlagworte vergeben werden. Somit wird der Artikel in verschiedenen E-Dokumentensammlungen auffindbar. Auch ist der gesamte digital gesammelte Bestand texterkannt und kann im Volltext durchsucht werden. Dies erschließt den Bestand auf völlig neue Art und Weise, worauf weiter unten näher eingegangen wird. Aktuell umfasst der Bestand an „born digital" Zeitungsausschnitten ca. 180.000 Artikel. Pro Jahr kommen zwischen 25.000 bis 30.000 Artikel neu dazu.

Ende 2020 wurde die Clipping-Software abgelöst. Das SWA arbeitet seit 2021 mit der Schweizerischen Mediendatenbank SMD, die alle Artikel der teilnehmenden Zeitungsverlage enthält und Journalisten und Journalistinnen als Arbeitsinstrument dient. Die Artikel liegen bereits digital, texterkannt, layouterkannt und metadatiert vor. Auf diese Weise kann von den dort vorhandenen („Original"-)Daten und Metadaten bruchlos profitiert werden. Das mühsame „Clippen", also Schneiden der Artikel, und das nachträglich händische Metadatieren entfallen. Die zur Verfügung stehende Software ist deutlich ergonomischer in der Bedienung. Wie bis anhin werden die Artikel lokal gespeichert und lokal langfristig archiviert.

Und ein weiterer Automatisierungsschritt ist in Arbeit. In einem Projekt der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) Winterthur wurde mit Hilfe von Maschinellem Lernen (Deep Learning) ein Algorithmus entwickelt, der die intellektuelle Beschlagwortung der einzelnen Zeitungsartikel unterstützt. Das vorhandene digitale Textmaterial der mit der Clipping-Software bearbeiteten Artikel wurde in Beziehung zu den verwendeten Schlagworten gesetzt. Der Algorithmus ist nun fähig, mit neuem Textmaterial aufgrund der vorhandenen Daten potenzielle Schlagworte vorzuschlagen.[5]

Die Online-Zurverfügungstellung barg jedoch die Herausforderung der Einhaltung des Urheberrechts. Da die neuen Artikel nicht frei im Internet zugänglich gemacht werden dürfen, wurde eine Login-Funktion programmiert.[6]

5 Siehe https://www.zhaw.ch/no_cache/de/forschung/forschungsdatenbank/projektdetail/projektid/2248 (Zugriff am 29.06.2022).

6 Unbestritten ist, dass Angehörige von Universitäten und Schulen geschützte Inhalten online nützen dürfen. Uneinigkeit herrscht darin, ob eine kontrollierte Nutzung (z. B. durch bewusste Freischaltung von Nutzerinnen und Nutzern) einzelner Artikel (nicht einer ganzen Ausgabe einer Zeitung) urheberrechtskonform sei. Das SWA hat ein Login für Angehörige von Hochschulen und Schulen programmiert (Angehörige der Universität Basel nutzen VPN), um keine Urheberrechtsverletzung zu begehen. Alle anderen müssen die Sammlung vor Ort nutzen.

Abb. 2: Das alte E-ZAS Portal, basierend auf einer kommerziellen Softwarelösung für die Medienbeobachtung.[7]

7 Siehe Fußnote 1. Eine Anmeldung um einen Zugang auf E-ZAS – Elektronische Zeitungsausschnittsammlung zur Schweizer Wirtschaft zu erhalten wird hier erklärt: https://ub-easyweb.ub.unibas.ch/de/news/details/das-zeitungsausschnitt-portal-des-swa/ (Zugriff am 29.06.2022).

Nach der Umstellung des analogen auf den digitalen Sammlungsprozess nahm das SWA ein Anschlussprojekt in Angriff: Die historische Sammlung über 2,7 Millionen Zeitungsausschnitte wird retrodigitalisiert. Drei Hauptziele werden damit verfolgt: Erstens entsteht durch die Digitalisierung eine Schutz- und Sicherungskopie. Ein Aspekt, der gerade für das fragile Zeitungspapier erheblich ist. Zweitens wird die Sammlung damit online zugänglich, digital durchsuchbar und auf neue Weise präsentierbar. Drittens wird damit der entstandene Medienbruch zur digitalen Sammlung ab 2012 geschlossen. Das Projekt läuft über mehrere Jahre und wird zu erheblichen Teilen mit Drittmitteln finanziert. Bis anhin ist die Hälfte der Artikel retrodigitalisiert worden. Auch diese Artikel haben Texterkennung durchlaufen und sind metadatiert. Damit steht nun ein großer Korpus an digitalen Zeitungsartikeln in hoher Qualität zur Verfügung. Und nun beginnt das Träumen über Volltextsuchen über das gesamte Material, über virtuelle Dokumentensammlungen, über neue Präsentations- und Auswertungsformen und über Weiternutzungen der Daten in anderen Umgebungen.

4 Aufbruch zur digitalen Forschungsinfrastruktur

4.1 Leitgedanken

Schon bei der Umstellung auf E-ZAS zeigten sich die Möglichkeiten der digitalen Transformation, war es doch nun möglich, gezielt von zuhause in den Volltexten zu recherchieren. Doch das Durchblättern größerer Treffermengen war aufwendig, musste doch jedes PDF einzeln aufgerufen werden. Interoperabilität, dass also die Artikel auch außerhalb der E-ZAS-Oberfläche nachgenutzt und eingebunden werden konnten, war nicht gegeben.

Die Entwicklungen in Richtung Infrastruktur war zu Beginn des Projekts weniger von einem detaillierten Konzept geprägt, sondern vielmehr ein Lernprozess, der besonders durch die begrenzten Projektmittel für die Präsentationskomponenten, aber auch durch die zeitliche Möglichkeit in der Entwicklung innezuhalten und zu reflektieren, geprägt war.

Und auch der Zufall spielte eine Rolle. Da durch den Umzug des bisherigen Clipping Dienstleisters nach Indien die Zusammenarbeit äußerst schwierig wurde, musste überhaupt erst eine eigenständige Lösung außerhalb des alten E-ZAS Portals gefunden werden. Know-how und Infrastruktur bzgl. Suchtechnologie war in der Universitätsbibliothek Basel (UB) als Trägerorganisation des

SWA insbesondere durch swissbib[8] vorhanden und auch ein flexibles Frontend zur Präsentation von heterogenen Daten bereits in Planung.

Nur für Digitalisate existierte keine Lösung, da diese bislang auf den großen Schweizer Plattformen e-rara[9], e-manuscripta[10] und e-codices[11] publiziert wurden. Mit IIIF (International Image Interoperability Framework) als Standard zur Präsentation von Digitalisaten waren aber bereits erste Erfahrungen gesammelt worden. Gerade die Möglichkeiten zur Nachnutzung von IIIF-Komponenten, die nicht selbst programmiert werden mussten, wie etwa einem Viewer, boten hier aufgrund der begrenzten Projektressourcen eine interessante Perspektive.

Langsam formten sich daraus grundlegenden Ideen, was bei der Überführung dieser analogen Sammlung in den digitalen Raum und bei der Verknüpfung mit E-ZAS zu berücksichtigen wäre, um für die nächsten Jahre anschlussfähig zu bleiben. Rückwirkend betrachtet haben die folgenden Thesen die Entwicklung des ZAS-Projekts maßgeblich geprägt:

- **Infrastruktur vor Portal**: das Projekt definiert sich stark über die Interoperabilität seiner IIIF APIs und nicht primär über seine Oberfläche.
- **Nachnutzung durch Standards und Interoperabilität**: durch die Standards profitiert das Projekt von einer großen Community, die Applikationen rund um IIIF entwickelt bzw. können andererseits die ZAS-Daten von dezentralen Forschungsumgebungen genutzt werden
- **Easy2Use API**: da im IIIF Standard die API auf die Präsentation der Daten ausgerichtet ist, wird die technische API Spezifikation, gerade bei der Präsentation im Viewer, für den Endnutzer nahezu unsichtbar. Das vereinfacht auch für technische Laien die Interaktion mit der API.
- **„Virtuelle Dokumentensammlungen“**: die Möglichkeit über 800.000 Artikel zu individuellen Datensets zusammenstellen zu können, ist ein zentraler Mehrwert der spezifischen Ausformung der IIIF APIs in diesem Projekt.
- **Authentifizierung**: da es sich bei der Sammlung primär um urheberrechtlich geschützte Materialien handelt, spielt die Authentifizierung eine zentrale Rolle.

8 swissbib war bis Januar 2021 der Metakatalog und Datenhub aller Schweizer Hochschulbibliotheken: https://www.swissbib.ch (Zugriff am 29.06.2022).
9 https://www.e-rara.ch/ (Zugriff am 29.06.2022).
10 https://www.e-manuscripta.ch (Zugriff am 29.06.2022).
11 https://www.e-codices.unifr.ch/de (Zugriff am 29.06.2022).

4.1.1 Exkurs: IIIF APIs

Viele dieser Leitgedanken trägt der IIIF Standard schon genuin in sich. Das Spezifikum dieses Projekts ist aber die hohe Flexibilität in der Generierung der IIIF-Datensets, wodurch diese Aspekte noch verstärkt werden. IIIF ist ein Set von API-Spezifikationen zur Präsentation von Bild und in Zukunft auch AV- bzw. 3D-Objekten.[12] Rund um den Standard hat sich in den letzten Jahren eine starke Community gebildet, die von zentralen Akteuren im Bibliotheks- und Archivbereich getragen wird. So unterstützen in der Schweiz und weltweit mehrere hundert Institutionen und Portale diesen Standard und stellen zig Millionen Objekte interoperabel zur Verfügung.

IIIF standardisiert diverse APIs, deren Bedeutung für das Projekt kurz erläutert werden soll. Eine Besonderheit ist auch, dass die APIs erst als stabil gelten, sobald es zumindest zwei Umsetzungen gibt. IIIF ist also ein funktionaler Standard. Folgend werden die zentralen APIs und deren Bedeutung für das Projekt kurz definiert. So existiert eine **IIIF Image API**[13], die von einer Vielzahl an Bildservern unterstützt wird.[14] Darüber können performant unterschiedliche Größen von Bildern ausgegeben werden, wie etwa Thumbnails für Voransichten. Es kann aber auch bis ins letzte Detail gezoomt werden. Eine **IIIF Search API**[15] ermöglicht es, Volltext durchsuchbar zu machen und Text und Bild im Viewer durch Highlighting miteinander zu verknüpfen. Durch die **IIIF Authentication API**[16] können selbst in dezentralen Tools von berechtigten Personen Inhalte angezeigt werden. Dieser Standard kann mit unterschiedlichen Authentifizierungskomponenten gekoppelt werden. All diese APIs inkorporiert die **IIIF Presentation API**[17], die diese Schnittstellen noch mit Metadaten und Strukturinformationen zum digitalen Objekt verknüpft, damit diese in einem Viewer angezeigt werden können.

Zentral ist, dass diese Standards alle interoperabel gedacht sind, sprich, dass die Daten, insbesondere die Bilddaten, auch in anderen Kontexten und von anderen Institutionen weltweit eingebunden werden können.

---

12 Für weitere Informationen zu IIIF für Einsteiger, aber auch Experten siehe https://iiif.io/community/#how-to-get-involved-1 (Zugriff am 29.06.2022).

13 https://iiif.io/api/image/3.0/ (Zugriff am 29.06.2022).

14 Im Projekt wird der vom DHLab in Basel entwickelte IIIF Bildserver Sipi genutzt, wobei hier von der engen Zusammenarbeit mit dem DHLab profitiert wird: https://github.com/dasch-swiss/sipi (Zugriff am 29.06.2022).

15 https://iiif.io/api/search/1.0/ (Zugriff am 29.06.2022).

16 https://iiif.io/api/auth/1.0/ (Zugriff am 29.06.2022).

17 https://iiif.io/api/presentation/3.0/ (Zugriff am 29.06.2022).

4.1.2 Infrastruktur vor Portal

Durch den Fokus auf IIIF steht also auch weniger die Präsentation im Portal, sondern die Interoperabilität der Daten im Mittelpunkt des Projekts. Soll es doch der Kreativität der Forschenden überlassen sein, wie die Daten genutzt und in Forschungstools eingebunden werden. Daher baut die Präsentation der Digitalisate im Portal auch auf Schnittstellen auf, die nach außen zugänglich sind. Im Gegensatz zu anderen Portalen ist die IIIF Presentation API nicht nur ein weiteres Exportformat, sondern auch die Schnittstelle für das Portal zur Präsentation der digitalen Objekte.

Für die wenigsten Forschenden ist eine einzige Datenquelle ausreichend. Beschäftigt man sich etwa mit der Schweizer Wirtschaftsgeschichte, wird auch die Perspektive von außen, etwa in internationalen Medien, interessant sein. Daher ist es zentral, dass Daten aus unterschiedlichen Infrastrukturen in einem standardisierten Format wie IIIF zur Integration und Nachnutzung vorliegen, um in externen Tools zusammengeführt werden zu können. Die Datensammlung des ZAS-Portals versteht sich als ein Baustein einer digitalen Forschungsinfrastruktur.

4.1.3 Nachnutzung durch Standards und Interoperabilität

Der im Portal für die Präsentation genutzte Viewer ist eine IIIF Applikation.[18] Für die Zeitungsausschnittsammlung wurden mehrere Viewer getestet und die Entscheidung fiel auf den Universal Viewer[19], da dieser viele wichtige Funktionalitäten wie die integrierte Volltextsuche am besten abdeckt.

In der Austauschbarkeit der Komponenten zeigt sich aber gerade das Potenzial der digitalen Nachhaltigkeit der IIIF-Schnittstellen. So kann etwa der funktional sehr beeindruckende, aber optisch möglicherweise schon etwas veraltet wirkende Universal-Viewer einfach durch einen neuen Viewer[20] ersetzt

18 Dieser wurde unter anderem von der Agentur digirati für die Wellcome Library, die British Library und die National Library of Wales (weiter)entwickelt. Durch das einfache Einbinden ist es möglich, auf all den hier eingeflossenen Erfahrungen dieser Institutionen zur Präsentation von Digitalisaten aufzubauen.

19 Siehe https://universalviewer.io/ (Zugriff am 29.06.2022).

20 So wird etwa bald eine neue Version des Mirador-Viewers(https://projectmirador.org/, Zugriff am 29.06.2022) veröffentlicht, der auch gemeinsam von einer Vielzahl an großen Kultureinrichtungen entwickelt wird und weitere Funktionalitäten zur Präsentation von Digitalisaten enthält.

werden und damit die zentrale Präsentationskomponente im Portal aktualisiert werden.

Doch die Einbindung der Daten in einen Viewer ist nur der augenscheinlichste Anwendungsfall. Genauso gut können Forschungstools, die auf externen Servern laufen, die IIIF-Ressourcen aus dem ZAS-Projekt einbinden. Für Forschungsprojekte kann daher IIIF die Möglichkeit bieten, dass Tools nicht neu programmiert werden müssen, sondern nachgenutzt werden können. Durch die Interoperabilität steht auch weniger im Mittelpunkt, dass ein Tool alle Anforderungen erfüllen muss, sondern je nach Anwendungsfall können unterschiedliche, auf IIIF basierende Applikationen genutzt werden.

Unter anderem gibt es zum Beispiel recogito, eine Webanwendung, welche den DH Award 2018 gewonnen hat und Forschende bei der Annotation und Auswertung von Karten, die in IIIF verfügbar sind, unterstützt.[21] Möchte man seine Daten hingegen in einer kurzen Ausstellung präsentieren, würde man wohl auf storiiies[22] von cogapp zurückgreifen oder auf Inseri[23], eine kollaborative Arbeits- und Präsentations-Umgebung.

Es muss aber erwähnt werden, dass diese IIIF-Applikations-Infrastruktur für Forschende bis auf Einzelbeispiele noch nicht so einfach zu nutzen ist, wie dies etwa bei den Präsentationskomponenten (Viewern) der Fall ist. Betrachtet man allerdings die Dynamik, mit der IIIF voranschreitet, dürfte dies nur eine Frage der Zeit sein, bis eine niederschwellige Nutzung gang und gäbe ist.

Hier kann auch eine gewisse Trennlinie zwischen Daten-Infrastruktur und Forschenden gesehen werden. So sind Bibliotheken und Archive oft auch im digitalen Raum die Infrastruktur, über welche Quellen interoperabel zur Verfügung gestellt werden. Die Tools bzw. dadurch auch die Methoden, mit denen die Daten bearbeitet und präsentiert werden, könnten allerdings primär von den Forschenden oder Forschungs-Projekten kommen.

Wobei diese Trennlinie zwischen Infrastruktur und Tools/Methoden früher möglicherweise einfacher zu ziehen war. Heute sind die Methoden abhängig davon, welche Möglichkeiten die Schnittstellen bieten. IIIF ist hier ein erster Ansatz, aber es wird wahrscheinlich nicht die letzte Schnittstelle bleiben, da damit nur Inhalte in einer bestimmten Form abfragbar sind, in deren Mittelpunkt das digitale Objekt steht.

21 https://recogito.pelagios.org/ (Zugriff am 29.06.2022).
22 https://storiiies.cogapp.com/ (Zugriff am 29.06.2022).
23 https://github.com/nie-ine/inseri (Zugriff am 29.06.2022).

4.1.4 Easy2Use API

Eine große Stärke von IIIF ist, dass es ein Standard zur Präsentation von Daten ist. Das bedeutet, dass die API-Komponente nicht abstrakt bleibt, sondern der Nutzer durch die Darstellung der IIIF-Daten in einem Viewer sieht, welches Datenset er zusammengestellt hat. In diesem kann er auch im Volltext mit Highlighting suchen. Ist das Datenset für seine Forschung relevant, ist es ausreichend, die Daten per Link in ein anderes IIIF-fähiges Tool einzubinden.

4.2 Konkrete Umsetzung

Was bedeuten diese Leitgedanken aber nun konkret für die digitale Aufbereitung der Sammlung? Im Mittelpunkt stand die Frage, wie man das Konzept der analogen Dokumentensammlung in den digitalen Raum übertragen und um die digitalen Möglichkeiten erweitern konnte.

Der Schritt zur elektronischen Zeitungsausschnittsammlung 2012 war ein erster wichtiger Schritt: Der Zugriff von Zuhause, die Zuordnung von Artikeln zu mehreren Dokumentensammlungen und die so wichtige Volltextsuche war nun möglich. Im Gegensatz zu den analogen Mappen konnte man aber nicht mehr in einer Dokumentensammlung rasch von einem Artikel zum nächsten blättern und sich innerhalb kurzer Zeit einen Überblick verschaffen.

Schon aufgrund der Möglichkeit, dass in E-ZAS die Artikel den Dokumentensammlungen nicht mehr eindeutig zugeordnet werden mussten, aber das Dossier-Prinzip einen zentralen Charakter der Sammlung bildete, musste eine Lösung gefunden werden, die weiterhin die Abbildung der Dokumentensammlungen sowie die gemeinsame Darstellung der retrodigitalisierten und digital born Inhalte ermöglichte.

Es war auch klar, dass die starre Zuordnung zu einem Dossier nicht die Forschungsrealität abbilden konnte. Können die Forschungsfragen doch quer zu den Dokumentensammlungen verlaufen, auf Volltextsuchen beruhen oder nur einen bestimmten Zeitraum betreffen. Perspektivisch betrachtet lässt es sich auch nicht ausschließen, dass etwa in Zukunft maschinelle Verfahren dynamisch Dokumentensammlungen nach neuen Kriterien bilden könnten.

4.2.1 Virtuelle Dokumentensammlungen

Hier setzt das Konzept der virtuellen Dokumentensammlungen an, das zumindest unserem Wissen nach hinsichtlich der Flexibilität ein Alleinstellungsmerkmal des

Projekts ist.[24] Die Dokumentensammlungen können über die IIIF Presentation API des Portals komplett dynamisch gebildet werden. Alle Suchmöglichkeiten im Portal können dafür genutzt werden, sich eine virtuelle Dokumentensammlung zusammenzustellen und gleichzeitig persistent verfügbar zu machen. Ob nun die virtuelle Dokumentensammlung anhand der ursprünglichen Zuordnung zum Thesaurus durch die Dokumentarin bzw. den Dokumentar erfolgte, die Dokumentensammlungen zukünftig von einer KI gebildet wird oder es eine einfache Volltextsuche im Portal ist, spielt dabei keine Rolle.

Die virtuellen Dokumentensammlungen stellen also den Versuch dar, die Vorteile aus der digitalen und analogen Welt zu vereinen: einerseits rasch durch große Datenmengen blättern zu können, andererseits aber nicht länger durch die Grenzen der traditionellen Dokumentensammlungen limitiert zu werden und die Daten auch über die API in andere Umgebungen „mitnehmen" und einbinden zu können (siehe Abb. 3).

Diese Präsentationsform eignet sich, die Treffer aus den diversen Sucheinstiegen beispielsweise über die Volltexte, aber auch über den Standard Thesaurus Wirtschaft (STW) zu kombinieren. Der große Vorteil beim Einstieg über die Themen-Deskriptoren des STW liegt etwa darin, dass auch alternative Suchformen gefunden werden.

Wie schon erwähnt ist etwa der Begriff „Tourismus" ein relativ neuer Begriff. Sucht man danach in den 1930er-Jahren, erhält man fast keine Treffer. Hingegen sind dem Sachdeskriptor „Tourismus" in den 1930er-Jahren 130 Artikel zugeordnet und ermöglichen einen viel breiteren Einstieg in dieses Thema.

Volltextsuche und Deskriptoren können aber gewinnbringend gemeinsam in der Suche genutzt werde. Gibt man etwa den Begriff „Fremdenverkehr" in der Volltextsuche ein, erhält man eine Vielzahl an Treffern. Überraschenderweise sind die meisten aber nicht dem Deskriptor W13 „Tourismus" zugeordnet, sondern W11 „Gastgewerbe" (mit der Unterkategorie „Hotellerie"), wodurch man einen weiteren Sucheinstieg erhält, ohne die Struktur des STW vorab bereits bis ins letzte Detail kennen zu müssen. So kann durch die Nutzung von Volltextsuche und STW die Suche immer weiter verfeinert oder auch erweitert werden und ermöglicht auch eine einfachere Nutzung des Thesaurus für Laien (Abb. 4).

24 Ähnliche Zugänge auf einer IIIF-Meta-Ebene existieren etwa von Intranda im goobi Viewer (https://iiif.io/news/2020/03/31/newsletter, Zugriff am 29.06.2022), ÖNB Labs (https://labs.onb.ac.at/en/tool/sacha, Zugriff am 29.06.2022) und Biblissima (https://iiif.biblissima.fr/collections, Zugriff am 29.06.2022). Bei diesen Beispielen müssen aber jeweils manuell mehrere Manifeste zu einer IIIF-Collection hinzugefügt werden.

Abb. 3: Ein Ausschnitt, wie 50 Meter aneinandergereihte Artikel in einem IIIF Viewer aussehen und einen raschen Überblick über große Datenmengen ermöglichen.
Copyright: Universitätsbibliothek Basel, SWA

4.2.2 Digitaler Zugriff

Bislang war der Zugriff primär auf den IP-Range der Uni Basel beschränkt.[25] Das Schweizerische Urheberrecht erlaubt aber sogar eine kontrollierte Nutzung für den allgemeinen universitären Gebrauch. Auch hier kann Swiss Edu-ID[26], ein zentrales Authentifizierungssystem für den Schweizer Hochschulbereich, eingesetzt werden. Allerdings muss diese Anbindung an IIIF erst umgesetzt werden.

Dabei zeigen sich aber auch die Herausforderungen und Abhängigkeiten für digitale Infrastruktur. Änderungen im Browser Chrome könnten nämlich in Zukunft dazu führen, dass die reibungslose Interoperabilität, also das Einbinden der Daten in andere Applikation nicht mehr so einfach funktioniert, sofern eine Authentifizierung notwendig ist.[27]

25 Auch anderen Nutzenden kann der Zugang gewährt werden. Sie müssen sich anmelden und frei geschaltet werden. Diese Lösung muss für das neue Portal ebenfalls erst integriert werden.

26 Siehe https://projects.switch.ch/eduid/ (Zugriff am 29.06.2022).

27 Siehe dazu etwa der Blog-Artikel Access Control, Interoperability and Cultural Heritage von Tom Crane vom 31.1.2021, siehe: https://tom-crane.medium.com/access-control-interoperability-and-cultural-heritage-4d3c6241f63f (Zugriff am 29.06.2022).

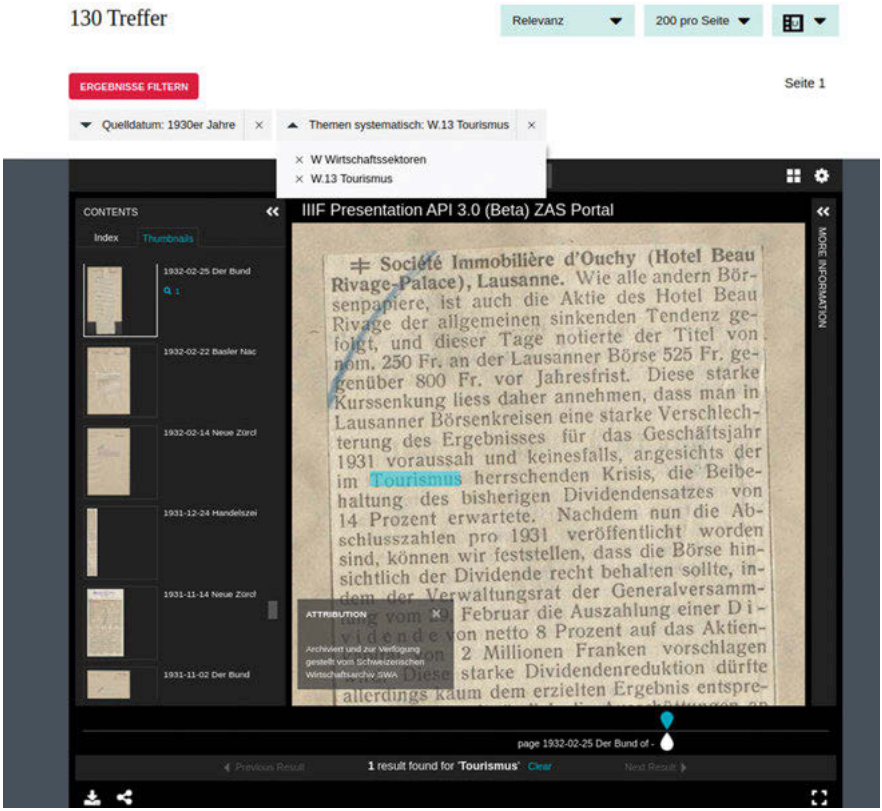


Abb. 4: Über die Freitextsuche und die Facetten des Portals können individuelle Subsets der insgesamt über 800.000 Artikel (retrodigitalisiert und digital born) im Viewer angezeigt und im Volltext durchsucht werden. Hier wurde eine Auswahl über den STW und die zeitliche Zuordnung getroffen und innerhalb des Subsets eine Volltextsuche durchgeführt. Copyright: Universitätsbibliothek Basel, SWA.

4.3 Herausforderungen

Das ZAS-Portal bzw. vielmehr seine API versteht sich nur als kleines Puzzleteil einer dezentralen IIIF Infrastruktur, die von vielen Institutionen wechselseitig bereitgestellt und genutzt werden kann. Vieles ist dabei jedoch noch in Bewegung. Auch die Herausforderungen und Versprechungen, die damit verknüpft sind, werden uns sicherlich noch die nächsten Jahre begleiten und sind nicht mit der Publikation der Daten gelöst.

4.3.1 Persistenz/Versionierung

Im Gegensatz zum Zettelkatalog, als eine mögliche Form des verlängerten Gedächtnisses des Forschenden, sind es in einer interoperablen Infrastruktur oft nur mehr flüchtige Links,[28] die Quellen repräsentieren. Gerade wenn wir von Interoperabilität sprechen, ist daher die Persistenz die erste Bedingung, die erfüllt werden muss, um zu garantieren, dass dezentrale Systeme langfristig funktionieren können.

Oder um es konkret zu formulieren: Wenn wir als UB Basel unser Katalogisierungssystem wechseln, müssen die IDs und Links weiterhin stabil bleiben. Aber auch auf praktischer Ebene stellen sich bei virtuellen Dokumentensammlungen Fragen hinsichtlich der Persistenz, sofern eine Sammlung noch nicht abgeschlossen oder fertig digitalisiert ist. Soll etwa ein zukünftiger Aufruf der virtuellen Dokumentensammlung „Wirtschaftskrise“ nur jene Artikel beinhalten, die zum Zeitpunkt X der Publikation digitalisiert waren oder doch alle, die zum heutigen Zeitpunkt Y verfügbar sind? Für den Forschenden haben diese zwar digital nicht existiert, analog wären diese aber vorhanden gewesen.

4.3.2 Zuverlässigkeit

Neben der Persistenz ist beim Zugriff auf externe Daten zentral, dass diese rasch ausgeliefert werden und die Schnittstellen zuverlässig ohne Unterbrechungen funktionieren. Nicht umsonst definieren sich Services von großen Anbietern durch ihre praktisch hundertprozentige Verfügbarkeit.

4.3.3 Nachhaltigkeit

Es stellt sich auch die Frage, wie man Persistenz und Zuverlässigkeit mit dem Projektcharakter vieler Initiativen verknüpfen kann. Ist es doch nicht mehr bloß das Portal, das nach der Projektlaufzeit sanft entschlummern könnte, sondern verlieren Forschende den Zugriff auf ihre Daten, wenn gewisse Infrastrukturen nicht mehr funktionieren oder nur teilweise gewartet werden würden?

---

28 Im Idealfall sind es Persistent Identifier, wie urn, dois oder ark, die durch ein Resolving-System selbst bei URL-Änderungen stabil bleiben. Aber auch diese bleiben nur bei aktiver Pflege stabil.

Aber auch hinsichtlich der notwendigen Schnittstellen für Forschende werden die Entwicklungen nicht stehen bleiben. Momentan spielt IIIF als Schnittstelle eine wichtige Rolle, um (Bild-)Daten nachnutzbar zu machen. Welche Schnittstellen und Standards benötigten aber etwa in Zukunft Forschungsprojekte, die komplexe Volltext-Analysen durchführen wollen? Werden deren Bedürfnisse in Zukunft berücksichtigt werden können oder ist es möglicherweise sogar einfach besser, die „Rohdaten" zur Verfügung zu stellen?

4.3.4 Transparenz

Transparenz gilt es auf unterschiedlichen Ebenen zu gewährleisten. In Zukunft wird nicht nur der Forschende seine digitalen Methoden zur Quellenanalyse reflektieren müssen, sondern auch die Institutionen, die die Daten zur Verfügung stellen. Es können einfache Fragen sein wie, welcher Teil der Sammlung bereits digital vorhanden ist und welche Auswirkungen das auf die Nutzung der Quellen hat. Wahrscheinlich ist, dass jene im Volltext zugänglichen Teile der Sammlung eine viel stärkere Nutzung erfahren. Die Strategie, welche Teile digitalisiert werden und die Transparenz darüber, spielt dann eine zentrale Rolle.

Aber auch Fragen hinsichtlich der Datenqualität müssen adressiert werden. Was bedeutet es etwa für einen Teil des Quellenkorpus, wenn die Fraktur-OCR bei nicht geglätteten Zeitungsartikeln schlechtere Resultate liefert? Solche Qualitätskennzahlen zur OCR sind für ZAS-Volltexte jedoch nicht vorhanden. Andere Informationen, wie die vom OCR-Prozess erkannte Sprache, die starke Auswirkungen auf die Qualität hat, existieren hingegen schon.

Dabei muss zwischen Nutzerfreundlichkeit und Transparenz abgewogen werden. So könnte der Hinweis auf die von der OCR erkannte Sprache (zum Beispiel OldGerman oder GermanNewSpelling) für den normalen Nutzer, der sich nicht mit methodischen Fragen zur OCR-Qualität auseinandersetzt, zuallererst Verwirrung stiften.

Es dürfte überhaupt eine Gratwanderung hinsichtlich der Nutzererwartungen sein, wie komplex oder intuitiv die Suchmöglichkeiten sind. Soll es sich bei diesen Infrastrukturen um Expertentools handeln, bei denen man die ganzen Möglichkeiten eines Suchservers auf einer Oberfläche abbildet oder soll die Einstiegshürde möglichst niedrig gehalten werden?

Doch nicht nur die OCR birgt ihre Geheimnisse, auch maschinelle Verfahren, seien es etwa standardisierte Prozesse für das Suchranking oder Machine-Learning zur Aufbereitung bzw. Strukturierung der Daten, beeinflussen, wie sich Forschende den Quellen annähern. Bis zu einem gewissen Punkt formen wir dadurch auch die „Forschungs-Realität" für die Wissenschaft. Reicht hier

eine Mischung aus guter alter Quellenkritik und Digital Literacy oder braucht es doch mehr? Wieviel Transparenz müssen die Institutionen an den Tag legen und wieviel können sie überhaupt mit ihren Daten und Ressourcen leisten? In der drastischsten Variante bedeutet das hinsichtlich der Quellen „program or be programmed". Die Leerstellen im Code oder die pragmatischen Entscheidungen bei der Datenaufbereitung können wohl, wenn überhaupt, nur jene mit ausreichendem technischen Grundverständnis nachvollziehen.

Es stellt sich aber die Frage, wie transparent die Aufbereitung und Selektion von Quellen früher war. Weshalb etwa eine Dokumentarin einen Artikel einer bestimmten Dokumentensammlung zugeordnet hatte, war zu einem gewissen Teil ja eine persönliche Entscheidung, die sich zudem über die Zeit gewandelt haben konnte oder durch einen personellen Wechsel Veränderung erfahren hat. Aber zumindest konnte man bis zu einem gewissen Grad das Gespräch suchen, die Algorithmen hingegen schweigen (zumindest noch).

5 Resümee

Für Forschende bietet das neue ZAS-Portal einen wertvollen Baustein in ihrer Recherche zur Schweizer Wirtschafts- und Sozialgeschichte. So wurde unlängst auf einer Veranstaltung darauf hingewiesen, dass die gesammelten Zeitungsausschnitte für Historiker von großem Wert sind, da Zeitungsartikel eine Verdichtung des Diskurses aus unterschiedlichsten Perspektiven zu einem Thema darstellen. Noch besser als in Zeitungsarchiven hat man bei Ausschnittsammlungen durch die manuelle Auswahl und Erschließung die Möglichkeit, diesen Diskurs einfach nachzuverfolgen.

Dass hier eine Zeitungsausschnittsammlung und keine Zeitung digitalisiert wurde, ist ebenfalls ein Spezifikum. Gibt es doch im deutschsprachigen Raum unseres Wissens nur wenige digitalisierte Sammlungen[29], wobei die Sammlung des SWA die einzige sein dürfte, die tatsächlich neben ihren Klassifikationskriterien im Volltext und anhand von Facetten durchsuchbar ist.

Gerade diese Suchmöglichkeiten bieten in Verknüpfung mit den IIIF Präsentationskomponenten neue Einstiegspunkte, die zusätzlich zur bisherigen Samm-

29 So existiert das Innsbrucker Zeitungsarchiv (https://iza-server.uibk.ac.at/adb/index.jsp, Zugriff am 29.06.2022) und die Pressemappe des 20. Jahrhunderts der Zentralbibliothek Wirtschaft (http://webopac.hwwa.de/pressemappe20, Zugriff am 29.06.2022). Gerade die Pressemappe des 20. Jh. zeigt auch noch das Potential für die ZAS-Sammlung hinsichtlich der Möglichkeiten der weiteren Integration von Linked Open Data (LOD).

lungsbildung verlaufen und den Forschenden viele effiziente neue Möglichkeiten zur Quellenanalyse und -sammlung und digitalen Nachnutzung geben und die Besonderheiten einer manuell kuratierten Zeitungsausschnittsammlung auch digital in den Vordergrund rücken.[30]

Die Überführung der Zeitungsausschnittsammlung in den digitalen Raum zeigt aber auch die Bedeutung der manuellen Erschließung über Thesauri und Normdaten. So streben immer mehr Projekte die Verschlagwortung mit ML-Verfahren an, um weitere Einstiegspunkte neben den Möglichkeiten (aber auch Grenzen) der Volltextsuche zu bieten. Dafür braucht es hochqualitative Trainingssets, wofür die Sammlung als Grundlage dienen kann. Andererseits bietet die Verschlagwortung für die Forschenden auch ein Korrektiv zur reinen Volltextsuche. Existiert doch durch die Zuordnung zu einer Dokumentensammlung ein zusätzlicher Sucheinstieg, über den verwandte Artikel gefunden werden können, die über eine Volltextsuche nicht auffindbar wären.

Das Aufbauen dieser Infrastruktur für Zeitungsausschnitte ist aber nur ein erster Schritt. Auch gemeinsam mit der Forschung muss noch das Potential hinsichtlich der Interoperabilität der Quellen und die Vorzüge einer offenen Forschungsinfrastruktur erprobt werden, in der Quellen online konsultiert, gesammelt, zitiert und möglicherweise schlussendlich auch im Sinne eines Digital Life Cycle Managements selbst hinzugefügt und annotiert werden können.

Die Kenntnis von archivischer Erschließung und guter wissenschaftlicher Arbeit wird dann auch durch Digital Literacy ergänzt werden müssen. Das Verständnis von Suchalgorithmen, Standards und damit auch die Auslassungen in diesem Prozess dürfte wohl für die Forschenden ein zentrales Handwerkszeug werden, um diese Angebote effizient zu nutzen und auch deren Grenzen und Möglichkeiten beurteilen zu können. So wie es früher Einführungen zum guten wissenschaftlichen Arbeiten, zur Recherche und zum Zitieren gab, könnte es in Zukunft Kurse zu Technologie-Standards geben, um diese neuen Infrastrukturen beurteilen und Quellenkritik hinsichtlich der Grenzen und Möglichkeiten üben zu können.

Gerade IIIF und die Interoperabilität der Daten ist aber auch eine Chance, in gewissen Bereichen Rechercheprozesse im digitalen Raum zu vereinfachen. Wo früher diverse Portale oder Interfaces konsultiert werden mussten, bietet IIIF die Möglichkeit, Digitalisate aus verschiedensten Quellen in einem einheitlichen Interface (etwa dem Universalviewer) anzeigen zu lassen, um sich wieder weniger auf die unterschiedlichen Oberflächen als auf die Inhalte konzentrieren zu können.

30 Einerseits werden die Artikel in einer Ausschnittssammlung dekontextualisiert, da diese aus der Zeitung ausgeschnitten werden und von den restlichen Artikeln getrennt werden, andererseits werden sie gerade kontextualisiert, da diese in den Sammlungen neben Artikel aus anderen Zeitungen zum gleichen Thema gestellt werden können.

Der Schritt in Richtung digitale Forschungsinfrastruktur bietet für Kultureinrichtungen und auch den Forschenden viele neue Möglichkeiten, doch gibt es noch immer eine Reihe an Herausforderungen. So werden wir in einigen Bereichen von den Erfahrungen anderer lernen, aber auch unsere eigenen Fehler machen müssen, um voranschreiten zu können. Daher wird es sicherlich noch eine Weile dauern, bis wir ohne schlechtes Gewissen das Beta in der Version entfernen können. Aber bei all dem darf nicht vergessen werden, dass nun eine Sammlung von momentan gut 1.000.000 der 2,7 Millionen Zeitungsausschnitte und gut 180.000 Born-Digital-Artikel im Volltext online interoperabel verfügbar ist und sich dadurch für die Forschung hoffentlich ganz neue Möglichkeiten und Perspektiven ergeben.

Wir sind daher gespannt, in welche Richtung uns der weitere Austausch und die Zusammenarbeit mit den Forschenden leiten wird und ob unsere Thesen zum Umgang mit digitaler Forschungsinfrastruktur rückblickend gehalten haben, was sie für uns momentan versprechen.

Claudia Resch

Volltextoptimierung für die historische *Wiener Zeitung*

Mit einem Anwendungsszenario aus der germanistischen Sprachgeschichte

Abstract: For several years now, archives and libraries have been working closely with information technology experts to advance the digitisation of historical newspapers and to provide full-text search capabilities on their textual transcriptions, acquired automatically by optical character recognition (OCR). In practice, there are significant differences in the quality of these transcripts, ranging from poor or even "dirty" OCR to manually enhanced, almost error-free OCR (comparable to the gold standard corpora used in linguistics). The more reliable the transcripts, the more likely it is that researchers from different disciplines will use large newspaper corpora to answer their research questions. Drawing on experience gained from the digitization of the "Wiener Zeitung", the oldest Austrian newspaper published since 1703, this chapter discusses measures that can contribute to improving the automatic character recognition rate of newspaper texts printed in Gothic script and demonstrates the impact of the transcription quality with a research question from historical linguistics, that of the gradual implementation, by the "Wiener Zeitung", of new writing standards emerging during the 18th century.

Keywords: Text recognition for Fraktur script, corpus of the "Wiener Zeitung", historical linguistics

1 Vorbemerkungen zur Volltextgenerierung historischer Zeitungen

Seit mehreren Jahren kooperieren Forschende und an Archiven oder Bibliotheken Verantwortliche mit IT-Expert*innen, um die Digitalisierung historischer Zeitungen auf nationaler und internationaler Ebene voranzutreiben. Infolgedessen lässt sich beobachten, dass ein überwiegender Teil des neu hinzukommenden digitalisierten Materials aus Zeitungs- und Zeitschriftensammlungen stammt: „newspapers, magazines and periodicals constitute the vast majority of newly digitised material" (Nicholson 2013, 64). Bei der Mittelvergabe drängen Förderinitiativen inzwischen sowohl auf die

https://doi.org/10.1515/9783110729214-005

Zurverfügungstellung von Bilddigitalisaten als auch auf die Generierung von Volltexten. Somit stellen Zeitungsprojekte ohne Volltextgenerierung inzwischen eher eine Ausnahme dar, die der gesonderten Begründung bedarf.[1] Das Angebot an verfügbaren Volltextseiten für historisch Forschende ist dadurch signifikant gestiegen.

Obwohl einzelne Disziplinen unterschiedliche Forschungsansätze verfolgen und – „abhängig von den Anforderungen der jeweiligen wissenschaftlichen Fragestellung" (Müller und Hermes-Wladarsch 2017, 45) – gegebenenfalls auch eine reine Bilddigitalisierung als ausreichend bewertet werden kann, bleibt die Erschließung von Texten auf Zeichenebene ein erstrebenswertes, gemeinsames Ziel all jener, die mit größeren historischen Zeitungskorpora arbeiten. Schließlich verbessert die Verfügbarkeit bestmöglicher Volltexte bekanntermaßen nicht nur die maschinelle Suche, sondern bedeutet generell einen entscheidenden Zuwachs an Möglichkeiten, mit umfangreichen digitalen Sammlungen zu verfahren.

Um den beschriebenen Erwartungen seitens der Fördergeber und Forschenden zu entsprechen, stellen größere wie kleinere Portale heutzutage neben Bilddigitalisaten standardmäßig auch Volltexte zur Verfügung – die Qualität dieser Texte auf Wort- und Zeichenebene bleibt für Benutzer*innen allerdings oft intransparent. So bestehen in der Praxis deutliche Qualitätsunterschiede, welche von automatisch generierten und daher fehleranfälligen Daten („dirty OCR") bis hin zu manuell verbesserten und nahezu fehlerfreien Volltextversionen reichen, die in ihrer Akkuratheit mit sprachwissenschaftlich nutzbaren Goldstandard-Korpora[2] zu vergleichen sind. Letzteres stellt im Fall von historischen Zeitungskorpora jedoch die Ausnahme dar, da der überwiegende Teil aufgrund der enormen Textmengen überhaupt nur automatisch erstellt werden kann, wobei erfahrungsgemäß nach wie vor gravierende Probleme auftreten. In einer Synopse von Problemen und Risiken der Zeitungsdigitalisierung stellt Erik Koenen (2018, 543) daher unter anderem „unterschiedliche Qualitäten und Tiefen von Digitalisaten, Metadaten und OCR" fest. Die mangelnde OCR-Qualität betrifft Huub Wijfjes (2017, 19) zufolge „almost all texts produced before 1850". Auch Ian Milligan (2013, 561) thematisiert die ungünstigen Ausgangsbedingungen vor allem bei der Volltextgenerierung älterer Zeitungen: „In the worst-case scenario, that of seventeenth- and eighteenth-century digitized collections, we can see an accuracy

1 Vgl. etwa die Empfehlungen der Deutschen Forschungsgemeinschaft (DFG), die vor allem „eine Digitalisierung mit Volltextgenerierung" als förderfähig einstuft: https://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_zeitungsdigitalisierung.pdf, S. 2 (letzter Zugriff: 12. 12. 2020).

2 Vgl. etwa das Projekt zur „Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712–1848)" der Universität Paderborn in Zusammenarbeit mit dem Deutschen Textarchiv.

rate in the 40 per cent ballpark; others have estimated that in these conditions, perhaps even more than half of the information may be missed through keyword searching."

Große Zeitungsportale, wie Europeana Newspapers und andere[3], weisen die Benutzer*innen ihrer digitalen Sammlungen daher darauf hin, dass die Texterkennung insbesondere bei historischen Dokumenten ein fehleranfälliger Verarbeitungsschritt ist: „However", heißt es dort einschränkend, „Optical Character Recognition (OCR) – the process of having software automatically detect and recognize text from an image – is still a complex and error-prone task, especially for historical documents with their idiosyncrasies and wide variability of font, layout, language and orthography."[4] Dass auf diese Schwachstelle aufmerksam gemacht wird, ist für Benutzer*innen unerlässlich, zumal sie unmittelbare Auswirkungen auf deren Such- und Recherchevorgänge[5] hat. Ungeachtet dessen zeigt sich in der Praxis, dass User*innen, die bei der Verwendung ungenauer oder fehlerhafter Daten sonst sehr vorsichtig wären, zur Recherchezwecken dennoch die (mehr oder weniger verlässliche) Volltextsuche nutzen, was Milligan (2013, 560) kritisch kommentiert: „Historians need a deeper understanding of OCR and what it means."

3 Ausführliche Hinweise zur Qualität der Volltexte bietet u. a. das Portal Teßmann digital: „Bei der automatischen Volltexterkennung können einige Faktoren die korrekte Erkennung des Ausgangstextes erschweren […]. Die Ergebnisse der automatischen Texterkennung bieten demnach keine hundertprozentige Übereinstimmung mit den Texten der Originaldokumente, was auch die Treffergenauigkeit einer Volltextsuche in unserem Portal beeinträchtigen kann." (https://digital.tessmann.it/tessmannDigital/Information#ocr) Eine Erklärung findet sich weiters auch auf ANNO, dem Austrian Newspapers Online Portal der Österreichischen Nationalbibliothek, wo man Benutzer*innen in der Suchhilfe darüber informiert, dass der Volltext auf automatisch OCR-gelesenen Daten basiert, „weshalb es in manchen Texten zu einer sehr hohen Fehlerdichte kommen kann." (http://anno.onb.ac.at/suchhilfe.htm) Außerdem thematisiert auch das Portal Delpher Qualitätsschwankungen: „Wenn wir viel digitalisieren wollen", gibt man dort zu bedenken, „ist dies bei hoher Qualität nicht immer möglich." (https://www.delpher.nl/nl/platform/pages/helpitems?nid=372) (jeweils letzter Zugriff: 12. 12. 2020). Doch sind solche Informationen längst nicht auf allen Portalen auf den ersten Blick zu finden.

4 Vgl. https://pro.europeana.eu/page/issue-13-ocr (letzter Zugriff: 12. 12. 2020).

5 In Bezug auf die Burney collection der British Library konnte etwa festgestellt werden, dass durch die automatische Texterkennung nur etwa die Hälfte der Wörter korrekt transkribiert wurde – die andere Hälfte wäre nicht auffindbar. Dass die Sammlung dennoch genutzt wird, widerspreche, wie Tim Hitchcock (2011) provokant formuliert, dem kritischen Ansatz von Historiker*innen: „Only 48% of the significant words in the Burney collection of eighteenth-century newspapers are correctly transcribed as a result of poor OCR. This makes the other 52% completely unfindable. […] We use the Burney collection regardless – entirely failing to apply the kind of critical approach that historians have built their professional authority upon. This is roulette dressed up in scholarship."

Eine kürzlich von David A. Smith und Ryan Cordell (2018, 16) durchgeführte Befragung unter Forschenden kommt daher zu dem Ergebnis, „that communication about OCR errors should be improved." Hierbei stellt sich allerdings die Frage, wie die variable Qualität eines zur Verfügung gestellten Volltextes überhaupt adäquat kommuniziert werden kann, denn nicht immer erweist sich die Angabe einer durchschnittlichen Erfolgsquote für eine Textsammlung als aussagekräftig genug – vielmehr geht es auch um die Verteilung von Fehlern, zumal häufig einzelne Seiten mit schlechter Erkennungsquote (etwa durch ungenügende Bildqualität) für vorhandene Qualitätsschwankungen verantwortlich zeichnen.

Vor dem Hintergrund dieser einleitenden Bemerkungen will vorliegender Beitrag nun anhand eines konkreten Zeitungsdigitalisierungsprojektes zur historischen *Wiener Zeitung*[6] aufzeigen, mit welchen Maßnahmen auch kleinere Projekte dazu beitragen können, die automatische Erkennungsrate von historischen Zeitungen entscheidend zu verbessern. Daran anschließend soll am Beispiel einer konkreten Forschungsfrage aus der germanistischen Sprachgeschichte herausgearbeitet werden, dass die Qualität der Volltexte in diesem Fall ein wesentliches Kriterium für die Reliabilität der Ergebnisse darstellt und daher – trotz der beschriebenen Herausforderungen – nicht vernachlässigt werden darf.

2 Texterkennungstraining für die historische *Wiener Zeitung* mittels Transkribus

Die *Wiener Zeitung*, die 1703 als *Wiennerisches Diarium* gegründet wurde und von da an zunächst zweimal wöchentlich erschien, gilt heute als älteste existierende Zeitung der Welt.[7] Das für die historische Presseforschung bedeutende und von Forschenden häufig konsultierte Periodikum ist mit nur wenigen zeitlichen Lücken in ANNO, dem Austrian Newspaper Online-Portal der Österreichischen Nationalbibliothek, digital verfügbar, wodurch die Originale nicht mehr

6 Das Projekt „Das Wien[n]erische Diarium: Digitaler Datenschatz für die geisteswissenschaftlichen Disziplinen" (GD2016/16) wurde von 2017–2020 am Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) durchgeführt, wobei die Erstellung der Volltextversionen und die Anreicherung der Daten in Kooperation mit dem Institut für kunst- und musikhistorische Forschungen und der Gruppe Digitalisierung und Elektronische Archivierung am Institut für Germanistik bzw. dem Forschungszentrum Digital Humanities an der Universität Innsbruck erfolgte. Vgl. https://digitarium.acdh.oeaw.ac.at/ (letzter Zugriff: 12. 12. 2020).

7 Weiterführende Literatur zur historischen *Wiener Zeitung* haben kürzlich Anna Mader-Kratky, Claudia Resch und Martin Scheutz (2019, 110–113) zusammengestellt.

ausgehoben werden müssen (vgl. Resch 2018b, 185). Grundlage für diese umfangreiche Digitalisierung von mehr als 1,3 Millionen Seiten waren einerseits die bereits vorhandenen Mikrofilme und andererseits aufgelöste Exemplare, die in Sammelbänden zusammengefasst vorliegen (vgl. Rachinger 2003, 53).

2.1 Frakturschrift als besondere Herausforderung

Wie andere Zeitungen aus dem deutschsprachigen Raum wurde auch die historische *Wiener Zeitung* von 1703 bis 1940 in Frakturschrift gesetzt, was ihre automatische Texterkennung mit herkömmlichen OCR-Methoden wesentlich erschwert. Dass sich die Buchstabenerkennung bei gebrochenen Schriften besonders fehleranfällig zeigt, war auch eines der Ergebnisse eines kürzlich veröffentlichten Beitrags von Clemens Neudecker et al. (2019, 53): „The results from established OCR methods have so far been insufficient when it comes to the recognition of old printing type, especially Gothic types."

Da es also insbesondere bei automatisch eingelesenen Frakturtexten zu einer hohen Fehlerdichte kommen kann, die für Forschende (etwa bei der Suche) nur schwer einschätzbar[8] ist, führt mittelfristig kein Weg an der Optimierung der automatisch erstellten Volltexte vorbei. In dem oben genannten Projekt wurden daher erstmals neue, aus der Handwritten Text Recognition (HTR) stammende Ansätze für die historische *Wiener Zeitung* erprobt. Dabei wurden mehr als 300 Ausgaben, verteilt über das gesamte 18. Jahrhundert, ausgewählt[9], im Bestreben, verlässliche Volltexte zu generieren, um einerseits abschätzen zu können, mit welchem Aufwand eine qualitativ höherwertige Volltexterschließung verbunden wäre, und andererseits, um die sorgfältig transkribierten Volltexte als digitales Textkorpus nutzen zu können – wie etwa für historisch ausgerichtete, linguistische Fragestellungen (vgl. Abschnitt 3).

8 Suchergebnisse, die durch zahlreiche erzielte Treffer den Anschein von Vollständigkeit haben, können jedenfalls keine Auskunft darüber geben, welche und wie viele Belege den Suchenden aufgrund von OCR-Fehlern entgehen.

9 Das partizipative Auswahlverfahren („Call for Nominations") findet sich bei Resch (2018, 24) beschrieben. Auf dessen Basis wurde ein repräsentatives Korpus erstellt, in welchem jedes Erscheinungsjahr durch fünf Ausgaben vertreten sein sollte.

2.2 Ein Modell (nicht nur) für die *Wiener Zeitung*

Wie Milligan (2013, 563) vorschlägt, ist eine der Möglichkeiten, die Genauigkeit der Texterkennung zu steigern, der direkte verbessernde Eingriff durch Forschende („direct human intervention"). Da dies bei großen Textmengen zu zeitaufwändig wäre, könnte man selbstlernende Software einsetzen – aber auch das funktioniert nicht ohne intellektuellen Korrekturaufwand, wie Wijfjes (2017, 19) betont: „Technicians predict that self-learning software can solve the problem in the longrun, but this requires human input to „instruct" the software of what is correct and what is not." Da dieser Ansatz dennoch sehr aussichtsreich erscheint, soll er am Beispiel der *Wiener Zeitung* näher beschrieben werden:

Hierfür wurden die Volltextversionen mit Hilfe der inzwischen weit verbreiteten Software Transkribus[10] erstellt, mit welcher sich Handschriften, aber eben auch Druckschriften erkennen und sich Modelle mittels neuronaler Netze trainieren lassen. Dabei gilt: Je mehr verlässliche Trainingsdaten vorhanden sind, umso besser das Ergebnis. Im Fall der *Wiener Zeitung* wurden Tranchen zu jeweils 50 Ausgaben als Bilder in die Transkribus-Plattform geladen, innerhalb welcher Textregionen und Lesereihenfolge festgelegt und Grundlinien als Basis der Texterkennung markiert wurden. Die Ergebnisse dieses Workflows, den Dario Kampkaspar (2019, 132–134) im Detail beschreibt, waren automatisch generierte Volltexte, die durch das Projektteam und den Dienstleister Innsbruck University Innovations unter großem Zeitaufwand manuell nach den im Projekt entwickelten Transkriptionsregeln nachkorrigiert wurden, um wiederum als „Ground truth"[11] für das Trainieren neuer Modelle und die Volltextgenerierung weiterer Zeitungsausgaben verwendet werden zu können. Unter der Voraussetzung, dass die Bildvorlagen von guter Qualität sind, weist das derzeitige Modell, das nun in Transkribus öffentlich zur Verfügung steht (vgl. readcoop.eu/de/modelle/german-frakturr-18th-century/), eine Character Error Rate (CER) von unter 0,3 auf, was bedeutet, dass sich auf tausend Zeichen durchschnittlich nicht mehr als drei Fehler finden.

Dass für manche der ausgewählten Ausgaben im Rahmen des Projekts keine Volltexte generiert werden konnten, liegt an den Mikrofilmen, die häufig so blass sind, dass sie eine Erkennung verunmöglichen: „Die größte Fehlerquelle stellt

10 Vgl. https://transkribus.eu/Transkribus/ (letzter Zugriff: 12. 12. 2020) sowie Mühlberger et al. 2019, 954–976.

11 Der Begriff bezieht sich im maschinellen Lernen auf genaue, objektive Informationen. Im Fall von Transkribus wird diese „Ground truth"-Information durch das Training des Systems mit einer ausreichenden Anzahl von Daten erzeugt, um damit ein Modell zu erstellen, welches auf große Mengen derselben Schrift erfolgreich angewendet werden kann (vgl. Mühlberger et al. 2019, 969 Anmerkung 8).

damit nicht mehr die eigentliche Texterkennung dar, sondern die Qualität der Bilder" (Kampkaspar 2019, 133). So mag die Mikroverfilmung zwar vor Jahren ein geeignetes Speichermedium gewesen sein – als Vorlage zur Bilddigitalisierung verursacht sie bei der Texterkennung jedenfalls grundlegende Probleme, die auch von Smith und Cordell (2018, 12) thematisiert werden: „Finally, some digitized materials are of poor quality because of their source media – particularly microfilm, which is, for example, often the source for digitized newspaper collections – was itself of poor quality." Gerade im Fall der *Wiener Zeitung*, die aufgrund des kontinuierlich bestehenden Forschungsinteresses als eines der ersten Periodika schon vor vielen Jahren mikroverfilmt und 2012 gescannt wurde, wäre zur Abhilfe dieses Problems über ein technisch verbessertes Re-Imaging (d. h. ein nochmaliges Scannen der Originale) einiger Jahrgänge nachzudenken. Dies war jedoch nicht Aufgabe des dreijährigen Projektes, sondern müsste von Bibliotheken, welche die *Wiener Zeitung* im Original beherbergen, durchgeführt werden.

Die im Projekt bereits bearbeiteten Volltextausgaben der historischen *Wiener Zeitung*, vormals *Wien[n]erisches Diarium*, sind seit 2020 im sogenannten DIGITARIUM[12] (Resch & Kampkaspar 2019) online abrufbar. Dieser Prototyp bietet eine Übersicht über alle publizierten Nummern, welche in einer synoptischen Ansicht von Volltext und Digitalisat eingesehen und verglichen werden können. Die Rahmung der einzelnen Vorschaubilder gibt Auskunft über die Qualität der erstellten Volltexte, wobei dunkelgrün markierte Ausgaben mehrfach, türkis markierte Ausgaben zweifach und hellblau gerahmte Ausgaben einmal manuell überprüft worden sind. Ein weißer Rahmen kennzeichnet schließlich jene Ausgaben, die den automatisch erstellten Text enthalten und keinen manuellen Korrekturdurchgang durchlaufen haben. Dieses abgestufte Farbsystem wiederholt sich in den Balkendiagrammen, wo die bislang erreichte Qualität der Texte für das gesamte 18. Jahrhundert[13] oder pro Dekade visualisiert ist, und entspricht in etwa den von Smith und Cordell (2018, 16) vorgeschlagenen „categorial confidence levels, such as low, medium, and high."

Angesichts der insgesamt circa 10.000 Ausgaben der *Wiener Zeitung*, die im Verlauf des 18. Jahrhunderts erschienen sind, ließe sich nun einwenden, dass diese vergleichsweise kleine Edition dem periodischen Erscheinungscharakter der Zeitung bei Weitem nicht gerecht wird. Zweifellos wäre eine durchgängige und einigermaßen verlässliche Volltextverfügbarkeit das erstrebenswerte Ziel –

12 Vgl. digitarium-app.acdh.oeaw.ac.at/ (letzter Zugriff: 12. 12. 2020).

13 Die Abdeckung über die einzelnen Dekaden ist – aufgrund der oben erwähnten variablen Qualität der Bildvorlagen – weniger regelmäßig als anfänglich geplant.

allerdings ist auch das bisher Geleistete hierfür wesentliche Voraussetzung: Im Gegensatz zu großen Digitalisierungsunternehmen haben Projekte von kleinem Ausmaß immerhin die Möglichkeit, spezifische und an das Quellenmaterial perfekt angepasste Modelle zu trainieren, wodurch Erkennungsquoten signifikant gesteigert werden können. Im Fall der *Wiener Zeitung* etwa können weitere Seiten künftig verbessert automatisch eingelesen werden.

Ob das im Projekt erstellte Modell, welches anhand einer in Fraktur gedruckten historischen Zeitung des deutschsprachigen Raumes trainiert worden ist, auch bei anderen Periodika erfolgreich Anwendung finden kann – entweder zur Volltextgenerierung oder als Ausgangspunkt eines eigenen Modells –, wird zu prüfen sein. Im direkten Vergleich von *Wiener Zeitung* und zeitgleich erschienenen Periodika zeigen sich in Bezug auf Layout und Schrift große Ähnlichkeiten (wie in Abb. 1 zu sehen ist), was jedenfalls darauf hoffen lässt, dass von dem trainierten Modell auch andere Projekte profitieren werden können.

Abb. 1: Gegenüberstellung von Titelseiten der *Wiener Zeitung* und zeitgleich erschienener Periodika.

Nicht nur das Modell, sondern auch die erstellten Texte selbst könnten zudem als Testdaten zur Grundlage weiterer Trainings dienen. Smith und Cordell (2018, 19) gehen in ihren generellen Empfehlungen schließlich davon aus, dass die Texterkennung unter anderem von der Nutzung bereits vorhandener Editionen künftig substanziell profitieren könnte: „We propose that progress in OCR could be accelerated by exploiting existing digital editions to provide ground-truth training and test data." Hierin bestätigt sich also abermals, dass zeitlich und personell limitierte Vorhaben, wie das hier beschriebene, trotz oder gerade aufgrund deren intensiver Beschäftigung mit projektspezifischem Quellenmaterial maßgeblich dazu beitragen können, die bislang problematische Volltexterschließung von historischen Zeitungen zu verbessern.

3 Ein Anwendungsszenario aus der germanistischen Sprachgeschichte

In der Praxis stellt sich bei vielen Forschungsvorhaben die Frage, welche Bedeutung der Verlässlichkeit von Volltexten überhaupt zukommt beziehungsweise welche Fehlerraten (noch) akzeptabel wären. Nicht selten ist die Datenqualität und deren Auswirkungen aber nur schwer einzuschätzen,[14] weshalb sich oftmals nicht sagen lässt, wie stark Ergebnisse durch Fehler im Volltext beeinflusst sind. Um hierfür mehr Bewusstsein zu schaffen, soll im folgenden Abschnitt mit einem Anwendungsszenario aus der historischen Linguistik die Relevanz von qualitativ hochwertigen Volltexten bei der Beforschung von Zeitungen anhand eines konkreten Erkenntnisinteresses überprüft und unter Beweis gestellt werden.

Das Potenzial digital verfügbarer Zeitungen hat sich der historischen Linguistik erst in den letzten Jahren erschlossen, was darauf zurückzuführen ist, dass zum einen „Zeitungsbestände erst in jüngerer Zeit katalogisiert, faksimiliert und neuerdings digitalisiert wurden" und zum anderen „der sprachhistorische Aufschlusswert von Zeitungen in der älteren Literatur gar nicht erkannt oder jedenfalls nicht hervorgehoben wurde", wie der Sprachwissenschaftler Thomas Gloning (2017, 121) feststellt. Durch diese Umstände sind insbesondere die Zeitungen des 18. Jahrhunderts laut Britt-Marie Schuster und Manuel Wille (2017, 103) „aus sprachhistorischer und textsortengeschichtlicher Perspektive bisher nur ansatzweise beschrieben worden". Volker Bauer und Holger Böning (2011, X) wiederum unterstreichen die Bedeutung des Mediums für die Herausbildung und Stabilisierung der Schriftsprache: „Der Beitrag der Zeitungsschreiber zur Entwicklung der deutschen Sprache ist noch kaum gewürdigt und schwer zu überschätzen." Dieser Aspekt wird auch von Jörg Riecke (2016, 181, 202) als ein bisher „noch zu wenig beachteter Faktor für die sprachliche Standardisierung" bezeichnet, wobei er auf den Einfluss früher Zeitungen verweist, „die vor allem im norddeutschen Raum einen entscheidenden Beitrag zur Verbreitung der neuen Schriftsprache im Alltag geleistet haben." Ähnliches gilt auch für den oberdeutschen Raum, der nun seit Kurzem durch die hier vorgestellte *Wiener Zeitung* vertreten ist.

14 „Since scholars naturally are disinclined to devote time on data that turn out not to be useful, it is difficult to assess „how dirty is too dirty" for different tasks", so Smith und Cordell (2018, 15).

3.1 „Die Wienerische Zeitungen klingen nicht so wol“

Vor der Sprachreform des 18. Jahrhunderts und der Gründung der *Wiener Zeitung* hatten die Zeitungen in Wien aus Sicht des damals zeitgenössischen Sprachwissenschaftlers und Pressehistorikers Kaspar von Stieler (1695, 89f.) keinen allzu guten Ruf: „Von Regenspurg / wo teutsche Räthe und gesante versamlet seyn“, vermerkt er, „kommen wol die beste [Zeitungen]; wie auch von den Sächsischen Höfen: Die Wienerische Zeitungen klingen schon nicht so wol“. Eine konkrete Begründung für diese Einschätzung bleibt der Gelehrte leider schuldig, doch lässt der Verweis auf die Sächsischen Höfe vermuten, dass dieses Urteil mit den Unterschieden zwischen der Schreibsprache der nieder- und mitteldeutschen Regionen im Norden und jener des oberdeutschen, katholischen Südens in Zusammenhang stand. So hatte sich im habsburgischen Österreich eine oberdeutsche Schreibsprache herausgebildet, die von einflussreichen Orden zwar befördert, aus Sicht der fortschrittlicheren protestantischen Gelehrten jedoch als „rückständig und provinziell“ (Riecke 2016, 174) empfunden wurde. Dennoch waren zwei anerkannte Leitvarietäten entstanden: das ostmitteldeutsche Meißnisch-Obersächsische als Sprache des Protestantismus und das Bairisch-Oberdeutsche als Sprache des Katholizismus (vgl. Wiesinger 2014, 309).

Ab dem Jahre 1730 erwachte auch in Österreich Interesse an einer einheitlichen Schriftsprache. Die Tatsache, dass die nieder- und mitteldeutsche Sprache schon länger von den Einflüssen der Aufklärung geprägt war und höheren Geltungsgrad genoss, sollte dem Süden nicht zum Nachteil gereichen: Mit der beginnenden Aufklärung setzte sich zunehmend die Meinung durch, dass das Denken mit Hilfe von Sprache erfolge – „[w]o aber die Sprache schlecht und mangelhaft ist, dort könne auch nicht ordentlich gedacht und weder geistige Leistung hervorgerufen noch Wissenschaft und Fortschritt erzielt werden“ (Wiesinger 2014, 323). In dem daraus resultierenden innerdeutschen „Sprachenstreit“, an dem in der zweiten Hälfte des 18. Jahrhunderts unter anderen der Leipziger Sprachkritiker Johann Christoph Gottsched beteiligt war, ging es letztlich um die Definition einer allgemein verbindlichen deutschen Schriftnorm. Obwohl es sich bei der infolgedessen vollzogenen Reform um keinen Sprachwechsel im eigentlichen Sinn handelte, bedeuteten die orthografischen, morphologischen und syntaktischen Unterschiede, die Paul Rössler (2005, 359) und Peter Wiesinger (2014, 343–345) auflisten, doch eine bewusste Formanpassung der bis dahin gewohnten Schriftsprache, bei der das Oberdeutsche schließlich zugunsten des Neuhochdeutschen der ostmitteldeutschen Gebiete aufgegeben wurde.

3.2 Sprachreform in den oberdeutschen Gebieten: „seynd“ versus „sind“

Über den zeitlichen Verlauf und die Implementierung dieser Sprachreform ist – außer offiziellen Eckdaten – bislang wenig bekannt: In ihren Publikationen nehmen Martin Durrell et al. (vgl. 2009, 264) sowie Silke Scheible et al. (2011, 540) die Zeitspanne zwischen 1650 und 1800 in den Blick und beobachten dabei eine allmähliche Anpassung des Südens: „the more northerly standard originating in the Central German area was gradually adopted in the South.“ Heißt es bei Jörg Riecke (2016, 173), dass sich der Wechsel „nach und nach“ vollzogen hätte, so nennt Peter Wiesinger (2014 309 f., 342) als engeren Zeitraum der Sprachreform die Jahre zwischen 1730 und 1760, weist aber darauf hin, dass es „bis zu ihrem vollen Durchdringen rund vier Jahrzehnte, nämlich die Regierungszeiten von Maria Theresia bis 1780 und ihres Sohnes und Nachfolgers Kaiser Josephs II. bis 1790“ gedauert hätte. Anna Havinga (2018, 16) schränkt diese lange Zeitspanne weiter ein und geht in ihrer Dissertation davon aus, dass im Hinblick auf die Wirkung der Sprachnormierungsmaßnahmen die Jahre zwischen 1744 und 1774[15] wohl als entscheidende Übergangszeit, als „key transitional period“, zu bezeichnen sind.

Eine genauere zeitliche Eingrenzung der Reform scheint auch deshalb schwierig, weil es bei einigen sprachlichen Distinktionsmerkmalen eklatante Meinungsverschiedenheiten unter den damaligen Gelehrten gab. Ein Beispiel hierfür war die Bildung der 1. und 3. Person Plural Indikativ Präsens des Auxiliarverbs „sein“: Während das Ostmitteldeutsche die Variante „wir/sie sind“ präferierte, lautete die Form im Oberdeutschen „wir/sie seynd“. Johann Balthasar Antesperg bevorzugte in seiner 1747 erschienenen „Kayserliche[n] deutsche[n] Grammatick“[16] die ostmitteldeutsche Variante, schlug aber alternativ auch die oberdeutsche vor: „Wir sind oder seynd / sie sind oder seynd“ (1747, 116 f.), lautete seine Empfehlung. Der Sprachkritiker Johann Christoph Gottsched (1748, 297) hingegen stigmatisierte in seiner „Grundlegung einer deutschen Sprachkunst“ die oberdeutsche Variante, indem er schrieb: „Wir sind und Sie sind (nicht seyn.)“ – in der nächsten Auflage schließlich erwähnte er die oberdeutsche Variante gar nicht mehr (vgl. auch Havinga 2018, 67).

15 Das Jahr 1774 ist jenes, in welchem die allgemeine Schulpflicht in den habsburgischen Erblanden eingeführt und die Normierung als Schul- und Amtssprache verankert wurde.

16 Dieser ersten Auflage folgte 1749 eine zweite – beiden Auflagen blieb der erhoffte Erfolg jedoch verwehrt, was Wiesinger (2014, 369 und 331) damit begründet, dass Antesperg sich „vielfach nicht vom herkömmlichen oberdeutschen Sprachusus löste“ und sich „durch Akzeptierung von Doppelformen nicht zu einer gewünschten einheitlichen Norm durchringen konnte.“

3.3 „seynd“ und „sind“ in der historischen *Wiener Zeitung*

Die Frage, die sich vor diesem Hintergrund nun aufdrängt, ist also, wie ein Periodikum wie die *Wiener Zeitung* mit derartigen Vorgaben verfährt, wann diese auch für ihr Lesepublikum Gestalt annehmen und welche Rolle das Medium selbst im Sprachnormierungsprozess einnimmt. Um diese Fragen zur Gänze beantworten zu können, müssten alle von Paul Rössler (2005, 359) und Peter Wiesinger (2014, 343–345) beschriebenen Distinktionsmerkmale in der *Wiener Zeitung* abgefragt werden. Da dies im Rahmen der vorliegenden Untersuchung jedoch nicht vollumfänglich möglich ist und es in erster Linie darum geht, das Potenzial verlässlicher Volltexte für die historische Sprachwissenschaft aufzuzeigen, möchte die Verfasserin bei dem Beispiel von „seynd“[17] und „sind“ bleiben, welches – wie zu vermuten steht – in der *Wiener Zeitung* relativ frequent sein müsste.

Im Gegensatz zu anderen korpusbasierten Untersuchungen[18] ist die vorliegende digitale Volltextsammlung zur *Wiener Zeitung* mit nahezu vier Millionen Tokens beziehungsweise 334 Ausgaben sehr umfangreich. Dabei gilt es allerdings zu beachten, dass (aus den bereits genannten Gründen) nicht jeder Jahrgang zwischen 1703 und 1799 repräsentiert ist und dass die Tokenzahlen der einzelnen inkludierten Jahrgänge schwanken, da die Stärke der Ausgaben über das Jahrhundert betrachtet zunimmt. In der Untersuchung dürfen daher keine absoluten Werte verglichen werden, sondern relative Häufigkeiten, die für den Vergleich von „seynd“ und „sind“ heranzuziehen sind.

Eine erste Auswertung in der Open-Source-Version des Korpusanalysetools Sketch Engine[19] zu den Wortformen „seynd“ und „sind“ bildet deren Vorkom-

17 Die Schreibweise „seind“ kommt im untersuchten Korpus nur sechsmal vor, davon viermal im Jahr 1712; sonst sind durchgängig „seynd“ oder „sind“ gebräuchlich.

18 Kristina Schneider (2002, 53 f.) verwendet für ihre Untersuchung der „popular papers“ Samples von 20.000 Wörtern, wobei jedes Sample von mehr als zwei verschiedenen Zeitungen bestehen sollte, um daraus eine allgemeine Tendenz ableiten zu können. Hirofumi Hosokawa (2014, 29) stellt für ihre sozіopragmatischen Untersuchungen zur Sprache in Zeitungen um 1850 ein Korpus von 129.416 Wörtern aus sechs verschiedenen Zeitungen zusammen. Unter der Leitung von Martin Durrell (2017, 83) ist das GermanC-Korpus entstanden, in welchem in Form von Textauszügen zu jeweils 2.000 Wörtern auch Zeitungstexte aus fünf verschiedenen Textregionen und drei Zeitperioden enthalten sind. Anna Havinga (2018, 100), die neben der *Wiener Zeitung* noch andere Textgattungen beforscht, wählt insgesamt 19 Ausgaben im Abstand von fünf Jahren und analysiert darin jeweils die ersten (bis zu acht) Seiten.

19 Das zuvor beschriebene Interface des Prototyps DIGITARIUM ist vorrangig zur Lektüre und Suche in den Ausgaben vorgesehen. Um bei der Datenauswertung flexibler zu sein und Möglichkeiten der Visualisierung nutzen zu können, wurden die Textdaten in das Korpusanalysetool Sketch Engine überspielt, vgl. https://www.sketchengine.eu/ (letzter Zugriff am 12. 12. 2020).

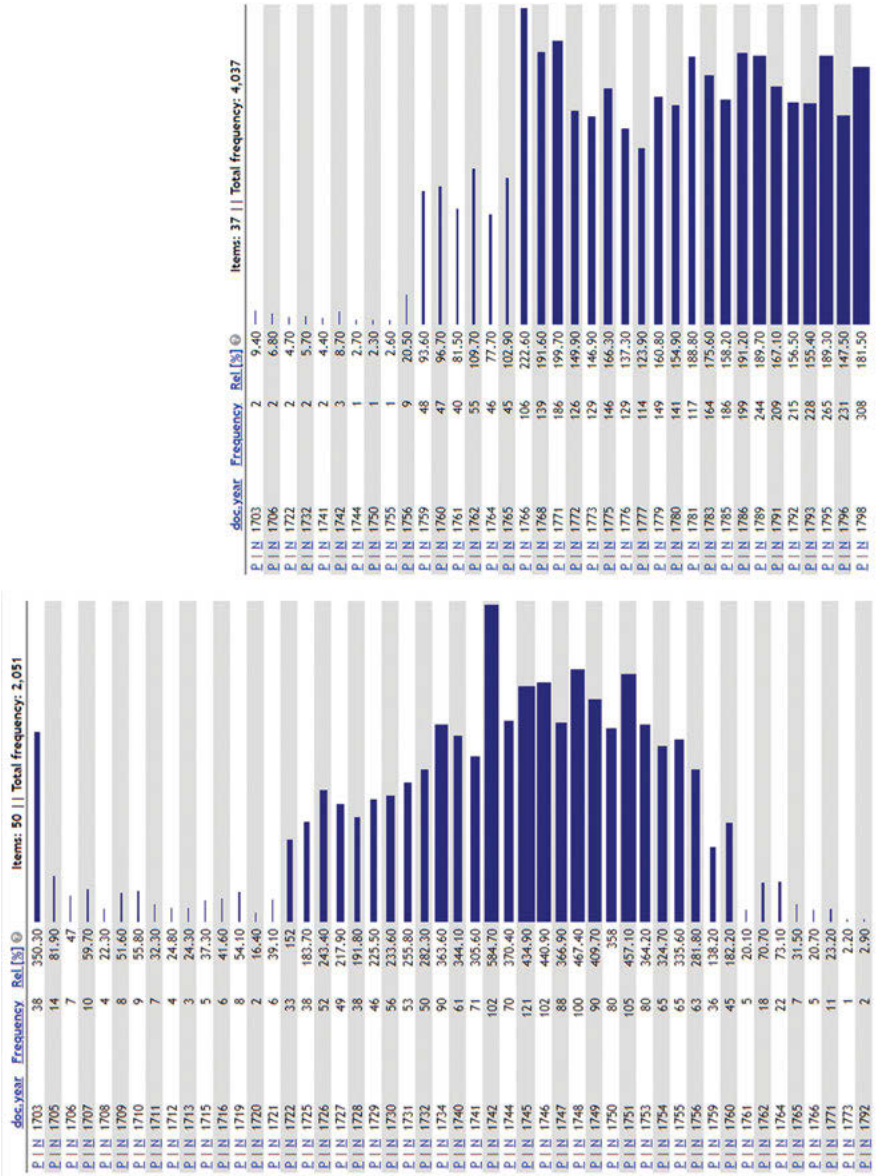


Abb. 2a und 2b: „seynd“ (a) und „sind“ (b), chronologisch geordnet mit Angaben zur absoluten und relativen Häufigkeit.

men chronologisch nach Jahren geordnet ab (vgl. Abb. 2, Jahre ohne Treffer scheinen dabei nicht auf):

Um diesen zeitlichen Verlauf noch auf andere Art und Weise darstellen zu können, wurden die Daten für eine zweite Auswertung in das am ACDH-CH entwickelte Tool CorpSum[20] überführt, das für Visualisierungen abermals auf die Open-Source-Version der Sketch Engine zugreift (vgl. Abb. 3).

Anhand dieser Visualisierung wird erkennbar, dass in den ersten Erscheinungsjahren der *Wiener Zeitung* nahezu ausschließlich die oberdeutsche Form „seynd" (in blau) verwendet wird. Ab der Mitte der 1750er Jahre jedoch erfährt die ostmitteldeutsche Variante „sind" (in orange) einen relativ steilen Anstieg, während die bis dahin gewohnte Form stark abfällt und nach dem Jahr 1771 so gut wie gar nicht mehr verwendet wird. Ein Nebeneinander beider Varianten ist demnach vor allem für das Jahrzehnt zwischen 1756 und 1766 beobachtbar, während davor und danach jeweils eine Variante deutlich präferiert und die andere marginalisiert scheint.

Am Beispiel dieser Trends für „seynd" beziehungsweise „sind" zeigt sich also, dass die *Wiener Zeitung* durch bewusste Entscheidungen im Sprachgebrauch oberdeutsche Varianten aussondert und eine einheitliche Schriftnorm verbreitet. Der Normenwechsel geschieht jedoch nicht übergangslos, weshalb insbesondere jene Belegstellen interessant zu betrachten sind, die gerade nicht dem jeweiligen Trend entsprechen. Ein Blick in das Quellenmaterial belegt etwa, dass einige der sehr seltenen Verwendungen von „sind" in den frühen Jahren des Periodikums aus Berichten über den Norden stammen, wie etwa untenstehendes Beispiel aus der ersten Dekade (vgl. Abb. 4).

Seltene Funde von „sind" lassen sich auch in Berichten aus „Naumburg in Sachsen" (30. Juni 1742) oder „Aus Dresden" (19. September 1742) beobachten. Da in anderen Nachrichten aus diesen Gebieten aber ebenfalls „seynd" verwendet wird, bleibt zu vermuten, dass in diesen wenigen Fällen möglicherweise darauf verzichtet oder vergessen wurde, die Meldungen dem oberdeutschen Schreibusus anzupassen.

Bei den späten Belegen von „seynd", daher nach 1766, zeigt sich umgekehrt, dass diese insbesondere in Inseraten auftreten. In untenstehender Ankündigung für Blumenzwiebeln aus dem Jahr 1771 (vgl. Abb. 5a) beispielsweise kommen „seynd"-Formen noch mehrfach vor, welche in einer späteren Formulierung desselben Angebots von 1789 (vgl. Abb. 5b) dann aber ebenfalls durch „sind" ersetzt werden.

Eine erste kursorische Durchsicht der Belege lässt vermuten, dass bei der Festigung der neuen Schriftnorm Unterschiede zwischen den Texten der Zei-

20 Vgl. Asil Çetin. CorpSum: Multi-Faceted Visual Corpus Analysis Tool. 2020.

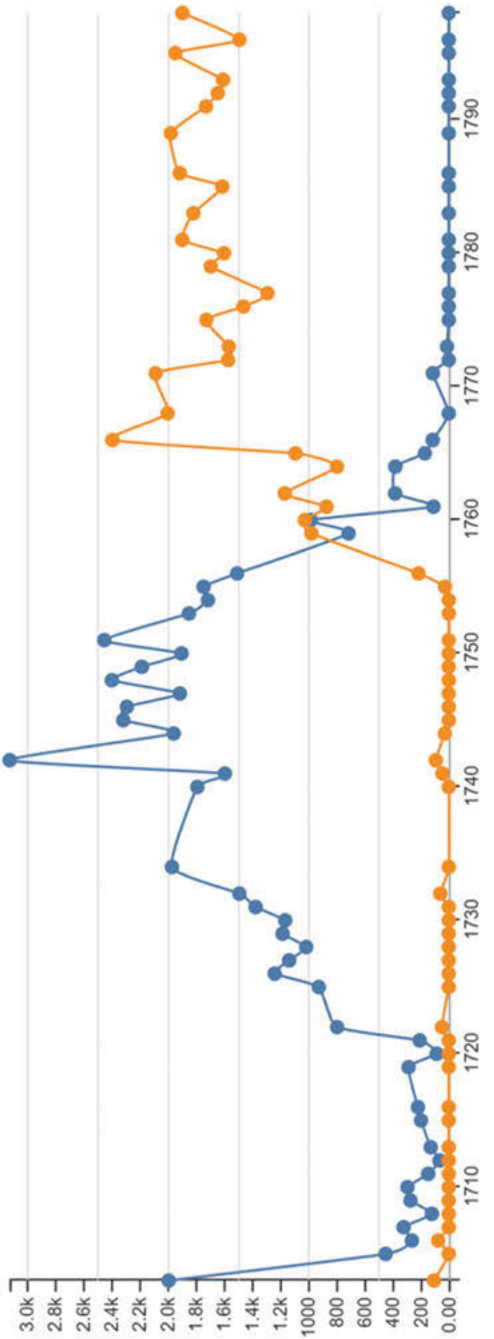


Abb. 3: Verlaufskurven von „seynd“ (blau) und „sind“ (orange).

Auß dem Hollsteinischen vom 26. *Julij* Jhro Hochfürstliche Durchl. deß Herrn Bischoffs zu Eutin Reichs=*Contingent*, wie auch deß Herrn Obristen von Osten Hollsteinische Dragoner=Regiment / so alle wol *montirt* und brave Leüthe / sind den 19. *hujus* auß den Eutinischen auffgebro= chen / und haben zu Schwartau ihr *Rendevous* gehalten / daselbst sie ins=

Abb. 4: Belegstelle aus der allerersten Ausgabe des „Wiennerischen Diarii“ von 8.–11. August 1703, 4.

NB . Es seynd hier wieder angekommen die Ge= brüdere Montani mit verschiedenen Blumenkielen wie folget :

Obgedachte Gebrüdere Montani seynd bey Hrn . Sebastian Lanser im Gewürzgewölb beym golde= nen Stern in der Wolzeil anzutreffen .

Abb. 5a: Inserat vom 16. März 1771, 16.

Blumenkiele und Saamen.

Bey Joh. Angelo Montano, burgl. Han= delsmann in seinem Gewürzgewölbe zum gol= denen Stern in der Wolzeil nächst dem Stu= benthor sind wieder angekommen, und zu haben die frischen Blumenzwiebel, wie folgt:

Abb. 5b: Inserate vom 23. Mai 1789, 1318.

tungsmeldungen, die der Sprachreform folgten, und den Texten des Anzeigenteils bestehen. Dass die oberdeutsche Form „seynd“ insbesondere in den Inseraten noch gelegentlich auftritt und akzeptiert wird, könnte darauf zurückzuführen sein, dass diese sich unmittelbar an das lokale Lesepublikum wenden und daher der „Alltagssprache“ (Hosokawa 2014, 24) angepasst sein wollten. So schafft diese besondere Kommunikationssituation möglicherweise eine „Sprache der Nähe“ (Koch und Oesterreicher 1985), weswegen (auch weitere) oberdeutsche Varianten häufiger in lokalen als in überregionalen Nachrichten dokumentiert sind, wie Havinga (2018, 176) beobachtet.

3.4 Fazit und Ausblick

Die periodische Erscheinungsweise der Presse und insbesondere der historischen *Wiener Zeitung*, deren ausgewählte Ausgaben des 18. Jahrhunderts nun in einem chronologisch relativ ausgewogenen Korpus vorliegen, lassen diese (und andere historische Zeitungssammlungen) für bestimmte Untersuchungen als besonders prädestiniert erscheinen: „the unique periodicity of the press makes it particularly well suited for studying continuity and change“, betont Bob Nicholson (2013, 64). Am gewählten Beispiel wird aber auch deutlich, dass man, um granulare Untersuchungen an historischen Zeitungen durchführen zu können, idealerweise eine grobe Klassifizierung der Texte nach Genre und Korrespondenzort

vornehmen müsste, um etwa Berichte von Inseraten unterscheiden zu können oder „Inländische Berichterstattung" von „Auswärtigen Nachrichten".

Zudem stellt sich am Ende dieses Vergleichs die Frage, inwieweit es zulässig ist, Zeitungen aufgrund ihres Erscheinungsortes einer Sprachlandschaft zuzuordnen und die *Wiener Zeitung* uneingeschränkt als „oberdeutsches Periodikum" zu bezeichnen, oder ob sie aufgrund der Vielzahl an heterogenen Texten nicht tendenziell eher als „supra-regional" verstanden werden müsste (vgl. Havinga 2018, 173 und 176). Wie Hirofumi Hosokawa (2014 57) betont, ist die Sprache der Zeitungen eher „eine Zusammensetzung aus verschiedenen Schreibstilen" und auch Peter von Polenz (2013, 398) geht von einem „Konglomerat mehrerer Schrifttraditionen" aus. Mit dem Hinweis „newspapers include both locally produced texts and material from press agencies" macht Marianne Hundt (2008, 179) abermals darauf aufmerksam, dass Texte in Zeitungen in Bezug auf ihre Herkunft meist nicht homogen seien.

Gerade im Fall der *Wiener Zeitung* weiß man bislang generell noch viel zu wenig darüber, wie mit von auswärts einlangenden Berichten verfahren wurde[21] und ob und in welchem Ausmaß Texte (wohl auch unter Zeitdruck) redigiert und der lokalen Varietät angepasst wurden. Die Ergebnisse zum Auxiliarverb „sind", die sehr eindeutig (zugunsten der einen oder anderen Form) ausfallen, lassen jedoch darauf schließen, dass in der Redaktion sehr wohl normierend und redigierend eingegriffen wurde, um Varianten nach den damals neuen, überregional geltenden Regeln zu verbessern und die zunehmend einheitliche Schriftsprache des Deutschen in die Praxis zu implementieren. Wie Wiesinger (2014, 335) zu Recht betont, „darf dabei auch der Anteil der Verleger und ihrer Setzer nicht übersehen werden, die vor allem die Sprachform mit Orthographie und Flexionen bestimmten, indem sie die nun fehlerhaft erscheinenden oberdeutschen Schreibungen und Formen nach den neuen Regeln verbesserten."[22]

Um die Implementierung und Datierung der Sprachreform in der *Wiener Zeitung* noch eingehender untersuchen zu können, müsste man das Beispiel von „seynd" und „sind" freilich im Verbund mit weiteren sprachlichen Markern und Distinktionsmerkmalen sehen, die erst in Summe die „Sprachreform" ausmachen

21 In diesem Zusammenhang sind Einzelstudien über die Beziehungen der historischen *Wiener Zeitung* zu zeitgleich erschienenen Periodika und anderen Quellen (wie etwa zur Flugpublizistik des 18. Jahrhunderts) geplant.

22 Dass die Herausgeberschaft 1755 von Johann Peter van Ghelen auf seinen Sohn Johann Leopold van Ghelen überging und nach dessen Tod 1760 abermals zu seinen Erben wechselte (Frank und Frimmel 2008, 77 f.), ist bei der Interpretation der Daten ebenso zu bedenken wie die Herkunft mancher Verleger und Drucker aus nördlicher gelegenen Sprachregionen (vgl. Wiesinger 2014, 366).

und mit textexternen Faktoren (wie beispielsweise mit den in Anmerkung 22 beschriebenen Herausgeberwechseln) ausgewertet werden sollten. Die Datenbasis hierfür ist durch das beschriebene Projekt jedenfalls gelegt und das exemplarische Anwendungsszenario dokumentiert, dass nun deutlicher als bisher einzuschätzen ist, in welchen Zeiträumen sich sprachliche Veränderungen in Bezug auf das untersuchte Phänomen vollzogen haben.

Zugleich ist an dieser Fallstudie zu sehen, dass die generierten Abfragen zu bestimmten Wortformen immer der Deutung und der Einbettung in bestehende Forschungskontexte und -traditionen bedürfen. Computer und digitale Textmethoden entlasten Forschende nicht von der Interpretation, bestätigt Koenen (2018, 549) und präzisiert: „Vielmehr fordern sie sie im kontinuierlichen Wechselspiel permanent zum Interpretieren heraus, um den vom Computer entdeckten Textmustern Sinn zu verleihen". Dieses Wechselspiel beschreiben Pim Huijnen und Melvin Wevers (2015) als „a constant to and fro between distant reading and close reading": So zeigt auch diese Studie, dass das Erkenntnisinteresse sich sowohl den quantitativen Überblicksauswertungen als auch den einzelnen Vorkommen der individuellen Wortform in ihrem Kontext zuwenden muss, um zu validen Deutungen zu gelangen. Dass die Erstellerin des Korpus zugleich auch deren Analystin[23] ist und die vorhergehenden Datenprozesse und -transformationen kennt, ist hier von Vorteil – wenn User*innen aber historische Zeitungskorpora anderer benutzen, ist es umso wichtiger, dass sie ein Verständnis dafür entwickeln, welche Interaktionen zuvor stattgefunden haben, die nun ihre Sicht auf die Daten formen (vgl. etwa Koolen et al. 2019, 379).

4 Automatisch generierte, hochwertige Volltexte als Ideal

Abschließend lässt sich festhalten, dass Zeitungen insbesondere für historisch arbeitende Sprachwissenschaftler*innen ein wertvoller, aber reflektiert zu nutzender Datenschatz sind: Je verlässlicher die erstellten Volltexte, umso eher werden sich aber auch Forschende anderer Disziplinen dafür begeistern können, sich mit ihren Fragestellungen auf größere Textkorpora einzulassen. Am Beispiel des digitalen Korpus zur *Wiener Zeitung*, für deren Frakturschrift eigens trainierte und nachnutzbare Modelle eingesetzt wurden, konnte zudem gezeigt werden, welche

23 Für diese Konstellation hat Lynne Flowerdew die treffende Bezeichnung „compiler-cum-analyst" geprägt (vgl. 2005, 329).

Anwendungsszenarien eine bestmögliche[24] (keine perfekte!) Volltexterkennung etwa zu Fragen des Sprachwandels und der Sprachgebrauchsgeschichte bietet – wobei freilich noch zahlreiche weitere sprachhistorisch relevante Fragestellungen denkbar wären.

Zugleich stellt sich die Frage, inwieweit diese Anwendungsszenarien nicht auch unter „schlechteren" Bedingungen, d. h. anhand von rein automatisch erstellten Volltexten, zufriedenstellend erprobt werden hätten können. Eine der Voraussetzungen hierfür wäre sicherlich, dass Forschende imstande sind, sich selbst einen Eindruck von der Qualität der verwendeten Volltexte zu verschaffen. Ob die beispielhaft gewählte Untersuchung zu „seynd" oder „sind" in der *Wiener Zeitung* auch mit den in ANNO verfügbaren automatisch generierten digitalen Volltexten zu annähernd vergleichbaren Ergebnissen geführt hätte, wäre interessant zu erproben. Einerseits wären dort enorme Textmengen vorhanden, andererseits gingen gerade bei dieser Fragestellung wohl auch mehrere fälschlich als „feynd" und „find" erkannte „seynd" und „sind" für die Auswertung verloren, weil zwischen dem langen „s" und dem „f" im Frakturdruck große Ähnlichkeiten bestehen, wodurch diese Buchstaben häufig nicht richtig erkannt werden.[25]

In einer kürzlich veröffentlichten Studie versuchen Mark J. Hill und Simon Hengchen die Auswirkungen schlechter OCR auf die Analyse historischer Texte aus dem 18. Jahrhundert zu quantifizieren. Dabei nehmen sie an, dass Text, der zu 80% korrekt ist, keine deutlich schlechteren Ergebnisse liefert als 100% richtiger Text – allerdings gilt das nicht für alle Arten von Analysen: Während schlechte Textqualität sich auf Verfahren des Topic Modellings und der Autorschaftsattribuierung nur wenig auswirkt, zeigt sich etwa bei Kollokationsanalysen ein deutlich negativer Effekt, weil statistisch signifikante Informationen verloren gehen (vgl. 2019, 833). Klar ist, dass manche Methoden trotz schlechter OCR-Qualität[26] angewandt werden können, während der Erkenntniswert anderer Verfahren unter hohen Fehlerraten leidet beziehungsweise bei genauerem Hinsehen seine Glaubwürdigkeit einbüßt.

24 Paul Conway (2013, 27) unterscheidet in seiner Untersuchung kleinere Fehler, die bei größeren Textsammlungen durchaus akzeptiert würden, von gravierenden Fehlern, welche seiner Meinung nach die Vertrauenswürdigkeit von Portalen gefährden („fatal error compromises the integrity of large-scale digitization and threatens the long-term trustworthiness of repositories").

25 Wörter, die ein langes „s" enthalten, sind statistisch häufiger unter jenen Tokens, die falsch erkannt werden (vgl. Hill und Hengchen 2019, 829). Bei hochfrequenten Phänomenen würde dies – gemessen an den Datenmengen – möglicherweise nicht allzu stark ins Gewicht fallen beziehungsweise müsste man versuchen, diese falsch erkannten Formen zu identifizieren, um sie in der Auswertung berücksichtigen zu können.

26 Was dringend gebraucht und von Hill und Hengchen auch angekündigt wird, sind Modelle, welche die Genauigkeit von OCR-gelesenen Texten abschätzen können (vgl. 2019, 837).

Bei der Arbeit mit manuell nachkorrigierten historischen Zeitungstexten aus dem oberdeutschen Raum zum Thema der Sprachreform im 18. Jahrhundert hat sich jedenfalls bestätigt, dass qualitativ hochwertige Volltexte für bestimmte Forschungsintentionen (wie die einfache Suche nach Wortformen) eine zentrale und unverzichtbare Voraussetzung sein können, um Informationen zu lokalisieren und den Gebrauch und die Entwicklung von Sprache nachzuvollziehen. Die Tatsache, dass es bei der Distinktion von ostmitteldeutschen und oberdeutschen Varianten mitunter auf die Existenz oder Nicht-Existenz eines einzigen (zusätzlichen) Zeichens[27] ankommen kann, lässt eine möglichst zeichengetreue, verlässliche Texttranskription nach wie vor als erstrebenswertes und vielversprechendes Ideal erscheinen.

Bibliographie

Böning, Holger. 2011. Handgeschriebene und gedruckte Zeitung im Spannungsfeld von Abhängigkeit, Koexistenz und Konkurrenz. In: Volker Bauer und Holger Böning (Hg.): Die Entstehung des Zeitungswesens im 17. Jahrhundert. Ein neues Medium und seine Folgen für das Kommunikationssystem der Frühen Neuzeit. (Presse und Geschichte – Neue Beiträge 54). Bremen, S. 23–56.

Conway, Paul. 2013. Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust. In: Preservation, Digital Technology & Culture 42/1, S. 17–30. http://hdl.handle.net/2027.42/99522.

Durrell, Martin. 2017. Zeitungssprache und Literatursprach bei der Ausbildung standardsprachlicher Normen im Deutschen im 17. und 18. Jahrhundert. Ein Vergleich anhand eines repräsentativen Korpus. In: Oliver Pfefferkorn, Jörg Riecke und Britt-Marie Schuster (Hg.): Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung. Berlin, Boston, S. 81–98.

Durrell, Martin, Astrid Ensslin und Paul Bennett. 2009. Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert. In: Zeitschrift für deutsche Philologie 127, S. 263–79.

Flowerdew, Lynne. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. In: English for Specific Purposes 24, S. 321–332.

Frank, Peter R. und Johannes Frimmel (Hg.). 2008. Buchwesen in Wien 1750–1850. Kommentiertes Verzeichnis der Buchdrucker, Buchhändler und Verleger. (Buchforschung. Beiträge zum Buchwesen in Österreich 4). Wiesbaden.

27 Beispiele, in welchen die Existenz oder Nicht-Existenz des Buchstabens „e“ (wie in „seynd“ versus „sind“) zur Unterscheidung der ostmitteldeutschen von der oberdeutschen Schriftsprache ausschlaggebend ist, sind unter anderen: oberdeutsch „ihme“ versus ostmitteldeutsch „ihm“ oder oberdeutsch „nachdeme“ versus ostmitteldeutsch „nachdem“ – umgekehrt ist es bei oberdeutsch „Seel“ versus ostmitteldeutsch „Seele“, oberdeutsch „ersucht“ versus ostmitteldeutsch „ersuchet“ oder oberdeutsch „sah“ versus ostmitteldeutsch „sahe“.

Gloning, Thomas. 2017. Alte Zeitungen und historische Lexikographie. Nutzungsperspektiven, Korpora, Forschungsinfrastrukturen. In: Oliver Pfefferkorn, Jörg Riecke und Britt-Marie Schuster (Hg.): Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung. Berlin, Boston, S. 121–147.

Havinga, Anna D. 2018. Invisibilising Austrian German. On the effect of linguistic prescriptions and educational reforms on writing practices in 18th-century Austria. Berlin, Boston.

Hill, Mark J. und Simon Hengchen. 2019. Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study. In: Digital Scholarship in the Humanities 34/4. Oxford, S. 825–843.

Hitchcock, Tim. 2011. Academic History Writing and ist Disconnects. In: Journal of Digital Humanities 1. http://journalofdigitalhumanities.org/1-1/academic-history-writing-and-its-disconnects-by-tim-hitchcock/ (letzter Zugriff am 12. 12.2020).

Hosokawa, Hirofumi. 2014. Zeitungssprache und Mündlichkeit. Soziopragmatische Untersuchungen zur Sprache in Zeitungen um 1850. (= Kieler Forschungen zur Sprachwissenschaft Band 4). Frankfurt am Main.

Huijnen, Pim und Melvin Wevers. 2015. Digital Deconstruction. The Digital Turn and the use of news media as sources for historical research. https://www.officinadellastoria.eu/it/2015/09/14/digital-deconstruction-the-digital-turn-and-the-use-of-news-media-as-sources-for-historical-research/ (letzter Zugriff am 12. 12.2020).

Hundt, Marianne. 2008. Text corpora. In: Anke Lüdeling und Merja Kytö (Hg.): Corpus Linguistics. An International Handbook. Band 1 (Handbooks of Linguistics and Communication Science 29/1). Berlin, Boston, S. 168–187.

Kampkaspar, Dario. 2019. Das DIGITARIUM – Volltexterstellung und Nutzungsmöglichkeiten. In: Das Wien[n]erische Diarium im 18. Jahrhundert. Digitale Erschließung und neue Perspektiven (Teil I). Wiener Geschichtsblätter (74. Jahrgang, Heft 2), S. 131–135.

Koch, Peter und Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36, S. 15–43.

Koenen, Erik. 2018. Digitale Perspektiven in der Kommunikations- und Mediengeschichte. In: Publizistik 63/4, S. 535–556. Springer Link, doi:10.1007/s11616-018-0459-4.

Koolen, Marijn, Jasmijn van Gorp und Jacco van Ossenbruggen. 2019. Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice. In: Digital Scholarship in the Humanities 34/2, S. 368–85. academic.oup.com, doi:10.1093/llc/fqy048.

Mader-Kratky, Anna, Claudia Resch und Martin Scheutz. 2019. Das Wien[n]erische Diarium im 18. Jahrhundert. Neue Sichtweisen auf ein Periodikum im Zeitalter der Digitalisierung. In: Das Wien[n]erische Diarium im 18. Jahrhundert. Digitale Erschließung und neue Perspektiven (Teil I). Wiener Geschichtsblätter (74. Jahrgang, Heft 2), S. 93–113.

Milligan, Ian. 2013. Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010. The Canadian Historical Review 94/4, S. 540–69.

Mühlberger, Günter et al. 2019. Transforming scholarship in the archives through handwritten text recognition. In: Journal of Documentation 75/5, S. 954–976. https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html (letzter Zugriff am 12. 12.2020).

Müller, Maria Elisabeth und Maria Hermes-Wladarsch. 2017. Die Digitalisierung der deutschsprachigen Zeitungen des 17. Jahrhunderts – ein Projekt mit Komplexität! In: Oliver Pfefferkorn, Jörg Riecke und Britt-Marie Schuster (Hg.): Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung. Berlin, Boston, S. 39–59.

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Kay-Michael Würzner, Matthias Boenig, Elisa Herrmann, Volker Hartmann. 2019. OCR-D: An end-to-end open source OCR framework for historical printed documents. Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019) ACM, New York, S. 53–58. https://dl.acm.org/doi/10.1145/3322905.3322917.

Nicholson, Bob. 2013. The Digital Turn. Exploring the Methodological Possibilities of Digital Newspaper Archives. In: Media History 19/1, S. 59–73.

Polenz, Peter von. 2013. Deutsche Sprachgeschichte vom Mittelalter bis zur Gegenwart. Band II: 17. und 18. Jahrhundert. 2. Auflage, bearbeitet von Claudine Moulin unter Mitarbeit von Dominic Harion. Berlin, Boston.

Rachinger, Johanna. 2003. „Alles Denckwürdige, so von Tag zu Tag … ". In: 300 Jahre Wiener Zeitung. Eine Festschrift mit einem Begleitteil zur Ausstellung „Zeiten auf Seiten" in der Österreichischen Nationalbibliothek. Wien, S. 53.

Riecke, Jörg. 2016. Geschichte der deutschen Sprache. Stuttgart.

Resch, Claudia. 2018a. „Zeitungs Lust und Nutz" im digitalen Zeitalter. In: medien & zeit. Kommunikation in Vergangenheit und Gegenwart 2. Wien, S. 20–31.

Resch, Claudia. 2018b. Historische Zeitungen im digitalen Medium. In: Ingo Börner, Wolfgang Straub und Christian Zolles (Hg.): Germanistik digital: Digital Humanities in der Sprach- und Literaturwissenschaft. Wien, S. 183–198.

Resch, Claudia und Dario Kampkaspar. 2019. DIGITARIUM – Unlocking the Treasure Trove of 18th Century Newspapers for Digital Times. In: Thomas Wallnig, Marion Romberg und Joelle Weis (Hg.): Digital Eighteenth Century: Central European Perspectives. Wien, Köln, Weimar, S. 49–64.

Rössler, Paul. 2005. Schreibvariation – Sprachregion – Konfession. Graphematik und Morphologie in österreichischen und bayerischen Drucken vom 16. bis ins 18. Jahrhundert. (Schriften zur deutschen Sprache in Österreich 35). Frankfurt am Main, Bern, Wien, Paris, New York.

Scheible, Silke, Richard J. Whitt, Martin Durrell und Paul Bennett (2011): Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In: Marek Konopka, Jacqueline Kubczak, Christian Mair, František Šticha und Ulrich H. Waßner (Hg.): Grammatik und Korpora 2009. Tübingen, S. 539–547.

Schneider, Kristina. 2002. The development of popular journalism in England from 1700 to present: corpus compilation and selective stylistic analysis. Rostock.

Schuster, Britt-Marie und Manuel Wille. 2017. Die Volltextdigitalisierung der „Stats- und gelehrten Zeitung des Hamburgischen Correspondenten" und ihrer Vorgänger (1712–1848) und ihr Nutzen: Befunde zur Genese und zum Wandel von Textmustern. In: Oliver Pfefferkorn, Jörg Riecke und Britt-Marie Schuster (Hg.): Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung. Berlin, Boston, S. 99–119.

Smith, David A. und Ryan Cordell. 2018. A Research Agenda for Historical and Multilingual Optical Character Recognition. https://repository.library.northeastern.edu/downloads/neu:m043qk202?datastream_id=content (letzter Zugriff am 12. 12. 2020).

Stieler, Kaspar von. 1695. Zeitungs Lust und Nutz/ Oder: derer so genanten Novellen oder Zeitungen/ wirckende Ergetzlichkeit/ Anmut/ Notwendigkeit und Frommen […] Hamburg.

Wiesinger, Peter. 2014. Das österreichische Deutsch in Gegenwart und Geschichte. 3., aktualisierte und neuerlich erweiterte Auflage. (= Austria: Forschung und Wissenschaft. Literatur- und Sprachwissenschaft Band 2). Wien.

Wijfjes, Huub. 2017. Digital Humanities and Media History. A Challenge for Historical Newspaper Research. In: Tijdschrift Voor Mediageschiedenis 20/1, 4–24. www.tmgonline.nl, doi:10.18146/tmg20277 (letzter Zugriff am 12.12. 2020).

Verwendete Daten und Tools

Çetin, Asil. 2020. CorpSum: Multi-Faceted Visual Corpus Analysis Tool (letzter Zugriff am 12.12.2020).
Resch, Claudia und Dario Kampkaspar (Hg.). 2020. Wienerisches DIGITARIUM. https://digitarium.acdh.oeaw.ac.at (letzter Zugriff am 12.12.2020).
Sketch Engine. 2020. https://www.sketchengine.eu/ (letzter Zugriff am 12.12.2020).
Transkribus. 2020. https://transkribus.eu/Transkribus/ (letzter Zugriff am 12.12.2020).

Claire-Lise Gaillard

Feuilleter la presse ancienne par gigaoctets

Abstract: Digitisation and digital methods are profoundly reshaping our approach to and use of historical newspapers. The online availability of major daily titles not only facilitates the work of historians, but also makes it more systematic. These developments offer new ways of looking at the breadth of historical newspapers — from the best-known title to the least-known — and the immediacy of the archive makes it possible to write new parts of history that have remained in the shadows until now, as the author has experienced in working on the history of the matchmaking market. Despite the exciting new doors it opens, we must be careful of the biases this paradigm shift induces in our research. This article proposes to identify these limitations and to consider some precautions to avoid them.

Keywords: digitised newspapers, interface critique, matchmaking market

Le recours à Gallica est désormais un des réflexes primaires de tout·e·s historien·ne·s.[1] Dans les bibliothèques numériques, la presse ancienne a occupé une place pionnière au sein des programmes de numérisation. Elle a par exemple été au cœur des premiers efforts de la Bibliothèque nationale de France en la matière[2]. L'accès en ligne à n'importe quel numéro des principaux quotidiens du XIX^{e} et XX^{e} siècles rend l'utilisation des archives de la presse ancienne non seulement plus simple mais plus systématique dans nombre de travaux d'historien·ne·s. Le corpus Europeana Newspaper[3] propose par exemple de télécharger en masse les principaux quotidiens européens. Cette immédiateté redéfinit notre goût de l'archive car elle remodèle en profondeur notre ergonomie de travail.

1 Ce chapitre est une réécriture du chapitre du même titre publié en ligne sur *Le goût de l'archive à l'ère numérique*: Frédéric Clavert et al., eds. (2017). Le goût de l'archive à l'ère numérique. https://gout-numerique.net/.

2 Moreux, Jean-Philippe, « Livre numérique accessible et numérisation de masse à la BnF: retour d'expérience », 9^{e} Forum européen de l'accessibilité numérique, 8 juin 2015, Paris.

3 https://www.europeana.eu/fr - Voir à ce propos Neudecker, Clemens and Apostolos Antonacopoulos, « Making Europe's Historical Newspapers Searchable » 2016 12th IAPR Workshop on Document Analysis Systems (DAS), IEEE, 2016 ; mais aussi Pfanzelter, Eva, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais et Stefan Hechl, « Digital interfaces of historical newspapers: opportunities, restrictions and recommendations » 2020 ⟨hal-02480654v4⟩.

https://doi.org/10.1515/9783110729214-006

1 D'un geste à l'autre : l'ergonomie du clic

La numérisation de la presse ancienne modifie en profondeur le geste même de découverte des documents, en particulier parce qu'elle supprime l'intermédiaire de l'archiviste au profit de l'interface numérique. Cette reconfiguration donne le sentiment que tout est déjà là, sous la main, à disposition.

Cette disponibilité quasi immédiate des titres de presse ancienne *via* différents portails numériques redéfinit notre rapport aux revues et journaux qui d'abord se déplient, se feuillettent, en somme se manipulent. Dans le cadre de mes recherches sur l'histoire du marché de la rencontre, j'ai eu à faire ressortir des fonds de la Bibliothèque nationale de France (BnF) bien des titres de presse matrimoniale de la fin du XIX^{e} siècle jusqu'à l'entre-deux-guerres. Quiconque a eu affaire à ce genre de presse sait le rôle de leur matérialité, le papier utilisé par la presse d'entre-deux-guerres, particulièrement fragile, est la cause de bien des refus de consultations. L'encre s'efface facilement et colore les doigts au fil de la lecture. Sur les plateformes numériques, nous perdons nécessairement cette dimension sensible de l'archive. Et dès lors nous n'avons pas la même expérience de lecture que celle des acteurs et actrices que nous étudions. On pourra m'objecter avec raison que notre expérience de lecture est nécessairement guidée par notre posture de recherche – et c'est juste. Lorsque je compile les annonces matrimoniales, que je les recopie une par une, j'en ai une lecture toute différente de celle des célibataires qui y ont cherché leur conjoint, en ciblant d'emblée les pages et profils les plus intéressants. Néanmoins il me semble que les outils numériques dont nous disposons nous éloignent encore davantage de la matérialité que nous pouvons partager avec les primes usagers de nos sources.

Pour Arlette Farge, l'archive est « difficile dans sa matérialité [...] parce que démesurée, envahissante comme les marées d'équinoxes, les avalanches ou les inondations »[4]. Dans le cas de la presse ancienne en ligne, c'est au contraire l'immatérialité de l'archive qui nous submerge. Le numérique semble dans un premier temps donner plus d'emprise sur les choses, puisque les mots-clés, les tris par pertinence ou par chronologie sont autant de manières d'agripper le texte, de faire ressurgir un sujet. Impossible il y a vingt ans d'entreprendre, comme je le fais, l'histoire des annonces et des agences matrimoniales. Et c'est probablement pour cette raison que cette page d'histoire était restée assez blanche. Pour faire ressortir de la presse ancienne un sujet si marginal, il aurait fallu, sans l'aide du numérique, dépouiller de façon systématique quelques titres pour espérer, sans aucune garantie de succès, tomber sur les récits que

4 Farge, Arlette, *Le goût de l'archive,* Éditions du Seuil, 1989, p. 10.

certains journalistes font de leur expérience des agences, ou lister les publicités matrimoniales à la quatrième page des journaux. Et pourtant, bien que le numérique nous donne les outils pour mesurer la profondeur des fonds, rien n'est plus simple que de s'y perdre.

La métaphore marine s'impose pour Arlette Farge afin d'évoquer le sentiment de profondeur par lequel on entre dans les archives. Il me semble que l'archive numérique aplanit cette perception des choses, notamment parce que la reconnaissance plein texte – disponible sur de plus en plus de titres de presse – nous permet de passer outre les subdivisions des fonds et d'aller directement au paragraphe qui nous intéresse sans avoir été introduits par un inventaire, une introduction, puis un chapitre. La technique de reconnaissance optique des caractères (en anglais *optical character recognition* ou OCR) permet en effet aux plateformes de presse ancienne en ligne de superposer le texte à l'image. Les sources ne sont plus de simples photos, elles se nourrissent d'une nouvelle couche d'informations puisque l'on peut les interroger par le texte, au-delà de la simple notice descriptive. Aussi est-il aisé de travailler par mots-clés, et de s'en constituer tout un répertoire qui puisse refléter au mieux les contours du sujet étudié. Dans le cadre de mon travail cette liste de mots-clés s'est constituée au fil de la lecture des résultats et n'a cessé d'évoluer : dans le cadre de ma recherche, « agence matrimoniale » et ses dérivés s'imposent, mais on constate rapidement dans les résultats des termes cooccurrents comme « hymen », « mariages riches », « industrie matrimoniale », « courtage matrimonial », « grandes relations », « faciliter mariages », et la liste est longue. Dans Gallica par exemple, la recherche simple « agence matrimoniale[5] » fait ressortir 939 documents, qui traitent très inégalement du thème: il peut aussi bien s'agir de la publicité d'une agence, d'une mention au passage d'un roman feuilleton, que d'une enquête journalistique dédiée au sujet (voir figure 1). C'est le mot-clé qui, surligné en jaune, conditionne la lecture du document concerné. Si bien qu'on a parfois l'impression qu'il se suffit à lui-même, alors que c'est précisément le contexte dans lequel il s'insère qui est signifiant.

Les bases de données ont d'ailleurs cet intérêt d'afficher l'information voulue en un clic (ou presque) sans avoir à passer par les ramifications interminables de métadonnées afin d'avoir directement accès au contenu. Gallica nous fait facilement oublier les profondeurs de ses ressources en faisant une sélection spécifique à notre sujet. Travailler sur les résultats d'une requête Gallica nous amène nécessairement à penser les documents qui ressortent les uns par rapport aux

5 https://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve&version=1.2&query=%28gallica%20adj%20%22agence%20matrimoniale%22%29&lang=fr&suggest=0.

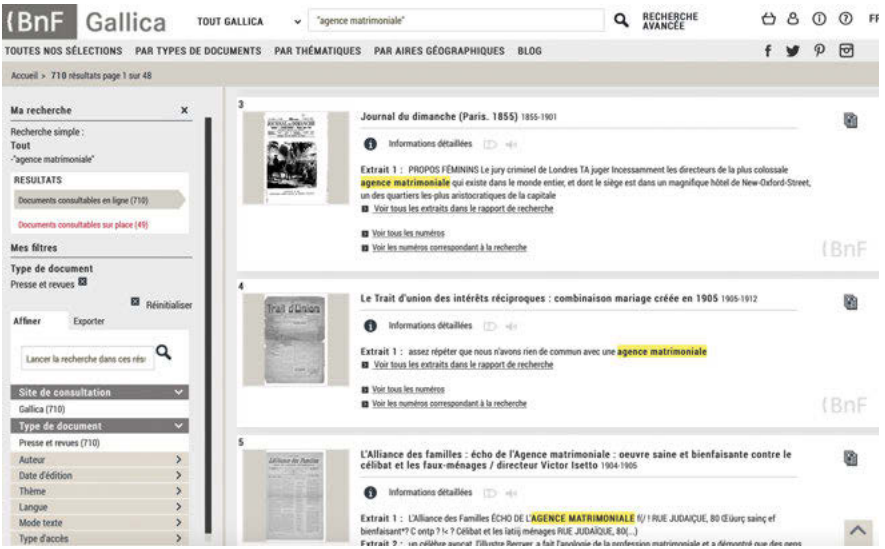


Fig. 1: Capture d'écran des résultats de la recherche « agence matrimoniale » sur Gallica. URL : https://gallica.bnf.fr [Consulté en Décembre 2020]

autres, or, ce lien de coprésence dans un même panel de résultats est parfois le seul lien que les documents entretiennent entre eux. Ce corpus de documents n'existe que parce que la requête a été faite, il n'a pas d'autre raison d'être que notre recherche. La porte d'entrée par mots-clés court-circuite l'entrée thématique que l'on pourrait avoir dans un inventaire d'archives. Le gain de temps est énorme pour l'historien·ne, mais cette efficacité a pour conséquence de changer la voie d'accès aux sources – ce qui doit être gardé en tête. Cette entrée « par le bas » de la presse ne peut donc jamais se suffire à elle-même. Il nous revient toujours de recontextualiser ces occurrences, de les comprendre dans un genre de presse spécifique, un rythme de publication, un type de lectorat, etc.

Par ailleurs, l'existence ou non d'une numérisation conditionne en partie les choix de nos corpus. Lorsque vient le moment de choisir les journaux sur lesquels travailler, la presse ancienne en ligne offre des avantages non négligeables par rapport aux titres qui ne sont disponibles qu'en papier ou sur microfilms. Le confort de travail se comprend aisément: Gallica permet de travailler hors des murs de la Bnf, n'importe où pourvu qu'une connexion internet soit garantie. À intérêt scientifique égal, on privilégie forcément pour un traitement sériel le journal numérisé à celui qui ne l'est pas. À plus forte raison si l'on prévoit de passer les images à l'OCR, une numérisation vaut mieux que des photographies prises

manuellement. Car on peut ajouter un second degré de discrimination entre les journaux: ceux qui sont passés par un logiciel de reconnaissance de caractères ressortent nécessairement davantage dans les résultats de recherche que les autres pour lesquels seules les images sont disponibles. En d'autres termes, l'accessibilité de la source façonne nos corpus de recherche, et donc nos résultats[6].

Ce constat n'est pas une fatalité. Il y a évidemment des garde-fous à poser. Le premier est précisément d'avoir cette posture réflexive qui consiste à lever cette illusion d'immédiateté de l'archive. Les requêtes sur les portails numériques ne permettent pas de délimiter les corpus de recherche, mais elles nous indiquent où chercher. C'est par exemple la recherche par mots-clés dans Gallica qui m'a aiguillée vers *La vie Parisienne*[7], que j'ai ensuite dépouillé entièrement, en alternant recherche plein texte et recherche manuelle pour y recenser les publicités d'agences matrimoniales sur toute la période de publication de la revue[8]. Au moment de rendre compte des résultats de l'enquête historique, il faudra donner aux lecteurs les clés pour retracer les étapes de cette co-construction des corpus. Ce principe de transparence scientifique n'est pas propre aux recherches qui ont recours aux humanités numériques, mais nous ne pouvons d'autant moins en faire l'économie, qu'une partie de la démarche dépend des fonctionnalités du portail numérique.

2 Le flux et le stock : voracité et infobésité

On se perd sur Gallica comme on se perd sur internet : de lien en lien, de page en page, un document en amène un autre dans le grand réseau que représente la base de données. Beaucoup de trouvailles sont faites au passage, en cherchant tout autre chose, ou au détour du lien d'un·e collègue sur Twitter. Les informations passent et puis se perdent (dans un carnet de notes, dans un post-it numérique ou manuscrit). D'où l'obsession du rongeur : il faut stocker et organiser ce qui est vu « au passage », ou « pour plus tard », tout en sachant que l'information reste disponible en ligne. Aussi, sans planification préalable, il

6 Putnam, Lara, « The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast », *The American Historical Review*, Volume 121, Issue 2, April 2016, Pages 377–402, https://doi.org/10.1093/ahr/121.2.377.

7 https://gallica.bnf.fr/ark:/12148/bpt6k1253730h/f712.image.r=%22agence%20matrimoniale%22?rk=21459;2.

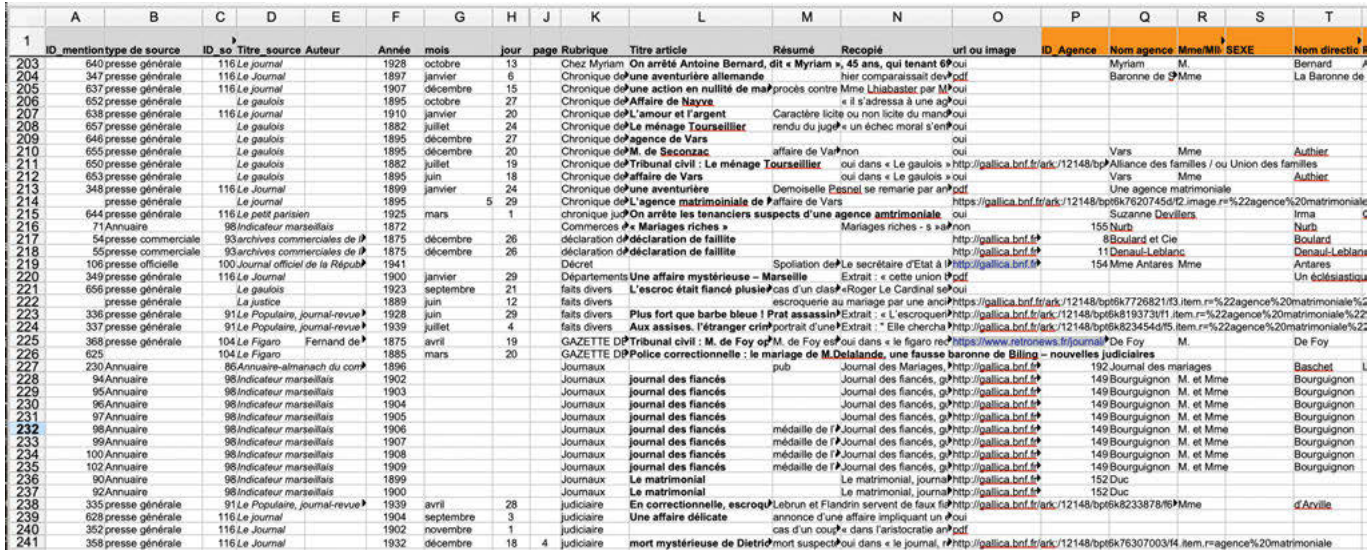
8 https://gallica.bnf.fr/ark:/12148/cb328892561/date.r=%22agence%20matrimoniale%22%20la%20vie%20parisienne.

est facile d'ouvrir frénétiquement des fichiers textes, des dossiers d'images, de stocker des informations dans Zotero, pour plus tard – un plus tard toujours un peu hypothétique et parfois illusoire. À titre personnel, j'ai tenté de parer à cet éparpillement en consignant, dans un tableur, toutes les occurrences qui passent sous mes yeux. L'idée n'est évidemment pas de tout traiter sur le même plan. Pour chaque occurrence je précise s'il s'agit d'un recensement exhaustif, échantillonné ou au contraire ponctuel (voir figure 2). Le tableur permet de garder le lien avec la source (document téléchargé ou URL de Gallica), mais surtout il permet de faire des rapprochements entre les différentes occurrences: par type de presse, par thème, par année, par rubrique dans le journal (à l'aide par exemple de tris dans le tableur ou de tableaux croisés dynamiques).

La facilité d'accès à l'information augmente considérablement la part de la presse dans nos corpus de sources (et des sources imprimées numérisées de façon générale). On est ainsi parfois confronté à l'hypertrophie de la presse dans les corpus ; il faut alors penser le biais qu'elle représente et ne pas surestimer les représentations journalistiques par rapport à celles que véhiculent d'autres productions culturelles, ni même mesurer l'ampleur d'un phénomène culturel et social par son empreinte médiatique. Le cas des agences matrimoniales du XIX^{e} siècle est symptomatique de ce point de vue-là. En approchant le sujet par la presse, j'ai d'abord été frappée par sa récurrence dans la presse de la fin du siècle, au point de s'ériger parmi les sujets de société. Mais en confrontant cet imaginaire médiatique à d'autres archives, j'ai constaté une disproportion des discours face à la marginalité numérique des agences. Cette inflation médiatique relève ici de l'histoire des imaginaires, mais pour la percevoir il faut d'abord interpréter la profusion des résultats de requête des portails numériques.

En effet, selon le sujet traité, il n'est pas rare qu'en interrogeant Gallica on soit confronté à une surabondance de résultats. La presse, comme d'autres types de sources d'ailleurs, traite à peu près de tous les sujets ; aussi faut-il se prémunir de l'effet loupe que produit la recherche lexicale. En l'état actuel des outils disponibles sur Gallica, il est délicat d'estimer l'importance médiatique d'un sujet. Les rapports de recherche exportables que propose le site sont certes bien utiles, mais seulement s'il y a moins de 50 résultats. Le site Retronews[9] propose pour sa part de calculer une « fréquence du terme » sur un nombre de titres de presse choisi, mais la courbe qui est dessinée représente le nombre de pages dans lequel apparaît le terme : dans la majorité des cas, elle montre davantage l'évolution du nombre de pages des journaux sur la période concernée que l'importance du sujet en question. Sans accès à la table de données servant

9 https://www.retronews.fr/.



	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T
1	ID_mention	type de source	ID_so	Titre_source	Auteur	Année	mois	jour	page	Rubrique	Titre article	Résumé	Recopié	url ou image	ID_Agence	Nom agence	Mme/MlleSEXE		Nom directic
203	640	presse générale	116	*Le journal*		1928	octobre	13		Chez Myriam	**On arrêté Antoine Bernard, dit « Myriam », 45 ans, qui tenant 6**		oui			Myriam	M.		Bernard
204	347	presse générale	116	*Le Journal*		1897	janvier	6		Chronique de	**une aventurière allemande**		hier comparaissait dev	pdf		Baronne de S	Mme		La Baronne de
205	637	presse générale	116	*Le journal*		1907	décembre	15		Chronique de	**une action en nullité de ma**	procès contre	Mme Lhiabaster par M	oui					
206	652	presse générale		*Le gaulois*		1895	octobre	27		Chronique de	**Affaire de Nayve**		« il s'adressa à une ag	oui					
207	638	presse générale	116	*Le journal*		1910	janvier	20		Chronique de	**L'amour et l'argent**	Caractère licite	ou non licite du mand	oui					
208	657	presse générale		*Le gaulois*		1882	juillet	24		Chronique de	**Le ménage Tourseillier**	rendu du juge	« un échec moral s'en	oui					
209	646	presse générale		*Le gaulois*		1895	décembre	27		Chronique de	**agence de Vars**			oui					
210	655	presse générale		*Le gaulois*		1895	décembre	20		Chronique de	**M. de Seconzac**	affaire de Var	non	oui		Vars	Mme		Authier
211	650	presse générale		*Le gaulois*		1882	juillet	19		Chronique de	**Tribunal civil : Le ménage Tourseillier**		oui dans « Le gaulois »	http://gallica.bnf.fr/ark:/12148/bp		Alliance des familles / ou Union des familles			
212	653	presse générale		*Le gaulois*		1895	juin	18		Chronique de	**affaire de Vars**		oui dans « Le gaulois »	oui		Vars	Mme		Authier
213	348	presse générale	116	*Le Journal*		1899	janvier	24		Chronique de	**une aventurière**	Demoiselle Pesnel	se remarie par ar	pdf		Une agence matrimoniale			
214		presse générale		*Le journal*		1895		5	29	Chronique de	**L'agence matrimoniale de**	affaire de Vars		https://gallica.bnf.fr/ark:/12148/bpt6k7620745d/f2.image.r=%22agence%20matrimoniale					
215	644	presse générale	116	*Le petit parisien*		1925	mars	1		chronique juc	**On arrête les tenanciers suspects d'une agence amtrimoniale**		oui			Suzanne Devillers			Irma
216	71	Annuaire	98	*Indicateur marseillais*		1872				Commerces d	**« Mariages riches »**		Mariages riches - s »	non	155	Nurb			Nurb
217	54	presse commerciale	93	*archives commerciales de l*		1875	décembre	26		déclaration d	**déclaration de faillite**			http://gallica.bnf.fr	8	Boulard et Cie			Boulard
218	55	presse commerciale	93	*archives commerciales de l*		1875	décembre	26		déclaration d	**déclaration de faillite**			http://gallica.bnf.fr	11	Denaul-Leblanc			Denaul-Leblanc
219	106	presse officielle	100	*Journal officiel de la Répub*		1941				Décret		Spoliation de	Le secrétaire d'Etat à l	http://gallica.bnf.fr	154	Mme Antares	Mme		Antares
220	349	presse générale	116	*Le Journal*		1900	janvier	29		Départements	**Une affaire mystérieuse – Marseille**		Extrait : « cette union	pdf					Un éclésiastiqu
221	656	presse générale		*Le gaulois*		1923	septembre	21		faits divers	**L'escroc était fiancé plusie**	cas d'un clas	«Roger Le Cardinal se	oui					
222		presse générale		*La justice*		1889	juin	12		faits divers		escroquerie au	mariage par une anci	https://gallica.bnf.fr/ark:/12148/bpt6k7726821/f3.item.r=%22agence%20matrimoniale%2					
223	336	presse générale	91	*Le Populaire, journal-revue*		1928	juin	29		faits divers	**Plus fort que barbe bleue ! Prat assassin**		Extrait : « L'escroquer	http://gallica.bnf.fr/ark:/12148/bpt6k819373t/f1.item.r=%22agence%20matrimoniale%22					
224	337	presse générale	91	*Le Populaire, journal-revue*		1939	juillet	4		faits divers	**Aux assises. l'étranger crin**	portrait d'une	Extrait : " Elle chercha	http://gallica.bnf.fr/ark:/12148/bpt6k823454d/f5.item.r=%22agence%20matrimoniale%22					
225	368	presse générale	104	*Le Figaro*	Fernand de	1875	avril	19		GAZETTE DE	**Tribunal civil : M. de Foy or**	M. de Foy est	oui dans « le figaro re	https://www.retronews.fr/journal		De Foy	M.		De Foy
226	625		104	*Le Figaro*		1885	mars	20		GAZETTE DE	**Police correctionnelle : le mariage de M.Delalande, une fausse baronne de Billing – nouvelles judiciaires**								
227	230	Annuaire	86	*Annuaire-almanach du com*		1896				Journaux		pub	Journal des Mariages,	http://gallica.bnf.fr	192	Journal des mariages			Baschet
228	94	Annuaire	98	*Indicateur marseillais*		1902				Journaux	**journal des fiancés**		Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
229	95	Annuaire	98	*Indicateur marseillais*		1903				Journaux	**journal des fiancés**		Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
230	96	Annuaire	98	*Indicateur marseillais*		1904				Journaux	**journal des fiancés**		Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
231	97	Annuaire	98	*Indicateur marseillais*		1905				Journaux	**journal des fiancés**		Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
232	98	Annuaire	98	*Indicateur marseillais*		1906				Journaux	**journal des fiancés**	médaille de l'	Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
233	99	Annuaire	98	*Indicateur marseillais*		1907				Journaux	**journal des fiancés**	médaille de l'	Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
234	100	Annuaire	98	*Indicateur marseillais*		1908				Journaux	**journal des fiancés**	médaille de l'	Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
235	102	Annuaire	98	*Indicateur marseillais*		1909				Journaux	**journal des fiancés**	médaille de l'	Journal des fiancés, g	http://gallica.bnf.fr	149	Bourguignon	M. et Mme		Bourguignon
236	90	Annuaire	98	*Indicateur marseillais*		1899				Journaux	**Le matrimonial**		Le matrimonial, journa	http://gallica.bnf.fr	152	Duc			
237	92	Annuaire	98	*Indicateur marseillais*		1900				Journaux	**Le matrimonial**		Le matrimonial, journa	http://gallica.bnf.fr	152	Duc			
238	335	presse générale	91	*Le Populaire, journal-revue*		1939	avril	28		judiciaire	**En correctionnelle, escroq**	Lebrun et Flandrin	servent de faux fi	http://gallica.bnf.fr/ark:/12148/bpt6k8233878/f6	Mme				d'Arville
239	628	presse générale	116	*Le journal*		1904	septembre	3		judiciaire	**Une affaire délicate**		annonce d'une affaire impliquant un	oui					
240	352	presse générale	116	*Le Journal*		1902	novembre	1		judiciaire			cas d'un coup	« dans l'aristocratie ar	pdf				
241	358	presse générale	116	*Le Journal*		1932	décembre	18	4	judiciaire	**mort mystérieuse de Dietri**	mort suspect	oui dans « le journal, r	http://gallica.bnf.fr/ark:/12148/bpt6k76307003/f4.item.r=agence%20matrimoniale					

Fig. 2: Capture d'écran du tableur qui me permet de recenser les occurrences relevant du marché de la rencontre dans la presse ancienne. [Capture Décembre 2020]

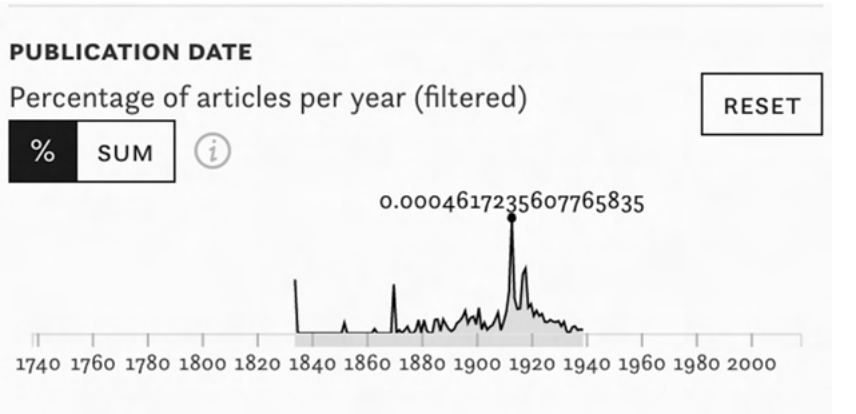


Fig. 3: Capture d'écran de l'interface *impresso* : occurrences de l'expression « annonces matrimoniales » rapporté au nombre d'articles que compte la plateforme par année. URL : https://impresso-project.ch/app [Consulté le 9 Décembre 2020]

à dessiner ce graphe, impossible de faire apparaître de véritables proportions. La plateforme Impresso[10], par exemple, propose de représenter dans le temps le nombre d'articles concernés par l'occurrence recherchée, rapporté au total des articles disponibles dans le portail (voir figure 3)[11].

Il faudra probablement attendre la mise en place de véritables outils statistiques qui allient les possibilités de Gallica et de la textométrie pour avoir des résultats sur un corpus français de grande ampleur[12]. C'est d'ailleurs à ce travail

10 https://impresso-project.ch/app/.

11 La plateforme est transparente sur son mode de calcul et explique : « SUM displays the absolute frequencies of result hits, whereas % shows the relative number of hits, compared to all articles from the impresso corpus. The highest result per year is displayed in number. The impresso interface groups the results per article, so the displayed frequency is the count of all articles containing at least one occurrence of the keyword. In other words, it is not the raw number of hits for the searched key-word, but the number of hits grouped by articles, as shown in the search summary. For instance, a query for 'Einstein' returns 8 967 articles. These 8 967 articles contain 13 535 individual hits of 'Einstein'. An item is included in the search when it contains at least one character, this means that the search is conducted also within advertisements and tables when they contain text. » https://impresso-project.ch/app/faq#relative-vs-absolute-year-graph.

12 L'outil *Gallicagram*, sur le modèle du logiciel *Ngram* Viewer de Google, permet désormais de produire des visualisations chronologiques des fréquences d'usages d'un mot sur l'ensemble des ressources de Gallica. Contrairement à *Ngram Viewer*, l'outil garantit la maîtrise et la transparence du corpus. Cf Azoulay, Benjamin & de Courson, Benoît. (2021). Gallicagram : un outil de lexicométrie pour la recherche. 10.31235/osf.io/84bf3.

que s'attelle Pierre-Carl Langlais dans le cadre de l'ANR Numapresse[13]. L'idée de ce travail collectif est d'analyser la presse ancienne à grande échelle avec des méthodes automatisées de fouilles de données[14] pour faire apparaître des dynamiques invisibles à l'œil nu comme la vitalité des contenus médiatiques par exemple.

Bien que tentante, l'exhaustivité des résultats est donc illusoire. Il faut alors trouver des parades, constituer des échantillons et, bien entendu, garder en tête qu'il est difficile de faire confiance aux recherches d'occurrences dans un corpus exhaustif, dans la mesure où l'OCR laisse toujours des erreurs[15]. Si le logiciel de reconnaissance de caractères indique « matrimonia1e » au lieu de « matrimoniale », le document n'apparaîtra pas dans les résultats de ce mot-clé. De même si le mot est coupé par un tiret dans la mise en page du journal.

3 Lorsque l'œil ne suffit plus : de nouvelles pratiques de la presse ancienne

Lorsque vient le moment de rassembler les choses, de faire émerger un propos, d'autres questions se posent. La facilité d'accès suscite également la possibilité de ne jamais réellement se confronter à la matérialité de la source : elle est là, sous la main, dans un dossier, à portée de clic, sagement référencée par un système de tags que l'on aura pris soin de définir à l'avance.

Aussi est-il possible d'avoir rassemblé et organisé un corpus de sources important sans savoir vraiment de quoi celui-ci est fait ; et ce en faisant confiance

13 https://anr.fr/Projet-ANR-17-CE27-0014.

14 Moreux, Jean-Philippe, « Approches innovantes pour la presse ancienne numérisée: fouille et visualisation de données », Carnet de la recherche à la Bibliothèque nationale de France, ISSN 2493–4437, 30 décembre 2016. Disponible en ligne : https://bnf.hypotheses.org/208 (consulté le 8 décembre 2020).

15 Magallon, Thibault, Frédéric Béchet et Benoit Favre, « Détection d'erreurs dans des transcriptions OCR de documents historiques par réseaux de neurones récurrents multi-niveau », 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, France. Voir également : Chiron, Guillaume, Jean-Philippe Moreux, Antoine Doucet, Mickaël Coustaty et Muriel Visani, « Erreurs OCR et biais d'indexation: impact sur les usages », 17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computationnel, Jan 2017, Grenoble, France, p. 69–73 ⟨hal-01455763⟩. Hill, Mark J. et Simon Hengchen, « Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study », in: *Digital Scholarship in the Humanities*, vol. 34, no 4, Oxford Academic, décembre 2019, p. 825–43 (academic.oup.com), DOI :10.1093/llc/fqz024.

aux mots-clés, aux occurrences, à l'OCR – tous ces adjuvants qui rendent à la fois la recherche plus facile et plus lointaine.

Lorsque le corpus de presse rassemblé est trop important, il est possible de le faire passer par des logiciels d'analyse pour y faire émerger des dynamiques tantôt invisibles à l'œil nu, tantôt que l'on peut seulement pressentir, mais non objectiver. Les méthodes d'analyse de la presse numérisée se renouvellent et se complexifient. La textométrie[16] en est un bon exemple, elle permet de dessiner le lexique d'un texte, de comprendre quels sont les termes les plus fréquemment utilisés, quels champs lexicaux dominent, mais aussi d'analyser les co-occurrences de termes clés. On peut également penser au *topic modeling*, une méthode statistique qui permet de prédire la thématique d'un texte et dont les résultats peuvent faciliter la recherche sur de grandes masses de données (par ex. via des filtres sur la base des *topics*) ou contribuer à l'analyse d'une question spécifique[17], ou encore à la reconnaissance d'entités nommées (noms propres) dans les textes, permettant un point d'entrée 'référentiel' et non plus uniquement lexical, sur les collections[18].

Avec tous ces outils, on peut avoir l'illusion momentanée que le logiciel nous dispense de l'analyse puisque nous sommes dépassés par la masse. Aussi est-il important de suspendre ces moments de pêche frénétique de l'information (l'envie de tester tel mot-clé, telle rubrique, telle fonctionnalité), pour prendre le temps de la lecture, sans adjuvant, sans logiciel, sans mise en gras des mots recherchés qui guident tout de suite l'œil sur la rubrique du journal que l'on cherche. Laisser l'œil se perdre, c'est aussi laisser l'œil reprendre le rythme de la lecture, tranquille, parfois monotone, mais plus seulement chasseur. Ce tra-

16 http://textometrie.ens-lyon.fr.

17 Je vous renvoie ici au billet de Pierre-Carl Langlais sur le carnet de Numapresse (https://numapresse.hypotheses.org/11), mais aussi à la fonctionnalité de filtre sur la base de topics implémentée dans l'interface impresso (https://impresso-project.ch/app), ainsi qu'aux nombreux travaux sur le topic modelling appliqué aux archives, par ex. Hengchen, S., M. Coeckelbergs, S. van Hooland, R. Verborgh et T. Steiner, « Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the European Commission », 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, p. 3245–3249 ; Jockers, Matthew et David Mimno, « Significant Themes in 19th-Century Literature », Poetics 41.6 (2013), p. 750–769 ; Yang, Tze-I, Andrew J. Torget et Rada Mihalcea, « Topic Modeling on Historical Newspapers », Conference Proceedings of the Association for Computational Linguistics Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011), June 2011, p. 96–104, pour n'en citer que quelques-uns.

18 Voir à ce sujet la récente vue d'ensemble par Ehrmann, M., M. Romanello, A. Flückiger, et S. Clematide, « Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers », CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2020, vol. 2696, p. 38, DOI : 10.5281/zenodo.4117566.

vail de va-et-vient entre la lecture et la requête informatique est nécessaire pour que l'analyse se nourrisse de ces deux niveaux de compréhension de sources.

En somme, la dimension numérique de l'archive est à utiliser comme un indice ou comme un outil et non comme un résultat de recherche. Les requêtes sur les portails numériques ne rendent donc pas le rapport à la source plus immédiat, mais rajoutent au contraire une couche d'informations qu'il faut interroger pour ne pas être dupe de leur construction, mais qui élargissent considérablement le champ des analyses possibles. L'immatérialité de l'archive en fait donc un nouveau palimpseste numérique : les pixels, l'OCR, les métadonnées, les algorithmes des portails de recherche ou de fouille de données sont autant de niveaux d'écriture qui se superposent. Il est indispensable d'en faire le feuilletage et de les rendre visibles pour en tirer le meilleur profit.

Bibliographie

Azoulay, Benjamin et Benoît de Courson. 2021. "Gallicagram : Un Outil De Lexicométrie Pour La Recherche." SocArXiv. doi: 10.31235/osf.io/84bf3.

Chiron, Guillaume, Jean-Philippe Moreux, Antoine Doucet, Mickaël Coustaty et Muriel Visani, « Erreurs OCR et biais d'indexation: impact sur les usages », 17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computationnel, janvier 2017, Grenoble, France, p. 69–73 ⟨hal-01455763⟩.

Ehrmann, M., M. Romanello, A. Flückiger, et S. Clematide, « Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers », CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2020, vol. 2696, p. 38, DOI : 10.5281/zenodo.4117566.

Farge, Arlette, *Le goût de l'archive*, Éditions du Seuil, 1989, p. 10.

Hengchen, S., M. Coeckelbergs, S. van Hooland, R. Verborgh et T. Steiner, « Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the European Commission », 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, p. 3245–3249.

Hill, Mark J. et Simon Hengchen, « Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study », in: Digital Scholarship in the Humanities, vol. 34, no 4, Oxford Academic, décembre 2019, p. 825–43 (academic.oup.com), DOI :10.1093/llc/fqz024.

Jockers, Matthew et David Mimno, « Significant Themes in 19th-Century Literature », Poetics 41.6 (2013), p. 750–769; Yang, Tze-I, Andrew J. Torget et Rada Mihalcea, « Topic Modeling on Historical Newspapers », Conference Proceedings of the Association for Computational Linguistics Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011), June 2011, p. 96–104,

Magallon, Thibault, Frédéric Béchet et Benoit Favre, « Détection d'erreurs dans des transcriptions OCR de documents historiques par réseaux de neurones récurrents multi-

niveau », 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, France.

Moreux, Jean-Philippe, « Approches innovantes pour la presse ancienne numérisée: fouille et visualisation de données », Carnet de la recherche à la Bibliothèque nationale de France, ISSN 2493-4437, 30 décembre 2016. Disponible en ligne : https://bnf.hypotheses.org/208 (consulté le 8 décembre 2020).

Moreux, Jean-Philippe, « Livre numérique accessible et numérisation de masse à la BnF: retour d'expérience », 9e Forum européen de l'accessibilité numérique, 8 juin 2015, Paris.

Neudecker, Clemens and Apostolos Antonacopoulos, « Making Europe's Historical Newspapers Searchable » 2016 12th IAPR Workshop on Document Analysis Systems (DAS), IEEE, 2016;

Pfanzelter, Eva, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais et Stefan Hechl, « Digital interfaces of historical newspapers: opportunities, restrictions and recommendations » 2020 ⟨hal-02480654v4⟩.

Putnam, Lara, « The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast », *The American Historical Review*, Volume 121, Issue 2, April 2016, Pages 377–402, https://doi.org/10.1093/ahr/121.2.377

Sarah Oberbichler, Eva Pfanzelter

Tracing Discourses in Digital Newspaper Collections

A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage

Abstract: Based on a case study on return migration – defined as the movement of emigrants, refugees, prisoners of war, and others back to their country of origin – this paper deals with a historical-critical approach to digital newspaper collections and interfaces. Using discourse driven historical research questions to examine return migration, it highlights the usability and relevance of digital source criticism (OCR issues, digitization process, newspaper title selection, etc.), digital query criticism (selection of keywords or parts of keywords, consequences of length, polysemy, etc.), and interface criticism (possibilities to search, view, select, collect, visualize, contextualization-efforts, metadata, etc.). We argue that due to the complexity of language coupled with intransparent functionalities of newspaper interfaces, distorted, misunderstood or misinterpreted results can occur. Some inherent biases in the digital collections themselves as well as missing metadata and context can lead to distortions that only at a first glance seem trivial and banal.

Keywords: digitized newspapers, return migration, digital hermeneutics, digital history, newspaper user interfaces

1 Introduction

The digitisation of historical newspapers and their access through online interfaces provided by national or regional libraries offers many opportunities for different user groups: researchers, students, teachers, lay historians, and other interested people can use online interfaces to research, for example, specific topics, their family or community history, events in the press or changes of interpretations over time. Institutions such as the Austrian National Library (ONB), the Bibliothèque nationale de France (BnF) or the National Library of Finland, to name only a

Acknowledgment: This work has been supported by the European Union's Horizon 2020 research and innovation program under grant 770299 (NewsEye). We thank our colleague Stefan Hechl from the NewsEye project for proofreading.

https://doi.org/10.1515/9783110729214-007

few, have engaged and invested heavily in newspaper digitisation. With the Europeana Newspaper's collection,[1] there is also a pan-European newspaper portal available that includes digitized newspapers from 20 countries, dating from 1618 to 1996. Practical considerations such as the availability of thousands of issues of newspapers in libraries as well as non-existent copyright restrictions for material older than 100 years were decisive factors for digitizing newspaper collections. Digitization was also fostered by the appreciation of the potential of newspapers as rich sources for research on history, society and language. For libraries, making newspapers digitally available had and still has several implications. On the one hand, they reach out to a potentially global audience, and on the other they can (usually) preserve the fragile original printed versions in their archives or reserve access to limited in-house users.[2]

Before digitisation, working with newspapers oftentimes was a daunting task due to the amount of material and the time that had to be invested in order to find information that goes beyond reactions to a specific event or a known date. It included spending long hours at library desks and going through large numbers of documents without inventories or any kind of structure.[3] With the advances in digitisation and text recognition, digital approaches have gained in relevance for humanities research and the social sciences, linguistics or history, while historical corpora have become an interesting field of investigation for computer scientists. Interdisciplinary projects which bring together cultural heritage institutions with researchers from the humanities and from computer science are multiplying. For example, the importance of context and meaning is slowly gaining ground in big data research and use. Equally, a critical approach to digital searching and text mining enables researchers to explore and interpret large amounts of newspaper data in novel ways, to visualize quantitatively and to perform individual qualitative analysis.[4] As a result, newly developed methods and tools promise to change the way digital heritage data is accessed, (re)searched, used and analysed.

---

1 https://www.europeana.eu/de/collections/topic/18-newspapers.

2 Natasha Stroeker and René Vogels (2012). *Survey Report on Digitisation in European Cultural Heritage Institutions 2012*. Tech. rep., p. 25. URL: http://enumeratedataplatform. digibis.com/reports/core-survey-i-final-report/detail; Clemens Neudecker and Apostolos Antonacopoulos (2016). "Making Europe's Historical Newspapers Searchable." In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. Santorini, Greece: IEEE, pp. 405–410. DOI: 10.1109/DAS.2016.83; Eleni Galiotou (2014). "Using digital corpora for preserving and processing cultural heritage texts: a case study." In: *Library Review* 63.6/7, pp. 408–421. DOI: 10.1108/LR-11-2013-0142.

3 Richard Abel (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1, pp. 1–10. DOI: 10.2979/filmhistory.25.1–2.1.

4 Adrian Bingham (2010). "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." In: *Twentieth Century British History* 21.2, pp. 225–231.

Advances in digitization also call for an update of the historical-critical method, the hermeneutical process which guides the investigation of the past through three main steps: heuristics, source criticism and interpretation.[5] With regards to the second step, the demand for digital source criticism (and it often goes hand in hand with tool criticism) by now is of long standing.[6] Other parts of this historical method, like the relevance and critical assessment of adequate corpus creation in the heuristic research step, have so far not been considered enough. For the third methodological phase, the interpretation of digital sources and corpora, we only begin to understand the possibilities offered by computer science and the far reaching consequences automation has on both the humanities' disciplines and on society. To use digitized newspaper corpora to dig deeper into the historical-critical method in the digital age is therefore a logical consequence.

In order to understand social and historical developments based on digital newspaper corpora, however, critical readers and interpreters are needed.[7] Many users are not aware of the biases that come along with the processing and datafication of historical newspapers as well as their access via interfaces.[8] Interfaces steer what users can learn from digitised newspapers (e.g., selective curation[9]) and they influence workflows by offering specific, pre-selected functions and tools. The

5 Peter Leyh (1977). *Johann Gustav Droysen: Historik. Bd. 1: Rekonstruktion der ersten vollständigen Fassung der Vorlesungen (1857). Grundriß der Historik in der ersten handschriftlichen (1857/58) und in der letzten gedruckten Fassung (1882)*. Stuttgart-Bad Cannstatt. 532 pp.

6 Andreas Fickers (2020). "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" In: *Zeithistorische Forschungen/Studies in Contemporary History* 17.1, pp. 157–168. DOI: 10.14765/ZZF.DOK-1765; Marijn Koolen, 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; Eva Pfanzelter (2010). "Von der Quellenkritik zum kritischen Umgang mit digitalen Ressourcen." In: *Digitale Arbeitstechniken für Geistes- und Kulturwissenschaften*. Ed. by Martin Gasteiner and Peter Haber. UTB M (Medium Format) 3157. Stuttgart, pp. 39–50; Pascal Föhr (2017). "Historische Quellenkritik im Digitalen Zeitalter." Thesis. University of Basel. DOI: 10.5451/unibas-006805169.

7 James Mussell (2012). *The Nineteenth-Century Press in the Digital Age*. London: Palgrave Macmillan UK.

8 Elizabeth Toon (2019). "The tool and the job: Digital humanities methods and the future of the history of the human sciences." In: *History of the Human Sciences* 32.1, pp. 83–98. DOI: 10.1177/0952695119834152; Andrew Hobbs (2013). "The Deleterious Dominance of *The Times* in Nineteenth-Century Scholarship." In: *Journal of Victorian Culture* 18.4, pp. 472–497. DOI: 10.1080/13555502.2013.854519.

9 Jon Coburn (2020). "Defending the digital: Awareness of digital selectivity in historical research practice." In: *Journal of Librarianship and Information Science*, pp. 1–14. DOI: 10.1177/0961000620918647.

possibilities to perform source criticism in this sense is affected by interfaces that select, capture, sort, and sometimes gather additional (outside) information.[10] Some studies indicate that interfaces do not always concur with the needs of different user groups. Users comment on missing metadata (e.g., digitization strategy, metrics of newspapers and issues, publication frequency, geographical area covered), request contextualization of the source material (e.g., additional information on newspapers, on political orientation, explanation of the search-results, historical context of both newspapers and results, etc.) and ask for references to the limitations of collections (e.g., missing issues, absent time-periods, bad paper quality and subsequently bad Optical Character Recognition [OCR]). Additionally, in these studies a demand for adequate tools to deal with newspapers and search results as well as for material for educational use was clearly identified.[11]

As the role of digital tools when working with newspapers grows, it is important for scholars to understand the biases inherent in digitization processes, digital interfaces and search strategies. While these are sometimes acknowledged as an issue, they are rarely discussed by means of concrete examples. Bearing these findings in mind, the goal of this chapter is to use one case study – the topic of the return of emigrants to their place of origin between 1850 and 1950 – to show both opportunities and limitations that come when working with digital newspapers via interfaces. It is argued that, in addition to still highly relevant issues about the quality of the source material and text recognition, the lack of transparency regarding both digital newspaper collections and interfaces introduces several biases. We also show how newspaper interfaces (e.g., contextualization, explanation and documentation) could effectively counter misunderstandings and misinterpretation. In order to underline our arguments we introduce some advanced text mining functionalities as methods of analysis and interpretation that have the potential to support academic or expert approaches, for example, when creating specific

10 Johan Jarlbrink 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. DOI: 10.1108/JD- 09- 2016- 0106; Ludmilla Jordanova (2014). "Historical Vision in a Digital Age." In: *Cultural and Social History* 11.3, pp. 343–348. DOI: 10 . 2752 /147800414X13983595303237; Arn Keeling and John Sandlos (2011). "Shooting the Archives: Document Digitization for Historical-Geographical Collaboration: Shooting the Archives." In: *History Compass* 9.5, pp. 423–432. DOI: 10.1111/j.1478–0542.2011.00771.x.

11 Maud Ehrmann, Estelle Bunout, and Marten Düring (2019). "Historical Newspaper User Interfaces: A Review." In: URL: https://zenodo.org/record/3404155; Sarah Oberbichler et al. (2019). *Online research of digital newspapers of three national libraries: A survey*. URL: https://www.newseye.eu/blog/news/online-research-of-digital-newspapers-of-threenational-libraries-a-survey-by-sarah-oberbichler-stef/.

corpora for in-depth research or for the classification of search results. This means we try to give answers to the question of how distortions and one-sided interpretations can be reduced already during the heuristic process. Thus the chapter underlines the importance of digital source, query and interface criticism by pointing out the biases that can arise when dealing with large quantities of digital data and it therefore is intended as a contribution to discussions about digital hermeneutics.

In this paper, the Austrian newspaper collections accessible trough the platform AustriaN Newspapers Online (ANNO)[12] was used to investigate the topic of return migration in the second half of the 19th century and first half of the 20th century. ANNO is the user interface for digital newspaper corpora run by the Austrian National Library. The platform offers free access to 21 million Austrian newspaper and magazine pages that were published between 1568 and 1950.

2 Discourses on Return Migration

2.1 Return Migration as a Topic of Historical Research

The definition of remigration or return migration is not consistent within research literature. However, all of the explanations of the term have in common that they define return migration as 'cross-border migration to the country of origin',[13] or as the return of a person to the country of origin after spending a longer period of time abroad. The return of emigrated people has always been part of every migration movement – including recent times. Examples cover the return from overseas in the 19th and 20th century, the return of war veterans or the repatriation of war refugees during and after the First World War, or the return and repatriation of prisoners of war, refugees, exiles, concentration camp survivors, etc., during and after the Second World War.

Despite knowing about these return movements, migration was often viewed as a one-way process especially in academic studies, beginning with the 'uprooting' of people at the point of origin and ending with 'assimilation' into their adopted culture and country.[14] Still, recent research indicates that many

12 http://anno.onb.ac.at/.

13 Edda Currle (2006). "Theorieansätze zur Erklärung von Rückkehr und Remigration." In: *Sozialwissenschaftlicher Fachinformationsdienst soFid* Migration und ethnische Minderheiten 2006/2, pp. 7–23.

14 Peterson Glen (2013). "Return migration." In: *The Encyclopedia of Global Human Migration*, pp. 1–5. DOI: 10.1002/9781444351071.wbeghm453.

people left their home countries with the notion of returning home at a certain point not so far in the future. This is also true for those who left their home voluntarily to resettle in other countries. People who emigrated by their own choice and later decided to come back to their homelands had different reasons for their decisions to return: success, failure, homesickness, rejection in the new country, changes in the economic or socio-political conditions in their native countries, or perhaps family members asked them to return.[15] For involuntary emigrants, some of these reasons may also apply, but their return was often supported, organized, or even forced by the country of origin. Although many voluntary and involuntary emigrants returned home disillusioned, with empty bags and ruined health, many of them also brought financial resources, new ideas, knowledge and skills. At the same time, those who had never left their native land often saw their return critically. The returnees were not always welcomed with open arms. There was mistrust of, for example, changed (political) ideas of return-migrants. Their return could increase problems on an already tight labour market, or be an additional burden on housing shortages. Those who left could also be seen as traitors who had abandoned those who had remained, or, just as well, as bearers of (lost) hope who brought promises of a better future with them.[16] Historical research on return migration, which is still rare, considers such patterns of, and motivations for, return migration,[17] examines historical, political, sociological or economical backgrounds,[18] or deals with autobiographical questions.[19]

So far, the complexity and heterogeneity of sources about return migration (e.g., letters, photos, reports, statistical material on return, snippets of personal

15 Mark Wyman (2001). "Return migration – old story, new story." In: *Immigrants & Minorities*. Historical Studies in Ethnicity, Migration and Diaspora 20.1, pp. 1–18. DOI: 10.1080/02619288.2001.9975006.

16 Annemarie Steidl, Wladimir Fischer-Nebmaier, and James W. Oberly (2017). *From a multiethnic empire to a nation of nations: Austro-Hungarian migrants in the US, 1870–1940*. Transatlantica (Innsbruck, Austria) volume 10. Innsbruck: StudienVerlag.

17 Mark Wyman (2001). "Return migration – old story, new story." In: *Immigrants & Minorities*. Historical Studies in Ethnicity, Migration and Diaspora 20.1, pp. 1–18. DOI: 10.1080/02619288.2001.9975006.

18 Kristina E. Poznan (2017). "Return Migration to Austria-Hungary from the United States in Homeland Economic and Ethnic Politics and International Diplomacy." In: *The Hungarian Historical Review* 6.3, pp. 647–667; Claudia Olivier (2013). "Brain Gain oder Brain Clash? Implizites transnationales Wissen im Kontext von Rückkehr-Migration." In: *Transnationales Wissen und Soziale Arbeit*. Ed. by Désirée Bender et al. Beltz Juventa, pp. 181–205; Marjory Harper (2012). *Emigrant homecomings: The return movement of emigrants*, 1600–2000. Studies in Imperialism MUP. Manchester: Manchester University Press.

19 Katharina Prager and Wolfgang Straub, eds. (2017a). *Bilderbuch-Heimkehr? Remigration im Kontext*. Arco Wissenschaft Band 30. Wuppertal: Arco Verlag.

documentation, interviews, newspaper reporting) has complicated a structured analysis of the topic of return migration. The historical circumstances which brought a homecoming about were so diverse that generalizations are difficult to make and thus inherently faulty. What adds to the difficulty is the lack of empirical evidence. For example, it can still only be roughly estimated how many people who have emigrated voluntarily or involuntarily returned to their home countries.[20] The compilation of a corpus is therefore still an essential factor if the topic of re-migration is to be researched. This is not only true for global or transnational studies, but also for national and regional studies.

2.2 Discourse Analysis of Newspaper Reporting

Newspapers have an impact on the thinking, speaking and acting of a society.[21] For Norman Fairclough, newspapers are the 'predominant social field' for the creation of information, beliefs or arguments, which are necessary 'for establishing and sustaining economic, social and political systems and orders'.[22] Given the significance of news reporting for societies, it should come as no surprise that the discourses of newspaper reporting are essential objects of investigation.[23] There is no uniform definition of discourses or discourse analysis. Michel Foucault, who decisively shaped the concepts of discourse, defined a discourse as a 'group of statements that belong to a single system of formation'.[24] Based on this, humanities researchers have adapted the approach in various ways. For Michael Stubbs, analysing discourses means to 'study larger linguistic units such us conversational exchanges or written texts' whereby he defined discourse as 'language above sentences or above the clause'.[25] Jürgen Link described discourse as 'an institutionally consolidated concept of speech in as much as it determines and

20 Katharina Prager and Wolfgang Straub (2017b). "Die Rückkehr zur Remigration. Zur Einleitung." In: *Bilderbuch-Heimkehr? Remigration im Kontext*. Ed. by Katharina Prager and Wolfgang Straub. Arco Wissenschaft Band 30. Wuppertal: Arco Verlag, pp. 9–19.

21 Niklas Luhmann (1995). *Die Realität der Massenmedien*. Wiesbaden: VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-663-16287-2.

22 Norman Fairclough (2013). *Critical discourse analysis: the critical study of language*. Second Edition. New York: Routledge.

23 John E. Richardson (2007). *Analysing Newspapers: An Approach from Critical Discourse Analysis*. New York: Macmillan International Higher Education.

24 Michel Foucault (1969). *The Archaeology of Knowledge*. London: Routledge.

25 Michael Stubbs (1983). *Discourse analysis: the sociolinguistic analysis of natural language*. Language in society 4. Chicago: University of Chicago Press; Oxford: Blackwell.

consolidates action and thus already exercises power',[26] and Ruth Wodak and Katharina Köhler, the front figures of critical discourse analysis, define discourse as social practices within a historical, political, socio-economic and cultural context.[27] (Critical) discourse analysis 'examines how texts represent and construct reality within a specific ideological system through implicit messages based on what is said and left unsaid'.[28] Discourse analysis therefore can help scholars to understand how social interaction shapes topics and concepts and how those change over time, whereas conceptual history, by comparison, focuses on the etymology and change of the meaning of words. Historical-semantic discourse analysis thus can open up semantic aspects and elements of knowledge that could escape a purely word-oriented history of meaning.[29]

2.3 Studying Discourses on Returnees in Newspapers and How to Get There

Reading newspapers, we can find different types of news coverage on return migration. The most frequent findings are reports on remigration and repatriation often written by journalists or repatriates. But also letters from returnees printed in newspapers can give interesting glimpses into return migration, as they provide a good insight into the thoughts of returnees. The following excerpt from the *Wiener Kurier* is part of a letter from a Jewish refugee talking about the fear of coming home to post-war Austria in 1945:

> Many may have gone away as Jews – but whoever comes back, comes as an Austrian! But this does not diminish the fear of those who return. Over seven years lie between us – and those experiences! Will it be possible to bridge this gap? Will we speak the same language? Will we not reproach each other, unconsciously?[30]

26 Jürgen Link (1983). "Was ist und was bringt Diskurstaktik." In: *kultuRRevolution* 2, pp. 60–66.

27 Ruth Wodak and Katharina Köhler (2010). "Wer oder was ist »fremd«?: Diskurshistorische Analyse fremdenfeindlicher Rhetorik in Österreich." In: *Sozialwissenschaftliche Studiengesellschaft* 2010.1, pp. 33–55.

28 Susana de los Heros (2009). "Linguistic pluralism or prescriptivism? A CDA of language ideologies in Talento, Peru's official textbook for the first-year of high school." In: *Linguistics and Education* 20.2, pp. 172–199. DOI: 10.1016/j.linged.2009.01.007.

29 Dietrich Busse (2008). "Begriffsgeschichte – Diskursgeschichte – Linguistische Epistemologie. Bemerkungen zu den theoretischen und methodischen Grundlagen einer Historischen Semantik in philosophischem Interesse anlässlich einer Philosophie der ‚Person'." In: *Diskurse der Personalität*. Ed. by Nikolaj Plotnikov and Alexander Haardt. Wilhelm Fink Verlag, pp. 115–142. DOI: 10.30965/9783846744321_009.

30 "Wir sind quitt" (Oct. 13, 1945). In: *Wiener Kurier*, p. 4.

Next to letters, appeals and calls for help represent another type of coverage on return migration. Especially in the news coverage on returning prisoners of war, appeals for help were printed regularly in newspapers. Finally, a substantial part of the reporting on returnees consists of small advertisements. Here, returnees tried to sell land (often abroad), to find a partner for life, and asked for donations or a job. To describe oneself as a 'returnee' in these advertisements (Figure 1) seemed to be significant:


RÜCKWANDERER

verkauft 36 ha bestes Siedlungsland an Bahnstation,

BRASILIEN

Zahlung hier. Reichsbankgenehmigung erforderlich. Auskünfte unter „Parana 400" an das Ankündigungs-Büro dieses Blattes.

HEIRAT

Heimkehrer, Fünfzig., mit Beruf, wünscht Anschluß an nette gebildete Frau über 40 Jahre. Unter „Verlassen" an die Anz.-Annahme, Platzl 2.

8

Fig. 1: Small advertisements from returnees in Austrian newspapers from the 'Neue Freie Presse', 29 May 1938, p. 31 (left, 'Rückwanderer') and the 'Salzburger Blatt', 5 March 1946, p. 8 (right, 'Heirat'), retrieved from the ANNO platform (accessed in Dec 2020 at https://anno.onb.ac.at/).

Newspapers are therefore an interesting and available source to research return migration. They enable longitudinal studies, create a starting point for discourses on return migration, point to further primary sources and allow for an exploration of the discourses on how returning people were talked about in their country of origin: were they welcomed or perceived as a burden and threat? Can differences between groups of returning migrants be identified and were there changes over the time? How did returnees depict themselves?

Especially in reports, letters and appeals, we find several characteristic discourses on return migration, which include (implicit) arguments to support, promote, regulate or prohibit the return of people to their country of origin. Three main discourses can be identified: discourses on encouragement, authorisation or prohibition of return; discourses on motives of return like disappointment or deception; and discourses on benefits and threats of returnees for the country of origin.

Many examples of the discourse on encouragement of return can be found in 1907, when the governments of Austria-Hungary (and the Hungarian government in particular) became an active promoter of return migration to further an

economic and nationalist agenda.[31] The newspaper *Das Vaterland*, for example, printed the following paragraph in 1907:

> The Hungarian Farmers' Union, in association with the Hungarian Agricultural Association, has undertaken an action to promote the return of those who emigrated from Hungary to America and to facilitate their colonisation at home.[32]

While the return of emigrants from America was encouraged by the government, more critical voices called for control and prohibition of certain groups of returnees. In the following example from the *Salzburger Chronik für Stadt und Land* in 1907, it is suggested that only those people who were able to support themselves should be allowed to return:

> Through the leaders of the cooperatives, the trusted representatives in America are given the strictest instruction to encourage only those to return who have so much money in the domestic savings banks that they can acquire some land in their home country.[33]

Return migrants could help or endanger the government of the country of origin, both economically and politically. A return migrant might be someone who failed in the country he had emigrated to or someone who brought back skills, political ideas, or capital to invest in the homeland economy. Discourses about benefit or threat appear in many different historical and political contexts and usually come with specific (political or economical) intentions. Acceptance of returnees and a positive mood in the majority society was usually asked for because a political or economic benefit of this return was expected. On the other hand, there was rejection on the basis of assumed 'liberal' thoughts by those returning to Austria-Hungary, as the following excerpt from the *Salzburger Chronic für Stadt und Land* (November 1907) shows:

> In political terms, however, return migration will lead to a noticeable strengthening of the anti-Magyar nationalist movement, since the returnees understandably bring the American idea of freedom with them and carry it into the smallest towns and into the rural population.[34]

---

31 Kristina E. Poznan (2017). "Return Migration to Austria-Hungary from the United States in Homeland Economic and Ethnic Politics and International Diplomacy." In: *The Hungarian Historical Review* 6.3, pp. 647–667.

32 "Kolonisierung von Auswanderern" (n.d.). In: *Das Vaterland* (), p. 11.

33 "Ungarn-Die Rückwanderung aus Amerika" (Nov. 29, 1907). In: *Salzburger Chronik für Stadt und Land*, p. 2.

34 "Die Rückwanderung aus Amerika" (Nov. 29, 1907). In: *Salzburger Chronik für Stadt und Land*, p. 2.

So, while digital methods can help us to find texts where return migration is talked about in one way or the other, they cannot (yet) label the discourses themselves: it is still the human eye that has to identify these discourses as being about encouragement or discouragement, motives, and advantages or disadvantages. This is 'partly because the way in which computers work is not automatically compatible with the way human brains work' as Hinke Piersma and Kees Ribbens phrased it.[35] As edmond noted, tools that focus on text do not reflect the practice of historians, which is based on broader information gathering and finding meaning in historical sources.[36] Nevertheless, digital methods can support discourse-analytical approaches,[37] especially when working with historical newspapers, because corpora containing discourses and arguments must be found before they can be identified, interpreted, and contextualized. Computational methods can strongly assist with the creation of adequate newspaper corpora for specific research questions and topics or when trying to find discourses. Without technical help, it is difficult to find, organise, and structure thousands of newspaper clippings. However, 'simple' procedures such as keyword searches, i.e. searching for specific articles using single or combined search terms, can be challenging because of the complex nature of language, changes in word meaning over time and polysemy. The following sections give an insight into the tricky challenges faced when creating a corpus on return migration using advanced keyword searches and frequency graphs.

3 The Keyword Challenge

3.1 Return Migration – Hard to Trace by Keywords

As emphasized in Section 2, discourse analyses of historical newspaper items depends on corpora that are at the basis of investigation. Interfaces can play a crucial role for the creation of corpora with search and collection building functions and easy-to-use download features. The process of collection building usually starts with complex search strategies designed to find relevant newspaper issues,

35 Hinke Piersma and Kees Ribbens (2013). "Digital Historical Research: Context, Concepts and the Need for Reflection." In: *BMGN – Low Countries Historical Review* 128.4, pp. 78–102.

36 Jennifer Edmond (2018). "How Scholars Read Now: When the Signal Is the Noise." In: *Digital Humanities Quarterly* 12.1.

37 Andrew Ravenscroft and Colin Allen (2019). "Finding and Interpreting Arguments: An Important Challenge for Humanities Computing and Scholarly Practice." In: *Digital Humanities Quarterly* 13.4.

pages and articles given a research topic. Furthermore, due to the design and technology of interfaces, for historians working with digital newspapers, keywords are still key.[38] At the same time, keyword search goes hand in hand with some serious weaknesses for its use in historical research, which is caused by the complexity of language: alternative spellings, abbreviations, polysemy, changing word usage, idioms, misspellings or omissions of parts of words as well as sentences.[39] But this is only one part of the challenge, since for some topics we do not have simple keywords, as we will explain in the next section.

3.1.1 Concepts Do Not Equate to Single Words

Even when taking spelling variations and word changes into account, many search requests are difficult to define conceptually and hard if not impossible to trace by a simple keyword search alone. For the topic on return migration, there are a few specific German keywords that can be used in order to find relevant articles. These keywords mostly are variations of the word 'returnees', such as '*Heimkehrer*' [homecomers], '*Rückkehrer*' [returnees], '*Rückwanderer*' [return migrants], '*Heimgekehrten*' [repatriates].

If the (relative) frequency of these keywords is compared to the frequency of other migration-related keywords in ten Austrian newspapers, the result shown in Figure 2 (page 137) is obtained. The orange color shows the frequency of the variations of 'returnees', compared with variations of 'emigrants' (blue), 'refugees' (purple), and 'immigrants' (green).

The graph provides a rough overview of the evolution of the terms used to trace migration in Austrian newspapers between 1820 and 1920. These terms do not reflect the entire spectrum of reporting on migration, but they show the trickiness and complexity when using terms in connection with migration movements. Even if the terms used for return migration (orange line) seem to appear less frequently in the corpus than the words indicating emigration (blue line) or refugee movements (purple line), it is nevertheless not correct to deduce from the graph that return migration was less frequently reported on than the other issues. For example, many of the emigrants or refugees mentioned here are people returning to their communities of origin: They would be emigrants/refugees who return

38 Eva Pfanzelter et al. (2021). "Digital interfaces of historical newspapers: opportunities, restrictions and recommendations." In: *Journal of Data Mining and Digital Humanities*. DOI: https://doi.org/10.46298/jdmdh.6121

39 Sheila Bair and Sharon Carlson (2008). "Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship." In: *Journal of Library Metadata* 8.3, pp. 249–262.

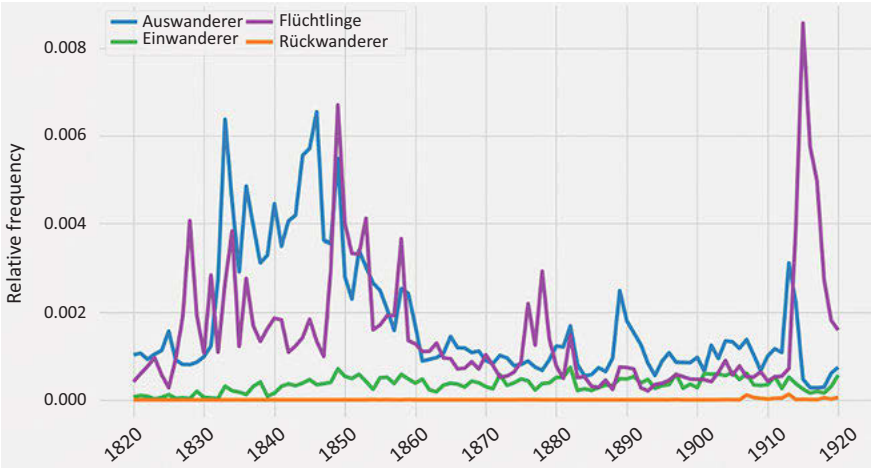


Fig. 2: Frequency graph of the terms *auswander**/*ausgewander** (*emigration/emigrated*), *einwander**/*eingewander**(*immigration/immigrated*), *flüchtling**/*geflüchtete** (*refugee/people who fled*), *rückwanderer**/*rückkehrer**/*heimkehrer** (*return migrant/returnee/homecomer*) in ten Austrian newspapers from 1820 to 1920.

home. The terms needed to identify the right group of people (the returnees) and/or the right discourses (those about return migration) would therefore be a combination of the words emigrants/immigrant/refugees and return/returning/coming home ('Auswanderer/Einwanderer/Flüchtlinge' and 'heimkommen/heimkehren/zurück kommen' and similar). This example – and there are many similar ones – shows that using single keywords or even bi-grams, if they were available in an interface search, is not enough when looking for re-migration issues.

3.1.2 Word Flexions

Combining keywords, as it is done when specific distances between words are taken into consideration (the so-called distance search) can help to find linguistic constructions such as 'refugees ... return home'. The ANNO interface enables users to find keyword combinations with defined word distances, a feature that is extremely helpful. However, word combinations do not always solve the problem. First, identifying the necessary terms to find the sought-for data is time-consuming and error-prone. In ANNO, this kind of search is even made more difficult because wildcards cannot be used in combination with the distance search and users are therefore obliged to try out many different word flexions and spelling

variants. For example, it is not possible to search for '*heimat *rückkehr**'~20 ('home *return**'~20), which would include different tenses of the verb 'to return' as well as the noun 'the return'.

3.1.3 Ambiguous Keywords

In addition, because of language ambiguity, many of the combined keywords, but also more ambiguous single keywords lead to results that have nothing to do with the original search request. For researchers eliminating these irrelevant results manually is not possible if a search request includes several thousands of hits. One of the ambiguous words for return migration is '*Rückwanderung*' ('return migration'), which can occur in very different contexts, as Figure 3 clearly shows:

‚Rückwanderung' relevant	‚Rückwanderung' not relevant
‚Allein in New=York drängen sich nun Tausende von Opfern der Krise, deren Mittellosigkeit ihnen die Rückwanderung bisher verwehrt hat, in die Vorzimmer der beteiligten Konsulate, um hier ihre Absicht zu erklären, „zu den Fahnen zu eilen", bei kostenfreier Ueberfahrt und zehn Heller Löhnung den Tag – soweit es sich um österreichisch= ungarische Reservisten handelt.'	'Die mächtigen Dämme, die in den letzten Jahren vom Lido aus auf Kilometerlänge ins Meer hinausgebaut wurden, zwingen die Fluth zu rascherem, mächtigerem Eindringen in das Wirrsal der Lagunencanäte, und auf der Rückwanderung zum Urgewässer nimmt sie den aufgewühlten Schlamm mit sich und trägt ihn hinaus ins Freie.'
(In New York alone, thousands of victims of the crisis, whose pennilessness has so far prevented them from returning home, are now crowding into the antechambers of the consulates involved to declare their intention to 'rush to the flag', with free passage and then Heller a day – as far as Austrian-Hungarian reservists are concerned.)	(The massive dams that have been built in recent years from the Lido over a distance of several kilometers into the sea force the tide to penetrate the chaos of the lagoon channels more quickly and more powerfully, and on its way back to the primeval waters it takes the stirred up mud with it and carries it outside.)

Fig. 3: Extracts from newspaper articles with the keyword '*Rückwanderung*', on the left an example of a relevant article and on the right a non-relevant article in connection with return migration.

Both the use of unique keywords and the combination of keywords make it difficult to create appropriate newspaper corpora. Unique keywords lead the researcher to only a restricted selection of articles, while the combination of keywords influences the search on the basis of the researcher's prior knowledge and oftentimes trigger irrelevant results.

Newspaper interfaces make language phenomena more apparent without actually creating them in the first place. Interfaces steer query and search processes and can therefore be seen as tools for source criticism but also as tools

that need to be reviewed and used critically. Interfaces can profit from some basic methods and tools, as the following sections show.

3.2 The Disappearance of the Second World War

One important first step when creating sub-corpora is to look at the frequency of words within a corpus. This can uncover the basic use and distribution of terms over time. They can give information on when a word or term was used and they can help to find entry points to corpora as well as to restrict the analysis to a specific period of time, if needed. With the help of diachronic frequency analyses, relative and absolute frequencies of pages, articles or words can be counted and thus information about the relevancy of texts over time or the distribution of certain linguistic patterns can be provided.[40] The ANNO platform offers no possibilities to create diachronic frequency graphs. The interface provides the number of pages where a keyword can be found and allows searching for specific years, however, the distribution of keywords over time is currently not traceable.

Still, even the amount of pages where a keyword appears can be helpful to gain a first impression of the search results. Using the absolute frequency of the keyword '*Heimat*' in combination with '*rückkehr/zurückkehren/heimkehren/zurückgekehrt*' ('home' combined with 'returning/return/return home/returned'), the graph in Figure 4 can be created:

Taking historical events into account, a closer look at the graph immediately reveals a discrepancy: The number of hits, not unexpectedly, rises sharply during and immediately after the First World War, roughly between 1915 and 1920. Very surprisingly, however, such a peak does not occur during the Second World War, between 1939 and 1945/46. This inconsistency and unlikely result also becomes apparent when using other search terms that should show a rising number of hits in the period after 1933. The question of why the results of the frequency analyses are distorted leads to a consideration and subsequent examination of the following questions:

- Were important search terms overlooked?
- Was there a change in the use of terms? Were new terms used after the beginning of the Second World War?

40 Gregor Wiedemann, Matthias Lemke, and Andreas Niekler (2013). "Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949–2011." In: *ZPTh – Zeitschrift für Politische Theorie* 4.1.

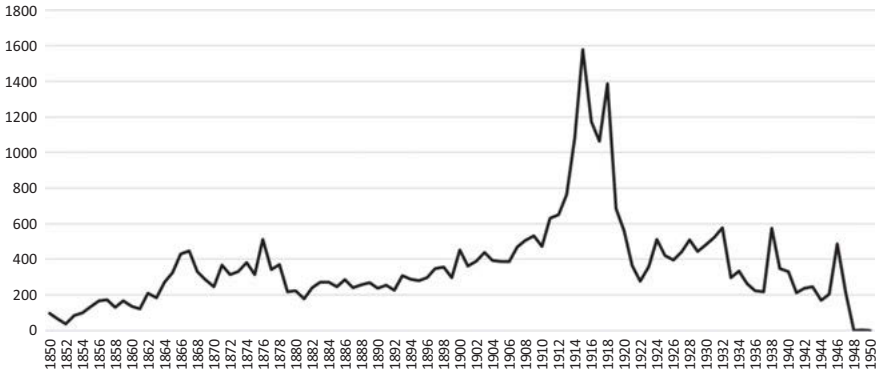


Fig. 4: Absolute frequency graph of '*Heimat*' ('home') in combination with '*rückkehr/zurückkehren/heimkehren/zurückgekehrt*' ('returning/return/return home/returned') (1850–1950) within a distance of 20 words. The distance search function in ANNO has been used to combine the terms.

- Is there a large discrepancy between the number of newspapers published during the First World War and those published during the Second World War?
- Is there a discrepancy between the number of found pages from the First World War and the Second World War that are available in digitised form?
- Is there a change in the quality of the digitised newspapers or the OCR for certain newspapers?

Careful browsing and comparison of two selected newspapers for this search and their OCR'd versions ultimately revealed that the OCR quality between 1938 and 1945 is extremely variable across newspapers and greatly differs from the one observed for time period 1915–1920. Figure 5 (page 141) shows an article on return migration in the newspaper *Neue Freie Presse* from 1938, with a page facsimile on the left and its barely decipherable OCR on the right. Apart from the headline, almost no combination of letters, numbers or signs is readable and many of the combined characters shown are not German words but seemingly random collections of letters and symbols.

3.3 Bringing Back the Second World War

Having worked with many newspaper interfaces, this outcome should not have been surprising. There have been ongoing discussions about the impact of the

Rückkehr in die befreite Heimat

Donnerstag mittag trafen in Reichenberg die ersten in die Heimat zurückgekehrten Flüchtlinge ein. Es handelt sich um 16 Männer aus dem Bereich der SA-Gruppe Mitte (Magdeburg), die während ihres Aufenthaltes unter dem Schutze des Altreiches von der Gruppe SA-gemäß ausgebildet, vollkommen eingekleidet und jetzt in ihre Heimat zurückgeleitet wurden. Zum Empfang versammelte sich die gesamte Bevölkerung auf dem Marktplatz, wo gegen Mittag sturmweise die Männer mit gepacktem Tornister, voran die Fahne, auf dem Adolf-Hitler-Platz aufmarschierten. Besonders stürmisch begrüßt von den Heimkehrern wie von den Reichenbergern wurde bei seinem Erscheinen der Reichskommissar. Nach der Begrüßung durch einen Vertreter der Stadt und nach einer Ansprache von SA-Obergruppenführer Kob, erklärte Konrad Henlein: Alle, die hier standen, seien bereit gewesen, ihr Leben für die Heimat in die Schanze zu schlagen. Daß sie nicht eingesetzt zu werden brauchten, dankten sie dem Führer. Konrad Henlein gedachte dann der Toten, die um Sudetendeutschlands Freiheit gefallen waren. Sein Dank galt den Betreuern der Flüchtlinge im Reich. Die Kameradschaft und begeisterte Aufnahme im Schutze des starken Reiches sollten die Männer durch ihre stete Einsatzbereitschaft danken. Nach dem Sieg-Heil auf den Führer zogen die Formationen unter Vorantritt des Musikkorps und des Spielmannszuges der Gruppe Mitte an Konrad Henlein und den Gruppenführern vorüber.

Rückkehr in die befreite Heimat
®onnerStag mittag trafen in IReidhcubcrg bie
e r fte n in bie .§cimat zurückgekehrten f5tkrhi^iitgo ein.
h®>^bclt sich 16 SJlänner auS bem Sereidt ber @91»
©nippe HUitte (ünagbeburg), bie währenb il)reS 9lufenthalte§
unter bem ©d)uhe bcS 9lltreid)e§ oon ber ©ruppc '^»gemiä^
ausgebilbet, oollkommen eingekleibet unb je|t in ihre .^eimat
Znrückgeleitet würben. ©mpfang oerfammelte sich bie
gesamte ©eoöllterung auf bem iDlarktplah, Wo gegen Sllittag
ftnrmweife bie SRäuner mit gepadrtem Tornister, Uoran bic
f^ahnc, auf bem 9tbolf»,§itler»ißlah aufmarfdperten. ®e=
fonberS stürmisch begrüfd non ben .^eimkehrern wie oon ben
ikeidjenbergern würbe bei feinem ©rfcheinen ber fRcichiS»
kommiffar. 9ladji ber Segrü^iing burd) einen 9?ertreter bet
©tabt unb itod) einer 9tnfpta<he toon ©2l--Dbergruppenfül)rer
,t 0 b, erklärte tonrab § e n l e i n : 9ll[c, bie hier fton^n,
feien bereit gewesen, ihr Sehen für bie ^cimat in bie ©chanze
ZU fd)lagen. ®a^ sie nidjt cingefeht zu werben brausten,
bankten sie bem ü l) r e r. tonrab §entein gebachte bann
ber Soten, bie um ©iibeteiibeutfchlanbS gefallen
waren. Sein ®ank galt ben tBetreuem bet f^tüchtlinge tim
Ilieich. ®ie tamerabfdjaft unb begeisterte 9lufnahmc im
©cljiuho beS starken i){eid)e§ sollten bie üUännor burch ihoo
stete ®infahbereitfd)iaft banken. 9foch bem ©ieg=.g>cil ans ben
f^ührer Z'ogou bic f^onuotionen unter 35orantritt be§ SOluftk»
korpg unb be§ ©pielmaiinzugeä ber (%uppe SOlittc on tonrab
^nlcin unb ben ©ruppenführern borübet.

Fig. 5: *Neue Freie Presse*, 15 October 1938 (fac-simile on the left, OCR text on the right).

OCR quality on research findings.[41] Still, the extent of the influence of poor OCR quality in combination with missing metadata (which will be discussed below) is striking. The following paragraphs will therefore introduce some methods, tools and functions that can help increase the reliability of the findings.

3.3.1 Data and OCR Quality

The OCR quality of the newspaper collections in ANNO not only changes significantly over time, but also varies considerably between different newspapers. Confronted with the findings above, researchers should become extremely cautious and skeptical about the results they get using digital newspaper interfaces and the functionalities implemented there. These irregularities lead to the frustrating conclusion that the full-text search results in ANNO – and in many other newspaper interfaces – are only part of what is in the data and that they are distorted by faulty OCR and missing data. With the OCR quality as poor as it is in the example above, not even good frequency analysis tools could help to

41 Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman (2015). "Impact Analysis of OCR Quality on Research Tasks in Digital Archives." In: *Research and Advanced Technology for Digital Libraries*. Ed. by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Vol. 9316. Cham: Springer International Publishing, pp. 252–263.

create an interpretable picture. However, having a graph of relative frequencies next to the search results could point users to possible OCR flaws or similar problems in the dataset or the digitised material. This is important, because for researchers, faulty OCR does not necessarily have to be accepted as an unchangeable fact. Tools like Transkribus[42] can help to re-OCR documents and reach a text-quality that minimizes the impact of OCR on further analysis. Figure 6 on the right shows the improved, almost flawless OCR from the example shown in Figure 5 (*Neue Freie Presse* from 1938). The improved OCR was created with Transkribus using the downloaded, original scan from the ANNO platform. Interface providers can support this step by providing suitable download options (e.g., the possibility to download a whole newspaper issue).

OCR from the ANNO platform (ÖNB) (*Neue Freie Presse*, 15 October 1938)	**Improved OCR with Transkribus** (*Neue Freie Presse*, 15 October 1938)
Rückkehr in die befreite Heimat ®onnerStag mittag trafen in IReidhcubcrg bie e r fte n in bie .§cimat zurückgekehrten f5tkrhi^iitgo ein. h®>^bclt sich 16 SJlänner auS bem Sereidt ber @91» ©nippe HUitte (ünagbeburg), bie währenb il)reS 9lufentholte§ unter bem ©d)uhe bcS 9lltreid)e§ oon ber ©ruppc '^»gemiä^ ausgebilbet, oollkommen eingekleibet unb je|t in ihre .^eimat Znrückgeleitet würben. ©mpfang oerfammelte sich bie gesamte ©eoöllterung auf bem iDlarktplah, Wo gegen Sllittag ftnrmweife bie SRäuner mit gepadrtem Tornister, Uoran bic f^ahnc, auf bem 9tbolf»,§itler»ißlah aufmarfdperten.	Rückkehr in die befreite Heimat Donnerstag mittag trafen in Reichenberg die ersten in die Heimat zurückgekehrten Flüchtlinge ein. Es handelt sich um 16 Männer aus dem Bereich der SA= Gruppe Mitte (Magdeburg), die während ihres Aufenthaltes unter dem Schutze des Altreiches von der Gruppe SA=gemäß ausgebildet, vollkommen eingekleidet und jetzt in ihre Heimat zurückgeleitet wurden. Zum Empfang versammelte sich die gesamte Bevölkerung auf dem Marktplatz, wo gegen Mittag sturmweise die Männer mit gepacktem Tornister, voran die Fahne, auf dem Adolf=Hitler=Platz aufmarschierten.

Fig. 6: *Neue Freie Presse*, 15 October 1938 (OCR retrieved from ANNO in December 2020 on the left, OCR by Transcribus on the right).

3.3.2 Relative and Absolute Frequency Graphs

One of the key features used in humanities research, as already mentioned above, is frequency analysis. Frequencies are used to gain insights into language use and to create balanced corpora for adequate analysis.[43] Absolute frequencies, as seen in Figure 4, can produce an unrealistic picture, as they only count the number of times a term appears in a corpus without putting the results in relation, e.g., with the total number of words, articles or pages for a certain time

42 https://readcoop.eu/transkribus/.

43 Jani Marjanen (2019). *What's the frequency, Kenneth?* URL: https://www.newseye.eu/ blog/news/what-s-the-frequency-kenneth/.

period or a certain newspaper. If many more newspapers were available during the First World War than during the Second World War, which is in fact the case in the ANNO platform, then absolute frequencies will always show a more frequent use of the term during the First World War, even if actually, in relation, many more articles were written about people returning to their country of origin during the Second World War.

Relative frequencies, on the other hand, can solve this problem by taking these relations into account. Using ANNO, relative frequencies cannot be generated automatically, as required metadata such as word counts or number of digitised pages/newspapers per year are not available. Metadata (e.g., pages per issue per year), however, is available at ONB Labs,[44] the Austrian National Library's digital laboratory. This makes a manually created frequency calculation (pages per year) possible, as can be seen in Figure 7:

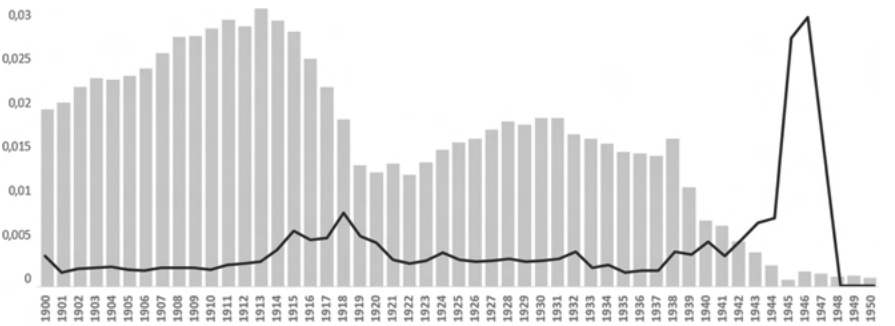


Fig. 7: Relative frequency graph (pages per year) of 'Home' combined with 'returning/return/return home/returned' (1900–1950).

As Figure 7 clearly shows, the number of hits obtained when researching remigration during the Second World War is significantly higher than the one during World War One. This outcome, which meets the expectations of historians, shows how important it is to critically question the information given on newspaper interfaces. It also underlines that such methods can be used to test and prove existing hypotheses. Metadata about digital collections provided on interfaces consulted by users is therefore very important. Missing metadata, as shown in Figure 4, as well as the lack of basic tools to analyse and evaluate findings affect

44 *ONB Labs – Static Plotly Graphs* (n.d.). URL: https://labs.onb.ac.at/en/tool/sprachenin-anno/.

general faceting and browsing and therefore hinder the preparation of conclusive frequency analyses, which in turn could point users to some of the above-mentioned flaws.

3.3.3 Contextualization

Another important feature concerns contextualization of historical sources. Contextual information is essential for researchers in order to describe, explain, compare, or evaluate them. Interfaces could offer different kinds of context to users: Context on newspapers themselves, which sometimes is available already, can situate the document in time and space by including information on publication periods, publication places, editors, publishers, imprint details, political orientation as well as possible predecessor or successor newspapers. Frequency graphs (as can be seen in the example of return migration) can sometimes give context to the query and they can point to faulty data, keywords or dataset issues. Additional information on the newspapers and on datafication processes can help to understand and interpret the results better. Contextualization of the newspaper content, on the other hand, could be given through the recognition and linking of events or named entities (person, names, organizations or places) to external knowledge bases such as Wikidata. Automatic linking of newspaper content, however, can again create biases. Named entity recognition, for example, can create demographic biases (gender, ethnicity or race),[45] or biases caused by noisy text,[46] similarities in names or places, etc.

3.3.4 Transparency on Data and Tools

The call for perfectly clean data and appropriate tools should be preceded by the call for more transparency. This includes the traceability of:

- The search functions: are stemmed or exact words/word combinations being looked for and what does that imply? How does the search function work and what does it show?

---

45 Shubhanshu Mishra, Sijun He, and Luca Belli (2020). "Assessing Demographic Bias in Named Entity Recognition." In: *arXiv:2008.03415 [cs]*. URL: http://arxiv.org/abs/2008. 03415.

46 Vinh-Nam Huynh, Ahmed Hamdi, and Antoine Doucet (2020). "When to Use OCR Post-correction for Named Entity Recognition?" In: *Digital Libraries at Times of Massive Societal Transition*. Ed. by Emi Ishita, Natalie Lee San Pang, and Lihong Zhou. Vol. 12504. Cham: Springer International Publishing, pp. 33–42. DOI: 10.1007/978-3-030-64452-9_3.

- The search results: what do the results show? Do hits represent the total number of matches or the number of pages/articles in which the matches occur and is there more than one match within one article?
- The data visualisations: are visualized results based on the total number of matches or pages/articles in which the matches occur? Are relative or absolute numbers visualized and how can they be interpreted?
- The text mining tools: what is the tool doing and what not? How are results being created and how can they be interpreted?

In order to give users a better command and a better understanding of the search process and search results, it would be an important feature to always give users the possibility to compare results with the original material. Again, transparency about datasets, search functionalities and data visualization could make gaps and empty spaces in the available collection visible and would enable the necessary source criticism.

3.3.5 Help Files, Information Pages and Best Practice Examples

In addition to the issues mentioned before, instructions, explanations and best practice examples are needed to guide users through the collections, the interfaces and digital tools. The better search functions, faceted searches, and use of interfaces are explained with different approaches, the better users will know how to interpret the results and how to use the interfaces to their advantage. We also found that creating educational material addressing the above-mentioned pitfalls and challenges for the libraries' use on the interfaces or in conjunction with their respective educational settings can help raise awareness to critical issues about data, interfaces and tools and attract new user groups.

4 The potential of Text Mining

While the above-mentioned methods seem to be fairly easy to be implemented in existing newspaper interfaces, there are several strong indicators that support calls for more elaborate user interfaces, tools and methods. While methods to achieve satisfactory search results including simple keyword searches and frequency graphs can help to get a general idea about a topic, more sophisticated techniques such as bi-gram searches or keywords in context can increase search outcomes significantly. Even simple word clouds chosen for specific time slices can point to changing

discourses within a researched topic. In addition, keyword suggestions based on word embeddings as provided by the *impresso* and *NewsEye* interfaces can help to find keywords that are semantically related to the users' query[47] and therefore open up new perspectives and possibilities.

As Figure 8 shows, the keyword suggestion function included in the beta version of the NewsEye platform, a project aiming at improving digital approaches to historical newspaper corpora, found many matches that improve the tedious manual search in ANNO and hinted at other issues that were not discovered previously, such as the combination of 'homecomers' ('*Heimkehrer*') with 'internees' ('*internierte*'), 'reservists' ('*reservisten*') and 'convalescent' ('*rekonvaleszent*'). All of these could indeed indicate discourses worth pursuing when working on issues of return migration.

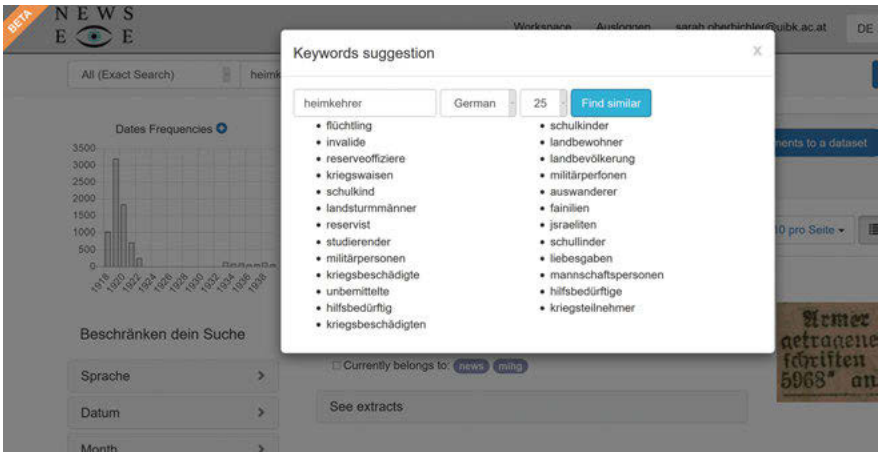


Fig. 8: Suggested keywords (set to 25) for the term 'homecomer' ('*heimkehrer*') produced by the NewsEye keyword suggestion tool as implemented in the beta version of the NewsEye platform (accessed in Dec 2020, https://platform.newseye.eu).

Also, text mining methods such as topic modeling, named entity recognition/linking or word embeddings can provide significant advantages when creating specific corpora and have the potential to become imperative for critical digital hermeneutics. For example, Federico Nanni, Simone Ponzetto and Laura Dietz used named entities and word embeddings to automatically build entity-centric

47 Saar Kuzi, Anna Shtok, and Oren Kurland (2016). *Query Expansion Using Word Embeddings*. Indianapolis Indiana USA.

event collections.[48] Their goal was to address the needs of humanities' and social sciences' scholars who work quantitatively on specific topics and events. With a similar aim in mind, researchers at the University of Helsinki are experimenting with approaches to adapting topic models and word embeddings on multilingual corpora.[49]

Text mining methods are also especially promising for the structuring, organisation and classification of text. For example, Pierre-Carl Langlais used a supervised topic modeling approach for genre classification, which allows the refining of searches to a specific genre.[50] Sarah Oberbichler, on the other hand, used methods that support article-based corpus building by taking the context (full content of the article) of keywords into account.[51] As discussed in Section 3, the corpus building for the topic of return migration is complicated by the fact that keywords are often either too narrow and leave part of the discourse in newspapers unseen, or too broad and thus lead to too many irrelevant results. Text classification can help group and subsequently filter articles that are actually relevant for the research question. This means that even ambiguous words can be used for the search without having to combine them with other terms, making the search less influenced by the researcher's prior knowledge. Using an approach combining Latent Topic Modeling (LDA) and the Jensen-Shannon Distance (JSD) method, articles about return migration and those on other, non-migration-related topics (as in Fig. 9) can be successfully and automatically grouped into and separated from each other.[52]

We do not argue that all text mining methods can or should be integrated in user interfaces, since it is difficult to find a 'one size fits all' approach for different user needs and specific research questions. However, interfaces that give access to huge historical datasets could greatly benefit from giving information on possibilities of further analysis options for the found material and by providing

48 Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz (2017). "Building Entity-Centric Event Collections." In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Toronto, ON, Canada, pp. 1–10.

49 Elaine Zosa and Mark Granroth-Wilding (2019). "Multilingual Dynamic Topic Model." In: *International Conference on Recent Advances in Natural Language Processing ((RANLP 2019))*. URL: https://zenodo.org/record/4153232#.YyIQoN9CTy0.

50 Pierre-Carl Langlais (2020). *Numapresse/TidySupervise*. URL: https://github.com/Numapresse/TidySupervise.

51 Sarah Oberbichler (2020). *Using LDA and Jensen-Shannon Distance (JSD) to group similar newspaper articles*. URL: https://zenodo.org/record/3887193#.YyIvEt9CTy0.

52 Sarah Oberbichler and Eva Pfanzelter (2021). "Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods". In: Journal of Digital History, 1.1. DOI: 10.1515/JDH-2021-1003.

	Relevant_Text	3	Non_Relevant_Text	0
0	Ein Knirps unterm Tropenhelm\n Auf dem Anlegep...	3	„Ober wos denkst denn, so long ausbleibn!"\nFü...	0
1	Wenn trotzdem süber die Behandlung der Heimkeh...	3	Innsbruck, 31. Mai. Auf einer Pressefahrt der ...	0
2	Was die materialle Versorgung der Heimkehrer\n...	3	Nachdem Gauleiter Fritz Wächtler die Meldung d...	0
3	Die Austria wird diesmal mit einer starken Elf...	3	Der Stand der Sonne erinnerte Henriette daran,...	0
4	Walz war im Frühjahr 1920 von der öster\nreich...	3	Die äußeren Ursachen sind bekannt: Die Rückwan...	0
5	Das ist alles recht schön, aber jedenfalls\nei...	3	Millionen Soldaten, die im großen Weltkrieg fü...	0
6	Ein Fall auf dem Dampfer\nTriest, 6. Juni. Ges...	3	Repertoire des Deutschen Volkstheaters. Samsta...	0
7	Tomaschek war nach Kriegsausbruch eingerückt, ...	3	Seit Harald Paulsen das Theater am Nollendorfp...	0
8	Daß die Versorgung der Heimkehrenden in der\ne...	3	säule. Zugleich aber fiel das Barometer, und s...	0
9	Das gesamte deutsche Volk, vorab wir im Gau Ti...	3	.Vom 15. bis 22. Mai veranstaltete die Kreisle...	0
10	Rasch und doch sicher löste sie das Blatt\naus...	0	(Turistenunglück auf der Hohen Wand.)\nAus Wie...	3

Fig. 9: Articles automatically grouped into relevant (*Relevant Text*) and non-relevant articles (*Non Relevant Text*) for the topic on return migration. For testing, articles were manually assigned to numbers (number 3 for relevant and number 0 for non-relevant articles) which proves whether the articles were selected correctly.

users with various export possibilities (.csv, .json, .xlsx) in order to allow them to apply advanced digital methods. Users, on the other hand, would greatly profit from such functionalities. It would sensitize them even more to faults and concerns, limitations and possibilities, and again, train them in source as well as tool criticism.

5 Conclusion

New digital tools and methods to support research, analysis and corpus building are currently being created, tested and evaluated for example by research projects such as '*impresso* – Media Monitoring of the Past'[53] or NewsEye[54]. Their aim is to enrich and connect historical newspapers of regional or national libraries with more advanced tools. Keyword suggestions, suggestions of related articles or topics, tools to build datasets and download options are powerful ways to support research with historical newspapers. The value of digitized newspapers for humanities increases significantly with these new developments. The project's interfaces take a first step towards collection building supported by machine

53 https://impresso-project.ch/.
54 https://www.newseye.eu/.

learning methods. They show how enriched data could ultimately improve access to historical newspaper collections. These interfaces, like some others, also aim at offering a variety of download options in order to facilitate further processing, e.g., the qualitative analysis of the created collections or further processing of the data outside the interface.

As we have tried to show, such efforts take digital hermeneutics seriously. A focus was put on pitfalls and challenges that come with corpus creation and thus with digital heuristics. To critically use, test and reflect the tools and outcomes remains an integral part of the research process. While individual users can often hardly understand how tools and methods work (and do not have to), they have to be given the information and tools to understand and critically question the outcomes. Therefore, digging into digital newspaper corpora continues to be an exciting field for interdisciplinary research. The contributions of humanities' critical approaches to the fields of computer science or digital cultural heritage curation are important, but their implications at this point cannot be grasped fully yet.

Bibliography

Abel, Richard (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1, pp. 1–10. DOI: 10.2979/filmhistory.25.1-2.1.

Bair, Sheila and Sharon Carlson (2008). "Where Keywords Fail: Using Metadata to Facil- itate Digital Humanities Scholarship." In: *Journal of Library Metadata* 8.3, pp. 249–262.

Bingham, Adrian (2010). "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." In: *Twentieth Century British History* 21.2, pp. 225–231.

Busse, Dietrich (2008). "Begriffsgeschichte – Diskursgeschichte – Linguistische Epistemologie. Bemerkungen zu den theoretischen und methodischen Grundlagen einer Historischen Semantik in philosophischem Interesse anlässlich einer Philosophie der ‚Person'." In: *Diskurse der Personalität*. Ed. by Nikolaj Plotnikov and Alexander Haardt. Wilhelm Fink Verlag, pp. 115–142. DOI: 10.30965/9783846744321_009.

Coburn, Jon (2020). "Defending the digital: Awareness of digital selectivity in historical research practice." In: *Journal of Librarianship and Information Science*, pp. 1–14. DOI: 10.1177/0961000620918647.

Currle, Edda (2006). "Theorieansätze zur Erklärung von Rückkehr und Remigration." In: *Sozialwissenschaftlicher Fachinformationsdienst soFid* Migration und ethnische Minderheiten 2006/2, pp. 7–23.

"Die Rückwanderung aus Amerika" (Nov. 29, 1907). In: *Salzburger Chronik für Stadt und Land*, p. 2.

Edmond, Jennifer (2018). "How Scholars Read Now: When the Signal Is the Noise." In: *Digital Humanities Quarterly* 12.1.

Ehrmann, Maud, Estelle Bunout, and Marten Düring (2019). "Historical Newspaper User Interfaces: A Review." In: URL: https://zenodo.org/record/3404155.

Fairclough, Norman (2013). *Critical discourse analysis: the critical study of language*. Sec- ond Edition. New York: Routledge.

Fickers, Andreas (2020). "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" In: *Zeithistorische Forschungen/Studies in Contemporary History* 17.1, pp. 157–168. DOI: 10.14765/ZZF.DOK-1765.

Föhr, Pascal (2017). "Historische Quellenkritik im Digitalen Zeitalter." Thesis. University of Basel. DOI: 10.5451/unibas-006805169.

Foucault, Michel (1969). *The Archaeology of Knowledge*. London: Routledge.

Galiotou, Eleni (2014). "Using digital corpora for preserving and processing cultural heritage texts: a case study." In: *Library Review* 63.6/7, pp. 408–421. DOI: 10.1108/LR- 11-2013-0142

Glen, Peterson (2013). "Return migration." In: *The Encyclopedia of Global Human Migration*, pp. 1–5. DOI: 10.1002/9781444351071.wbeghm453.

Harper, Marjory (2012). *Emigrant homecomings: The return movement of emigrants, 1600–2000*. Studies in Imperialism MUP. Manchester: Manchester University Press;

Heros, Susana de los (2009). "Linguistic pluralism or prescriptivism? A CDA of language ideologies in Talento, Peru's official textbook for the first-year of high school." In: *Linguistics and Education* 20.2, pp. 172–199. DOI: 10.1016/j.linged.2009.01.007.

Hobbs, Andrew (2013). "The Deleterious Dominance of *The Times* in Nineteenth-Century Scholarship." In: *Journal of Victorian Culture* 18.4, pp. 472–497. DOI: 10.1080/13555502.2013.854519.

Huynh, Vinh-Nam, Ahmed Hamdi, and Antoine Doucet (2020). "When to Use OCR Post-correction for Named Entity Recognition?" In: *Digital Libraries at Times of Massive Societal Transition*. Ed. by Emi Ishita, Natalie Lee San Pang, and Lihong Zhou. Vol. 12504. Cham: Springer International Publishing, pp. 33–42. DOI: 10.1007/978-3-030-64452-9_3.

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. DOI: 10.1108/JD-09-2016-0106.

Jordanova, Ludmilla (2014). "Historical Vision in a Digital Age." In: *Cultural and Social History* 11.3, pp. 343–348. DOI: 10.2752/147800414X13983595303237.

Keeling, Arn and John Sandlos (2011). "Shooting the Archives: Document Digitization for Historical-Geographical Collaboration: Shooting the Archives." In: *History Compass* 9.5, pp. 423–432. DOI: 10.1111/j.1478-0542.2011.00771.x.

"Kolonisierung von Auswanderern" (Nov. 6, 1907). In: *Das Vaterland*, p. 11.

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.

Kuzi, Saar, Anna Shtok, and Oren Kurland (2016). *Query Expansion Using Word Embeddings*. Indianapolis Indiana USA.

Langlais, Pierre-Carl (2020). *Numapresse/TidySupervise*. URL: https://github.com/Numapresse/TidySupervise.

Leyh, Peter (1977). *Johann Gustav Droysen: Historik. Bd. 1: Rekonstruktion der ersten vollständigen Fassung der Vorlesungen (1857). Grundriß der Historik in der ersten handschriftlichen (1857/58) und in der letzten gedruckten Fassung (1882)*. Stuttgart-Bad Cannstatt. 532 pp.

Link, Jürgen (1983). "Was ist und was bringt Diskurstaktik." In: *kultuRRevolution* 2, pp. 60–66.

Luhmann, Niklas (1995). *Die Realität der Massenmedien*. Wiesbaden: VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-663-16287-2.

Marjanen, Jani (2019). *What's the frequency, Kenneth?* URL: https://www.newseye.eu/blog/news/what-s-the-frequency-kenneth/.

Mishra, Shubhanshu, Sijun He, and Luca Belli (2020). "Assessing Demographic Bias in Named Entity Recognition." In: *arXiv:2008.03415*. URL: https://arxiv.org/abs/2008.03415.

Mussell, James (2012). *The Nineteenth-Century Press in the Digital Age*. London: Palgrave Macmillan UK.

Nanni, Federico, Simone Paolo Ponzetto, and Laura Dietz (2017). "Building Entity-Centric Event Collections." In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Toronto, ON, Canada, pp. 1–10.

Neudecker, Clemens and Apostolos Antonacopoulos (2016). "Making Europe's Historical Newspapers Searchable." In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. Santorini, Greece: IEEE, pp. 405–410. DOI: 10.1109/DAS.2016.83.

Oberbichler, Sarah (2020). *Using LDA and Jensen-Shannon Distance (JSD) to group similar newspaper articles*. URL: https://zenodo.org/record/3887193#.YyIvEt9CTy0.

Oberbichler, Sarah, Stefan Hechl, Barbara Klaus, Minna Kaukonen, Tuula Pääkkönen, and Marion Ansel (2019). *Online research of digital newspapers of three national libraries: A survey*. URL: https://www.newseye.eu/blog/news/online-research-of-digital-newspapers-of-three-national-libraries-a-survey-by-sarah-oberbichler-stef/.

Oberbichler, Sarah and Eva Pfanzelter (2021). "Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods". In: *Journal of Digital History*, 1.1. DOI: 10.1515/JDH-2021-1003

Olivier, Claudia (2013). "Brain Gain oder Brain Clash? Implizites transnationales Wissen im Kontext von Rückkehr-Migration." In: *Transnationales Wissen und Soziale Arbeit*. Ed. by Désirée Bender, Annemarie Duscha, Lena Huber, and Kathrin Klein-Zimmer. Beltz Juventa, pp. 181–205.

ONB Labs – Static Plotly Graphs (n.d.). URL: https://labs.onb.ac.at/en/tool/sprachen-in-anno/.

Pfanzelter, Eva (2010). "Von der Quellenkritik zum kritischen Umgang mit digitalen Ressourcen." In: *Digitale Arbeitstechniken für Geistes- und Kulturwissenschaften*. Ed. by Martin Gasteiner and Peter Haber. UTB M (Medium Format) 3157. Stuttgart, pp. 39–50.

Pfanzelter, Eva, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl (2021). "Digital interfaces of historical newspapers: opportunities, restrictions and recommendations." In: *Journal of Data Mining and Digital Humanities*. DOI: https://doi.org/10.46298/jdmdh.6121.

Piersma, Hinke and Kees Ribbens (2013). "Digital Historical Research: Context, Concepts and the Need for Reflection." In: *BMGN – Low Countries Historical Review* 128.4, pp. 78–102.

Poznan, Kristina E. (2017). "Return Migration to Austria-Hungary from the United States in Homeland Economic and Ethnic Politics and International Diplomacy." In: *The Hungarian Historical Review* 6.3, pp. 647–667.

Prager, Katharina and Wolfgang Straub, eds. (2017a). *Bilderbuch-Heimkehr? Remigration im Kontext*. Arco Wissenschaft Band 30. Wuppertal: Arco Verlag.

Prager, Katharina and Wolfgang Straub (2017b). "Die Rückkehr zur Remigration. Zur Einleitung." In: *Bilderbuch-Heimkehr? Remigration im Kontext*. Ed. by Katharina Prager and Wolfgang Straub. Arco Wissenschaft Band 30. Wuppertal: Arco Verlag, pp. 9–19.

Ravenscroft, Andrew and Colin Allen (2019). "Finding and Interpreting Arguments: An Important Challenge for Humanities Computing and Scholarly Practice." In: *Digital Humanities Quarterly* 13.4.

Richardson, John E. (2007). *Analysing Newspapers: An Approach from Critical Discourse Analysis*. New York: Macmillan International Higher Education.

Steidl, Annemarie, Wladimir Fischer-Nebmaier, and James W. Oberly (2017). *From a multiethnic empire to a nation of nations: Austro-Hungarian migrants in the US, 1870–1940*. Transatlantica (Innsbruck, Austria) volume 10. Innsbruck: StudienVerlag.

Stroeker, Natasha and René Vogels (2012). *Survey Report on Digitisation in European Cultural Heritage Institutions 2012*. Tech. rep., p. 25. URL: http://enumeratedataplatform.digibis.com/reports/core-survey-i-final-report/detail.

Stubbs, Michael (1983). *Discourse analysis: the sociolinguistic analysis of natural language*. Language in society 4. Chicago: University of Chicago Press; Oxford: Blackwell.

Toon, Elizabeth (2019). "The tool and the job: Digital humanities methods and the future of the history of the human sciences." In: *History of the Human Sciences* 32.1, pp. 83–98. DOI: 10.1177/0952695119834152.

Traub, Myriam C., Jacco van Ossenbruggen, and Lynda Hardman (2015). "Impact Analysis of OCR Quality on Research Tasks in Digital Archives." In: *Research and Advanced Technology for Digital Libraries*. Ed. by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Vol. 9316. Cham: Springer International Publishing, pp. 252–263.

"Ungarn-Die Rückwanderung aus Amerika" (Nov. 29, 1907). In: *Salzburger Chronik für Stadt und Land*, p. 2.

Wiedemann, Gregor, Matthias Lemke, and Andreas Niekler (2013). "Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949–2011." In: *ZPTh – Zeitschrift für Politische Theorie* 4.1.

"Wir sind quitt" (Oct. 13, 1945). In: *Wiener Kurier*, p. 4.

Wodak, Ruth and Katharina Köhler (2010). "Wer oder was ist »fremd«?: Diskurshistorische Analyse fremdenfeindlicher Rhetorik in Österreich." In: *Sozialwissenschaftliche Studiengesellschaft* 2010.1, pp. 33–55.

Wyman, Mark (2001). "Return migration – old story, new story." In: *Immigrants & Minorities*. Historical Studies in Ethnicity, Migration and Diaspora 20.1, pp. 1–18. DOI: 10.1080/02619288.2001.9975006.

Zosa, Elaine and Mark Granroth-Wilding (2019). "Multilingual Dynamic Topic Model." In: *International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. URL: https://zenodo.org/record/4153232#.YyIQoN9CTy0.

Christoph Hanzig, Martin Munke, Michael Thoß

Digitising and Presenting a Nazi Newspaper

The example *Der Freiheitskampf*

Abstract: The Saxon Nazi newspaper *Der Freiheitskampf* has been digitised and presented online in order to fill in the gap left by the significant lack of sources about the Saxon National Socialist German Workers' Party (NSDAP). The editors are systematically indexing the content of its articles as well as compiling and categorising the Saxon-related articles in a relational database. This paper presents a joint project of the Hannah Arendt Institute for Research on Totalitarianism (HAIT) and the Saxon State and University Library Dresden (SLUB). It will address several aspects of using digitised newspapers as a historical source, namely how to offer a more critical approach to digital print sources by means of content-based indexing. It will also address Optical Character Recognition (OCR) and how to contextualise a newspaper collection by using standardised data and engage with legal and moral questions relating to the digital presentation of an ideologically highly tainted source collection. Tackling these issues from the perspective of a research institution and of a provider of a research infrastructure allows the discussion of different points of views on the mass-digitalisation of historical newspapers.

Keywords: saxony history, digitised newspapers, local national socialism history

1 Introduction

Historians researching the National Socialistic past in Saxony are repeatedly confronted with a massive lack of sources, since a large part of the relevant files were lost during the course of the war and through the destruction of documents ordered by the Nazis during the last weeks of the war. This particularly concerns documents of the Saxon NSDAP, its divisions and actors, as well as the state's regional authorities.

https://doi.org/10.1515/9783110729214-008

Daily newspapers were the main source of information at the time, reflecting the political, economic and cultural life of a nation.[1] They can be seen as "mirrors of past societies".[2] The staff of the Hannah Arendt Institute for Research on Totalitarianism (HAIT), therefore, decided in 2009 to investigate the extent to which information from the Nazi daily newspaper *Der Freiheitskampf* could be recovered for the establishment and consolidation period of National Socialism in Saxony. The database project we present here offers an important contribution to the scientific infrastructure, as it makes an additional source on National Socialism in Saxony available by means of in-depth indexing of the newspaper's content. In this way, the aforementioned gaps in tradition are to be at least partially closed. The basis for the project is the digitisation of the still existing issues of the newspaper by the Saxon State and University Library Dresden (SLUB). The first version of the database, which made it possible to search for the newspaper volumes 1930 to 1934, was published in January 2017. In the meantime, the volumes up to and including 1937 have been indexed in the database.

In the following sections, we will shed light on the project in its various dimensions. First, the *Freiheitskampf* will be presented in more detail as a historical source. Then we will explain the various functionalities of the database and the underlying processes of data collection. This will then be followed by an outlook on future projects, especially with regard to technical and legal issues. Finally, the possibilities of the project for scientific research and political education will be summarised and discussed in the broader context of the digitisation of daily newspapers.

2 The *Freiheitskampf* as a Source

The daily newspaper *Freiheitskampf* was the official party organ of the NSDAP in the Gau Sachsen, which was published from August 1, 1930 to May 8, 1945 (Fig. 1), making it the longest-running press product of Nazi Germany up to the last days of World War Two.[3] It emerged from the *Sächsischer Beobachter*, which had been

1 See: Oron J. Hale, *Presse in der Zwangsjacke 1933–1945* (Düsseldorf 1965), p.11.

2 Maud Ehrmann, Estelle Bunout, Maren Düring, "Historical Newspaper User Interfaces: A Review." In: *Libraries: Dialogue for Change. World Library and Information Congress*, 85th IFLA General Conference and Assembly, August 24–30 2019, Athens, p. 1, URL: http://library.ifla.org/2578/1/085-ehrmann-en.pdf.

3 On the history of the *Freiheitskampf* see Markus Fischer, "Neue Perspektiven auf die sächsische NS-Presse. Eine Aufarbeitung des NSDAP-Organs *Der Freiheitskampf*." In: Neues Archiv für sächsische Geschichte, 84 (2013), pp. 281–291.


DRESDNER ZEITUNG

Der Freiheitskampf

AMTLICHE ZEITUNG DER NSDAP. AMTLICHES BLATT DER BEHÖRDEN

Nr. 105 15. Jahrgang | Dienstag, 8. Mai 1945 | Preis 10 Rpf.

Vergeßt niemals

Das Reich sieht nun auf uns!

Keine falschen Folgerungen aus dem Abschluß der Waffenruhe an der Nordwest- und Südfront ziehen: Der Kampf gegen die Sowjets geht weiter

Auch Dr. Goebbels gefallen

Waffenruhe im Nordwesten und Süden

Die Berichte des Oberkommandos der Wehrmacht vom Sonnabend und Sonntag

Ruhe und Ordnung im Protektorat

Fig. 1: Front page of the issue of 8 May 1945 (License: free access, rights reserved).

founded in 1929 by the brothers Gregor and Otto Strasser, and published by their own Berlin "Kampf-Verlag". After Otto Strasser left the party during the course of the infighting within the NSDAP in mid-1930, the Saxon NSDAP regional leadership around Gauleiter Martin Mutschmann and his deputy Karl Fritsch felt compelled to

publish their own daily newspaper. With the founding of the *Freiheitskampf* in July 1930, Saxony joined in a development that spanned the entire Reich, in the course of which the party papers that had previously only appeared weekly were then converted into daily newspapers in the Gaue between 1930 and 1933.[4]

Until the beginning of 1933, the *Freiheitskampf* was primarily aimed at the Saxon members of the NSDAP. In the style of a typical "Kampfblatt", it polemically attacked the political and economic conditions of the Weimar Republic, which was in a deep recession. The attacks focused mainly on the liberal system of the Weimar Republic with its democratic representatives, "Jews", and the political opponents on the left.[5] During the numerous election rallies in the years 1930 to 1933, which often turned into violent clashes, the *Freiheitskampf* specifically defamed the left-wing opponents as "red murderers" and in return presented violent excesses of NSDAP-members as legitimate self-defence.[6]

With the introduction of the National Socialist rule, the characteristics of daily newspapers increasingly changed, with the *Freiheitskampf* having the function of being a newsletter for all state authorities. As early as October 1932, Otto Dietrich, the Reich press officer of the NSDAP, remarked that the concept of the "drum press" would not be able to reach a larger readership in the long term. He therefore pleaded for an improvement in journalistic quality.[7] This is also perceivable in the *Freiheitskampf:* since 1933 at the latest, a moderation in rhetoric can be observed, in addition to a broader regional and thematic orientation. The newspaper no longer addressed only NSDAP-members and sympathisers. Instead, the editors tried to reach as much of Saxon society as possible, including those who were previously distanced from the Nazi movement, including the bourgeoisie, the intelligentsia and the working class. The aim was to convince them of the correctness of National Socialist policy and ideology. For this reason, the *Freiheitskampf* now offered a wide range of different sections, which were

4 See: Hale, *Presse in der Zwangsjacke*, p.58; on the development of the Gaupresse see *ibid*., pp. 57–67.

5 See: Markus Fischer, "Neue Perspektiven", p. 279.

6 On the reception of violence against political opponents in the *Freiheitskampf* see Josephine Templer, "Rezeption von politischer Gewalt und ihrer Funktion in der sächsischen Presse zwischen 1930 und 1933. 'Der Freiheitskampf' und die 'Arbeiterstimme' im Vergleich." In: Gerhard Lindemann, Mike Schmeitzner (eds.), *... da schlagen wir zu. Politische Gewalt in Sachsen 1930–1935*, Göttingen (2020), pp. 21–52, DOI: 10.14220/9783737009348.21; as well as Christoph Hanzig and Michael Thoß, "'Rotmord' vor Gericht. Politisch motivierte Tötungsdelikte in Sachsen im Spiegel der NS-Tageszeitung 'Der Freiheitskampf' von 1931 bis 1936." In: ibid., pp. 193–230, DOI: 10.14220/9783737009348.193 .

7 See: Stefan Krings, *Hitlers Pressechef. Otto Dietrich (1897–1952). Eine Biografie*, Göttingen (2010), p. 129.

directed at different groups (youth, women, civil servants), or reported on specific topics (world affairs, regional affairs, culture, economy, sport). The fact that the newspaper enjoyed increasing popularity in Saxony can be seen from the steadily growing number of advertisements and rising circulation figures. While the circulation was 5,000 copies in August 1930, it had risen to around 58,000 by January 1933.[8] From this point, the *Freiheitskampf* was able to gain additional readers with its new function as an official announcement newspaper as well as with new potentials as a result of the repressive press policy of the National Socialists, such as the takeover of modern, formerly leftist printing plants and a preferential treatment of the Nazi press in the distribution system. Thus, the circulation was increased to 100,000 copies by April 1933. By 1936 the circulation had settled at 200,000.[9] Due to its broader thematic scope and local differentiation, the *Freiheitskampf* not only has "a multidimensional significance for both the history of National Socialism and for everyday and social history regarding the Nazi's enforcement of the dictatorship in society, but also for the organizational history of the NSDAP in Saxony and a look into the party's anchoring in the region".[10]

After extensive research in Saxon regional archives, the holdings of the newspaper are now available almost in their entirety, and have been digitised by the SLUB, based on the back-up copies on microfilm made in the 1990s.[11] It comprises a total of more than 66,000 sheets, which were brought together from the holdings of the SLUB and the Dresden City Archive. Several gaps in the total stock could be closed after research in the regional editions in the stock of the Freiberg City Archive. Smaller additions were made from the state libraries in Berlin and

8 For the exact circulation figures of the *Freiheitskampf* see: Die Geschichte der sächsischen NS-Presse von der Gründung 1930 bis Juni 1936, o. D, o. O, p. 10 [Bundesarchiv Berlin NS36/1013, unpag.]

9 *Ibid.*

10 Thomas Widera, Martin Munke, and Matti Stöhr, "'Der Freiheitskampf' – Digitalisierung und Tiefenerschließung einer NS-Zeitung." In: *Relying on News Media. Long Term Preservation and Perspectives for Our Collective Memory*. IFLA News Media Section Satellite conference 2017, August 16th-18th, 2017. Dresden 2017, URN: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-164012. All online sources have been last viewed on 24.02.2020.

11 Digitisation was carried out as part of a pilot project for newspaper digitisation in Germany, funded by the German Research Foundation (DFG). See Thomas Bürger, "Zeitungsdigitalisierung als Herausforderung und Chance für Wissenschaft und Kultur." In: *Zeitschrift für Bibliothekswesen und Bibliographie*, 63 (2016) 3, S. 123–132, DOI: 10.3196/186429501663332. On the Dresden subproject see Thomas Bürger, Sebastian Meyer, "Schlagzeilen im Binärcode. Fortschritte und Herausforderungen bei der Digitalisierung historischer Zeitungen." In: *BIS. Das Magazin der Bibliotheken in Sachsen* 9 (2016) 3, 139–141. URN: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-77780.

Munich; missing issues are also being digitised from private collections.[12] Due partly to the insufficient quality of the digital copies of the microfilm, and the software for the recognition of fractured writing, which was not yet sufficiently developed at the time of digitisation, full text recognition by means of Optical Character Recognition (OCR) was not initially carried out. The project staff therefore decided to index the source in depth. Compared to the quantitative method of full text recognition, this qualitative approach offers many advantages for the later users of the database. For example, the resulting "finding book" shows a way through the amount of data and, with the help of the content classification, enables a targeted search for specific facts, persons, places and propaganda content, which does not necessarily have to be associated with exact keywords. This is a very labour-intensive, but more sustainable method – especially since previous research has shown that "the accuracy of Optical Character Recognition (OCR) technologies considerably impacts the way digital documents are indexed, consulted and exploited".[13]

After the exploratory phase in 2009, the project could initially only be continued to a limited extent with the institute's own funds. With funding from the Saxon Ministry of Science and Art (2017–2019) within the framework of the joint project "Virtuelle Archive für die geisteswissenschaftliche Forschung" (Virtual Archives for Research in the Humanities), the indexing of the daily newspaper was accelerated and the database was linked to other projects, such as those of the joint project, by means of various indicators.

12 A common problem with the digitisation of newspapers, the "problem of incompleteness", could therefore be successfully solved with this project. However, this was only possible as the newspaper has a comparatively short publication period of only 15 years. See: Huub Wijfjes, "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." In: *Tijdschrift voor Mediageschiedenis*, 20 (2017) 1, pp. 4–24, here: 15–17, quotation: 15, DOI: https://doi.org/10.18146/2213-7653.2017.277.

13 See: Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, Jean-Philippe Moreux, "Impact of OCR Errors on the Use of Digital Libraries. Towards a Better Access to Information." In: *2017 ACM/IEEE Joint Conference on Digital Libraries* (JCDL), June 19–23 2017, Toronto, p. 1, DOI: 10.1109/JCDL.2017.7991582; Wijfjes, 'Digital Humanities and Media History', pp. 18–20.

3 The Database

3.1 Data Collection Process

Since national and international events were mostly reported in other, larger, and already digitised newspapers of the National Socialist (NS) press, or have already been extensively discussed in various publications, only articles with a direct reference to Saxony are included in the database. The editors evaluate the newspaper articles while browsing, record not only the metadata (article ID, date, localization in the newspaper) but also the rough content of the selected articles, the places mentioned and all persons involved. Relevant actors are captured by means of an additional database of persons, which is interlinked with the article database and was created as a by-product of data generation. There, information taken from the *Freiheitskampf* about the respective persons are collected. It currently comprises over 1,800 relevant Saxon officials and personalities. About 400 of them could also be assigned a GND number, an identifier of the Integrated Authority File (Gemeinsame Normdatei, GND) of the German National Library (DNB), as named entities. For a more precise registration of the location, the corresponding district authority is assigned on the one hand, and on the other hand the location of the event is linked to the likewise interlinked location database, which includes all Saxon locations with more than 2,000 inhabitants. This, in turn, is linked to the Historical Gazetteer of Saxony (Historisches Ortsverzeichnis von Sachsen, HOV) of the Institute for Saxon History and Cultural Anthropology, in which all Saxon towns are listed and provided with additional information such as geographical location, population figures, administrative affiliation, etc.[14]

The most important step in the article recording is the categorization of the content. For this purpose, the project staff developed a three-stage category thesaurus in numerous consultations with, among others, NS researchers at the HAIT. The six main categories (Ideology, War, NSDAP, Organizations, Political Institutions, and Regional History) are divided into further subcategories which are assigned to a total of 92 main topics in the last stage (e.g. anti-Semitism, public event NSDAP regional, Hitler Youth, Justice, Culture, Women in NS, or Economy). The editors use these 92 specifics to classify the content of the article thematically. Articles can also be assigned multiple categories.

The comments field summarizes the main content of the article and also contextualizes some of it. As a rule, the National Socialist style of the document is not adopted and if so, it is only marked as a quotation to emphasize

14 See: https://hov.isgv.de.

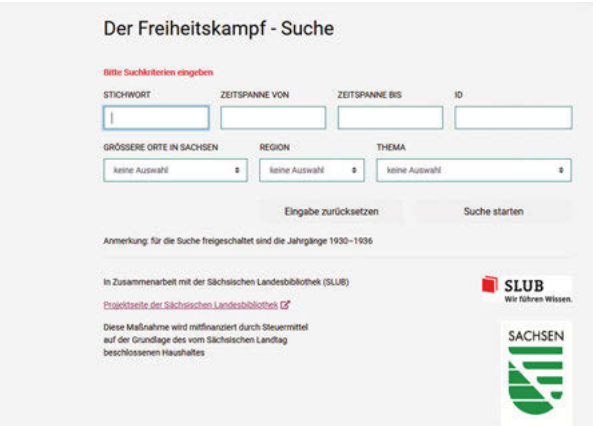


Fig. 2: Search fields for structured searches in the database, https://hait.tu-dresden.de/ext/forschung/der-freiheitskampf.asp, last accessed 20.06.2022.

propagandistic stylistic devices. In addition, all actors or places mentioned but not recorded in the database of persons or places are noted in the comments field. Thus, these items can be found again later via a search in the remarks field. Names of persecuted persons are not completely noted for ethical reasons, e.g. persons who have been publicly accused of "racial defilement". On the basis of these work steps, about 60 percent of the inventory (1930–1937 and 1943–1945) has been processed and over 26,500 newspaper articles have been recorded.

3.2 Research Options

On the homepage of the *Freiheitskampf* project,[15] the contributions of the years 1930 to 1937 that are recorded in the database are currently freely searchable. Various entry points have been created for this purpose. When accessing the database, you can search for a specific date or period in a search mask (Fig. 2). In addition, it is possible to output all articles of a certain issue. It is also possible to search for geographical places or regions by restricting the search result to a specific district council, or by entering the municipalities with more than 2000 inhabitants recorded in the local database. The search can also be narrowed down additional by selecting one of the 92 categories. Moreover, the search results can be further refined by searching for a keyword in the comment fields. Last but not

15 See: https://hait.tu-dresden.de/ext/forschung/der-freiheitskampf.asp.

least, the keyword function also allows a search for persons and places that are not recorded in the respective databases but in the comments field.

The access point "Persons" enables the user to obtain all articles linked to a person in the personal database of the *Freiheitskampf* by entering the family name or the GND number in chronological order. In addition, users receive some basic information (if known) about the person. For example, if the individual searched for has been provided with a GND identifier, links to the corresponding result pages of the DNB and the BEACON service are displayed.[16] By using this function of the German Wikipedia community, it is possible to take into account the findings of other institutions about the searched person. Here, all web pages of projects that participate in the BEACON service and use the same GND number are linked, such as the corresponding article in Wikipedia. The library of HAIT has been authorized by the DNB to create new records for persons in the GND, as it is also done by employees of the SLUB. This provides the possibility to "enrich" the data pool of the GND with further relevant actors from Saxony in the future and to further advance the connection of the research results. This is all the more important as the search for named entities such as persons and places is one of the most frequently used search strategies in digitised source collections and should therefore be prepared with particular care.[17] The relevant persons are thus also included and made searchable in the database of Saxon persons in the regional portal Saxorum, which is operated by SLUB.[18]

The entry point "Topics" enables a search along the three-level category thesaurus (Fig. 3), whereby the results can also be filtered according to keyword, period and location as described above. In most cases, the articles have been assigned to several categories, which allows the narrowing down of the hits by choosing another of the now 91 additional topics. The project team is currently discussing the possibility of linking thematic fields by using existing or newly created GND records for corporations and keywords, and therewith establish a connection with our own free keyword assignment and an existing thesaurus.

When clicking on the desired article, a results page is displayed, which lists the linked topics, places and persons, in addition to the summary of content (Fig. 4). There are also two links to the corresponding digitized newspaper. Here the article is not displayed individually, but via the link in the context of

16 See: https://meta.wikimedia.org/wiki/Dynamic_links_to_external_resources; https://github.com/gbv/beaconspec.

17 See e.g.: Chiron et al., "Impact of OCR Errors", p. 3; Wijfjes, "Digital Humanities and Media History", p. 17.

18 See: https://www.saxorum.de/index.php?id=10178.

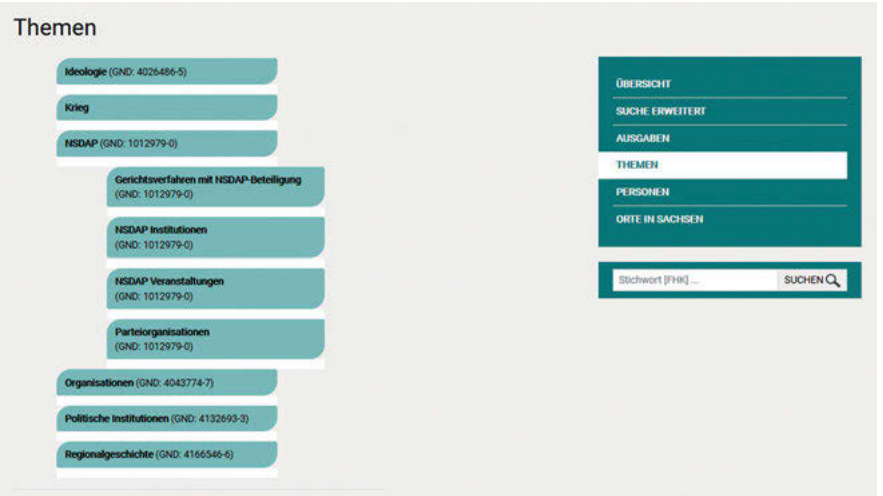


Fig. 3: Topics as a browsing entry point to the dataset, https://hait.tu-dresden.de/ext/forschung/der-freiheitskampf-themen.asp, last accessed 20.06.2022.

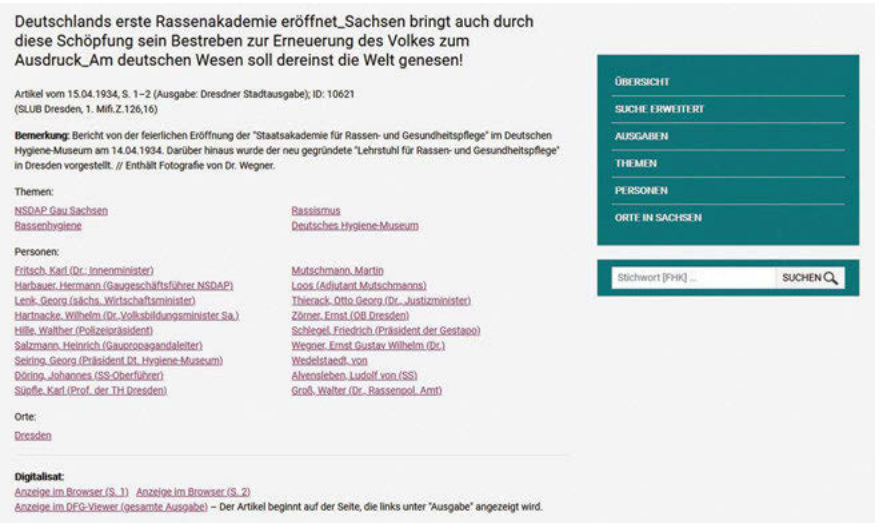


Fig. 4: Details page for a newspaper article, https://hait.tu-dresden.de/ext/forschung/der-freiheitskampf-artikel.asp?id=10621, last accessed 20.06.2022.

the single newspaper page, or via the DFG Viewer,[19] in the entire issue, which the user can browse through freely.

3.3 Restriction of Use

The database is freely accessible to all users via the above-mentioned internet address, and currently enables research for the years 1930–1937. An analysis of the website's access figures has shown that an average of 175 search queries are made daily. However, the usability of the database has been limited to the extent that there are ethical and legal concerns about making the actual content of the digitised versions of a Nazi newspaper accessible to everyone. Until November 2021, these could only be viewed at special workstations in the HAIT library and in the SLUB so that scientists had to travel to Dresden to be able to read the newspaper articles. What were the reasons for these restrictions?[20]

With regard to inciting propaganda and the inhuman National Socialist language, there are concerns that right-wing extremists could abuse these decontextualised contents for propaganda and radicalisation. These moral and ethical questions of publishing an ideologically highly tainted source collection are currently under intensive discussion, due to the resurgence of right-wing parties throughout Europe. A recent conference held in Vienna on precisely the question of how to deal with digitised content from the National Socialist era has shown that there are conflicting opinions between actors from research institutions, libraries, archives, museums, media and the civil society. Only a broad discussion can lead to widely accepted approaches. The impotance of, and interest in dealing with these issues is also evident from the large number of technical reports that have already appeared about the event.[21]

19 See: https://dfg-viewer.de/.

20 Since then the *Freiheitskampf* has been published here: https://digital.slub-dresden.de/werkansicht/dlf/486933/1. See: Martin Munke, "NS-Geschichte digital erforschen: Tageszeitung 'Der Freiheitskampf' jetzt online verfügbar." In: SLUBlog, 04.11.2021, https://blog.slub-dresden.de/beitrag/2021/11/4/ns-geschichte-digital-erforschen-tageszeitung-der-freiheitskampf-jetzt-online-verfuegbar.

21 See: Christoph Mentschl, "Die Verantwortung von Bibliotheken, Archiven und Museen sowie Forschungseinrichtungen und Medien im Umgang mit der NS-Zeit im Netz." In: *o|bib* 7 (2020) 2, DOI: 10.5282/o-bib/5599; Markus Stumpf, "Ausgewählte Erkenntnisse aus der Enquete 'Nationalsozialismus digital' (Wien, 27.–29. November 2019)." In: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare,* 73 (2020) 1, pp. 147–151. DOI: 10.31263/voebm.v73i1.3479; Jutta Fuchshuber, "Nationalsozialismus digital. Die Verantwortung von Bibliotheken, Archiven und Museen sowie Forschungseinrichtungen und Medien im Umgang mit der NS-Zeit

In addition, there are still problems with the German copyright law: for example, allowing access to the digitalised material collides with the 70-year protection period under § 64 of the Copyright Act, which protects those contributions written up to 1945 by named authors who were still alive after the end of the Second World War. According to a scientific assessment, in the case of the *Freiheitskampf* this would affect about 25 percent of the articles included. However, identifying the authors or their heirs would be a disproportionate effort.[22] The alternative of not publishing these digital copies is not a solution either, as this would result in the loss of important information on National Socialism in Saxony, which would significantly dilute the purpose of the project. This is where politics is being called upon to create appropriate framework conditions. Although a regulation on the use of out-of-print works in § 13 ff. Copyright Administration Law, since 1 June 2016 in § 51 f. Verwertungsgesellschaftengesetz (VVG), printed works in the form of books, newspapers, and magazines published in Germany before 1 January 1966 and currently out of print can be digitised and made freely accessible after licensing by the Verwertungsgesellschaft Wort (VG Wort). However, a framework agreement with implementation regulations for such works in periodicals, such as newspapers, has not yet been concluded.

Nevertheless, we advocate general access to the digitised material so that researchers worldwide can work with this accessible source in order to further close research gaps in the history of Saxony under National Socialism and to make Saxony more visible in the field of historical research concerning this time. Additionally, the reappraisal of National Socialism through initiatives outside the academic-university milieu, in the regions and communities of Saxony, should benefit from the free use of the newly exploited source for means of political education.[23]

Compared with the opportunity to provide both researchers and citizen scientists[24] with an additional research option, we consider the risk of abuse by right-wing extremists to be low, as they have been independently covering their

---

im Netz." In: *H-Soz-Kult*, 31.03.2020, https://www.hsozkult.de/conferencereport/id/tagungsberichte-8710.

22 See: Anne Lauber-Rönsberg, "Urheberrechtliche Regulierung der Digitalisierung vergriffener Periodika aus den dreißiger Jahren." In: *Zeitschrift für geistiges Eigentum* 8 (2016) 1, pp. 48–83, DOI: 10.1628/186723716X14586350989542.

23 See: Thomas Bürger, "Heilsames Gift? Politische Aufklärung durch digitale Bereitstellung von NS-Zeitungen." In: *Zeitschrift für Bibliothekswesen und Bibliographie* 64 (2017) 3/4, pp. 145–157, DOI: 10.3196/1864295017643469.

24 On their role in historical research with the help of digital tools see Martin Munke, "Citizen Science/Bürgerwissenschaft. Projekte, Probleme, Perspektiven am Beispiel Sachsen." In: Jens Klingner; Merve Lühr (eds.), *Forschungsdesign 4.0. Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive*, Dresden (2019), pp. 107–124, DOI: 10.25366/2019.11.

need for National Socialist sources for years and decades. In the context of the progress of DNB, which, due to copyright problems, had in the meantime restricted access to its "Exilpresse digital"[25] service for German-language exile magazines between 1933 and 1945, did make it freely available again, SLUB has also recently considered presenting the *Freiheitskampf* in a less restrictive manner in future. Various providers of newspaper portals have for some time now began to freely present periodic publications up to 1945 on the internet[26] – a step which the SLUB followed by offering free access to the *Freiheitskampf* and other newspapers of the time from the end of 2021 onwards via its Digital Collections and the regional portal Saxorum mentioned above[27]. The online presentation is accompanied by a disclaimer, as well as additional texts explaining the character of the source and its role in the press landscape of the time. In light of the importance of the *Freiheitskampf* not only as a source for historical research but also for political education, the residual copyright risk seems justifiable to us. However, the demand for an extension of the legal possibilities by updating existing framework agreements for periodicals as well remains unaffected.

4 Further Potentials of the Project

In addition to its function as a "finding book" of a newly tapped source for historical research, the database for *Der Freiheitskampf* contains much more potential, not only in Saxony, but especially in the field of political-historical education for pupils and teachers. In this area in particular, the working group has come up with new project ideas for the further use of the recorded data, for example, by developing didactic concepts for the use of the database in history lessons. In addition, a "slimmed-down" version of the database, specially tailored to educational work, would also be conceivable, with which pupils could use selected and

25 See: Sylvia Asmus, Dorothea Zechmann, "Exilpresse Digital und Jüdische Periodika aus NS-Deutschland. Zwei Digitalisierungsprojekte der Deutschen Nationalbibliothek." In: Paul Klimpel/ Ellen Euler (eds.): *Der Vergangenheit eine Zukunft – Kulturelles Erbe in der digitalen Welt, eine Publikation der Deutschen Digitalen Bibliothek,* Berlin (2015), pp 226–235.

26 See the Austrian online collection ANNO (AustriaN Newspapers Online), which includes for example the Vienna edition of the *Völkischer Beobachter* (http://anno.onb.ac.at/cgi-content/anno?aid=vob), or the digital collections of the university library of Heidelberg, which include the propagandistic women's magazine *NS-Frauen-Warte* (https://www.ub.uni-heidelberg.de/helios/digi/nsfrauenwarte.html).

27 See: https://digital.slub-dresden.de/kollektionen/143/, https://www.saxorum.de/index.php?id=11352.

extensively annotated articles on a wide range of topics to explore historical facts for themselves.

Due to the particularly dense tradition of NSDAP activities in the Dresden area and in the neighbouring Amtshauptmannschaft, an overview of the NSDAP local groups in this area was compiled in the course of data collection. Thus, the development of the local party organisation can be traced from the foundation of the first NSDAP local group in April 1924 in Dresden-Cotta, through the formation of the first sections in 1929, to a party unit with over 80 local groups in the Dresden city area alone, currently up to 1937. In this Excel database, in addition to the corresponding article IDs and important information (date of foundation, party membership numbers, divisions and mergers of local groups), the specified local group leaders, as well as the locations of the local party lines were noted. Since the comment fields of the main database contained the addresses of the seats of various party offices and structures in the Gau and local levels, as well as the locations of rallies and other party activities, the idea of a cartographic visualization of places of National Socialist rule, practice and participation in the Dresden urban area was maturing. Here, precisely those places could be entered in a geo-referenced 1938 Dresden city map and linked to the corresponding articles. The additional entry of locations of the Wehrmacht, armaments production or of concentration camps for forced labourers as well as places of other National Socialist crimes would make the presence of such places visually presentable and thus question the myth of the "innocent city".[28]

Further potential for the project results from the progress that automated text recognition has recently made, especially with the use of specifically trained OCR models.[29] This process step has only been a standard procedure in retro-digitisation at SLUB for a few years now and is handled by the commercial

28 What the intensive usage of newspapers can contribute to answer research questions like this is evident to some extent in the exhibition catalogue by Katrin Nitzschke, Johannes Wolff, "Stunde Null? Dresdner Tageszeitungen über Zusammenbruch und Neuanfang April bis August 1945." In: Dresden 2015, URN: urn:nbn:de:bsz:14-qucosa-164633. Another example is the paper of Ursula Fuchs-Materny, analysing the propagandistic accompaniment of the start of the war in 1939; see: Ursula Fuchs-Materny, "'Der Freiheitskampf' auf Kriegskurs. Dresdner Presse im Jahr 1939." In: *Dresdner Hefte* 11 (1993) 3, pp. 75–83, URL: https://digital.slub-dresden.de/id351372032/77.

29 See: Uwe Springmann, Anke Lüdeling, "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus." In: *Digital Humanities Quarterly* 11 (2017) 2, URL: http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html; Konstantin Baierer, Philipp Zumstein, "Verbesserung der OCR in digitalen Sammlungen von Bibliotheken." In: *027.7. Zeitschrift für Bibliothekskultur 4* (2016) 2, DOI: 10.12685/027.7-4-2-155.

solution of Abbyy Cloud.[30] This solution offers hardly any possibilities for post-correction within the open source Kitodo digitisation suite, which is in use at SLUB.[31] Currently, SLUB is one of the pilot libraries that want to implement the results of the OCR-D project funded by the German Research Foundation (DFG) and transfer them into regular workflows. In addition to the actual text recognition, the modular OCR-D software contains several pre- and post-processing steps, all based on free and open standard tools.[32] In the course of implementation, a subsequent OCR treatment of the *Freiheitskampf* is also to be carried out, and is currently already in progress. On the one hand, this will make life easier for indexing the articles, which up to now has been done manually. This also means that the qualitative project can be extended by a quantitative dimension. In addition to research based on "classical" historiographic methods, this would open up the analysis of the source in an interdisciplinary way for further projects in the field of Digital Humanities – for example, for linguistic studies on the diffusion of National Socialist ideology into everyday language.[33]

This also requires the possibility of contextualisation with other press products of the time. Corresponding to the *Freiheitskampf*, the SLUB will also extend the digitisation limit for other newspapers until 1945 in the context of the decision outlined above. Although the press landscape from 1933 onwards was increasingly under the control of the Nazis, some newspapers still offer different or at least complementary perspectives to those of a party organ. For Saxony, comparisons of the *Freiheitskampf* with the *Dresdner Neueste Nachrichten*, which by the end of 1943 had again become the highest-circulation newspaper in the state capital of Dresden, have shown this.[34] In the future, comparisons with the national and international press will be easier. An important approach is the juxtaposition of different newspapers via aggregators such as Europeana Newspapers[35] or the new national newspaper portal for Germany that has been developed in another joint

30 See: https://www.ocrsdk.com/.

31 See: https://www.kitodo.org/; https://github.com/Kitodo.

32 See: https://ocr-d.de/ (with a list of recent publications and presentations on the project); https://github.com/OCR-D.

33 The potentials of such approaches and methods in working with historical newspapers are hinted at in Erik Koenen, "Digitale Perspektiven in der Kommunikations- und Mediengeschichte. Erkenntnispotentiale und Forschungsszenarien für die historische Presseforschung." In: *Publizistik* 63 (2018), pp. 535–556, here: 548–550, DOI: 10.1007/s11616-018-0459-4.

34 See: Ralf Krüger, "Presse unter Druck. Differenzierte Berichterstattung trotz nationalsozialistischer Presselenkungsmaßnahmen. Die liberalen Dresdner Neueste Nachrichten und das NSDAP-Organ Der Freiheitskampf im Vergleich." In: Reiner Pommerin (ed.), *Dresden unterm Hakenkreuz,* Köln/Weimar/Wien (1998), pp. 43–66.

35 See: http://www.europeana-newspapers.eu/.

project with the participation of the SLUB.[36] Together with the aforementioned possibilities of contextualisation, an important but difficult source such as the *Freiheitskampf* can, in our view, be classified appropriately.

5 Conclusion

In view of the transformation of Saxony, from a stronghold of social democracy to a veritable Nazi model region in the 1930s,[37] research into the region is particularly interesting for historiography. This is why the loss of sources is all the more serious. With the indexing of the Saxon NSDAP regional organ *Der Freiheitskampf* and the provision of the database, HAIT and SLUB make an important contribution to the research infrastructure for the regional history of National Socialism. Due to the loss of primary sources on the NSDAP and its divisions, the daily newspaper of the NSDAP in Saxony provides a deep and comprehensive insight into the work of the representatives of National Socialism and the everyday life of the local population. Thus, the *Freiheitskampf* can already be regarded as the most important source for the establishment and consolidation of National Socialism in Saxony. In addition to automated full-text recognition, in-depth indexing of the content allows both a more precise search and a contextualisation of the article content, and with the help of the category structure, the places and the persons involved.[38] It is precisely these three characteristics – topics, persons and geographical locations – that are essential identifiers, offering networking potential with online content from other institutions and therewith connecting the data provided in the project with the world of Linked Data.[39] The first step has already been realised in part through the BEACON service and will be further expanded in the future. The combination of the different search approaches is an example of the profound changes in the handling of

36 See: Lisa Landes, "Ein Zeitungsportal für Deutschland." In: *Dialog mit Bibliotheken* 31 (2019) 2, pp. 12–14, URL: https://d-nb.info/1203670826/34; Reinhard Altenhöner, "Auf dem Weg zu einem nationalen Zeitungsportal. Eine materialspezifische Kooperation als Treiber eines neuen Dienstes für Wissenschaft und Forschung." In: Achim Bonte, Juliane Rehnolt (eds.), *Kooperative Informationsinfrastrukturen als Chance und Herausforderung. Festschrift für Thomas Bürger zum 65.* Geburtstag, Berlin (2018), pp. 144–160, DOI: 10.1515/9783110587524-019.

37 See: Claus-Christian W. Szejnmann, *Nazism in Central Germany. The Brownshirts in 'Red' Saxony*, New York (1999).

38 See: Koenen, "Digitale Perspektiven", pp. 543–544.

39 See: Tom Heath, Christian Bizer, *Linked Data. Evolving the Web into a Global Data Space*, San Rafael 2011, DOI: 10.2200/S00334ED1V01Y201102WBE001.

historical newspapers as sources that digitisation has brought with it. Conventional research based on manual searching oriented to publication schedule is being transformed into a parallel search across entire volumes, which, depending on the state of indexing of the material, will also be much more complete.[40]

The main obstacles to a productive use of the source are the fact that only half of the published volumes of the *Freiheitskampf* have been processed and indexed in the database so far, and the currently existing legal restrictions on the use of such works. Although a search in the database will produce a hit list of corresponding articles with a summary of their contents and linked categories, persons and places, the last step – namely access to the actual digitised versions of the newspaper articles – could only be carried out from two special workstations in the HAIT library and in the SLUB. As indicated, solutions are being worked on here that produced first results at the end of 2021.

Despite these previous limitations, the number of users of the database and the number of concrete inquiries to the project group already prove the relevance of the *Freiheitskampf* for historical research and for people interested in history. With the completion of the actual indexing, the potential of the database is far from exhausted. The HAIT is therefore considering further applications for the database, especially in the field of political-historical education. They concern in particular the development of didactic concepts for the use of the database in history lessons or a cartographic visualization of the places of National Socialist rule, practice and participation in a paradigmatic region. The continuation and further development of the project is, of course, dependent on long-term funding.

On the one hand, the example of the *Freiheitskampf* shows the importance of digitized daily newspapers as a source for historical science. On the other hand, it becomes clear that indexed, historical newspapers also create the basis for a variety of uses in political education or citizen science. The possibilities to inform people about the methods of undermining a democratic, constitutional state and the public communication strategies of the National Socialist movement as well as to show the criminal character of the National Socialist regime by using historical newspapers are significant. In this context, a local newspaper has the advantage of being able to show that the exclusion and persecution of different population groups did not just take place far away, but was instead visible in every community. The decoding of propaganda can also be illustrated by using such historical texts. Altogether, it outweighs the risk of misuse of single texts by right-wing extremists in the present.

40 See: Koenen, “Digitale Perspektiven”, pp. 541–542.

However, the easy accessibility of newspapers as sources is a prerequisite for their widespread use in research and educational work. The German legislature must create reliable legal conditions for implementing this. Until now, the legal situation regarding the provision of historical newspapers on the Internet has been unclear. As Huub Wijfjes stated some years ago, providing infrastructures for digital research with historical newspapers is also "about making a serious effort in solving the copyright problem by putting the interest of public consultation high on the agenda".[41] Therefore, the SLUB decided to expand its digitisation activities concerning newspapers until the publication year 1945, and to freely present the holdings affected in 2021. This includes the *Freiheitskampf*, which significantly enriches the functionality of online research in the HAIT database. Once OCR recognition has also been carrried out, the Nazi newspaper will be an easily accessible "El dorado" for researchers, teachers and citizens who are interested in the history of Saxony between 1930 and 1945.

Bibliography

Altenhöner, Reinhard, "Auf dem Weg zu einem nationalen Zeitungsportal. Eine materialspezifische Kooperation als Treiber eines neuen Dienstes für Wissenschaft und Forschung." In: Achim Bonte, Juliane Rehnolt (eds.), *Kooperative Informationsinfrastrukturen als Chance und Herausforderung*. Festschrift für Thomas Bürger zum 65. Geburtstag, Berlin 2018, pp. 144–160, DOI: 10.1515/9783110587524-019.

Asmus, Sylvia, and Dorothea Zechmann, "Exilpresse Digital und Jüdische Periodika aus NS-Deutschland. Zwei Digitalisierungsprojekte der Deutschen Nationalbibliothek." In: Paul Klimpel, Ellen Euler (eds.): *Der Vergangenheit eine Zukunft. Kulturelles Erbe in der digitalen Welt*, eine Publikation der Deutschen Digitalen Bibliothek, Berlin 2015, pp. 226–235, DOI: 10.17176/20180716-114912-0.

Baierer, Konstantin, and Philipp Zumstein, "Verbesserung der OCR in digitalen Sammlungen von Bibliotheken." In: 027.7. *Zeitschrift für Bibliothekskultur* 4:2 (2016), DOI: 10.12685/027.7-4-2-155.

Bürger, Thomas, "Heilsames Gift? Politische Aufklärung durch digitale Bereitstellung von NS-Zeitungen." In: *Zeitschrift für Bibliothekswesen und Bibliographie* 64 (2017) 3/4, pp. 145–157, DOI: 10.3196/1864295017643469.

Bürger, Thomas, "Zeitungsdigitalisierung als Herausforderung und Chance für Wissenschaft und Kultur." In: *Zeitschrift für Bibliothekswesen und Bibliographie*, 63 (2016) 3, S. 123–132, DOI: 10.3196/186429501663332.

Bürger, Thomas, and Sebastian Meyer, "Schlagzeilen im Binärcode. Fortschritte und Herausforderungen bei der Digitalisierung historischer Zeitungen." In: *BIS. Das Magazin der Bibliotheken in Sachsen* 9 (2016) 3, 139–141. URN: .urn: nbn:de:bsz:14-qucosa2-77780.

41 Wijfjes, "Digital Humanities and Media History", p. 21.

Chiron, Guillaume, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux, "Impact of OCR Errors on the Use of Digital Libraries. Towards a Better Access to Information." In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), June 19–23 2017, Toronto, DOI: .10.1109/JCDL.2017.7991582.

Ehrmann, Maud, Estelle Bunout, and Maren Düring, "Historical Newspaper User Interfaces: A Review." In: Libraries: Dialogue for Change. World Library and Information Congress, 85th IFLA General Conference and Assembly, August 24–30 2019, Athens, URL: http://library.ifla.org/2578/1/085-ehrmann-en.pdf.

Fischer, Markus, "Neue Perspektiven auf die sächsische NS-Presse. Eine Aufarbeitung des NSDAP-Organs *Der Freiheitskampf."* In: *Neues Archiv für sächsische Geschichte*, 84 (2013), pp. 281–291.

Fuchshuber, Jutta, "Nationalsozialismus digital. Die Verantwortung von Bibliotheken, Archiven und Museen sowie Forschungseinrichtungen und Medien im Umgang mit der NS-Zeit im Netz." In: *H-Soz-Kult*, 31.03.2020, www.hsozkult.de/conferencereport/id/tagungsberichte-8710.

Hale, Oron J., *Presse in der Zwangsjacke 1933–1945*, Düsseldorf (1965).

Hanzig, Christoph, Michael Thoß, "'Rotmord' vor Gericht, 'Politisch motivierte Tötungsdelikte in Sachsen im Spiegel der NS-Tageszeitung 'Der Freiheitskampf' von 1931 bis 1936." In: Gerhard Lindemann, Mike Schmeitzner (eds.), *. . . da schlagen wir zu. Politische Gewalt in Sachsen 1930–1935*, pp. 193–230, DOI: 10.14220/9783737009348.193.

Heath, Tom, Christian Bizer, *Linked Data. Evolving the Web into a Global Data Space*, San Rafael (2011), DOI: 10.2200/S00334ED1V01Y201102WBE001.

Koenen, Erik, "Digitale Perspektiven in der Kommunikations- und Mediengeschichte. Erkenntnispotentiale und Forschungsszenarien für die historische Presseforschung." In: *Publizistik* 63 (2018), pp. 535–556, DOI: 10.1007/s11616-018-0459-4.

Krüger, Ralf, "Presse unter Druck. Differenzierte Berichterstattung trotz nationalsozialistischer Presselenkungsmaßnahmen. Die liberalen Dresdner Neueste Nachrichten und das NSDAP-Organ Der Freiheitskampf im Vergleich." In: Reiner Pommerin (ed.), *Dresden unterm Hakenkreuz*, Köln/Weimar/Wien 1998, pp. 43–66.

Krings, Stefan, *Hitlers Pressechef: Otto Dietrich (1897–1952). Eine Biografie*, Göttingen (2010).

Landes, Lisa, "Ein Zeitungsportal für Deutschland." In: *Dialog mit Bibliotheken* 31 (2019) 2, pp. 12–14, URL:https://d-nb.info/1203670826/34.

Lauber-Rönsberg, Anne, "Urheberrechtliche Regulierung der Digitalisierung vergriffener Periodika aus den dreißiger Jahren." In: *Zeitschrift für geistiges Eigentum* 8 (2016) 1, pp. 48–83, DOI: 10.1628/186723716X14586350989542.

Fuchs-Materny, Ursula, "'Der Freiheitskampf' auf Kriegskurs. Dresdner Presse im Jahr 1939." In: *Dresdner Hefte* 11 (1993) 3, pp. 75–83, URL:https://digital.slub-dresden.de/id351372032/77.

Mentschl, Christoph, "Die Verantwortung von Bibliotheken, Archiven und Museen sowie Forschungseinrichtungen und Medien im Umgang mit der NS-Zeit im Netz." In: *o|bib* 7 (2020) 2, DOI: 10.5282/o-bib/5599.

Munke, Martin, "Citizen Science/Bürgerwissenschaft. Projekte, Probleme, Perspektiven am Beispiel Sachsen." In: Jens Klingner; Merve Lühr (eds.), *Forschungsdesign 4.0. Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive*. Dresden (2019), pp. 107–124, DOI: 10.25366/2019.11.

Munke, Martin "NS-Geschichte digital erforschen: Tageszeitung 'Der Freiheitskampf' jetzt online verfügbar." In: SLUBlog, 04.11.2021, https://blog.slub-dresden.de/beitrag/2021/11/4/ns-geschichte-digital-erforschen-tageszeitung-der-freiheitskampf-jetzt-online-verfuegbar.

Nitzschke, Katrin, and Johannes Wolff, "Stunde Null? Dresdner Tageszeitungen über Zusammenbruch und Neuanfang April bis August 1945", Dresden 2015, URN: urn: nbn:de:bsz:14-qucosa-164633.

Springmann, Uwe, Anke Lüdeling, "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus." In: *Digital Humanities Quarterly* 11 (2017) 2, URL:http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.

Stumpf, Markus, "Ausgewählte Erkenntnisse aus der Enquete 'Nationalsozialismus digital' (Wien, 27.–29. November 2019)." In: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 73 (2020) 1, pp. 147–151, DOI: 10.31263/voebm.v73i1.3479.

Szejnmann, Claus-Christian W., *Nazism in Central* Germany. *The Brownshirts in 'Red' Saxony*, New York (1999).

Templer, Josephine, "Rezeption von politischer Gewalt und ihrer Funktion in der sächsischen Presse zwischen 1930 und 1933. 'Der Freiheitskampf' und die 'Arbeiterstimme' im Vergleich." In: Gerhard Lindemann, Mike Schmeitzner (eds.), *. . . da schlagen wir zu. Politische Gewalt in Sachsen 1930–1935*, Göttingen (2020), pp. 21–52, DOI: 10.14220/9783737009348.21.

Widera, Thomas, Thomas, Martin Munke, and Matti Stöhr, "'Der Freiheitskampf' – Digitalisierung und Tiefenerschließung einer NS-Zeitung." In: *Relying on News Media. Long Term Preservation and Perspectives for Our Collective Memory*. IFLA News Media Section Satellite conference 2017, August 16–18 2017, Dresden, URN: urn: nbn:de:bsz:14-qucosa2-164012.

Wijfjes, Huub, "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." In: *Tijdschrift voor Mediageschiedenis*, 20 (2017) 1, pp. 4–24, DOI: 10.18146/2213-7653.2017.277.

Our sincere thanks go to Rhian Krische (SLUB Dresden) for her linguistic review of the paper.

François Robinet, Rémi Korman

Des usages des collections numériques de presse pour écrire l'histoire du génocide des Tutsi du Rwanda

Abstract: This paper is based on a confrontation of the uses of the written press, paper and digital, by two historians, Rémi Korman and François Robinet. The paper underlines the difficulties usually faced by historians. The Rwandan print media collections are dispersed and digital Rwandan press archives remain scarce. Although the French press is more easily accessible in print, it is not always available in digital format, that allow for instance full-text search. They therefore wonder about the feasibility of a systematic digitization of the Rwandan print media collections and its possible contributions to the writing of history.

Keywords: Rwanda, Genocide against the Tutsi in Rwanda, Digital archives, Print and digital media collections, Writing of history

1 Introduction

> For historians of political thought, African newspapers do more than merely enable us to identify key political actors and key events: they allow us to answer questions about how political ideas changed over time. Newspapers provide a crucial body of evidence of changing ideas and their circulation. Many important figures in the intellectual history of nineteenth and twentieth-century Africa were journalists or edited their own newspapers, and newspapers are thus a crucial source for their ideas. (Hunter 2018, 23)

Le constat posé par l'historienne Emma Hunter concernant la presse africaine s'applique très largement au cas rwandais, certains médias rwandais et journalistes extrémistes ayant joué un rôle criminel au cours du génocide des Tutsi. Dès lors, depuis 1994, chercheurs, militants, juristes ont collecté ce matériau et parfois contribué à rendre une partie des collections de presse écrite accessibles sous des formats numériques.

Du point de vue méthodologique, il est nécessaire de s'interroger sur le rôle de ces collections comme source pour l'écriture de l'histoire du génocide des Tutsi. Quels sont les apports spécifiques de ce type de matériau, en comparaison des archives ou des témoignages ? De quelle façon la presse a-t-elle été mobilisée par les universitaires ? Ces questionnements s'inscrivent dans un cadre largement

https://doi.org/10.1515/9783110729214-009

documenté (Kayser 1957; Trénard 1972)[1] considérant la presse comme une source majeure pour les historiens. Ces enjeux méthodologiques ont cependant été très largement renouvelés avec la numérisation massive de la presse écrite au cours des vingt dernières années (Clavert et Schafer 2018).

C'est dans ce contexte général de réflexion sur la presse comme source et objet d'histoire que s'inscrit ce travail[2]. Nous souhaitons discuter les effets possibles d'une numérisation massive de la presse écrite rwandaise, qui n'a pas encore eu lieu mais pourrait advenir au cours des prochaines années. Comment les sources de presse écrite rwandaise sont-elles mobilisées à ce jour alors que les collections numérisées restent rares ? De plus, quelle est la place accordée à la presse française et internationale dans l'écriture de l'histoire du génocide des Tutsi ?

Nous reviendrons d'abord sur l'histoire de la presse écrite rwandaise et sur les usages de celle-ci comme source par les historiens. Dans un deuxième temps, nous proposerons une première cartographie des collections de presse disponible au Rwanda et à l'étranger. Enfin, nous présenterons les enjeux et défis posés par un projet de numérisation de la presse écrite rwandaise. Nous verrons les apports possibles d'un tel effort de numérisation, mais aussi les questions politiques, éthiques et techniques qu'il soulève.

1 La réflexion des historiens, et des journalistes, sur ce sujet est ancienne : Pour une approche synthétique voir (Soulet 2012, 105-24).

2 Ce texte est le fruit de la confrontation des pratiques et usages de la presse écrite, papier et numérique, par deux historiens. Le premier, Rémi Korman, travaille sur la fabrique de la mémoire du génocide des Tutsi depuis 1994 quand le second, François Robinet, étudie la relation franco-rwandaise et la controverse politique, médiatique et académique suscitée par les engagements français au Rwanda entre 1990 et 1994. Aussi travaillent-ils à la fois sur les presses française et rwandaise, publiées avant, pendant et après le génocide. Ils font par ailleurs partie du groupe de coordination du réseau international de recherche RwandaMAP qui a pour ambition l'étude des traces archivistiques, mémorielles et patrimoniales du génocide des Tutsi. Dans ce cadre, ils réalisent actuellement une cartographie des ressources et des productions scientifiques (bibliographie collaborative ; annuaire des chercheurs ; inventaire des lieux d'archives) concernant le génocide des Tutsi. Voir le carnet de recherche du projet : https://rwandamap.hypotheses.org/ – Voir également (Robinet 2016; Korman 2014).

2 Usages passés et actuels des collections de presse écrite rwandaise dans l'historiographie

2.1 La presse rwandaise, un bel objet d'histoire

La période coloniale est marquée par le quasi-monopole de la presse missionnaire catholique (Bart 1980)[3]. Celle-ci se développe dès la fin de la Première Guerre mondiale avec les titres *Trait d'Union* (1917) et *L'Écho du Rwanda* (1923). D'abord destinée aux missionnaires, aux intellectuels et aux séminaristes, cette presse accompagne les missions d'éducation et d'évangélisation de l'Église (Lenoble-Bart 2005). L'absence de presse officielle et la proximité étroite entre l'administration belge, les autorités autochtones et les institutions catholiques font par ailleurs du principal titre de l'époque, le journal *Kinyamateka* (Fig. 1), créé en 1933, un relais efficace des décisions officielles (Munyakayanza 2013, 16). Dans les années 1950, certains titres comme *Kinyamateka* ou *Temps nouveaux d'Afrique* s'engagent dans les débats politiques et sociaux de leurs temps (Munyakayanza 2013, 35-36).

En 1962, l'indépendance du pays ouvre une deuxième période d'essor pour la presse écrite. Tout d'abord, la presse gouvernementale apparaît avec les publications du Ministère de l'Information, en particulier *Rwanda, Carrefour d'Afrique* (1963–1975), auquel succèdent à partir de 1975 les journaux de l'Office rwandais de l'information (Orinfor), *La Relève – Imvaho*, mais aussi des publications ministérielles. Au tournant des années 1960–1970 se déploient également de multiples bulletins diocésains, d'institutions religieuses ou paroissiennes, revues d'écoles secondaires, de communautés, de séminaires[4]. Enfin, de nouveaux titres catholiques se créent tels *Culture chrétienne* (1962–1964) et surtout, à partir de 1967, la revue bimestrielle *Dialogue* (Fig. 1) portée par des religieux dominicains (Bart 1992)[5]. Bien qu'elle reste très influente grâce à son excellent réseau de diffusion, la presse catholique perd donc son monopole (sur 65 titres recensés par Annie Bart en 1979, seuls 39 sont catholiques) et *Kinyamateka* se voit concurrencé par *Imvaho* (en 1984, le second titre tire désormais 23.000 exemplaires quotidiens contre 10.500 pour le premier) (Lenoble-Bart 2005, 475). Malgré l'essor du nombre de publications, jusqu'en 1987, les Première et Deuxième Républiques ne permettent pas le développement d'une presse privée indépendante du pouvoir.

3 Sur le parcours d'Annie Bart, voir (Dakhlia et Robinet 2016).

4 Citons entre autres la revue *Urunana* du grand séminaire de Nyakibanda, ou encore *Foi et Culture*, du diocèse de Nyundo.

5 Notons que la revue *Dialogue* est en partie accessible en ligne : https://www.genocidearchiverwanda.org.rw/.

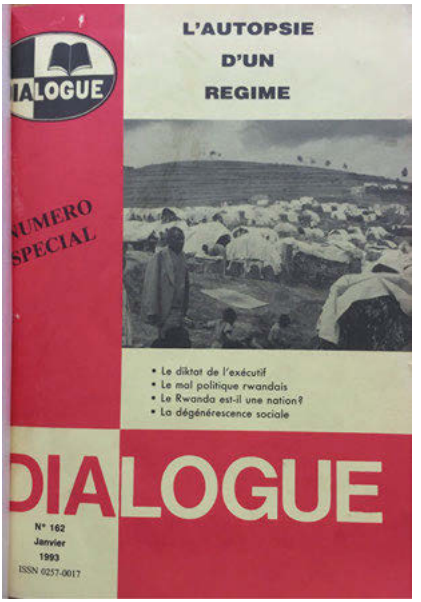


Fig. 1: Presse rwandaise catholique et intellectuelle.
Source: Couverture du journal *Kinyamateka*, n° spécial du Jubilé, 15 août 1975, Couverture de la revue *Dialogue*, n°162, janvier 1993
Copyright: Collections CEDOREP.

Cependant, dès 1988, le bimensuel *Kanguka* (Réveille-toi !), créé par Vincent Rwabukwisi, tente de faire entendre une voix critique face à la presse officielle. À partir de 1990 débute au Rwanda un mouvement d'ouverture politique, marqué par l'introduction du multipartisme ainsi que par la création de journaux, démocratiques ou extrémistes. Le rôle des médias extrémistes au cours de la guerre civile qui débute en octobre 1990 est très tôt discuté dans la sphère universitaire (Chrétien 1991), en particulier le cas du journal *Kangura* (Réveille-le !). La publication de l'ouvrage *Rwanda : Les médias du génocide* en 1995 a constitué un jalon important de la connaissance des acteurs de la presse rwandaise de cette époque et du rôle des médias dans l'essor du racisme anti-tutsi et de la culture de la violence (Chrétien et al 1995). L'écho rencontré par cet ouvrage a aussi indirectement conduit à des effets de sources, la violence politique de la période se trouvant dès lors le plus souvent discutée sous le seul angle des médias de la haine sans attention particulière accordée aux autres titres de presse (presse officielle, presse d'opposition . . .).

La presse rwandaise post-génocide n'a fait à ce jour l'objet que d'un nombre extrêmement limité de publications (Frère 2017). Les nouvelles autorités rwandaises reprennent les journaux distribués par l'Orinfor en changeant sim-

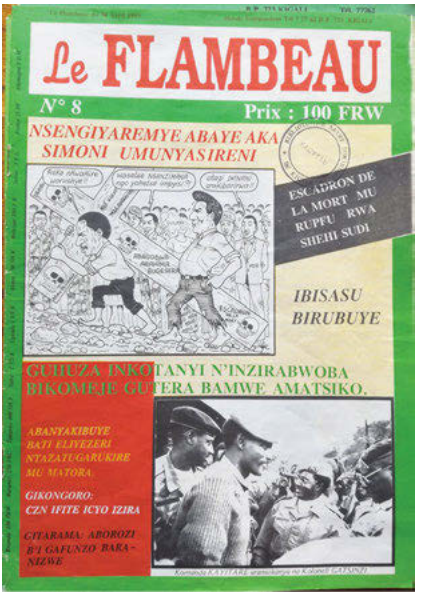


Fig. 2: Presse d'opposition et presse post-génocide.
Source: *Le Flambeau* n°8, 30 avril 1993, *Rwanda Libération* n°6, 14 avril-14 mai 1995.
Copyright: Collections CEDOREP.

plement le nom (*La Nouvelle Relève* et *Imvaho Nshya*). Certaines publications créées après 1990 continuent à exister, comme *Imboni* ou *Le Tribun du Peuple*, d'autres apparaissent après le génocide (*Goboka*, *Urwatubyaye*, *Rwanda Libération*, etc.) (Fig. 2). À cette presse principalement en kinyarwanda et français s'ajoutent enfin de nouveaux titres en anglais, principalement *The New Times* à partir de 1995. Apparaissent aussi de nouvelles revues, comme les *Cahiers Lumières et Société* portés par le dominicain Bernardin Muzungu, qui aborde des thématiques politiques, culturelles et religieuses. La revue *Dialogue* continue de son côté son travail « momentanément en exil » en Belgique puis au Rwanda[6].

2.2 La presse rwandaise : une source majeure pour l'écriture de l'histoire

Sans prétendre ici à l'exhaustivité, il est utile d'interroger l'évolution des usages de la presse pour l'écriture de l'histoire du génocide avant la numérisation

6 Sur l'évolution de la revue *Dialogue*, voir (Cristofori, 2009).

de la presse et l'essor de plateformes de consultation d'archives de presse numérisées[7]. Durant les premières années post-génocide, à l'heure où la presse en ligne en est à ses balbutiements, paraissent quelques ouvrages académiques proposant une première narration des faits. À partir de documents collectés durant les événements et de témoignages confrontés aux informations et aux récits des articles des presses belge et française, Gérard Prunier livre une étude fouillée centrée sur les quatre années de guerre et la préparation du génocide (Prunier 1995). Dans *Rwanda, le piège de l'histoire*, Jordane Bertrand utilise une large diversité de sources de presse françaises, belges et surtout rwandaises pour décrire l'évolution du Mouvement démocratique républicain et de l'opposition démocratique entre 1990 et 1994 (Bertrand 2000). Témoins directs des événements, Filip Reyntjens et André Guichaoua, mobilisent le même type de sources dans leurs études respectivement publiées en 1994 et 1995 (Reyntjens 1994; Guichaoua 1995).

Les usages des articles de presse écrite permettent à ces chercheurs de précocement consolider une chronologie et un récit des faits, de restituer certaines déclarations publiques des acteurs ou encore de décrire l'évolution du jeu diplomatique, de l'accord de cessez-le-feu de Mwanza en octobre 1990, jusqu'à la signature des accords d'Arusha en août 1993. La presse d'information belge, dont la couverture s'est avérée plus dense, régulière et précise que celle de la presse française, est mobilisée et confrontée aux articles de la presse écrite rwandaise[8]. Au regard de l'état des sources au Rwanda (pertes, dégradation, dispersion …) et de fonds clos ou non encore constitués en France, la presse écrite papier occidentale représente, durant ces premières années, une documentation immédiatement mobilisable.

De la fin des années 1990 à la fin des années 2000, la presse écrite reste une source importante. Une nouvelle phase de l'historiographie s'ouvre cependant durant laquelle les chercheurs – historiens mais aussi sociologues, politistes et anthropologues – privilégient désormais les enquêtes de terrain, les sources orales et les approches locales, parfois au détriment des sources issues de la presse écrite. En témoigne, par exemple, la faible mobilisation de ce type de sources dans l'ouvrage de référence *Aucun témoin ne doit survivre* (Des For-

7 Sur la conversion de la presse en ligne française au numérique : (Albert et Sonnac 2014; Delporte, Blandin, et Robinet 2016).

8 Un effort de collecte systématique des presses belges et françaises a été fait par le journaliste Oscar Gasarabwe sur la période avril 1994-avril 1995 : (Gasarabwe 1996). C'est également le cas des dossiers de presse et de la revue *Documentation réfugiés* édités de décembre 1986 à janvier 1995 par le Centre interassociatif francophone d'information et de documentation sur le droit d'asile et les réfugiés et accessibles en consultation à La Contemporaine à Nanterre.

ges 1999). Par ailleurs, les procès qui se tiennent au Tribunal pénal international pour le Rwanda (TPIR), lors des juridictions populaires Gacaca au Rwanda ou selon le principe de la compétence universelle produisent de nouvelles archives pour les historiens (Rovetta 2019, 252-63; Dumas 2014, 17-22). Aussi, si l'anthropologue Johann Pottier livre en 2002 une étude des récits médiatiques en lien avec la communication des acteurs (Pottier 2002), les productions de cette époque, prenant principalement pour source la presse écrite, sont quasiment exclusivement le fait de journalistes ou de militants[9].

Depuis la fin des années 2000, les archives de presse retrouvent une centralité dans certaines enquêtes historiennes où la presse est à la fois source et objet d'histoire et où les archives de presse papier et les archives numériques sont mobilisées pour étudier le regard porté par les contemporains sur l'événement lui-même. Les questions portent alors sur la mise en visibilité du génocide, sur le vocabulaire, les images et les discours produits avant, pendant et à l'issue de celui-ci. Elles portent également sur les liens entre les modalités de représentation de l'événement et les pratiques journalistiques ou encore sur le rôle de la presse comme filtre, relais ou amplificateur des discours produits sur le génocide. Ces études révèlent une masse documentaire conséquente, relativement aisément accessible, gérée et organisée pour partie au sein d'institutions d'archives et bénéficiant parfois désormais d'une numérisation. Quel fut dès lors le rôle de la numérisation dans l'essor de ces études et dans l'évolution du statut, de l'accessibilité et des usages des sources de presse ?

3 Quelles collections disponibles pour quelle accessibilité ?

3.1 La dispersion des collections de presse écrite rwandaise

Au Rwanda, les fonds de presse sont très largement morcelés entre diverses bibliothèques à travers le pays, principalement à Butare et Kigali[10]. En l'absence de catalogue ou d'inventaire, et du fait de l'éparpillement des collections, le dépouillement intégral de journaux est à ce jour un exercice particulièrement

9 Parmi ces productions certaines sont très proches des normes académiques : (Epelbaum 2005).

10 Voir en particulier les bibliothèques de l'Université nationale du Rwanda, de l'Académie rwandaise de la Culture et de la Langue, de la Bibliothèque nationale, du Centre dominicain de recherche et de pastorale et du Parlement rwandais.

complexe pour un chercheur isolé. À cet égard, les collections de presse ont été plus ou moins bien constituées et conservées selon les institutions. Le plus souvent, les journaux tels que *Imvaho Nshya* ou *Kinyamateka* ont été reliés par année. La méthode utilisée rend cependant très difficile une numérisation à plat (risque de perte d'informations si la marge de fond n'est pas suffisamment large ; risque de fragilisation de la reliure).

En dehors du Rwanda, il reste très difficile d'accéder aux collections de la presse rwandaise. Pour la France, par exemple, seules quelques collections privées déposées dans des bibliothèques ou centres d'archives permettent cet accès. Il faut aussi signaler les accès possibles à une cinquantaine de numéros de l'hebdomadaire *La Relève* en format papier à la Bibliothèque nationale de France (BnF), à une partie de la collection des *Rwanda, Carrefour d'Afrique* à La Contemporaine et à quelques numéros des *Études rwandaises* disséminés entre la BnF et les Archives nationales d'Outre-Mer d'Aix-en-Provence. À l'étranger, la British Library propose quelques numéros papier de la revue *Dialogue*, des *Etudes rwandaises* ou des *Cahiers du Centre Saint Dominique* tandis que la Library of Congress offre des collections – microfilmées et incomplètes – de titres comme *La Relève*, *La Nouvelle Relève*, *Kinyamateka* et *Imvaho*[11].

Bien qu'encore rares, les initiatives permettant d'accéder à certaines collections numérisées de presse rwandaise sont utiles (Fig. 3). Tel est par exemple le cas de l'initiative menée par The Aegis Trust qui a entrepris de collecter, de préserver, de numériser et de rendre accessibles de très nombreux documents sur le génocide des Tutsi, dont une vingtaine de numéros de *Kangura*, une vingtaine de numéros du quotidien d'opposition *Isibo* et quelques numéros épars de *Le Verdict*, *Le Partisan* (journal proche du FPR) ou *Zirikana* (journal proche des extrémistes de la Coalition pour la Défense de la République). Le Tribunal pénal international pour le Rwanda, dont une partie des archives est accessible en ligne[12], a aussi produit un effort de numérisation de certains journaux ou de certaines pages de journaux, sans souci d'exhaustivité, sans visée de production de sources pour l'historien et avec un objectif d'attestation de la preuve qui conduit à une surreprésentation de la presse extrémiste du début des années 1990.

Ainsi, l'incomplétude des collections, leur dispersion au Rwanda et la rareté des accès à des collections numérisées rend l'opération de reconstitution des collections difficile et chronophage pour l'historien. Celui-ci est en outre confronté à la question linguistique. Si le kinyarwanda est la langue nationale, parlée par

11 Recherche effectuée dans les catalogues en ligne de la Library of Congress (https://www.loc.gov/) et de la British Library (https://www.bl.uk/) le 22 mai 2020.

12 Pour un accès aux archives du TPIR : https://jrad.irmct.org/.

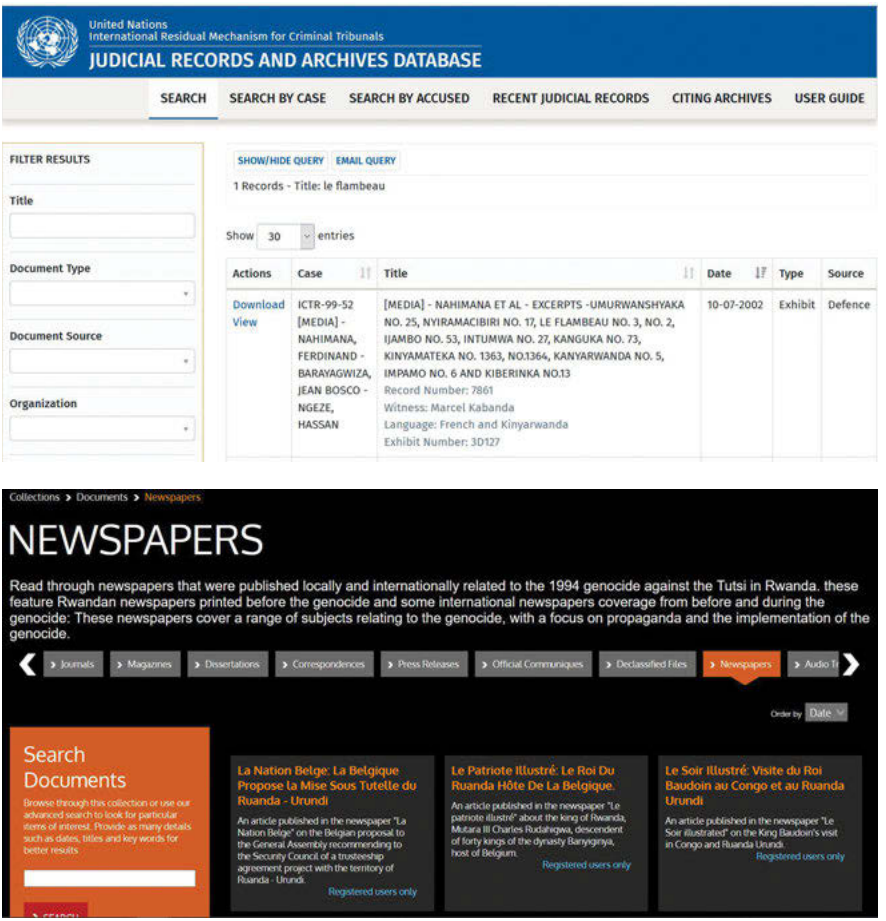


Fig. 3: Des premiers efforts de numérisation et de diffusion en ligne. En haut: Capture d'écran d'une recherche « Le Flambeau » effectuée sur le site des archives en ligne du TPIR « Judicial records and archives database », https://jrad.irmct.org/ [Consulté le 12 janvier 2021], Copyright: UN 2022. En bas: Capture d'écran de la rubrique « newspapers » du site « Genocide Archive of Rwanda », https://genocidearchiverwanda.org.rw/, [Consulté le 12 janvier 2021], Copyright: Aegis Trust 2022.

tous les Rwandais, le français a été la langue d'enseignement dans les établissements du secondaire et le monde universitaire jusqu'à la fin des années 2000, marquées ensuite par un tournant anglophone. Selon le public visé, mais aussi selon les rédacteurs, la presse rwandaise est principalement en kinyarwanda et en français. De fait, le dépouillement de la presse en kinyarwanda prend plus de temps et nécessite ensuite un long travail de traduction, un travail coûteux lorsque

l'aide d'un assistant de recherche est nécessaire, ce qui est le cas pour la majorité des chercheurs non rwandais. Cette situation explique la forte visibilité dans les publications universitaires de la revue *Dialogue*, qui rend compte de débats politiques, culturels et religieux rwandais en français.

3.2 Presse française et occidentale: profusion, fragmentation, diversité

L'historien qui souhaite travailler sur le génocide des Tutsi du Rwanda à partir de copies numérisées d'articles de la presse écrite française et/ou de pays occidentaux doit également s'attendre à être confronté à de nombreuses difficultés. Il ne trouvera pas de plateforme numérique aussi riche et bien référencée que Gallica pour la presse écrite de la fin du XVIII^{e} siècle jusqu'à 1948. Il devra aussi faire face à une historiographie encore émergente sur la presse des années 1990–2000 et sur le monde journalistique français de cette période. Enfin, il doit prévoir de faire des choix drastiques quant à son périmètre d'étude tant l'offre en presse écrite est pléthorique en termes de type de presse, de fréquence de parution ou de nature des productions. Profusion, fragmentation, diversité et difficulté d'accès seront donc au rendez-vous.

Si on la compare à celle de la fin des années 2000, la situation s'est pourtant améliorée puisque différentes bases permettent désormais de travailler à partir de la presse écrite française, de manière relativement efficace. Tout d'abord, les archives en ligne des rédactions se sont enrichies et permettent parfois la collecte des articles consacrés au génocide, à l'histoire du Rwanda avant 1994 ou aux retours qui se sont opérés sur l'événement durant la période post-génocide. Cependant, la diversité règne quant à la nature des matériaux accessibles (mode texte ou mode page, PDF océrisé ou page HTML . . .), quant à la qualité des moteurs de recherche proposés et quant à l'accessibilité des collections (l'accès complet aux archives en ligne étant généralement réservé aux clients abonnés). Pour les dépêches d'agences, la BnF propose par la plateforme multimédia AFP Forum un accès à l'ensemble des dépêches de l'AFP depuis 1994, un outil particulièrement efficace pour établir la chronologie de diffusion des dépêches, leur degré d'importance et leur contenu.

Cependant, l'outil désormais le plus utile est sans aucun doute la base en ligne Europresse qui offre une possibilité de recherche avancée parmi une collection de plus de 14.000 sources d'information (presse internationale, nationale, régionale, sites Web . . .), généralistes et spécialisées, françaises et étrangères. À ce stade, peu de rédactions autorisent un accès à l'année 1994, mais la base donne accès aux archives des grands quotidiens français comme *Le Monde* (depuis le 19 décembre 1944), *Libération* (2 janvier 1995), *La Croix* (1er septembre

1995), *Le Figaro* (31 octobre 1996) ou *L'Humanité* (16 novembre 1999) de même qu'à celles de certains hebdomadaires comme *L'Express* (depuis le 7 janvier 1993) ou *Le Point* (7 janvier 1995)[13]. La base offre aussi un accès à l'AFP (depuis le 18 mars 2001 pour les informations françaises et mondiales), à certains titres de la presse étrangère (*Le Temps*, *Le New York Times* . . .) et à certains titres de presse en ligne (*Rue89* d'avril 2010 à janvier 2017 par exemple). La recherche avancée a l'avantage d'autoriser les recherches en plein texte d'un nom, d'un lieu, d'un acteur sur une période précise, sur un ou plusieurs titres sélectionnés. Le matériel collecté est alors soit une page HTML seule, soit une page HTML associée au PDF de la page du journal. Des exports sont possibles aux formats PDF ou RIS, de même que l'export des métadonnées de chaque document dans Zotero. Le potentiel de cette plateforme reste en revanche très limité en termes de traitement quantitatif et elle a pour inconvénient majeur de ne pas offrir le même type de collecte pour tous les titres, certains proposant le PDF des versions imprimées, d'autres une simple page HTML. De ce fait, le chercheur souhaitant se constituer un corpus homogène pour travailler sur l'iconographie ou intégrer à ses analyses l'étude de la mise en page continuera de préférer se rendre en centre de documentation pour récupérer la page complète.

Pour ce qui est de la presse étrangère, Europresse ne permet pas d'accès aux collections en 1994 mais il est possible de travailler sur *Le Temps* (depuis le 27 mars 1998), *The New York Times* (1er janvier 2000), *Le Soir* (1er avril 2000), *The Daily Telegraph* (25 avril 2007), *la Libre Belgique* (29 août 2016). Pour accéder aux contenus de la presse étrangère de 1994, certaines bases comme ProQuest, Factiva ou Nexis peuvent être utiles tout comme les plateformes numériques de la British Library[14], de la National Library of Australia[15] ou de certains États ou Universités américaines comme la California Digital Newspaper Collection financée par l'U.S Institute of Museum and Library Services[16].

13 8.095.589 références au 23 mai 2020 depuis la base accessible sur la plateforme de la BU UVSQ Paris-Saclay (https://www.bib.uvsq.fr/bibliotheque-numerique). Sur les conditions d'accès : www.europresse.com/fr/.

14 Principalement des journaux régionaux et locaux publiés depuis 1800 avec, pour la période qui nous intéresse, des titres comme *Aberdeen Press and Journal* (né en 1747 !), *Newcastle Evening Chronicle* (1858), *Liverpool Echo* (1879). Pour l'accès aux collections en ligne de la British Library : https://www.bl.uk/collection-guides/newspapers#.

15 Avec un certain nombre de titres nationaux comme *Le Courrier australien* (1892) ou *The Canberra Times* (1926). Pour l'accès aux collections de la National Library of Australia : https://www.nla.gov.au/what-we-collect/newspapers.

16 California Digital Newspaper Collection : https://cdnc.ucr.edu/.

4 Écrire l'histoire du génocide par et avec la presse écrite numérisée : défis et perspectives

Plusieurs collections en ligne proposent donc de la presse rwandaise numérisée, mais celles-ci restent morcelées, lacunaires et non océrisées. Cette situation, peu confortable pour le chercheur, ne permet pas, pour le moment, un recours systématique à ce type de ressources dans les projets d'enquête scientifique. Elle invite en revanche à interroger la pertinence d'un effort collectif massif qui viserait à encourager les usages des archives de presse numérique dans l'écriture de l'histoire du génocide et plus largement du Rwanda[17].

4.1 Vers un projet de numérisation de la presse rwandaise ?

Au-delà de son travail sur la cartographie des sources du génocide des Tutsi, le projet RwandaMAP questionne les dynamiques passées, présentes et futures de la recherche sur le Rwanda. Dans ce cadre, une réflexion a été ouverte, avec certains acteurs publics et privés français et rwandais, sur la faisabilité et sur les enjeux du déploiement d'un vaste programme de numérisation de la presse rwandaise. La préciosité d'un tel projet ressort de ce questionnement, tout comme l'existence de difficultés importantes[18].

Le premier apport d'une telle numérisation serait d'assurer la conservation de collections papier le plus souvent stockées à l'air libre – parfois dans des boîtes Cauchard, avec des températures et une humidité relative peu favorables à la préservation du papier. Un tel projet aurait pour but de reconstituer des collections complètes de titres de presse et de les rendre accessibles aux chercheurs aujourd'hui confrontés à la dispersion des collections papier. Au regard des « numérisations sauvages » effectuées par les chercheurs[19] ou de certaines

17 Cette troisième partie se concentre sur la presse rwandaise. Il va de soi qu'un plus large accès aux collections de presse française post-1948 serait aussi souhaitable. Une telle réflexion dépasse cependant de loin le cadre de cet article. Sur les enjeux actuels de la numérisation des collections de presse : (Blandin et Garcin-Marrou 2018). Voir également le site du projet Numapresse (http://www.numapresse.org/).

18 Le recours aux archives numérisées et, plus largement, les évolutions des pratiques de recherche liées aux Humanités numériques ont fait l'objet de plusieurs temps de réflexion lors des workshops RwandaMAP2020 à Paris, le 20 avril 2018, puis à Bruxelles le 21 septembre 2018.

19 Les historiens que nous sommes doivent alors « bricoler » leurs propres collections de presse en photographiant les pages des éditions, en compilant dans un même fichier les pho-

numérisations de presse écrite évoquées précédemment, les enjeux de traçabilité devraient être considérés comme prioritaires. Les métadonnées proposées par les archives de presse numériques actuellement disponibles permettent en effet, au mieux, de connaître le numéro du journal, la date de parution, la langue et le nombre de pages. Les informations concernant la provenance de la version papier, le lieu d'édition, le format, la période de parution, l'origine de la collection, le rédacteur en chef, l'éditeur, le dépositaire, le nom des contributeurs, l'imprimeur, le nombre d'exemplaires imprimés sont généralement manquantes ce qui gêne l'opération de contextualisation des sources.

Les enjeux d'accessibilité sont tout aussi importants que ceux de conservation ou de traçabilité. Pour les chercheurs, la numérisation des collections de presse écrite rwandaise réduirait d'abord les processus de collecte actuels, coûteux en temps, en déplacements et en argent et permettrait aux chercheurs de concentrer leurs recherches sur des sources plus rares ou réellement difficiles d'accès. L'intérêt serait donc aussi de rendre ces collections accessibles aux non-chercheurs et, particulièrement, au public rwandais, au Rwanda ou en diaspora, ainsi qu'à celles et ceux qui enquêtent sur le sujet (journalistes, militants, juristes). Se pose ainsi la question des droits pour des titres qui peuvent être récents, parfois encore en activité, avec une nécessaire réflexion à conduire sur la création d'une plateforme permettant des accès différenciés selon l'ouverture des droits et les catégories de publics. L'enjeu des langues est tout aussi crucial et la mise en ligne de collections en kinyarwanda, tout autant que celles en anglais ou français, devrait être considérée comme une priorité. Cela rendrait ainsi possible le travail d'étudiants rwandais (généralement anglophones aujourd'hui) ainsi que la possibilité de développement d'usages nouveaux, hors des seules études sur le génocide.

4.2 Quel impact sur les usages historiens et académiques ?

Une numérisation des collections de presse rwandaise, fidèle au support original, sans sélection, et avec reconnaissance de caractère contribuerait à modifier les usages de la presse écrite par les historiens et par les académiques. D'abord, l'océrisation ouvre la voie à la recherche en plein texte et aux pratiques de fouilles

tographies de pages pour reconstituer l'édition complète puis en indexant correctement ces éditions dans un logiciel de gestion de références et de données comme Zotero ou Tropy (pour les JPG).

de données[20]. Bien sûr, le processus d'océrisation doit être considéré avec précaution, en prenant en compte le taux d'erreur de reconnaissance automatique (Milligan 2013). Une recherche textuelle de noms, auteurs ou encore de dates est cependant très largement facilitée. Cela encouragerait sans nul doute les études chronologiques, biographiques ou prosopographiques sur l'histoire rwandaise. Elle garantirait une attention plus soutenue aux enjeux de vocabulaire et de sémantique, offrirait la capacité de croiser plus facilement les sources de presse entre-elles et permettrait une mobilisation plus aisée des traitements statistiques de données textuelles. Ce type de numérisation pourrait aussi faciliter les opérations de traduction automatique.

Une des difficultés majeures pour le déploiement d'un tel projet est la nécessité d'associer différents types d'expertise assez poussés. Nécessaire est d'abord l'expertise historienne apportant maîtrise du contexte rwandais, connaissance de l'histoire de la presse écrite rwandaise et attention aux informations cruciales à recenser pour tout usage scientifique de la presse écrite. Cette connaissance des informations utiles aux usages scientifiques des documents sera croisée aux compétences de l'archiviste qui travaillera à la mise en place d'un ensemble de métadonnées assurant la bonne traçabilité de chaque document. L'archiviste devra aussi penser en amont le type d'outils à utiliser pour numériser un très grand nombre de pages, de manière professionnelle, avec océrisation, de même qu'il devra envisager les formes d'accès et de valorisation utiles aux chercheurs comme au grand public.

Si un tel outil existait, il contribuerait sans nul doute à donner un nouvel élan aux recherches sur le génocide des Tutsi et sur le Rwanda. On pourrait par exemple s'attendre à ce qu'il encourage une réflexion plus large sur l'histoire des médias, avec des recherches possibles par titre de presse, par journaliste, par nature de production (officielle, extrémiste, d'opposition . . .). Il serait aussi possible de proposer une réflexion portant sur des catégories spécifiques : les éditoriaux, le courrier des lecteurs, les caricatures de presse. De façon plus générale, l'étude de nombreuses thématiques cruciales pour la décennie 1990 (la place des femmes dans la société rwandaise ; la question des réfugiés ; la coopération internationale ; les relations avec les pays voisins ; la lutte contre le Sida) se verraient facilitées.

20 Pour obtenir une indexation plus fine, des opérations de traduction automatique ou des fouilles de texte à partir de l'indexation des contenus, les outils de traitement automatique des langues sont nécessaires : (Poibeau 2014).

4.3 Défis et difficultés

Un tel projet soulève en revanche de nombreuses difficultés scientifiques, juridiques voire politiques et éthiques. Au vu du contexte particulier de guerre civile tout au long des années 1990, la mise en ligne pose un certain nombre de questions, qui sont peu soulevées dans le cas rwandais. La recherche par mots-clés sur des moteurs de recherche facilite la collecte d'informations sur des individus précis. Depuis le début des années 2000, la numérisation et « redivulgation » en ligne d'articles de presse écrite a pourtant conduit à des débats juridiques sur l'équilibre entre le droit à l'information et le droit à l'oubli (ou au déréférencement). Cette situation pose un enjeu éthique pour les projets de mise en ligne ainsi que pour les chercheurs : « If those creating digital archives do not always recognize the potential for harm then this must surely place greater responsibility upon the researcher to carefully consider whether their disclosure of content has the potential to cause personal harm » (Crossen-White 2015, 114).

Cette situation est d'autant plus particulière que la presse rwandaise des années 1990 est marquée par des publications de nature très diverses. Si des journalistes proposent des articles d'analyse ou factuels, de nombreux écrits racistes furent publiés. Après 1994, on retrouve aussi de violentes accusations à l'encontre d'individus soupçonnés d'avoir participé aux massacres. Certains articles étaient cependant de nature diffamatoire, ce qui fit d'ailleurs l'objet de multiples controverses. La mise en ligne de ces journaux impliquerait nécessairement une explication claire du contexte de production de la presse de la décennie 1990 et, au vu de la dimension récente des évènements, la redivulgation s'accompagnerait inévitablement de risques judiciaires. La question se pose aussi pour les illustrations dans la presse, particulièrement violente à cette époque avec des photographies de cadavres ou encore de rescapés du génocide mutilés.

Notons aussi que la question du droit d'auteur a jusqu'ici été très largement éludée des débats sur la mise en ligne d'archives et documentation sur le génocide des Tutsi. La reconnaissance du génocide a toujours été considérée comme prioritaire face au droit d'auteur, la mise en ligne de documents ayant été réalisée par de multiples institutions (ONG, institutions rwandaises, Tribunal pénal international pour le Rwanda). La professionnalisation croissante des institutions rwandaises concernant les questions patrimoniales s'est accompagnée au cours des dernières années de nouvelles législations sur le patrimoine et la propriété intellectuelle. Les conséquences sur les projets de numérisation et de mise en ligne n'ont pas encore été à ce jour bien évaluées.

Une autre question délicate est la place qui sera conférée aux chercheurs dans ce processus et la nature des acteurs qui pourraient le développer. Doit-il s'agir d'une initiative publique de l'État rwandais ou d'initiatives privées ? Comment

concilier au mieux les compétences et les logiques respectives de chaque acteur ? Quels seraient les usages attendus de ces archives par les acteurs non scientifiques et quels enjeux soulèvent pour eux l'accessibilité à de telles collections ? S'il n'est pas du rôle des chercheurs de se substituer aux institutions rwandaises publiques en charge des questions archivistiques, de tels projets supposent l'implication des chercheurs qui auront l'usage de ce type d'outils ainsi que le support d'un ensemble de disciplines associées traditionnellement à la gestion de l'archive et/ou aux sciences du patrimoine[21]. L'identification des acteurs qui disposent de la compétence professionnelle sur l'ensemble de la chaîne – préparation des documents, numérisation, gestion des outils numériques – pose des problèmes de compétences, de financements et de formation qui ne peuvent être dépassés que par le montage de collaborations internationales ambitieuses.

Si numérisation et mise en ligne facilitaient le travail « hors-site » des chercheurs non rwandais, comment éviter cependant qu'elle ne s'accompagne d'une séparation encore plus grande entre les universitaires et la « communauté » rwandaise de recherche (archivistes, bibliothécaires), qui jouent un rôle majeur dans la mise en sens des sources (Abel 2013). Cette question est d'autant plus importante dans une situation de déséquilibre international de la production universitaire sur le Rwanda, en contexte postcolonial.

Enfin, un tel projet entrainerait inévitablement des effets de sources. Parmi les effets qu'il est d'ores et déjà possible d'anticiper, une presse rwandaise aisément accessible pourrait renforcer les déséquilibres de la recherche à l'échelle régionale, les presses burundaises ou congolaises n'étant pas d'un accès aisé. De même, il faudrait veiller à produire un effort conséquent de numérisation sur la presse des quatre périodes identifiées dans notre première partie afin d'éviter une focalisation de l'attention sur les périodes les plus contemporaines, au détriment par exemple de la période coloniale ou de la période 1962–1990. Donner une forte accessibilité à la presse rwandaise nécessite en outre de rester attentif au risque d'une recherche « hors-sol », déconnectée des sources rwandaises et de la littérature grise non numérisée. Il faut enfin souligner que l'accessibilité donnée à ce type de sources génèrera un fort déséquilibre avec des médias audiovisuels rwandais doublement sous-estimés (problème d'accès

21 De nombreuses institutions, conscientes de la valeur patrimoniale de la presse ancienne ou de certains titres rares, ont entrepris des projets de numérisation de leurs collections. Cette valeur patrimoniale tend aussi à être reconnue pour la presse en langue étrangère. Ainsi, la Fondation des Sciences du Patrimoine a soutenu le lancement du réseau transnational pour l'étude de la presse en langues étrangères (Transfopress). Celui-ci recense sur son site plusieurs banques de données numériques comme celle de la presse francophone d'Égypte (Centre d'Études Alexandrines): https://transfopresschcsc.wixsite.com/transfopress/services3.

aux archives audiovisuelles et problématique linguistique), alors même qu'il s'agit de médias qui avaient un public bien plus grand (Réra 2016).

5 Conclusion

Les collections de presse offrent, au final, une documentation inégale, difficile à exploiter, parfois trompeuse. Concernant l'histoire du Rwanda, elles forment un paysage dispersé avec une grande diversité de productions que l'historien doit traquer afin de parvenir à reconstituer des collections complètes et complémentaires. Il se trouve alors à devoir gérer une profusion de documents dont la traçabilité n'est pas toujours d'une grande lisibilité, ce qui pose quelques soucis au moment du périmétrage des corpus.

Malgré ses limites et les effets de source qu'elle peut générer, la numérisation des archives de presse rwandaises et françaises des années 1990 engendrerait des gains de temps conséquents pour les chercheurs d'aujourd'hui et de demain et elle leur permettrait de mobiliser plus aisément des approches ambitieuses (transmédiatiques ; démarches micro-macro ; croisements de fonds de différentes natures ; fouille de données) que la gestion de gros corpus non numérisés peut rendre délicates, voire irréalisables.

Rendre plus largement accessibles de telles archives pourrait aussi permettre des formes de médiations plus poussées sur les spécificités des archives de presse, trop souvent prises pour des reflets de la réalité, alors même qu'il s'agit sans doute d'archives qui doivent inviter le chercheur à une très grande prudence, tant il reste souvent délicat de déterminer leurs modalités et logiques de productions. À cette valeur documentaire réelle s'ajoute en outre une valeur testimoniale, mémorielle, voire patrimoniale qui donne du sens à l'accès renforcé de ces archives à un vaste public.

Bibliographie

Abel, Richard. « The Pleasures and Perils of Big Data in Digitized Newspapers ». *Film History*, vol. 25, n° 1–2, 2013, p. 1–10.

Albert, Pierre, et Nathalie Sonnac. *La presse française. Au défi du numérique*. La Documentation Française, 2014.

Bart, Annie. « Annuaire de la presse écrite rwandaise ». *Etudes rwandaises*, vol. 13, n° 2, 1980, p. 49–55.

Bart, Annie. *La presse au Rwanda; production, diffusion et lecture depuis le début du 20ème siècle*. Université de Bordeaux III, 1982.

Bart, Annie. « L'aventure d'une revue d'élites (1967–1992) ». *Dialogue*, n° 153, 1992, p. 13–26.
Bertrand, Jordane. *Rwanda, le piège de l'histoire: l'opposition démocratique avant le génocide, 1990–1994*. Karthala, 2000.
Blandin, Claire, et Isabelle Garcin-Marrou. « En quête d'archives. Bricolages méthodologiques en terrains médiatiques ». *Le temps long des archives de presse*, édité par Sarah Lécossais et Nelly Quemener, INA Editions, 2018, p. 43–50.
Chrétien, Jean-Pierre. « Presse libre et propagande raciste au Rwanda : Kangura et les 10 commandements du Hutu ». *Politique africaine*, n° 42, juin 1991, p. 109–120.
Chrétien, Jean-Pierre, éditeur. *Rwanda: les médias du génocide*. Karthala, 1995.
Clavert, Frédéric, et Valérie Schafer. « Chercher sur, autour et avec le numérique ». *Penser l'histoire des médias*, édité par Claire Blandin et al., CNRS Editions, 2018, p. 233–241.
Cristofori, Silvia. « A Kigali e momentaneamente in Belgio. "Dialogue": fran nuova idendità nazionale ed etnismo negazionista ». *Rwanda: etnografie del post-genocidio*, édité par Michela Fusaschi, Meltemi, 2009, p. 135–155.
Crossen-White, Holly L. « Using Digital Archives in Historical Research: What Are the Ethical Concerns for a 'Forgotten' Individual? » *Research Ethics*, vol. 11, n° 2, juin 2015, p. 108–19.
Dakhlia, Jamil, et François Robinet. « Médias d'Afrique subsaharienne : histoire, pouvoirs et mémoires. Entretien avec Annie Lenoble-Bart ». *Le Temps des médias*, n° 26, février 2016, p. 253–65.
Delporte, Christian, et al. *Histoire de la presse en France XX-XXIe siècles*. Armand Colin, 2016.
Des Forges, Alison. *Aucun témoin ne doit survivre: le génocide au Rwanda*. Karthala, 1999.
Dumas, Hélène. *Le Génocide au village. Le massacre des Tutsi au Rwanda*. Le Seuil, 2014.
Epelbaum, Didier. *Pas un mot, pas une ligne? 1944–1994? Des camps de la mort au génocide rwandais*. Stock, 2005.
Frère, Marie-Soleil. « Mutations de l'espace journalistique rwandais: les multiples facettes d'un système médiatique "post-génocide" ». *Rwanda 1994–2014. Récits, constructions mémorielles et écritures de l'histoire*, édité par Virginie Brinker et al., Les Presses du réel, 2017, p. 203–222.
Gasarabwe, Oscar. *Le génocide des tutsi du Rwanda et l'assassinat des opposants politiques vus à travers la presse belge et française 07. 04.1994–07.04.1995, Revue de presse en 5 volumes*. ASBL Bene Gihanga, 1996.
Guichaoua, André, éditeur. *Les crises politiques au Burundi et au Rwanda (1993–1994). Analyses, faits et documents*. Karthala, 1995.
Hunter, Emma. « Newspapers as Sources for African History ». *Oxford Research Encyclopedia of African History*, avril 2018, p. 1–34.
Kayser, Jacques. « L'historien et la presse ». *Revue historique*, vol. 2, n°218, 1957, p. 284–309.
Korman, Rémi. « L'État rwandais et la mémoire du génocide: Commémorer sur les ruines (1994–1996) ». *Vingtième Siècle. Revue d'histoire*, n°122, 2014, p. 87–98.
Lenoble-Bart, Annie. « Cinquante ans de presse catholique rwandaise (1933–1983) ». *Diffusion et acculturation du christianisme (XIXe-XXe s.)*, édité par Jean Comby, Karthala, 2005, p. 471–487.
Milligan, Ian. « Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010 ». *Canadian Historical Review*, vol. 94, n° 4, décembre 2013, p. 540–569.

Munyakayanza, Jean-François. *La presse catholique et son rôle dans la vie politique et sociale du Rwanda (1931–1961)*. Paris 1 Panthéon Sorbonne, 2013.

Poibeau, Thierry. « Le traitement automatique des langues pour les sciences sociales. Quelques éléments de réflexion à partir d'expériences récentes », *Réseaux*, vol. 188, n°6, 2014, pp. 25–51.

Pottier, Johan. *Re-Imagining Rwanda: Conflict, Survival and Disinformation in the Late Twentieth Century*. Cambridge University Press, 2002.

Prunier, Gérard. *The Rwanda Crisis: History of a Genocide*. Hurst & Co, 1995.

Réra, Nathan. « Les chantiers de la mémoire. Quelles archives audiovisuelles pour le Rwanda? » *Cahiers du CAP*, vol. 3, avril 2016, p. 199–229.

Reyntjens, Filip. *L'Afrique des grands lacs en crise. Rwanda, Burundi, 1988–1994*. Karthala, 1994.

Robinet, François. *Silences et récits. Les médias français à l'épreuve des conflits africains (1994–2015)*. INA Editions, 2016.

Rovetta, Ornella. *Un génocide au tribunal: le Rwanda et la justice internationale*. Belin, 2019.

Soulet, Jean-François. *L'histoire immédiate. Historiographie, sources et méthodes*. Armand Colin, 2012.

Trénard, Louis. « Chronique : Histoire et Presse ». *Revue du Nord*, vol. 54, n° 212, mars 1972, p. 69–90.

Unearthing New Artefacts: Digital Reshaping of Newspapers

Pierre-Carl Langlais

Classified News

Revisiting the History of Newspaper Genre with Supervised Models

Abstract: Automatic text classification is increasingly being used to explore and analyze the rapidly growing collections of digitized cultural heritage archives. *Numapresse*, a digital humanities project devoted to the historical study of French-speaking newspapers from 1800, has trained models to recognize major newspaper genres from political news to sports section or serial novels. The key output of this program has been the automated classification of all the major French dailies of the Interwar period, from 1920 to 1939, thanks to the comprehensive digitization of the French national library for this period. The first part of this paper presents a modeling strategy grounded in the perspective of cultural history and literary analysis exemplified by the building of historical-based models (spanning 20 years) and the reconceptualization of classification probabilities into potential tools to study intertextual discourses and genre hybridization. The second part showcases an exploration of the output data generated by the model through a method of *zoom reading*. The varying extent of newspaper genres produces regular patterns at different time scale such as weekly cycles based on thematic supplements, year cycles conditioned by large-scale cultural and social practices and decades trends displaying a long-term history of genre. Our conclusion stresses the promises of model transferability to build a new ecosystem of model reuse for research communities and libraries involved with large collection of cultural heritage archives.

Keywords: newspaper genre, digitized newspapers, supervised text classification, distant reading, literary analysis

1 Introduction

Research on French-speaking newspaper history has taken a textual and cultural turn at the start of the 21st century. Until the start of massive digitization programs, the history of the press was mostly a history of newspapers as political, economic and social actors that relied mostly on external archives: this approach is typically exemplified by the *Histoire générale de la presse française*

https://doi.org/10.1515/9783110729214-010

(1969–1978).[1] A new wave of literary, media and historical studies has increasingly focused on the newspaper as a literary, semiotic and editorial object, and on its cultural and social impact.[2] Consequently newspaper genre has become a major field of research. For the last 20 years numerous newspaper genres, sections and rubrics have been historicized, from the serial novel to courtroom debates and financial columns.[3] This dynamic of research has been expanded to non-textual content as well, such as illustrations.[4]

This research has been supported by the parallel development of a new research infrastructure: newspaper archives have been massively digitized. Beyond the remote accessibility of newspaper archives, full-text search greatly facilitates the search of textual forms and transforms historical research practices.[5] Yet, the very large scale of newspaper corpora makes it hard to get a structural outlook on the codification of newspaper genres. To get a perfect view of the global shift of media culture it would be 'absolutely necessary to read everything (...) and immerse oneself in a textual ocean to really perceive the cultural and writing dynamics'.[6] For now this total immersion has only been achieved on the small scale of specific case studies, covering, for instance, the interna-

1 Claude Bellanger et al., eds. (1969). *Histoire générale de la presse française. Tome II, De 1815 à 1871 / publiée sous la direction de Claude Bellanger, Jacques Godechot, PierreGuiral [et al.]* Presses Universitaires de France.

2 Dominique Kalifa et al., eds. (2011). *La civilisation du journal: Histoire culturelle etlittéraire de la presse au XIXè siècle*. Paris: Nouveau Monde Editions.

3 Corinne Saminadayar-Perrin (2007). *Les discours du journal: rhétorique et médias au XIXe siècle (1836–1885)*. Université de Saint-Etienne; Marie-eve Thérenty (2007). *La Littérature au quotidien. Poétiques journalistiques au XIXe siècle*. Français. Paris: Le Seuil; Adeline Wrona and Yves Jeanneret (2012). *Face au portrait: De Sainte-Beuve à Facebook*. Hermann. Paris: Hermann; Amélie Chabrier (2019). *Genres du prétoire: Lamédiatisation des procès au XIXe siècle*. French. Paris: Mare Martin; Pierre-Carl Langlais (2016). "La formation de la chronique boursière dans la presse quotidienne française (1801–1870)." PhD thesis. Université Paris 4 Sorbonne. URL: https://tel.archivesouvertes. fr/tel-01424740/document; Mélody Simard-Houde (2018). *Le reporter et sesfictions: Poétique historique d'un imaginaire*. Français. Presses Universitaires de Limogeset du Limousin.

4 Jean-Pierre Bacot (2005). *La Presse illustrée au XIXe siècle: Une histoire oubliée*. Limoges: Presses Universitaires de Limoges et du Limousin; Amélie Chabrier, Marie-Astrid Charlier, and Paul Aron, eds. (2018). *Coups de griffe, prises de bec: la satire dans la presse des années trente*. Bruxelles: Impressions Nouvelles.

5 James Mussell (2012). *The Nineteenth-Century Press in the Digital Age*. Springer.

6 Dominique Kalifa et al., eds. (2011). *La civilisation du journal: Histoire culturelle et littéraire de la presse au XIXè siècle*. Paris: Nouveau Monde Editions, p. 19.

tional circulation of well-defined sub-genres (like the *Mystères urbains*).[7] More ambitious programs would require new methods.

The textual turn of French-speaking newspaper studies has been recently extended by a global digital turn. At the beginning of the 2010s, innovative projects attempted to explore new technical and methodological opportunities made possible by the digitization of newspaper archives in the wider context of digital humanities and computational social sciences.[8] This emerging landscape started with *Europeana Newspaper* and includes the American project *ViralText*, the Swiss-Luxembourg project *Impresso – Media Monitoring of the Past*, the European project *Newseye*, the international initiative *Oceanic Exchanges* and, in France and Belgium, the (ANR) project *Numapresse*.[9] Each of these projects focuses on different technical and historical objectives. They aim to reconstruct more abstract metadata and historical knowledge beyond full-text search, such as the identification of newspaper editorial structures, the detection of wide networks of reprinted news, or the extraction and recognition of past persons, organizations and events.

Building up on previous qualitative research in a French-speaking context, *Numapresse* has largely focused on automated classification of newspaper genres. Several structural changes have recently created a favorable technical and scientific environment for large-scale genre classification. The implementation of open data and open content policy in digital libraries has made it possible to extract and analyze millions of cultural heritage documents. Text mining technologies have also become more readily available in the form of open-source software. Established modeling strategies used for several decades in industrial settings have been popularized thanks to their integration in accessible programming languages (like R and Python) and fitted to new purposes like literary research. In the United States, collaborations between libraries and researchers in computational humanities and cultural analytics has resulted in unprecedented classification projects. For instance, the *Page-Level Genre Metadata for English-Language Volumes in*

7 Marie-Ève Thérenty and Dominique Kalifa (2016). *Les Mystères urbains au XIXe siècle: Circulations, transferts, appropriations*. Médias19.

8 James Mussell (2012). *The Nineteenth-Century Press in the Digital Age*. Springer; Bob Nicholson (2013). "The Digital Turn." In: *Media History* 19.1, pp. 59–73. DOI: 10.1080/13688804.2012.752963; Dallas Liddle (2012). "Reflections on 20,000 Victorian Newspapers: 'Distant Reading' The Times using The Times Digital Archive." In: *Journal of Victorian Culture* 17.2, pp. 230–237. DOI: 10.1080/13555502.2012.683151.

9 See: http://www.europeana-newspapers.eu/, https://viraltexts.org/, https://impresso-project.ch/, https://www.newseye.eu/, https://oceanicexchanges.org/, http://www.numapresse.org/. Last accessed 2022-10-01.

HathiTrust, 1700–1922 dataset[10] contains the classification probabilities of each page of 854,476 volumes to belong to fiction prose, nonfiction prose, poetry, drama and paratext (such as ads or table of contents): fine-grained classification makes it possible to find incidental fictions as part of a periodic publication.

Beyond the practical utility of generated metadata for bibliographic research, theses new approaches have also contributed to the critical analysis of genre development and genre patterns. Model results can be "hacked" to indicate the strength of a classification. In *The Life Cycles of Genres*, Ted Underwood introduced a new experimental design to track the codification process of Science fiction, Detective fiction and Gothic fiction, using the probabilities of classification for a large corpus of thousands of novels published over two centuries (from the 1800s onward).[11]

This article presents the preliminary results of a large-scale classification project led by the *Numapresse* project on an all the daily national newspapers digitized by the French National Library from 1800 to 1945. While the digitization process remains a work in progress, the current corpus includes a significant share of the total number of national newspapers published. The Interwar period (1919–1939) is especially extensive and covers nearly half of all the national daily newspapers (roughly 25 out of 50) and all the major titles.

The *Numapresse* project has collected the available digitized archives in METS/ALTO format of 46 daily newspapers (3 million pages). These documents contain not only the raw text but also numerous supplementary metadata, such as the coordinates of the textual, visual or tabular objects in the original page and, for textual elements, the font size and the Optical Character Recognition (OCR) confidence rates. Figure 1 gives an overview of the time distribution of the titles imported by *Numapresse* from the French National Library. The *Numapresse* corpus comprises additional collections of weekly magazines and regional and foreign newspapers not displayed here, as they are not the focus of this study.

This article could be defined as a model paper: it aims primarily to describe a model in the same way as a data paper documents a dataset. As such it is neither a research paper in the humanities and social sciences (since results will not be our primary focus) nor a standard technical paper in the computer science (the modeling relies ultimately on established technologies). The main contribution will lie in-between: in the interaction between a corpus, a modeling infrastructure and a reconceptualization process transforming the model outputs into qualitative

10 https://figshare.com/articles/Page_Level_Genre_Metadata_for_English_Language_Volumes_in_HathiTrust_1700_1922/1279201.

11 Ted Underwood (2019a). *Distant Horizons: Digital Evidence and Literary* Change. Google-Books-ID: fQo5uwEACAAJ. University of Chicago Press.

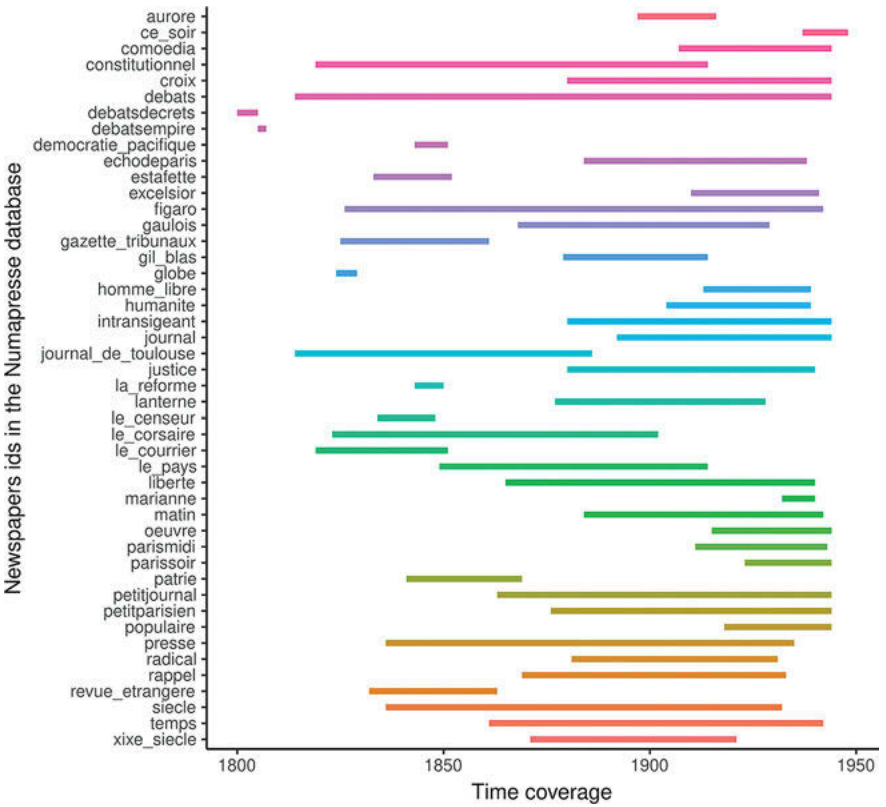


Fig. 1: Current holdings of daily newspapers in the Numapresse project (1800–1945).

interpretations. We aim to provide a reproducible modeling strategy for the identification of newspaper genre. Our interdisciplinary framework attempts to conciliate programming and statistical techniques with specific concepts and issues raised by newspaper historians and studies of journalistic discourses.

In comparison with other corpora, newspapers are better fitted for this approach, since news writing is highly standardized. The heterogeneous nature of press content means that each component should be easily identifiable by the reader either through visual patterns (e.g., lines and segments crossing the newspaper page), metadata (e.g. title, signature, dispatch heading, captions) or by the vocabulary. Nevertheless, any operationalization attempt remain by definition limited and imperfect. Previous research has extensively demonstrated that writing standards are not fixed. Therefore, our model architecture should incorporate the uncertainty and fluidity of actual writing practices.

This paper is organized in two parts. The first part presents a supervised classification model trained on 20 newspaper issues from the interwar periods. It details the epistemological implications of our modeling architecture and a potential strategy to combine automated classification with the vocabulary and purposes of qualitative research. The second part develops a preliminary exploration of classification data and describes noticeable patterns such as timelines (week-based, season-based or global trends) and the mapping of editorial policies. Our conclusion provides broader suggestions for reproducing this experiment on other cultural heritage collections and on different content (such as images).

2 The Making of a Newspaper Genre Model

Automated classification of news is an old practice. Seminal document processing and clustering techniques have been first applied to news corpus.[12] In 1990, the Reuters news agency released one of the leading dataset for computational linguistics, Reuters-21578, a collection of 21578 news wire sent in 1987.[13] Throughout this long history of news classification, newspaper corpora were not a research object but a testing sample, that served to build more general model. The diversity of written styles made these resources arguably closer to contemporary spoken languages than other publications (like literature or law).

These early choices are still affecting contemporary practices of text mining and natural language processing (NLP). Reuters-21578 is still integrated today for demonstration purposes in major text mining applications like the R TM Framework. Even with the growing usage of web-based corpora coming from Wikipedia, Twitter or Reddit, most of the linguistic models used by the leading NLP application, Spacy, remain based on manually annotated news articles.[14]

In the strictest sense, a (supervised) model is the output of a machine learning algorithm that encodes a set of features and weights learned from input data annotated such as exemplifying a specific phenomenon, and that can be

12 Fazli Can and Esen A. Ozkarahan (Dec. 1990). "Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases." In: *ACM Transactions on Database Systems* 15.4, pp. 483–517. DOI: 10.1145/99935. 99938; Louise Guthrie, Elbert Walker, and Joe Guthrie (1994). "Document Classification by Machine: Theory andPractice." In: *Proceedings of the Conference on Computational Linguistics,* pp. 1059–1063.

13 https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.

14 For instance, the French corpus, which uses, among other sources, excerpts from a well-known regional newspaper, *L'Est Républicain*: https://spacy.io/models/fr.

applied on new data to infer the presence of this phenomenon. In accordance with a general trend in computational humanities we will use the word *model* in a broader meaning, that stresses not so much the final end product rather than the process needed to get there. A model encompasses a workflow, that is a series of actions and interpretations that can only be partly automated: the creation and curation of a training corpus, the definition of classes and labels, the adjustment of model parameters, as well as the production and reuse of classification outputs. All these steps are common to any supervised classification project although they may be more or less visible depending on professional contexts and academic fields. Recent research on *digital labor* and the sociology of machine learning has stressed that the preparation of training data is frequently outsourced to invisible work forces.[15] Similar processes occur in libraries and cultural heritage institutions in the wider context of archive digitization, transcription and annotation.[16]

In what follows, we detail and justify our modeling strategy, based on a supervised support vector machine (SVM) model trained with our custom extension in R, TidySupervise. We describe the creation of the training corpus and estimate the accuracy of the model. The last section is a more reflexive discussion of the use of modeling in the humanities and the social sciences.

2.1 A Supervised Approach

Classic text classification techniques are structured into two approaches: supervised and unsupervised. Supervised methods rely on a typical artificial intelligence process: the model is trained on an annotated corpus (or "ground truth") and evaluated on a test corpus. Unsupervised classification methods (or "clustering methods") produce automated classifications based on formal consistencies that have to be interpreted *ex-post*.

Classification projects in digital humanities have usually relied on unsupervised methods. In 2006, David J. Newman and Sharon Block published a seminal analysis of the probabilistic topic decomposition in a corpus of eighteenth-century

15 Trebor Scholz (2013). *Digital Labor: The Internet as Playground and Factory*. Routledge. 274 pp; Sorin Adam Matei, Nicolas Jullien, and Sean P. Goggins (2017). *Big Data Factories: Collaborative Approaches*. Springer. 141 pp; Catherine D'Ignazio and Lauren F. Klein (2020). *Data Feminism*. Cambridge, Massachusetts: The MIT Press. 328 pp.

16 Mathieu Andro (2018). *Digital Libraries and Crowdsourcing*. John Wiley & Sons. 234 pp.

American newspapers. They make a case for Latent Dirichlet Allocation (LDA), which was at the time a rather experimental technique. Generated topics from LDA are presented as more objective tools than qualitative reading or supervised models, since they seem to reveal pre-existent discursive constructs that are independent from subjective appreciation: 'there is no a prior designation of topics – in fact there are very few "knobs to turn" in the method – historians do not need to rely on fallible human indexing or their own preconceived identification of topics'.[17]

Additionally, unsupervised classification creates an easy path for interdisciplinary collaboration. The corpus analysis is made by an engineer and/or a statistician, and then the end results have to be analyzed and contextualized by an expert historian. This approach quickly disseminated in the nascent digital humanities circles. The excellent *Index of Digital Humanities Conference*[18] makes it easy to recover this circulation of LDA and other unsupervised classification techniques. Early experiments focused on authorship attribution and on the analysis of intertextuality based on the comparison of texts sharing similar topics.[19]

Renewed works in digital history and cultural analytics on the history of genres and other structural stylistic patterns have led to a resurgence of supervised models. Ted Underwood has pioneered the use of trained corpus to build controlled literary experiments by evaluating, for instance, the stylistic nature of literary prestige or the gradual development of codified genres like detective and science fiction.[20] While this methodology is more constraining than unsupervised topic modeling, it fits well with historical research on major cultural categories. Genres, for instance, are not simply a stylistic construct, recoverable in topic modeling clusters. They are also social definitions that are actively used in the

17 David J. Newman and Sharon Block (2006). "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." In: *Journal of the American Society for Information Science and Technology* 57.6, pp. 753–767. DOI: 10.1002/asi.20342, p. 766.

18 https://dh-abstracts.library.cmu.edu/.

19 Charles Cooney et al. (2008). "Hidden Roads and Twisted Paths: Intertextual Discovery Using Clusters, Classifi Cations, and Similarities." In: ADHO 2008 – Oulu; Matthew Jockers (2007). "Macro Analysis (2.0)." In: ADHO 2007 – Urbana-Champaign.

20 Ted Underwood (2015). "The Literary Uses of High-Dimensional Space." In: *Big Data & Society* 2.2, p. 2053951715602494. DOI: 10.1177/2053951715602494; Ted Underwood (2016). "The Life Cycles of Genres." In: *Cultural Analytics* 1.1. DOI: 10.22148/16.005; Ted Underwood (2019b). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

distribution and preservation of works.[21] In this context, it makes sense to not only look for the case where genres occur, but also where they do not occur although they might have happened to do so. Underwood introduced an influential framework of negative testing of genres models: classification results and classification probabilities are tweaked to test the *strength* of cultural categories.[22]

The classification of newspaper genres in *Numapresse* has been largely influenced by this supervised approach. It is based on SVM, an established technique developed in the 1990s which provided an appropriate compromise between accuracy, scalability and legibility.[23] While they rely on more complex matrix calculations than older methods (like logistic regression), SVM models remain reasonably transparent (especially in comparison with deep learning architectures). Documents are simply processed as bag-of-words, and each word within a class is associated to a fixed coefficient value. The matrix decomposition processes make the model more efficient than alternative strategies when there is a large number of potential coefficients to deal with (which is generally the case for texts: our term/document matrix usually includes more than 3000 occurrences). Yet, it is not necessary to know all these inner steps to have a general idea of how the model has been working and of the determining factors affecting the final output.

Following several reconceptualizations introduced by Underwood, we consider genre models as interdisciplinary objects, which entail several major semantic shifts:

- Any classification approximates historical cultural and discursive norms. Models operate so long as they rely on a series of regularities and patterns immersed in a specific textual or visual culture. Consequently, *contextual* models are usually preferable to *universal* models, that is models curtailed around a specific type of documents (in our case daily newspaper), a specific period (1920–1940 or 1840–1860) and, even, a specific society (Interwar French).
- Model categories should be informed by historical research and by a thorough *close reading* of primary sources. In practice, the definition of classes and labels and the annotation of a training corpus have to be done by researchers with expert knowledge, preferably involved in all stages of the classification work through an iterative process. The range of genres cannot

21 Professor Amy J. Devitt (2008). *Writing Genres*. Carbondale: Southern Illinois University Press. 268 pp.

22 Ted Underwood (2016). "The Life Cycles of Genres." In: *Cultural Analytics* 1.1. DOI: 10.22148/16.005.

23 TidySupervise is ultimately based on LIBSVM, a C library first developed in the early 2000s (Chang and Lin 2011).

be defined in advance and will generally be partly informed by the actual categorization. Any classification project in the humanities and social sciences should anticipate a *feedback* process where the preliminary results will recursively impact the set of objectives.
- Model accuracy is constrained by the spread and intensity of discursive norms. Highly codified and/or highly specialized writing practices are generally more predictable (as shown by the probability distribution in Fig. 5 p. 17)
- A text can have more than one category, or even no category (among the predefined set of categories). While writing norms can have strong standardization effects, notably in the context of cultural and media industries, they are not all-powerful. Genre hybridization and intertextuality are not an exception but a current practice that has actually contributed to the emergence of news genres.
- The classification output cannot be a binary judgment but a range of possibilities and probabilities, which can serve to locate multi-classification.

The *Numapresse* project has gradually developed its own modeling strategy to fit theses specific needs and objectives.

The classification relies on a custom application in R, TidySupervise.[24] Technically TidySupervise is just an overlay of the SVM program in C used by virtually every classification library in R or Python (and not rewritten since the early 2000s). The added value of the application lies in the simplification of the modeling process which has been redesigned in accordance with the principles of "tidy data". TidySupervise uses TF-IDF metrics by default rather than raw occurrences. This approach has the effect of downplaying highly frequent words such as function words (also called stop words). TF-IDF metrics marginally enhance the accuracy of modeling (usually in the order of 2–3%). More crucially, they make the model more legible: as seen in Fig. 3 (p. 14), the ranking of the word with the highest coefficient per classification is more meaningful and easier to interpret. Since the model is computed on TF-IDF scores there is no need to filter stop words in advance, since the prevalence of function words will be nearly null unless they can contribute to the classification. For instance in newspapers, first and second person pronouns occur preferably in certain contexts such as the serial novel.

TidySupervise strives to provide a more detailed output than raw classification. The default results are the probabilities of each class per document. This make it possible to deal with multiple possible classifications and intertextual

24 https://github.com/Numapresse/TidySupervise.

settings. Our usual retrieving process is therefore based on probabilities cutoff (more than 30% of probability or 50% depending on our methodological needs) rather than simply identifying the "best candidate" regardless of the underlying probabilities.

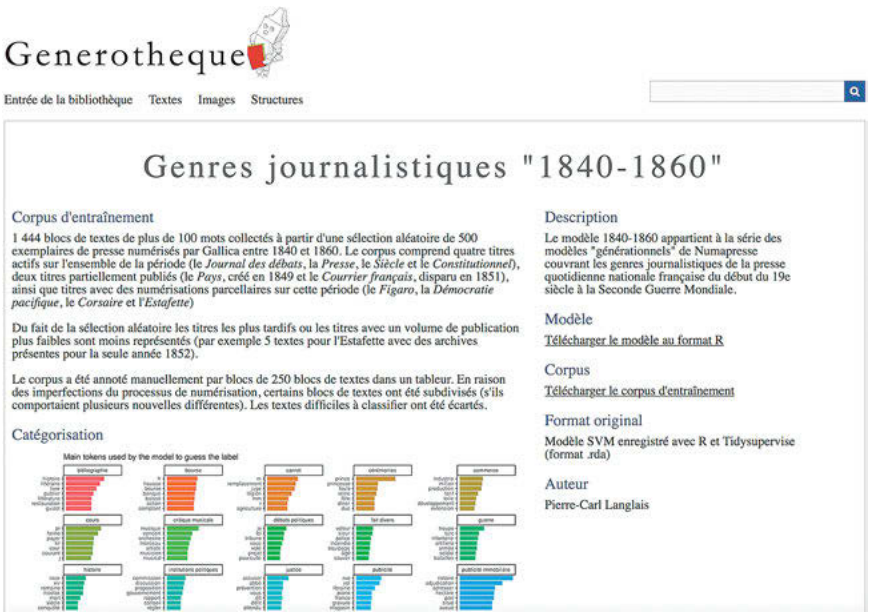


Fig. 2: An example of record in the library of models: the model of French daily newspapers published between 1840 and 1860.
Copyright: Numapresse project.

The models created with TidySupervise are shared on a new form of collaborative platform in digital humanities: a library of models called the *Générothèque* (a French portmanteau expression for *Library* (*Bibliothèque*) and *Genre*).[25] This concept is not new, *per se*: model sharing has become ubiquitous in the current AI environment given the fundamental role of model transfer in deep learning infrastructures. What distinguishes the *Générothèque* is the use of library-inspired designs as well as the enhanced documentation on training corpus and the definition of genres. These features will make AI more accessible and relevant to researchers working in the humanities and social sciences.

25 http://numapresse.org/generotheque/.

Figure 2 displays the record of the 1840–1860 model of French national daily newspapers.[26] The model itself is made available in the info box, alongside the training corpus and some basic metadata. The main section of the record details the selection process of the training corpus and justifies the categories used (by linking them, if possible, to already established historical news standards). Originating from the *Numapresse* project, the *Générothèque* is intended to become an autonomous project. Its repertoire will include, for example, models trained on literary genres or on historical scientific disciplines.

2.2 Defining the Genres

The *Numapresse* project aims to train supervised models for each 20-year period of the French daily press. This chronological approach roughly matches the global rate of development of news genres as well as several major breaks in the long-term history of French newspapers. For now, three models have been successfully trained and applied to the whole daily corpus of *Numapresse*. The models and their associated training corpus have been published on the *Générothèque*: the 1840–1860, the 1860–1880 and the 1920–1940 ("Interwar") models.[27] This final model was developed earliest, even before the creation of *TidySupervise*. It is by far the most used model in the various *Numapresse* projects, which is why it is the main focus of this article.

The Interwar model covers 20 different newspaper genres. It has been trained on 20 annotated issues from four major dailies of this time period, *Le Petit Parisien*, *Le Petit Journal*, *Le Matin* and *L'Intransigeant*. Each text block contained in the issue[28] has been assigned one genre, provided a non-ambiguous choice could be made (in some cases a text block belonged to several genres, was unclassifiable or was simply not relevant e.g., when there were too many OCR mistakes or when the "text block" was fixing different textual elements on the newspaper page). The final training corpus used by the model contains 2034 text blocks subdivided in 2729 text segments of at least one hundred words: longer text blocks with several hundreds of words were split into several segments.

26 http://numapresse.org/generotheque/items/show/4.

27 http://www.numapresse.org/generotheque/items.

28 A text block is the largest unit of digitization process (OCR): it usually matches any continuous textual sequences not interrupted by a newline like a paragraph, a title, or a row in a table. Of course this definition is mostly theoretical: the segmentation errors are frequent especially in archives digitized before 2013–2014. An article or a newspaper section is usually made of several text blocks.

Our subsequent model training was based on a different sampling strategy. The integration of the entire corpus of dailies digitized by the French National Library made it possible to retrieve a random corpus of text blocks with more than 100 words, rather than being focused on the entire content of single issues. The use of random selection mitigates potential biases of over-representation, especially for genres featured with a few long texts (which is typically the case for the serial novel).

The genres have been iteratively identified and redefined throughout the entire annotating process, building up on the historical sources and the preliminary results of the classification. For cultural historians and literary scholars, genres are never a straightforward category:

> Genres are ideal-typical discursive constructs. This means that textual manifestations do not always match the characteristics of these constructs perfectly. Deciding when articles that only partially comply with the textual characteristics of a certain genre can still be considered representative of that genre is challenging. In addition, articles might share characteristics with other genres and genres are dynamic constructions that change or fade away over time while new ones emerge.[29]

The particular nature of the sources significantly eased this ontological process: as the product of large-scale media industries with a massive circulation, leading national dailies are highly standardized, especially during the Interwar period. Just before the First World War, French newspaper had started to revolve on a turnover of weekly thematic supplements that included, for instance, sports sections on Monday, movie section on Friday or literary supplements in the weekend. This segmentation is largely similar across the entire newspaper ecosystem.

A careful examination of the preliminary results of the model revealed a few underlying genres that had a clear lexical identity although they were not properly identified in the newspaper. The large "advertising" genre was consequently cut off in two different genres: information ads (usually a formulaic announcement for a product or a price discount) and narrative ads (a potentially very elaborated story praising the effect of a product). These two genres are easily identifiable while reading the newspaper content. Bundling them in one large ad genre had the negative after-effect of *spreading* the lexical definition of ads, which had to cover a much larger range of discourse. All ad text blocks of the previous annotated issues have consequently been reassigned to the two

29 Marcel Broersma and Frank Harbers (2018). "Exploring Machine Learning to Study the Long-Term Transformation of News." In: *Digital Journalism* 6.9. Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2018.1513337, pp. 1150–1164. DOI: 10.1080/21670811.2018.1513337.

distinct ad genres. The modeling processing also had to cope with the available data. Several thematic supplements turned out to be too rare and specific to be included, such as lifestyle and travel articles, or music journalism.

The final selection comprises a mix of thematic genres (Sport, Movies, Theaters, Arts, Stock Exchange, Economics, Politics, International affairs ...), ad genres (informative ads, narrative ads and "*petites annonces*"), data structures (Theater programs, radio programs) and more stylistic genres which had nevertheless a strong semiotic identity in the newspaper's editorial structure (Serial Novel, "*faits divers*" and "*reportage*").

2.3 Evaluating the Model

The model is built upon a lexical *sketch* of each newspaper genre. It ranks 2756 common words according to their probability of belonging to each genre. Figure 3 shows the 10 likeliest words per genre. For most classifications, the results are not surprising. The movie sections focus on movies and stars, and the sports sections on games and championships. Stylistic genres yield more intriguing results: serial novels are defined primarily by their use of pronouns and objects from the daily life. In fact, there are no predefined topics for the novels published in the newspaper. Yet, they stand apart due to the presence of narratives in other sections of the newspaper, such as dialogues and first-person narratives.

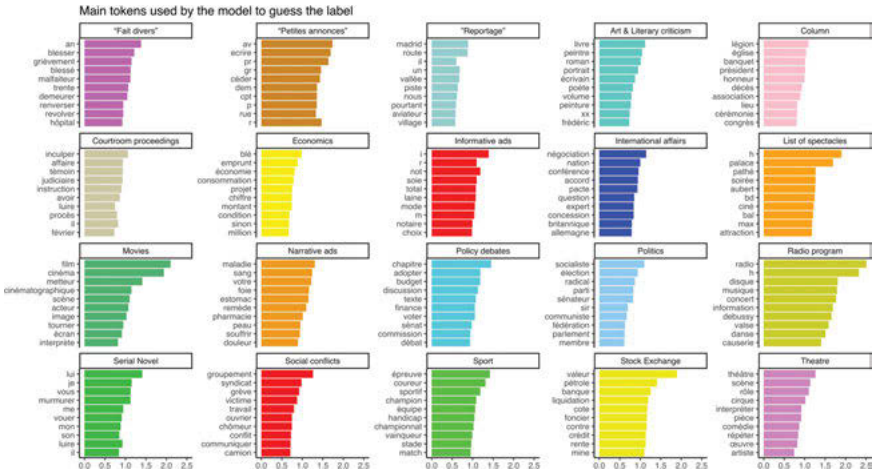


Fig. 3: Lexical sketch of the 20 genres in our Interwar model.

OCR mistakes are usually drowned in the ranking of significant words, and most often correspond to a long tail of hapaxes. They do occur in a few genres, such as ads or radio programs, where formulaic writing encourages the reiteration of a few OCR errors (like the "h" in radio programs, which is actually a digital misreading of the French preposition "à"). In addition, editorial layout can affect OCR quality: classified ads are often composed of tiny compressed letters, with each word paid for by advertisers, so newspapers have an incentive to create very dense pages with many words.

The model had a global accuracy of 70–75% on a random sample comprising 20% of the original corpus: roughly three out of four text blocks were correctly labeled. Obviously, as shown in Fig. 4 the quality of classification varies considerably across genres. Texts with a highly codified style such as classified ads ("*petites annonces*"), radio programs and show listings performed very well with F1 scores above 90. The model is also effective when the genre has a strong thematic or stylistic identity with highly specialized vocabulary or unusual syntactic constructions in the context of newspaper writing. Stock exchange sections, sports columns or, to a lesser extent, art and literary reviews are all characterized by the use of a specialized jargon. Serial novels use specific narrative forms.

Conversely, not all genres are easily defined by word counts. The French tradition of *reportage* proves rather elusive (F1 score < 50%). A *reportage* is not merely a genre but above all an enunciative situation: an event is narrated by a reporter which underlines his status of observer (hence the insistence in the lexical ranking on "looking" and being "there"):

> Reportage not only provides factual information, but also conveys the atmosphere and the experience of witnessing a certain event or issue, which can range from politics to war, sports and lifestyle.[30]

Unfortunately, our model has no contextual clues: articles are just processed as bags of words, so that only the overall themes and styles remain. Any semantic construct like, "I am writing from this place and I saw this", will be entirely diluted. Unsurprisingly, occasional *pastiche* exercises mimicking the phraseology of *reportages* will be highly rated. For instance, during a research program on the Second Italian-Ethiopian War of 1935, an historical retelling of a past historic event has been erroneously labeled as a reportage with a probability of 99%. While there was obviously no reporter on the field, the tone of the article is reminiscent of an actual *reportage*.

30 Marcel Broersma and Frank Harbers (2018). "Exploring Machine Learning to Study the Long-Term Transformation of News." In: *Digital Journalism* 6.9. Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2018.1513337, pp. 1150–1164. DOI: 10.1080/21670811.2018.1513337.

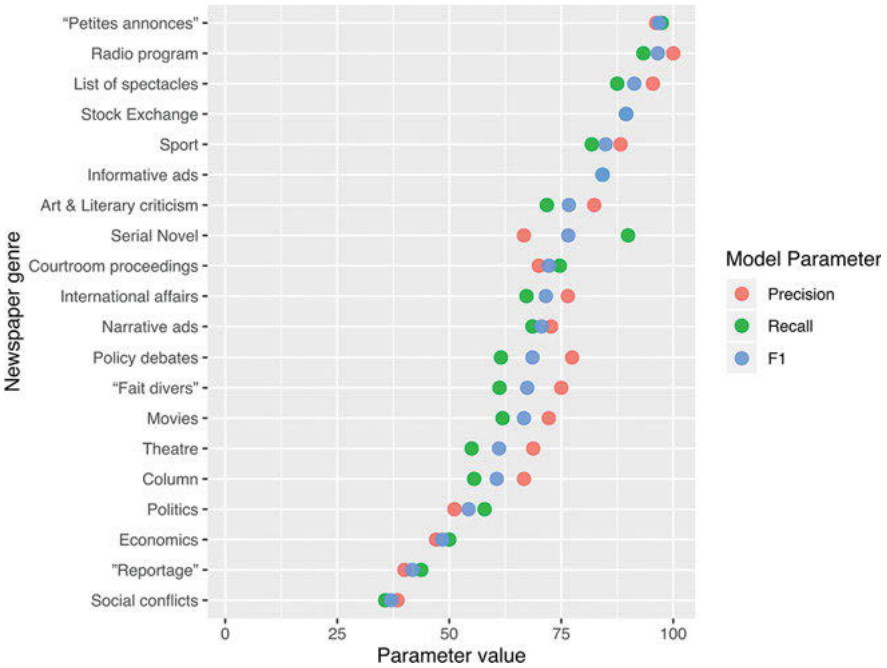


Fig. 4: Precision, recall and F1 score for the 20 genres in the model.

2.4 Model Hermeneutics as a Critical Method

This example shows that model evaluation is not just a crucial technical step for mass classification. It can actually contribute to genre analysis by indirectly producing some metrics of genre formalism. Beyond the precision, recall and F1 scores we tried to retrieve the probability density resulting from the application of the model on all issues of *Le Matin* of 1927 (Fig. 5).

This representation gives a complementary picture of the specificity of the genres. In contrast with the classic parameters used in Fig. 4, we no longer map binary results (which is the best genre candidate), but continuous probability distributions. Some genres have a "head" distribution, with a larger share of high probabilities like sports or literary criticism (or, even much more strongly, to the point they were not displayed here, classified ads and stock exchanges): most of the texts use a fairly distinct vocabulary which raises the confidence intervals. Other genres have a "flat" distribution like economics or movies sections: the vocabulary used is not always highly specific. For instance, articles on

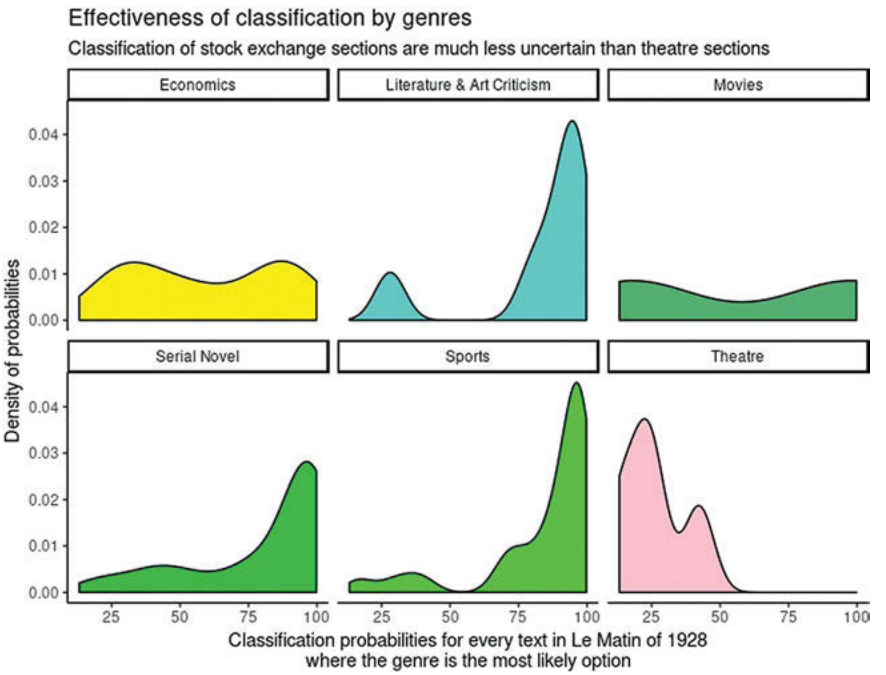


Fig. 5: Probability distribution of six genres in all the issues of *Le Matin* in 1928.

movies include not only specialized texts on film-making styles and techniques but also synopsis of recent releases that can be easily confused with other news (like *faits divers* for detective fiction), if not for a few distinctive twists. Finally, some genres have a "tail" distribution which suggests that their vocabulary is insufficiently unequivocal: Theatre news has many words common with Movie news, while the reverse is not true.

The comparative success of modeling projects can also yield potential clues on the degree of formalization of an entire corpus. For instance, the other large model developed by *Numapresse* for the press of the mid-19th century (1840–1860) achieves a much lower accuracy rate (62% instead of 75%) despite having virtually the same characteristics (same number of genres and a corpus of a comparable size.[31]) It is possible that, at this time, genres where less distinctive. In comparison

31 The only significant difference is that we used a random sample of the newspaper issues available in the *Numapresse* collection: we did not have this possibility at the time of the Interwar model as the corpora from the French National Library were still being integrated. This

with the Interwar model genres were more frequently "stylistic" rather than "thematic". Literary writing was also highly permeable and frequently used in other settings like crime news or political news (highly rhetorical discourses).

Model hermeneutics is not just a necessary step in model evaluation. It can actively contribute to the analysis of large-scale stylistic trends, such as comparing the degree of formalism of different genres or modeling through evolution in time. Nevertheless this meta-data has to be used carefully. In this approach, the model probabilities becomes an ambiguous metric, that both signals the uncertainty of the model (which may be caused by a lack of sufficiently distinctive texts in the training phase) and effective characteristics in the original corpus.

3 Methods of Zoom Reading

The relative efficiency of automatic classification of newspaper archives has encouraged the researchers of the *Numapresse* project to experiment with several new methodological practices that we have named *zoom reading*. Zoom reading is not exactly a hybrid of *distant reading* and *close reading*, but aims to encompass the wide range of possible distances that can be used to investigate a specific domain. Media corpus constantly articulates several timelines. While news are republished daily, giving it a seemingly ephemeral character, news writing evolves over longer and intricate time periods:

- Several weeks, to identify regularities in newspaper editorial choices and policies.
- Several months, or a year, to analyze the structural flow of reprints in the media ecosystem (6 months according to the Viral text project).
- Several years, for the inner sociological workings of the newsroom.
- Several decades, for the emergence and metamorphosis of news genre (hence the (roughly) 20-year length of the *Numapresse* models).

The following section presents an exercise of "de-zooming", starting with short-level patterns of days and weeks and ending with long-time, nearly secular, trends of cultural history. While obviously the use of larger time units precludes any attempt at actually reading the text, it does not mean that the interpretation of the classification output is bound to be solely quantitative. In fact we will show that very large-scale classification can be an efficient method to detect unusual occurrences.

---

approach may lower the formal accuracy rate while making the model more relevant since it draws from a more diverse set of documents.

3.1 Days and Weeks: Short-Term Patterns

The exploration of classification data can be initiated by an arbitrary sample: the issues of the *Le Matin* from March 1928. This is not a particularly interesting moment in the history of French newspapers, and this apparent randomness makes it actually easier to locate the ordinary monthly trends of newspaper genres. Figure 6 tracks the monthly trends of all the genres in our model.

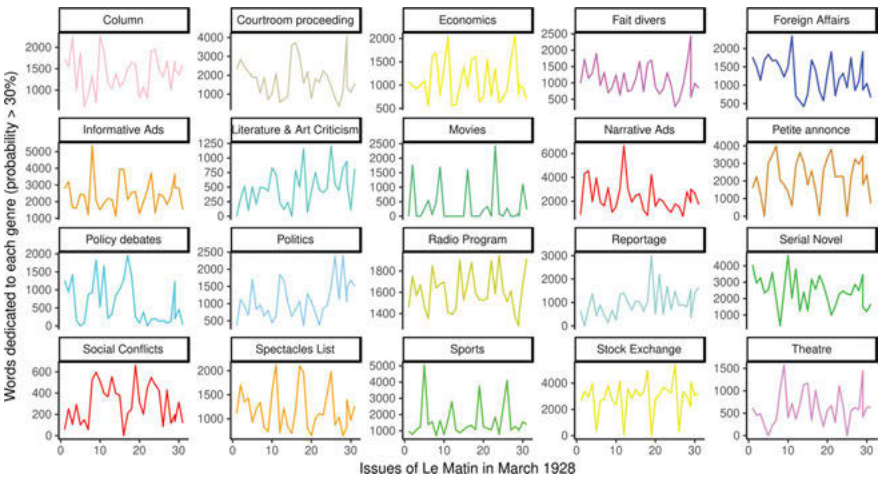


Fig. 6: Genre trends in *Le Matin* of March 1928.

The estimation is based on word counts. While classification is applied at the text block level, text blocks are an unreliable unit to quantify the extent of each genre in the newspaper. Some sections are broken into very small paragraphs, ads, and pieces of news, while others rely on long narratives and these editorial practices are more or less prevalent across the genres. Conversely, word counts neutralize layout preferences.

The estimation is based on a significant epistemological assumption: one text block can have more than one genre. We do not retain the most probable genre for each text, but all the propositions that passes a probability threshold (usually 20–30%). This method is better aligned with research practices in cultural history and literature studies: it anticipates the potential hybridization of genre in a newspaper context. While textual and semiotic standards are much more codified in a newspaper context than in other editorial productions, journalists can still take a distance from these unspoken norms. This freedom is

particularly noticeable in more authorial productions such as chronicles or personal columns. Obviously there is no way to encode idiosyncrasies into a classification model: by definition individual styles cannot be "counted" or even modeled. What is possible instead is to quantify the relative position of one text in regard to one or several writing standards. Genre probabilities can be cautiously used as a proxy for codified intertextualities. When the range of possibilities is highly fragmented, it may also suggest that the classification is entirely unknowable. Either the standards were not set or, even, there is no obvious standards where we could anchor this text.

This soft-clustering approach of classification data analysis should obviously remain very cautious. Once again, we are using probabilities as a multifaceted metric which reflects uncertainty, hybridity, and a measure of textual individuality. Future work of computational humanities may be able to disentangle these confused dimensions, but for the time being, we have to work our way through it (in the same way as we had to live with OCR noise).

The twenty small graphs shown in Fig. 6 draw highly divergent trends. To some extent each genre seems to live its own life with a particular sequence of peaks, slopes, and plateaus. Nevertheless some genres seem to follow a regular cyclical pattern, with an isolated jump interrupting dominant periods of residual activity. This is notably the case of sports and, to a lesser extent, of Movies, and Art and Literary Criticism.

The model has inadvertently rediscovered a core feature of French interwar dailies: weekly thematic supplements. Throughout this month, Monday is the day of the sports, Friday is the day of the movies and, to a lesser extent, Wednesday is the day of literature. On these occasions, the newsroom devotes an entire page to the subject, which acts as a small newspaper within the newspaper, with its own semiotic identity and editorial team. Sports news are covered by *La Vie sportive* (*Sporting Life*) and movie releases are summarized in *La Page du Cinéma* (*The Movie Page*), both on page 4. There is not a specific column for art and literature news, but museum exhibitions are usually described on Wednesday, page 4.

Obviously these patterns are not specific to March 1930, once again a rather unremarkable period in the history of French news. It simply reflects wide cultural arrangement regarding the media and, even, social structure of the calendar. Friday is the day of the movies simply because they were released on Friday. There is no financial column on Monday because the Parisian stock exchange is closed on Sundays. Whether this day was ultimately adopted due to media, industry, or public preferences can be seen as a chicken and egg problem. The fact remains that some newspaper genres are constrained by structural trends, either internal or external to the newsrooms, while others are influenced by a more chaotic flow of news and events.

3.2 Months and Years: Seasonal Trends

The week is not the only meaningful time dimension. Figure 7 extends the arbitrary sample to the entire run of issues of *Le Matin* published in 1928.

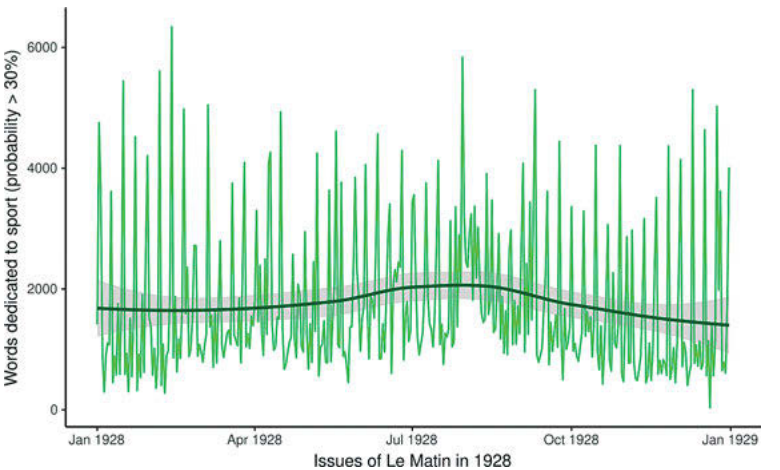


Fig. 7: Annual daily evolution of sports content in *Le Matin* in 1928.

This time, the selection focuses on one genre, namely sport, considering texts classified as such with more than 30% probability. The Monday spikes remain visible throughout the year. Yet, this pattern is supplemented by a larger trend made visible by the polynomial regression line (in dark green): sports contents are more prevalent in summer than in winter. In the French society of the first half of the 20th century, sport is largely synonymous with outdoor activities. This outdated definition is stressed by the early sports supplements appearing in the daily press in the 1910s, like *La Vie en plein air* (*Outdoor life*) in *Le Petit Journal*.

The "de-zooming" of classification data can be extended to years and decades. Figure 8 provides a time series decomposition of all text blocks labeled as sport (> 30% probability) of *Le Matin* for the interwar period (model 1920–1940).

The first plot displays the observed monthly changes, that can be broken down into three components:

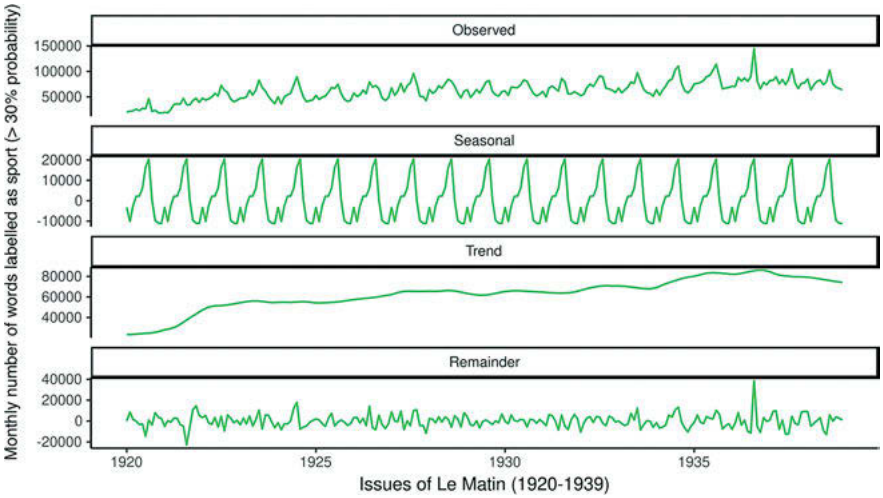


Fig. 8: Time series decomposition of sports content in *Le Matin* in the Interwar period (1920–1939).

- A seasonal pattern with more texts being published in the summer months (+ 16400 for July and + 20500 for August) and less text being published in the Autumn and Winter (roughly –10,000 for October, November, December and February).
- A global trend towards an increase in sports content, particularly in the early 1920s, as coverage of sports activities gradually expanded, especially when civilian life returned to normal after the First World War. For instance, the weekly supplement *La Vie sportive* was relaunched in November 1921 and sports news became more frequent on a daily basis in the late 1920s.
- A *remainder* which can be reframed in a newspaper context as large-scale events, disturbing any pre-defined editorial organization. A very clear-cut example is provided by the Summer Olympics of 1936 (the big isolated spike of the second half of the 1930s). While the summer season is already favorable to a growing share of sports content, the large-scale social and political impact of this event has arguably exceeded the model expectations.

The "de-zooming" of classification data makes it therefore possible to identify long-time regular trends that are relevant in the long-term perspective of cultural and social history but also, by contrast, intriguing anomalies that calls for a closer reading. While exceptional events, like the Summer Olympics of 1936, are an obvious example, it is also possible to look at a more complex stylistic phenomenon such as intertextual discourses. It may seem paradoxical since

the model was entirely built around the idea of "news standard" and of regular discursive structures, but the conservation of the full probability distribution of each document makes it possible to identify genre crossover which can be significantly linked to two different potential categories. For instance, by querying texts that significantly belong to both the sports news and serial novel genres, we spotted a heavily stylized account of fencing published in *Le Matin* on the 31 January 1922, which sounds like French historical novels of the 19th century (35% sport and 30% serial novel). An even stranger object was identified this time with the 1860–1880 model: a stock exchange section partly written in verse, as if it were a Baudelairian poem, in *Le Figaro* on the 13th May 1872.

These kinds of strange and unusual editorial experiments are definitely better fitted for a close qualitative analysis. Nevertheless, those outliers become identifiable in the classification data once we adopt a larger focal length on the scale of years. By definition, stylistic anomalies are rare and get most of their value by standing out of a perceived repertoire of "normal" textual practices.

3.3 Decades and Centuries: The Life Cycle of Newspaper Genres

Supervised classification is not only a tool which enables the study of one specific genre but, more globally, the process of genre-making. Figure 9 and 10 displays a long history of sports news on the core corpus of 25 French national dailies. While there are a few global phenomena (like the gradual emergence of sports columns by the last two decades of the 19th century, or the big drop during the First World War), the trends are highly differentiated: some dailies seem to have invested in the development of a large coverage of sports activities (*L'Écho de Paris*, *Le Figaro*, *Le Petit Parisien*, *Le Matin* and *Le Siècle*) while others seem to have neglected this new journalistic form (like *La Croix*, *Le Pays*, *La Lanterne*, *Le Radical*, *La Liberté* or *Le XIXe siècle*). Logically enough, sports news appear to be more preeminent in the newspapers with the largest circulation: nearly all the non-adopters are characterized by a smaller, less professional newsroom and a stronger political identity that may not have favored the coverage of non-political news.

Mass classification here demonstrates its utility as a "corpus inspector" by showcasing potential interesting phenomena or specific cases. For instance, a surprising finding has been the (comparatively) high level of sports content in two aging newspapers that both appeared in 1836 and disappeared before 1930, *Le Siècle* and, to a lesser extent, *La Presse*. In both cases, sports coverage seems to spike briefly before the First World War. The availability of the complete classification data at the text block level allows to reverse our

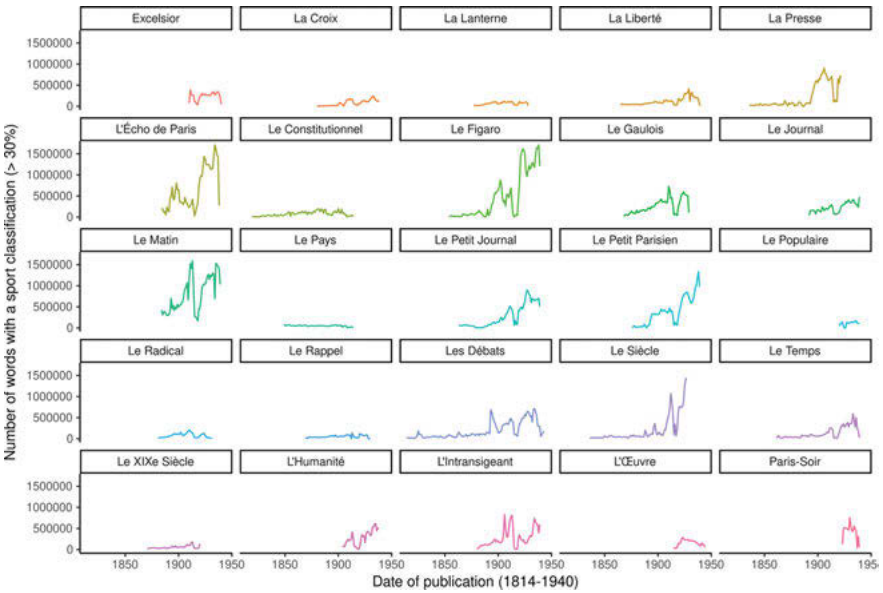


Fig. 9: Sports classification in the core corpus of Daily newspapers (1920–1940 model).

"de-zooming" process and to isolate potentially relevant sub-trends. It turned out that, for both titles, the newsroom had created a little-known weekly supplement and recruited professional sports reporters by the early 1910s. It appears that these trends were already identified by historians of sporting news: while the model did not make any surprising discovery, its output proved to be consistent with previous research.[32]

While the model output can be continuously checked by the practice of zoom reading, there is yet a significant caveat to the use of genre classification to analyze very long time trends: anachronism. To draw the complete life cycle of a genre from the early 1800s to the Second World War we have been forced to extend the classification way beyond its expected temporal window (1920–1940). What this means is that a model trained on four newspapers of the interwar period has been anachronistically applied to decades-older issues, despite the significant changes in journalism culture and, more broadly, in the French written language. For instance, the publication of thematic supplements is highly specific to 20th century newspapers. Through most of the 19th century, genre and thematic organization

32 Philippe Tétart and Collectif (2015). *La presse régionale et le sport: Naissance del'information sportive*. Rennes: PU Rennes.

remains rather fuzzy, except for a few localized editorial spaces like the *Feuilleton*. A large share of news was untidily piled up on pages 2 and 3, nicknamed the "soft belly of the newspaper", in contrast with the more regular political and foreign news coverage featured on page 1 and the advertising 'wall' of page 4. To interpret secular trends, we therefore have to deal not only with usual factor of uncertainties (such as the effectiveness of the model or the prevalence of intertextual productions), but also with the increasing discrepancy between genres learned by the model and the cultural practices at the time of the target corpus.

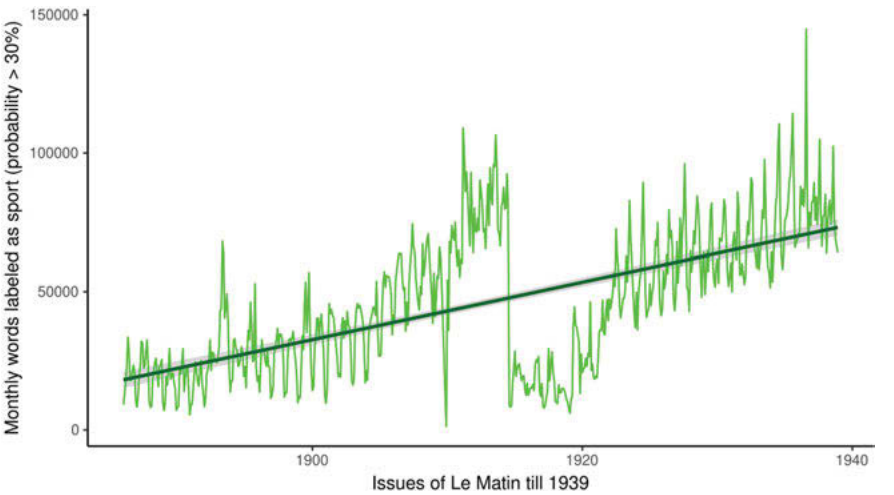


Fig. 10: Annual evolution of sports content in the entire digitized archives of *Le Matin*.

Since supervised models are fixed and transferable they can be used in a more oblique way: as an archaeological tool tracing the gradual codification of a genre or, more accurately, of a specific historical iteration of a genre. While we have stressed that our modeling strategy is contextual and grounded in a predefined editorial ecosystem, this does not mean that the model should be strictly limited to a relevant corpus or to a relevant period. Any extension beyond the original framework of the training corpus entails a radical change of methodology: we do not aim anymore to get straightforward results (minus the inevitable uncertainty of the model limits and genre hybridization) but some fugitive clues of the evolving proximity and distance toward a peculiar genre standard.

An anachronistic use of the supervised models can nevertheless start with the null hypothesis that the genre will gradually converge to the ideal definition of the genre abstracted by the model. This convergence process can take several

decades or result from sudden changes. It can also be heavily concentrated in a single newspaper, or be widely disseminated across the news ecosystem. In short the rhythmic patterns of anachronistic classification may be indicative of the wider processes of genre-making.

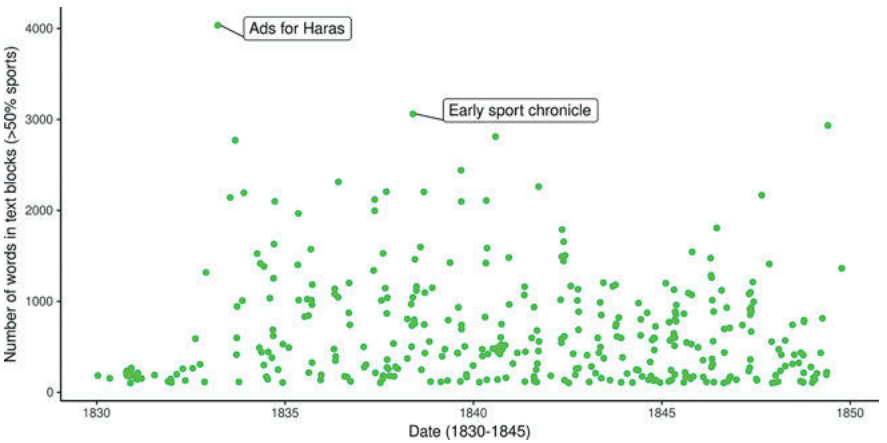


Fig. 11: Anachronistic classification of sports content in the *Journal des débats* (1920–1940 model).

Figure 11 shows the time distribution of the texts classified as sports (> 50% probability[33]) in the *Journal des Débats* between 1815 and 1845. Obviously in this period there is no regular sport section: the term "sport" itself appears in 1830 (initially as an *anglicism*), and enters the common language in the 1850s only. Yet, there are still occasional columns describing events and activities that will later be reframed as sport, such as horse racing. The two best article candidates for sports classification illustrate both the power and the limits of anachronistic classification. The first one is deceitful: it is caused by the association of several ads for Haras and horse selling (which are at least thematically linked to sport activities) as well as a classification mistake (a very long list of people's names). The second one is more interesting in terms of a long-term study of journalistic form: it is a detailed *feuilleton* published on May 23rd, 1838 recalling the *Course de Chantilly* in very vivid style that seems to foreshadow future developments of sport journalism.

33 Since we are dealing with the added uncertainty of anachronistic classification it seemed more reasonable to use a higher threshold for genre identification rather than 20–30%.

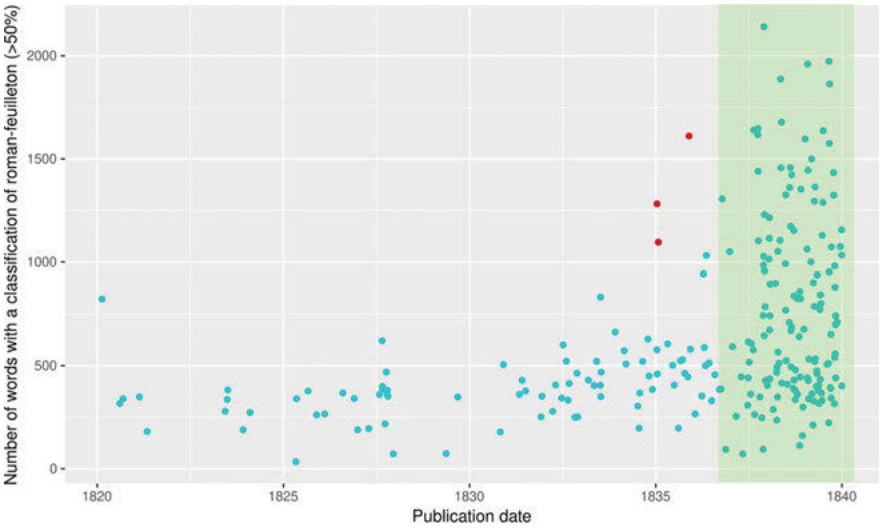


Fig. 12: Emergence of the serial novel in the first page of the *Journal des Débats* (1840–1860 model).

Figure 12 presents the gradual emergence of the serial novel (the *roman-feuilleton*) in the front page of the *Journal des Débats* between 1820 and 1840 using a model trained on the 1840–1860 period. The serial novel really became a staple genre of French newspapers by the end of the 1830s, following a series of successful editorial experiments in 1836–1837 including e.g., the first publication of Balzac's *La Vieille Fille* in *La Presse* in October 1836.

Unsurprisingly, the anachronistic model rediscovers this trend with a sharp increase in serial occurrences at the end of the period (highlighted in blue). However, the story is not complete: there is already a noticeable upward trend throughout the first half of the 1830s. The three red dots provide interesting hybrid experiments displaying an increasing presence of novelistic writing in the editorial space of the newspaper. They include a long excerpt of a novel adapted into a play, a fictionalized biography of a novelist, and a highly novelistic summary of a play.[34] While none of these are straight examples of serial novels, they are highly suggestive of a gradual acculturation of the *Feuilleton* to fictional literary narratives.

Clearly, output data from anachronistic classification should not be used at face value for quantitative analysis. As the model is applied on a corpus

34 *Journal des débats*, 23 Nov. 1835,13 Jan. 1835, 26 Jan. 1835, pp. 1–2.

significantly different from the training corpus, we frequently target a series of rare events, such as a genre that has not yet entered the standard journalistic repertoire and therefore appears erratically. From 1830 to 1850, only 1000 text blocks of the *Journal des Débats* are classified as sport with a probability higher than 50%. Since we are dealing with rare events, there is naturally an increasing risk of running into significant classification artifacts.

Nevertheless, anachronistic classification seems very promising as a tool for qualitative analysis, since it allows the detection of early consolidations of themes and/or journalistic writing that may have influenced the progressive codification of the genre. By definition, the model is able to decipher weak cocktails of words that would not be sufficient by themselves to discriminate a specific text, but, through their association are highly likely to connote a specific form of writing. Until now such examples of burgeoning genres have been frequently discovered through sheer luck: anachronistic classification has the potential to considerably enhance the findability of rare texts which are not documented by any external sources or association.

4 Conclusion: Toward a Model Ecosystem in the Humanities?

Mass automated classification is a promising method for cultural and literary history. This technique may embody a second revolution of *findability* after the implementation of full-text index: online search would not be limited to individual words and sentences but could be extended to abstract concepts like genres, thematics, styles or stereotypes.

Throughout this paper we have experimented with several uses of mass classification on a large and diverse corpus: French national dailies of the 19th century and the first part of the 20th century (roughly 1815–1940). Thanks to a recent wave of research on journalistic genres in several French-speaking countries, this was a controlled experiment: the output data corroborated trends that were already suspected in the literature. Theses experiments acted as a heuristic process: we attempted to acclimate several statistical and computational methods into the research practices of cultural history and media studies. Thus we reframed several standard metrics, like model accuracy, in connection with preexisting concepts relevant to literary analysis and cultural studies, such as intertextuality of genre hybridation.

The stress put on *zoom-* rather than *distant* reading aimed to circumvent longtime debates over the fitness of quantitative methods: mass classification

can contribute to highlighting unusual patterns and editorial experiments rather than dilutes them. The model provides a general baseline of news standards and, by contrast, makes it possible to find exceptional deviations from a repertoire of expected newspaper genres. These negative inquiries are heavily conditioned by the design of the classification process, given that they require the preservation of the detailed model output (with the probability density).

A promising feature of supervised models is their transferability: classification is not a one-time job that focuses on a specific corpus and ends with the completion of the task. Models can be stored, archived and reused. They can be re-purposed as a cost-effective solution to document a similar corpus. They can also be tweaked to design a wide range of classification experiments. We provided in this paper an example of anachronistic classification where a model trained on a given time period of the French daily press is applied to an earlier one to implement an archaeological inquiry of the development of newspaper genres at a moment they were not yet codified or identified.

Transferability is not limited to supervised approaches. Unsupervised models can be generalized to new corpora using inference methods. Besides, the impressive progress of deep learning infrastructures like BERT (Bidirectional Encoder Representations from Transformers)[35] tend to blur the distinction between supervised and unsupervised approach: while originally trained on a corpus, embeddings derived from a BERT model can be used to supplement topic modeling, by providing a better support for semantic representation.[36] For the time being, the high computational cost of these new models has prevented their implementation in the *Numapresse* project.[37]

Transfer learning may foster a model ecosystem among researchers and library communities working on cultural heritage collections. A storage infrastructure of shared models and training corpora would significantly reduce the time-consuming effort of producing annotated ground truths: similar corpora could be re-applied or marginally amended. For instance, it could be possible to rebuild a model for weekly magazines by adding a few genres specific to this format *on top* of one our newspaper models. The *Numapresse* project favors this

35 Jacob Devlin et al. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

36 Maarten Grootendorst (2022). "BERTopic: Neural topic modeling with a class-basedTF-IDF procedure." In: *arXiv preprint arXiv:2203.05794*.

37 Ted Underwood (2019b). *Do humanists need BERT? Library Catalog:* tedunderwood. com.

approach of model sharing. While initially envisioned as a specific hosting platform for newspaper genre model, the *Générothèque*[38] was revisited as an autonomous and generic project for text classification on cultural heritage archives. It acts as a collaborative library models where each model and its associated training corpus are extensively documented. In combination with TidySupervise, it aims to democratize the practice of modeling and its emerging creative uses in computational humanities for intertextual analysis or genre archaeology.

Bibliography

Andro, Mathieu (2018). *Digital Libraries and Crowdsourcing*. London: John Wiley & Sons. 234 pp.

Bacot, Jean-Pierre (2005). *La Presse illustrée au XIXe siècle: Une histoire oubliée*. Français. Limoges: Presses Universitaires de Limoges et du Limousin.

Bellanger, Claude, Jacques Godechot, Pierre Guiral, and Fernand Terrou, eds. (1969). *Histoire générale de la presse française. Tome II, De 1815 à 1871 / publiée sous ladirection de Claude Bellanger, Jacques Godechot, Pierre Guiral [et al.]* FR. Paris: Presses Universitaires de France.

Broersma, Marcel and Frank Harbers (2018). "Exploring Machine Learning to Study the Long-Term Transformation of News." In: *Digital Journalism* 6.9. Publisher: Routledge_eprint: https://doi.org/10.1080/21670811.2018.1513337, pp. 1150–1164. DOI: 10.1080/21670811.2018.1513337.

Can, Fazli and Esen A. Ozkarahan (Dec. 1990). "Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases." In: *ACM Transactionson Database Systems* 15.4, pp. 483–517. DOI: 10.1145/99935.99938.

Chabrier, Amélie (2019). *Genres du prétoire: La médiatisation des procès au XIXe siècle*. French. Paris: Mare Martin.

Chabrier, Amélie, Marie-Astrid Charlier, and Paul Aron, eds. (2018). *Coups de griffe, prises de bec: la satire dans la presse des années trente*. French. Bruxelles: Impressions Nouvelles.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines." In: *ACM Transactions on Intelligent Systems and Technology* 2.3, 27:1–27: 27. DOI: 10.1145/1961189.1961199.

Cooney, Charles, Russell Horton, Mark Olsen, Robert Voyer, and Glenn Roe (2008). "Hidden Roads and Twisted Paths: Intertextual Discovery Using Clusters, Classifi Cations, and Similarities." In: ADHO 2008 – Oulu.

D'Ignazio, Catherine and Lauren F. Klein (2020). *Data Feminism*. Cambridge, Massachusetts: The MIT Press. 328 pp.

Devitt, Professor Amy J. (2008). *Writing Genres*. Carbondale: Southern Illinois University Press. 268 pp.

38 http://www.numapresse.org/generotheque/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Grootendorst, Maarten (2022). "BERTopic: Neural topic modeling with a class-based TFIDF procedure." In: *arXiv preprint arXiv:2203.05794*.

Guthrie, Louise, Elbert Walker, and Joe Guthrie (1994). "Document Classification by Machine: Theory and Practice." In: *Proceedings of the Conference on Computational Linguistics*, pp. 1059–1063.

Jockers, Matthew (2007). "Macro Analysis (2.0)." In: ADHO 2007 – Urbana-Champaign.

Kalifa, Dominique, Philippe Régnier, Marie-Ève Thérenty, and Alain Vaillant, eds. (2011). *La civilisation du journal: Histoire culturelle et littéraire de la presse au XIXè siècle*. Paris: Nouveau Monde Editions.

Langlais, Pierre-Carl (2016). "La formation de la chronique boursière dans la presse quotidienne française (1801–1870)." PhD thesis. Université Paris 4 Sorbonne. URL: https://tel.archives-ouvertes.fr/tel-01424740/document.

Liddle, Dallas (2012). "Reflections on 20,000 Victorian Newspapers: 'Distant Reading'The Times using The Times Digital Archive." In: *Journal of Victorian Culture* 17.2, pp. 230–237. DOI: 10.1080/13555502.2012.683151.

Matei, Sorin Adam, Nicolas Jullien, and Sean P. Goggins (2017). *Big Data Factories: Collaborative Approaches*. Cham, Switzerland: Springer. 141 pp.

Mussell, James (2012). *The Nineteenth-Century Press in the Digital Age*. London: Springer.

Newman, David J. and Sharon Block (2006). "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." In: *Journal of the American Society for Information Science and Technology* 57.6, pp. 753–767. DOI: 10.1002/asi.20342.

Nicholson, Bob (2013). "The Digital Turn." In: *Media History* 19.1, pp. 59–73. DOI: 10.1080/13688804.2012.752963.

Saminadayar-Perrin, Corinne (2007). *Les discours du journal: rhétorique et médias au XIXesiècle (1836–1885)*. Saint-Etienne: Université de Saint-Etienne.

Scholz, Trebor (2013). *Digital Labor: The Internet as Playground and Factory*. New York: Routledge. 274 pp.

Simard-Houde, Mélody (2018). *Le reporter et ses fictions: Poétique historique d'un imaginaire*. Presses Universitaires de Limoges et du Limousin.

Tétart, Philippe and Collectif (2015). *La presse régionale et le sport: Naissance del'information sportive*. Rennes: PU Rennes.

Thérenty, Marie-eve (2007). *La Littérature au quotidien. Poétiques journalistiques au XIX e siècle*. Paris: Le Seuil.

Thérenty, Marie-Ève and Dominique Kalifa (2016). *Les Mystères urbains au XIXe siècle: Circulations, transferts, appropriations*. Médias19.

Underwood, Ted (2015). "The Literary Uses of High-Dimensional Space." In: *Big Data & Society* 2.2, p. 2053951715602494. DOI: 10.1177/2053951715602494.

Underwood, Ted (2016). “The Life Cycles of Genres.” In: *Cultural Analytics* 1.1. DOI: 10.22148/16.005.
Underwood, Ted (2019a). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.
Underwood, Ted (2019b). *Do humanists need BERT?* Library Catalog: tedunderwood.com.
Wrona, Adeline and Yves Jeanneret (2012). *Face au portrait: De Sainte-Beuve à Facebook*. French. Hermann. Paris: Hermann.

Melvin Wevers

Mining Historical Advertisements in Digitised Newspapers

Abstract: Historians have turned their focus to newspaper articles as a proxy of public discourse, while advertisements remain an understudied source of digitized information. This paper shows how historians can use computational methods to work with extensive collections of advertisements. Firstly, this chapter analyzes metadata to better understand the different types of advertisements, which come in a wide range of shapes and sizes. Information on the size and position of advertisements can be used to construct particular subsets of advertisements. Secondly, this chapter describes how textual information can be extracted from historical advertisements, which can subsequently be used for a historical analysis of trends and particularities. For this purpose, we present a case study based on cigarette advertisements.

Keywords: historical advertisements, text mining, digitized newspapers, digital history

1 Introduction

In recent years, we have seen an explosive growth of digitized historical newspapers. Innovations in natural language processing have extended the possibilities for historians to extract information from large corpora of digitized historical texts. National libraries and projects such as *impresso*, *Newseye*, and *Oceanic Exchanges* offer access to digitized archives of historical newspapers.[1]

For historians, newspapers provide a longitudinal understanding of public discourse.[2] Newspapers are not the only gateway to public discourse, since they do not capture public discourse in its entirety. In its function as a proxy, a newspaper operates as a transceiver; it is both the producer and the messenger of

1 https://impresso-project.ch/, https://www.newseye.eu/, https://oceanicexchanges.org/.

2 Neil Postman (Dec. 2005). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. en. London: Penguin; Frank van Vree (1989). *De Nederlandse pers en Duitsland 1930–1939: een studie over de vorming van de publieke opinie*. Groningen: Historische Uitgeverij.

Acknowledgment: The author acknowledges the National Library of the Netherlands (KB) for making its newspaper data available.

https://doi.org/10.1515/9783110729214-011

public discourse.[3] On a surface level, newspapers inform us about the views of journalists and people that were interviewed by these journalists. However, as Marshall claims, scholars can also uncover the "values, assumptions, and concerns, and ways of thinking that were a part of the public discourse of that time" by analyzing "the arguments, language, the discourse practices that inhabit the pages of public magazines, newspapers, and early professional journals."[4] With easier access to (and increased availability of) newspaper repositories, studies of the representation of ideas, values, and practices in public discourse have gained traction.[5]

Even though newspapers contain a considerable amount of advertisements, these remain an understudied source in computational studies of public discourse. This is surprising since advertisements are rich and varied carriers of information of the past. In his seminal work *Advertising the American Dream*, Marchand argues that adverts offer a lens on the past and provide "insight into the ideals and aspirations of past realities [...] they show the state of technology, the social functions of products, and provide information on the society in which a product was sold".[6] Others point out that, even though adverts provide perspectives on the past, this is a distorted one, as the content of ads is driven by commercial interest.[7] Marchand acknowledges this criticism and conceptualizes advertisements as distorted mirrors. He argues that despite the primary function of advertisements to sell products, they still communicate social and cultural values, albeit in a somewhat distorted manner.[8] Moreover, one could argue that, for ads to be successful, they need to resonate with their audience. As such, they have to be reflective of

3 Michael Schudson (1982). *The Power of News*. Cambridge: Harvard University Press, pp. 17–18.
4 Margaret Marshall (1995). *Contesting Cultural Rhetorics: Public Discourse and Education, 1890–1900*. Ann Arbor: University of Michigan Press, p. 8.
5 J. van Eijnatten and Ruben Ros (2019). "The Eurocentric Fallacy: A Digital-Historical Approach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990)." In: *International Journal for History, Culture and Modernity* 7. DOI: https://doi.org/10.18352/hcm.580; Joke Daems et al. (June 2019). "'Workers of theWorld'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940." In: *Journal of European Periodical Studies* 4.1, pp. 99–114. DOI: 10.21825/jeps.v4i1.10187.
6 Roland Marchand (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press.
7 Stephen R. Fox (1997). *The Mirror Makers: A History of American Advertising and Its Creators*. English. Urbana: University of Illinois Press.
8 Roland Marchand (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press; T. Jackson Lears (1994). *Fables of Abundance: A Cultural History of Advertising in America*. New York: Basic Books.

aspects of public discourse. Still, it is crucial to remain critical of the skewed representation of ideas and values in historical advertisements.

Advertisements are a fascinating and complex historical source, partly because of their multi-modal nature; they contain both visual and textual content. Since there often is an interplay between the visual and textual material, an analysis of only the textual content is somewhat limited. In recent years, advances in computer vision research have provided methods to examine digitized visual material at scale. While the use of computer vision is rapidly evolving and already offers promising methods of analysis, discussion of it falls outside of the scope of this chapter.[9] In this chapter, the topic of concern is the analysis of metadata and textual content.

The ability to extract information from large numbers of advertisements, both synchronically and diachronically, allows scholars to study cultural expressions on a macro-scale. This process has also been described as distant reading.[10] In this chapter, I refrain from using this much-debated term often used in a binary opposition to close reading, where quantitative approaches are incorrectly equated with distant reading and qualitative ones with a form of close reading. When switching between perspectives of analysis, there is more than merely the binary of close and distant. In my view, the use of computational methods can also offer fine-grained contextualized interpretations of particular expressions. More often than not, historical interpretations drawn from advertisements are based on a small selection of ads, which opens the door to cherry-picking. The ability to chart trends over time makes it possible to model whether findings in smaller subsets can be generalized, or whether specific expressions deviate from general patterns. On an intermediate, meso-scale, one can also more easily find variations of a single cultural expression.

While computational methods open up new modes of analysis of historical sources, we have to take into account that the quality of these digitized materials is often sub-optimal, despite continuous advances in Optical Character Recognition (OCR) software and natural language processing. It is especially challenging

9 Melvin Wevers and Thomas Smits (Apr. 2020). "The Visual Digital Turn: Using Neural Networks to Study Historical Images." In: *Digital Scholarship in the Humanities* 35.1, pp. 194–207. DOI: 10.1093/llc/fqy085; Taylor Arnold and Lauren Tilton (Mar. 2019). "Distant viewing: analyzing large visual corpora." In: *Digital Scholarship in the Humanities*. doi: 10.1093/digitalsh/fqz013. eprint: https://academic. oup.com/dsh/advance-article-pdf/doi/10.1093/digitalsh/fqz013/28082598/fqz013.pdf. URL: https://doi.org/10.1093/digitalsh/fqz013; Joon Son Chung et al. (2015). "Re-Presentationsof Art Collections." In: *Computer Vision – ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, pp. 85–100.

10 Franco Moretti (2013). *Distant Reading*. London: Verso.

for OCR software to recognize text in advertisements correctly. In addition to more generic factors such as the quality of paper or the printing technique, text in advertisements is often of varying sizes and fonts, or part of a heavily-stylized logo. OCR software regularly turns these forms of textual content into gibberish. The sub-optimal text quality of original documents makes it more challenging to study adverts than articles.

Despite these shortcomings, we can still extract meaningful information and patterns from the OCR-ed text. In this process, we have to be selective of the methods that we use, since their performance can be impacted in different ways by imperfect OCR.[11] In addition to the OCR-ed text, we can also learn about trends in advertising from metadata on the position and size of advertisements in newspapers. We can, thus, study multiple aspects of advertisements using computational means.

This chapter showcases how we can use computational techniques to study advertisements at scale in digitized newspapers. The first section shows how we can use metadata to examine trends in advertising. The second section gives examples of how text mining can be used to extract information from advertisements. This step is explained by demonstrating a case on product nationalities associated with cigarettes. In this case study, we rely on text mining to better understand changes and continuities in the associations with nationalities in advertisements for cigarettes.

2 Metadata Analysis: Advertisements in Almost all Shapes and Sizes

The digitization of newspapers has made it possible to study advertisements at scale using keyword searches. However, not all digitized newspaper collections have segmented the articles and advertisements, presenting users with full-page scans that contain multiple document types, for example, advertisements and articles. In cases where text blocks were segmented but not classified, text

11 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*. Malta: SCITEPRESS, pp. 484–496; Mark J. Hill and Simon Hengchen (Dec. 2019). "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." In: *DigitalScholarship in the Humanities* 34.4, pp. 825–843.

classifiers can be used to identify document types.[12] After training a text classifier with annotated data, the classifier can then predict metadata for unseen new data. Recently, there have been efforts to rely on computer vision to segment and classify the elements on a newspaper page (e.g. articles, tables, ads).[13]

Fortunately, *Delpher*, the digitized newspaper portal hosted by the National Library of the Netherlands (KB), includes segmented documents and metadata on the document type.[14] Access to this type of metadata allows researchers to filter for advertisements. For the twentieth century alone, we can quickly assess that *Delpher* holds over thirty million advertisements in national and regional newspapers. Browsing through the results, it quickly becomes clear that advertisements come in a wide range of shapes and sizes. We can find tiny, square-shaped advertisements; column-shaped classifieds advertisements; and full-page spreads (see Fig. 1 for an example of a column of classified ads). The content and target audiences of these ads vary considerably. How can we deal with this variation, without more specific metadata information?

This section explains how we can use metadata on the size and position of advertisements to cluster types of advertisements and subsequently filter out subsets. The analysis in this section is based on the metadata of the national newspaper *Trouw* (1946–1995).[15] Initial exploratory data analysis shows that this title contains instances of bad segmentation, in which small parts of advertisements appear as separate advertisements. To exclude these segmentation errors, we filtered out ads with a width or height smaller than 100 pixels. After removing

12 Aysenur Bilgin et al. (Oct. 2018). "Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history." In: *IEEE 14th International Conference on eScience, e-Science 2018*, pp. 486–496. DOI: 10.1109/eScience. 2018.00137.

13 Raphaël Barman et al. (Jan. 2021). "Combining Visual and Textual Features forSemantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & DigitalHumanities* HistoInformatics. URL: https://jdmdh.episciences.org/7097; Benjamin Charles Germain Lee et al. (2020). "The Newspaper Navigator Dataset: Extracting Headlinesand Visual Content from 16 Million Historic Newspaper Pages in Chronicling America." In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, pp. 3055–3062. DOI: 10.1145/3340531.3412767. URL: https://doi.org/10.1145/3340531.3412767; Bernhard Liebl and Manuel Burghardt (2020). "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18–20, 2020*. Ed. by Folgert Karsdorp et al. Vol. 2723. CEURWorkshop Proceedings. CEUR-WS.org, pp. 351–373. URL: http://ceur-ws.org/Vol-2723/long20.pdf.

14 http://www.delpher.nl. The metadata on document type was added manually.

15 For most of the examples in this chapter, direct access to the metadata and newspaper data is required. Researchers can contact KB for API access to the newspapers. Code is available on: https://github.com/melvinwevers/eldorado.


Te huur gevraagd

Te huur aangeb.

Pensions

Vakantie

Zaken en Huizen

Hypotheken enz.

Te koop gevraagd

Te koop aangeb.

Auto's, Motoren, Bromfietsen e.d.

Fig. 1: Classified adverts from *De Tijd* (19 January, 1957) Copyright: De Tijd.

these advertisements, *Trouw* contains about 1.31 million advertisements for the period 1945–1995. From Fig. 2, we can gauge that the total number of advertisements increased between 1946 and 1995, with a sudden increase in the early 1980s. Alongside this increase in the number of ads, we also see a sudden, albeit slight decrease in ads' size (Fig. 3).[16]

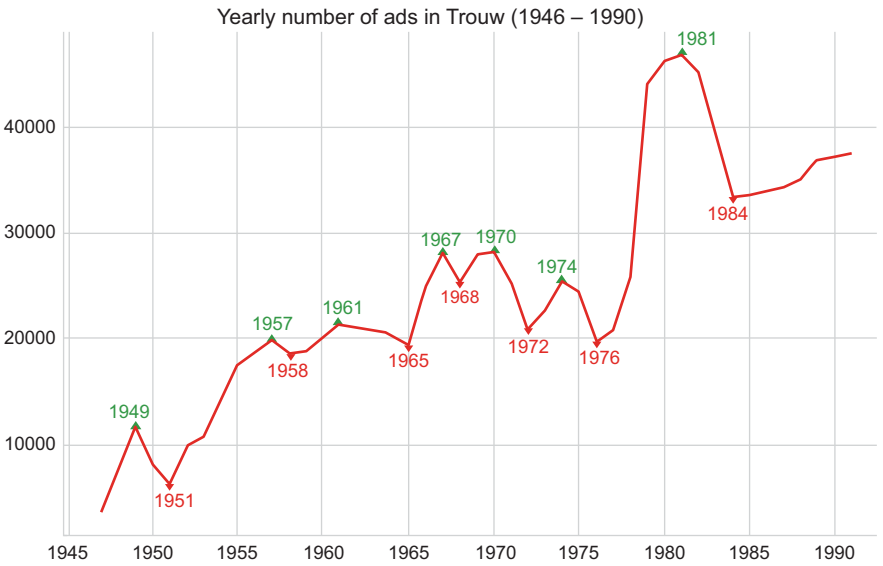


Fig 2: Yearly number of advertisements in *Trouw*. Green triangles indicate peaks and red triangle troughs.

In addition to their size, we can also examine where ads are positioned in the newspaper issue and where they can be found on the individual pages. Using the width and height information of advertisements, we created a heat map that shows which parts of the page are taken up by advertisements. There is a clear difference in pixel density between odd and even pages in *Trouw* (Fig. 4a and 4b). This can be explained by the fact that advertising on odd pages is more expensive than on even pages, as well as the common practice that new sections generally start on the odd pages. Moreover, we see that on the even pages, the upper-left corner, which is closer to the margin, is less populated than the lower-right side.

16 We log-transformed the mean size to visualize the rate of change better.

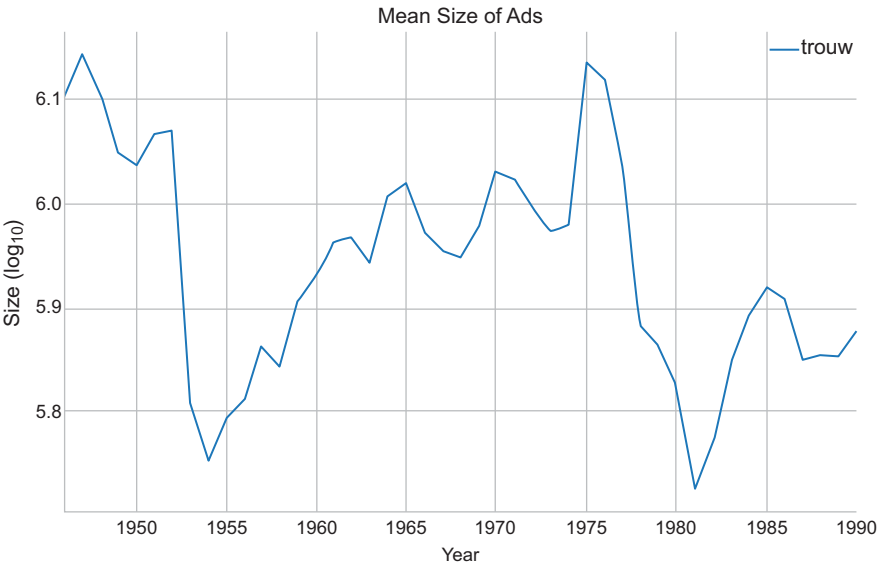


Fig 3: Mean size of advertisements in *Trouw*.

To get a better grasp of the groups of ads based on their size and their position in the newspaper, we can turn to unsupervised clustering methods. These methods can detect groups or clusters within a data set based on a set of features. We base this clustering on the features: width, height, and relative position of the ad in the issue. The latter indicates how close the ad is to the front page or the last page. The distribution of these features contains multiple peaks, indicative of multiple sub-distributions. To be able to capture the distributions of these sub-populations, we use Gaussian Mixture modeling, which can estimate the parameters of these mixtures.[17] Compared to K-means clustering, mixture models incorporate information on the co-variance structure of the data, making it possible to capture clusters with varying distributional shapes. We have applied the clustering separately to even and odd pages, since there seem to be different generative principles at play, as evinced by the heat maps in Fig. 4a and 4b. After estimating the optimal number of clusters for even and odd pages ($n = 8$ and $n = 7$), we fit a Gaussian Mixture model to the data.

17 Scikit-Learn offers the Gaussian mixture algorithm for the estimation of a mixture model.

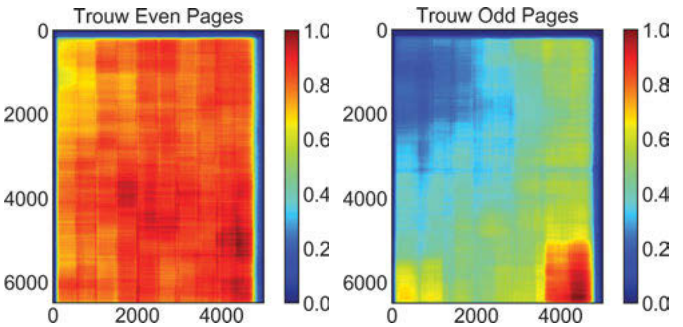


Fig. 4: Heat maps of pixel density in *Trouw*. (a) Pixel Density of Advertisements on Even Pages in *Trouw* (b) Pixel Density of Advertisements on Odd Pages in *Trouw*.

The results of this clustering are presented in Fig. 5a and 5b. These plots are based on a sample of 50,000 ads from even and odd pages. We see apparent differences in size and position in the newspaper between the odd and even pages. The odd pages are less structured, lack a clear signal, and show more noise, especially for ads with larger sizes. The ads on the even pages seem, by contrast, to be more structured. For some of these clusters, there is more variation, while others are closely clustered. A noteworthy cluster on the even pages is the brown cluster, which refers to full-page ads. We can also see that these ads appear toward the end of the newspaper. Newspapers are generally structured in columns, which is visible in the distinct structure present in the widths of advertisements (x-axis). The height of ads (y-axis) has more variance than the width, nevertheless, the algorithm distinguishes clusters based on their height. In terms of relative position, the number of advertisements increases towards the end of the newspaper issue (z-axis). Ads with a small width and a more considerable height, possibly classifieds ads (see Fig. 1 for an example), also appear toward the end of the paper. The information contained in these clusters can help us focus on or filter out particular subsets of advertisements. This method can help us determine the boundaries of these ads in terms of size and position.

We can also engineer other means to help us refine our subsets even further. A useful feature, for instance, is character proportion, which can be calculated by dividing the number of characters by the advert's size (width $\times$ height) of the adverts. We can use this metric to filter out textual advertisements or those that predominantly consist of visual material. A similar feature is the ratio of digits to characters. This feature can be used to distinguish ads that mostly list prices of imported goods or wholesale goods, from other ads. These numbers are dependent on the quality of the text extracted by OCR. Notwithstanding variations between newspapers and differences over time, these metrics allow us to filter

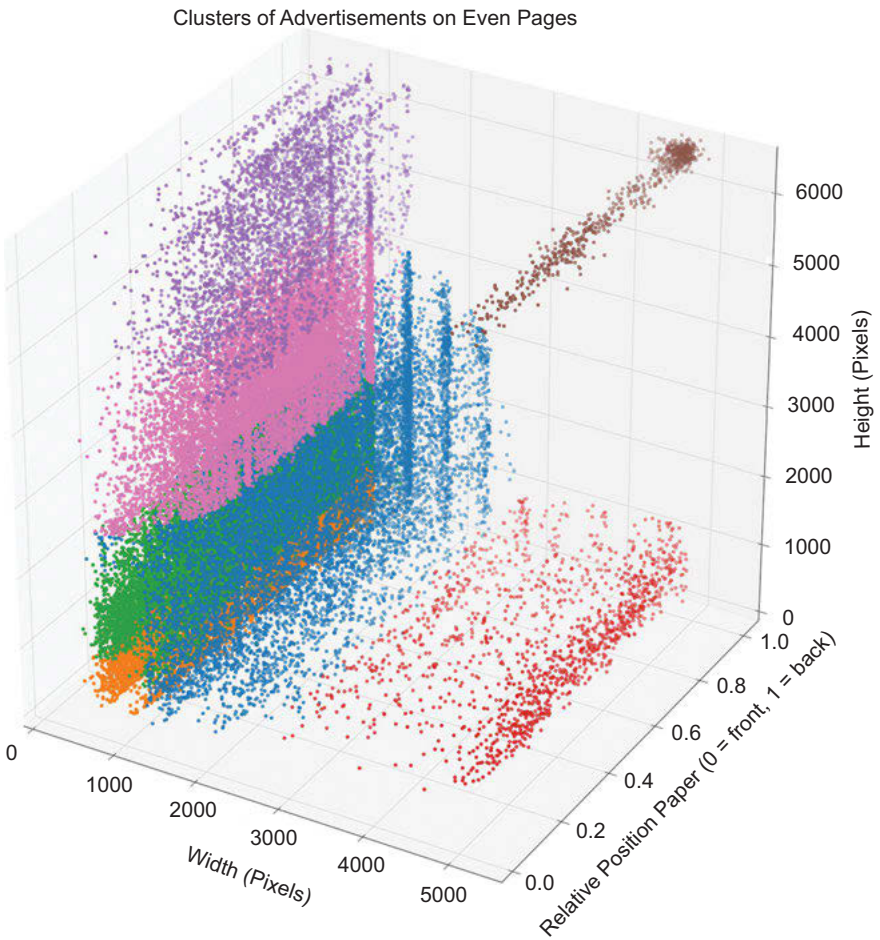


Fig. 5: Clustering of advertisements based on width, height, and relative position in *Trouw*. (a) Even pages (b) Odd pages.

particular types of advertisements. For example, we can filter out ads by inspecting particular clusters in the distribution of the character proportion. This form of exploratory data analysis can, for example, help to quickly filter out specific types of advertisements, such as classifieds. These advertisements are characterized by a higher than average height and smaller width, as well as a higher character proportion. We could also combine this metadata information with keyword searches to further filter out particular types of ads for selected products.

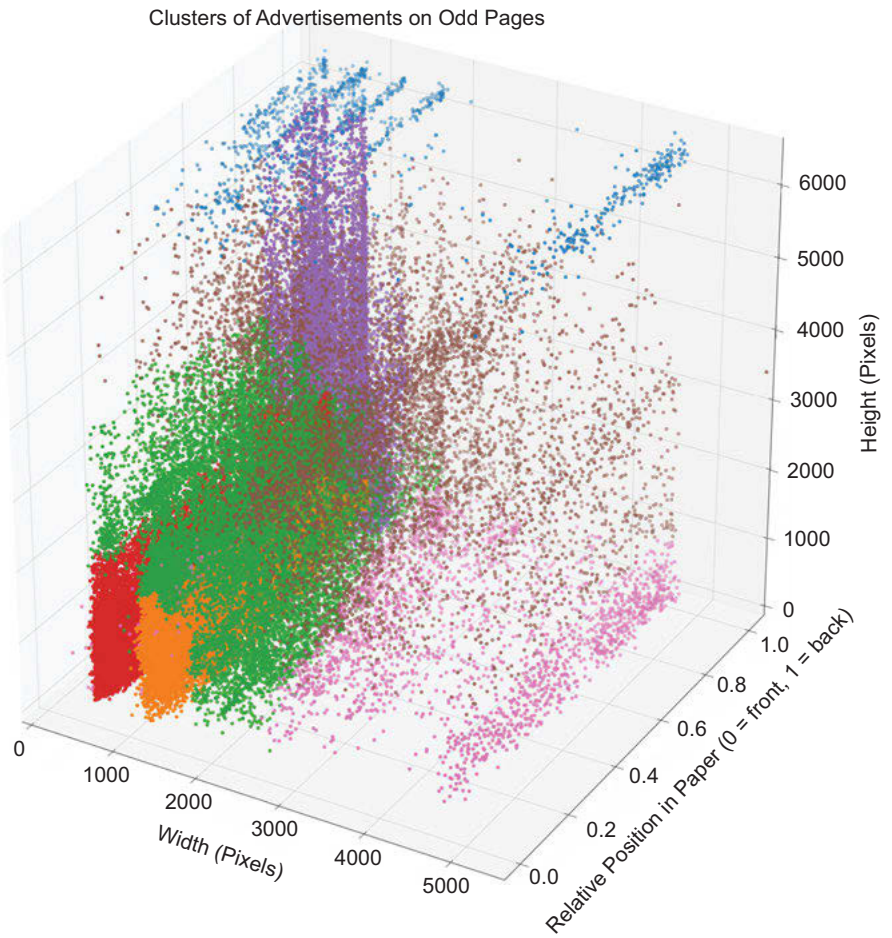


Fig. 5 (continued)

In this section, we have demonstrated a possible approach to using metadata to learn more about the structure and the organization of advertisements in historical newspapers. There are many more directions to explore. However, this section underscores that metadata should not be overlooked, and can be used in conjunction with text analysis to support further analysis.

3 Text Mining Advertisements

In this section, we show how we can extract textual information from historical advertisements, and how this textual information can be used for the analysis of trends and particularities. More specifically, we present a case study on the advertised nationalities of cigarettes. We constructed our corpus of cigarette advertisements from a larger corpus of advertisements from ten national newspapers spanning from 1890 to 1990 (Tab. 1).[18]

Tab. 1: Overview of selected national newspapers.

Newspaper	Period	Number of ads
Algemeen Handelsblad	1906–1970	979,312
Het Volk	1919–1945	191,626
Nieuwe Rotterdamsche Courant	1909–1929	472,536
NRC Handelsblad	1970–1990	460,996
Parool	1945–1990	1,626,204
Telegraaf	1893–1990	3,777,982
Trouw	1946–1990	1,154,746
Vaderland	1919–1945	317,440
Volkskrant	1940–1995	1,193,558
Vrije Volk	1945–1990	1,584,863

From the large corpus of advertisements, we extracted a sub-corpus of advertisements that contain the three most common spelling variants of cigarette: 'cigaret', 'sigaret', and 'cigarette'. These singular and plural variants are queried using the following regular expression: 'cigaret*\w+', 'sigaret*\w+'. This query yielded 43,781 advertisements. Figure 6 displays the distribution of the relative number of these cigarette advertisements. The trend line shows that the relative number of cigarette advertisements grew until the early 1920s, after which it decreased. In the 1950s, the relative number of ads again peaked, and after the 1960s, the number of advertisements that included variants of 'cigarette' dropped considerably, suggesting that after the 1960s, cigarette manufacturers advertised less in Dutch newspapers.

Plotting time series gives an overview of the temporal distribution of advertisements for a particular product. However, the search terms should be carefully selected. Using the different spelling variations, captured with the

18 Not all of these ten newspapers appeared throughout the entire period. Advertisements were selected using the metadata field article type.

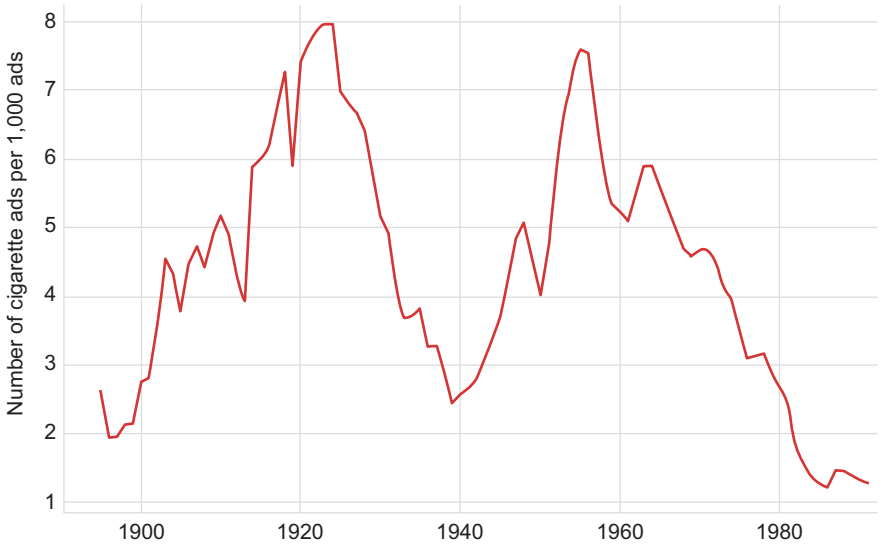


Fig. 6: Relative Number of Cigarette Advertisements 1890–1990.

regular expression, this query covers most of the historical variations in the twentieth century. In the case of cigarettes, the decrease in the last quarter of the twentieth century can in part be explained by the fact that advertisers stopped referencing cigarettes explicitly. Rather, they turned to ads that include brand names or logos, but not 'cigarette'. In such instances, computer vision could help to detect depictions of cigarettes or the occurrence of particular logos in advertisements. Many of these brand names are ambiguous, and querying them could also return advertisements for unrelated products. Moreover, the period during the Second World War should also be scrutinized more closely, as the newspaper landscape differs considerably, while the size and content of newspapers transformed drastically. Even when using relative frequencies to offset the changing number of newspapers, we cannot account for such factors.

3.1 Cigarettes and Product Nationalities

While text mining enables the study of large-scale trends, it can also help to explore more specific ways in which advertisers framed a product. One type of framing is the nationality of a consumer good, also known as 'product nationality'. The relationship between products and constructions of nationality has long

since been debated in consumer and marketing studies.[19] A key finding in these studies is that a strong connection to a favored country persuades consumers to spend more money on products with the same geographical connection.[20] The representations of product nationality are more than mere associations, they are "powerful narratives about the meanings and values transferred by products from their origin to their destination".[21]

In the case of the cigarette, the representation of product nationalities is an under-researched topic, which is surprising given that advertisers explicitly connected cigarettes to particular countries and regions. These product nationalities usually referred to their actual or perceived country of origin, but often also communicated particular characteristics associated with these countries.

The advertisement in Fig. 7 is a clear example of the association between product nationality and particular characteristics. This 1938 advert for the 'Buffalo' cigarette linked the brand to the United States in several ways. First, the brand name 'Buffalo' denoted the emblematic American prairie animal, as well as the city of Buffalo in upstate New York. This relationship between the brand and the United States was further enforced by a small print mentioning its producer: The Cumberland Company from Clarksville, USA. Secondly, in addition to these textual cues, the advertisement included a visual signifier: a background image of a giant cowboy bending over a Dutch tulip field. This picture of a cowboy – an exemplar of American culture – further substantiated Buffalo as an American cigarette. Furthermore, this image expressed the towering dominance of American products in the Netherlands. Thirdly, the ad presented the Buffalo cigarette as having an American product nationality by describing it as "the tastiest and spiciest American cigarette." The geographical association to the United States suggested the product's country of origin, but also signified a particular taste specific to American techniques of tobacco preparation.

19 Mrugank Thakor (1996). "Brand Origin: Conceptualization and Review." In: *Journal of Consumer Marketing* 13.3, pp. 27–42; Gordon Hull (Mar. 2016). "Cultural Branding, Geographic Source Indicators and Commodification." In: *Theory, Culture & Society* 33.2, pp. 125–145.

20 Luisa Menapace et al. (2011). "Consumers' Preferences for Geographical Origin Labels: Evidence from the Canadian Olive Oil Market." In: *European Review of Agricultural Economics* 38.2, pp. 193–212.

21 Søren Åskegaard and Güliz Ger (1998). "Product-Country Images: Towards a Contextualized Approach." In: *European advances in consumer research* 3.1, pp. 50–58.

Fig. 7: Advertisements for Buffalo cigarettes in *Limburger Koerier*, April 13, 1938.

The Buffalo ad is, of course, just one example. How does this particular advertisement compare to others, and is its messaging specific to cigarettes with an American product nationality? Using text mining, we can gather the information that helps us answer such questions.

3.2 Charting Product Nationalities

We can, for example, chart the nationalities most commonly associated with cigarettes in advertisements. There are two basic ways to establish product nationality. First, we could count the occurrences of bi-grams – two adjacent word tokens – that include an explicit reference to nationality and cigarettes. For example, to find out when and how often advertisers described cigarettes as American, one could count the advertisements that contained bi-grams such as 'American Cigarette' (*Amerikaanse sigaret*). Regular expressions can help to capture a wide range of possible spelling variations of such a bi-gram. However, a key flaw in counting the bi-gram 'American cigarette' is that we then only enumerate instances in which 'American' appeared directly to the left of 'cigarette'. Advertisers, however, also used other ways to relate cigarettes to a particular location. For instance, in the case of Egyptian cigarettes, advertisers relied on phrases such as "imported from Egypt." One possible yet time-consuming solution is to construct a list of possible strings that express a relationship to a specific nationality.

A second method is to count references to nationality, *i.e.* country names, that co-occur in the proximity of 'cigarette'. In this approach, one only counts words that co-occur within a specific span of words, rather than advertisements that contain the two words. The co-occurrence of words in one advertisement does not necessarily indicate a relationship between them. 'American' could appear in an advertisement for cigarettes without referring to the product nationality of the cigarette or one of its features. Nevertheless, word proximity is a good indicator of a semantic relationship between the given words. Hence, in what follows, we count references to America within a span of five words to the right or left of the keyword 'cigarette'.

An added benefit of proximity searches is that it helps to mitigate issues related to incorrect document segmentation. During digitization, the Optical Layout Recognition (OLR) did not always correctly segment advertisements. There are, for instance, cases where 'cigarette' and 'America' appeared in the actual newspapers in two separate advertisements, whereas after digitization, they were identified as one single advertisement (see Fig. 8). In Fig. 8, 'American' refers to the American hotel in Amsterdam, and 'cigarettes' *sigaretten* to an unassociated retailer advertising the product. The words appeared in one single advertisement, albeit separated by a large number of words. Therefore, the use of a span of five words would not have counted this as an instance in which the two words appeared together. Looking for words in proximity to each other reduces the impact of errors produced by composite advertisements.


NAAMLOOZE VENNOOTSCHAP:

AMERICAN HOTEL,

gevestigd te AMSTERDAM.

Opgericht bij Akte, dd. 29 Maart 1882.

UITGIFTE EENER 5 PCTS. OBLIGATIE-LEENING,

groot f 91,000,

aan te gaan kraehtens machtiging der Buitengewone Algemeene Vergadering van Aandeelhouders, dd. 30 December 1882. (1701).

De INSCHRIJVING op deze 364 Obligatiën ad f 250, à pari, met rente, ingaande 15 FEBRUARI 1883, is geopend op DONDERDAG den 18den JANUARI a. s., ten Kantore der Associatie-Cassa, van 's voormiddags 10 tot 's namiddags 4 uur.

PROSPECTUSSEN zijn te verkrijgen ten Kantore van de Heeren LEON WERTHEIM & Co., alhier.

Magazijn van Havana-Sigaren

C. BLOOKER Thzn.,

30, KALVERSTRAAT, 30. (1698).

OPRUIMING VAN DIVERSE SOORTEN SIGARETTEN.

Fig. 8: Example of an incorrect segmentation of advertisements. *Algemeen Handelsblad*, January 15, 1883.

For this section, we charted six product nationalities associated with cigarettes: American, British, Egyptian, Russian, Turkish, and Virginia.[22] Using regular expressions, we queried the singular and plural variants of these references as well as common spelling variations. The occurrence of these words in advertisements serves as a proxy for the popularity of cigarettes with varying product nationalities. Figure 9 shows the relative frequency of references to nationalities per 1,000 cigarette advertisements.

From 1890 to 1919, Egyptian, Turkish, and Russian cigarettes were the most popular. The popularity of these cigarettes mirrored the economic, cultural, and political power of the associated geopolitical entities. Before the First World War, the popularity of Russian and Turkish cigarettes mirrored the might of the Ottoman and Russian Empires. In the same period, the American cigarette industry was making its first forays into the European cigarette market.[23] Just after the First World War, Virginia and Egyptian cigarettes became the most popular cigarettes. After the First World War, the British cigarette industry, especially British American Tobacco (BATCO), played a prominent role in the production and dis-

22 We approach Virginia as a nationality since the term came to represent Britain.

23 Allan Brandt (2009). *The Cigarette Century*: *The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books.

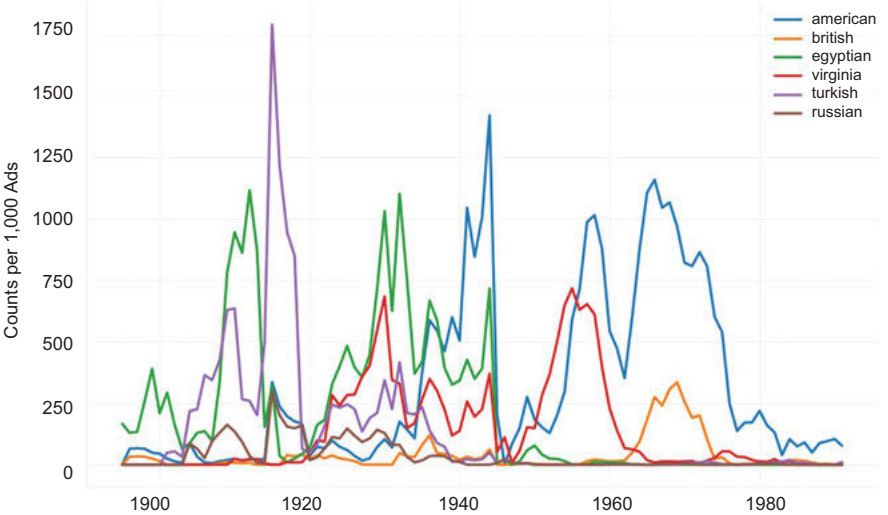


Fig. 9: Relative frequency of signifiers of nationality in cigarette advertisements, 1890–1990.

tribution of both Egyptian and Virginia cigarettes.[24] BATCO was also instrumental in disseminating Virginia cigarettes in the Netherlands.

The association with the United States gained prominence after the Second World War when both American and Virginia cigarettes towered above the other nationalities, which by that time had disappeared almost entirely. In the 1960s, when Virginia cigarettes lost their popularity, the American cigarette acquired sole dominance. A decrease in references to the United States characterized the subsequent decades.[25] This decline coincided with the growing sentiments of anti-Americanism in the Netherlands. Amid anti-American sentiments, advertisers might have refrained from associating their product with the United States. Furthermore, in 1964, the American Surgeon General Luther Terry published the Report on Smoking and Health, in which he presented the detrimental effects

24 Relli Shechter (2006). *Smoking, Culture and Economy in the Middle East: The Egyptian Tobacco Market 1850–2000*. London: Tauris, pp. 27–8.

25 Rob Kroes (2006). "European Anti-Americanism: What's New?" In: *Journal of American History* 93.2, pp. 417–432; Jessica C. E. Gienow-Hecht (Jan. 2006). "Always Blame the Americans: Anti-Americanism in Europe in the Twentieth Century." In: *The American Historical Review* 111.4, pp. 1067–1091.

smoking could have on one's health. The report led to a significant decrease in cigarette consumption in the Netherlands and the United States.[26]

As this section has shown, counting specific strings of text in digitized material is a relatively easy and fast way to gauge and compare the popularity of particular products. The trends in popularity in Dutch newspapers matched global trends. However, the interest in American cigarettes already declined before the Second World War, while academic literature on Americanization in the Netherlands commonly situates this process after the Second World War.

We will now examine how cigarettes with different product nationalities were presented in advertisements. In other words, were there distinct product features for cigarettes with particular product nationalities? Understanding the historical use of product features in advertising discourse can help us determine the cultural and technological impact of such products.

3.3 Product Features

In this section, we discuss two methods to examine the historical evolution of particular characteristics associated with cigarettes. First, we show how an unsupervised method that can detect noteworthy patterns, or 'bursts', in word use can be leveraged to extract words that typified a certain period. Second, we demonstrate how machine learning can determine whether cigarettes with a clear product nationality possessed features distinct from cigarettes with other nationalities.

3.3.1 Finding Trending Topics

When we want to study the use of particular words and examine how this use has evolved, one of the first tasks is to decide which words we are studying. One method is to rely on secondary sources to compile a vocabulary of words related to a subject or a particular period. A different approach is a data-driven one, in which we extract words that exhibit noteworthy use in the corpus using algorithms. One such algorithm is 'burst detection'.

26 Dietrich Hoffmann, Ilse Hoffmann, and Karam El-Bayoumy (2001). "The Less Harmful Cigarette: A Controversial Issue. A Tribute to Ernst L. Wynder." In: *Chemical research in toxicology* 14.7, pp. 767–790.

Burst detection is a modeling technique to detect *bursts of activity* in streams of data, for example, the sudden rise and fall in word frequency in serial publications. The trends for individual words cannot easily be compared. For example, for words with relatively little overall activity, a sudden, repetitive increase in use can signal a burst, whereas, for words with much activity, a different burst intensity might be required. Moreover, bursts of activity can also be nested within larger patterns. To be able to capture these bursts, Jon Kleinberg developed an algorithm that models the stream of information as an infinite-state automaton. The algorithm assigns costs to state transitions, which makes it possible to distinguish between short bursts and long burst even while the overall rate of transmission changes over time.[27] We use Kleinberg's algorithm to detect whether and when specific words exhibited "bursty" behavior in advertising discourse. In other words, we use it to identify *trending topics* in advertising discourse.

We model the 'burstiness' for a subset of words. This subset includes the 500 most distinctive determined adjectives and nouns – determined using tf-idf – from the corpus of cigarette advertisements. We then apply Kleinberg's algorithm using the default settings on monthly frequency counts of these 500 words. For each of these words, we get information on whether they bursted and how intense the burst was, as well as the duration of the burst. Figure 10 displays the top 50 bursty nouns and adjectives and when they bursted. We can, for example, see the appearance of the key term 'filter' around 1965, when the Surgeon General's report also appeared.

Moreover, the debate in the late 60s and early 1970s shifted toward the amount of nicotine in cigarettes. The appearance of the term 'health' (*gezondheid*) in the mid-1920s stems from advertisements that promoted the health benefits of smoking Virginia cigarettes. The figure also contains phrases such as 'job application' (*sollicitatie*) and 'human resources' (*personeelszaken*). Upon closer inspection, these words appeared in job ads placed by cigarette companies, a category that researchers might want to prune from their corpus.

Using this method, we can quickly gauge when particular words displayed 'bursty' behavior or which topics were trending at particular moments in time. This information can subsequently be used for closer examination of the texts in which these words appeared. Such techniques can also help distinguish a particular subset that one might want to remove, or treat as a separate corpus.

---

27 Jon Kleinberg (2002). "Bursty and Hierarchical Structure in Streams ⋆." In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 25.

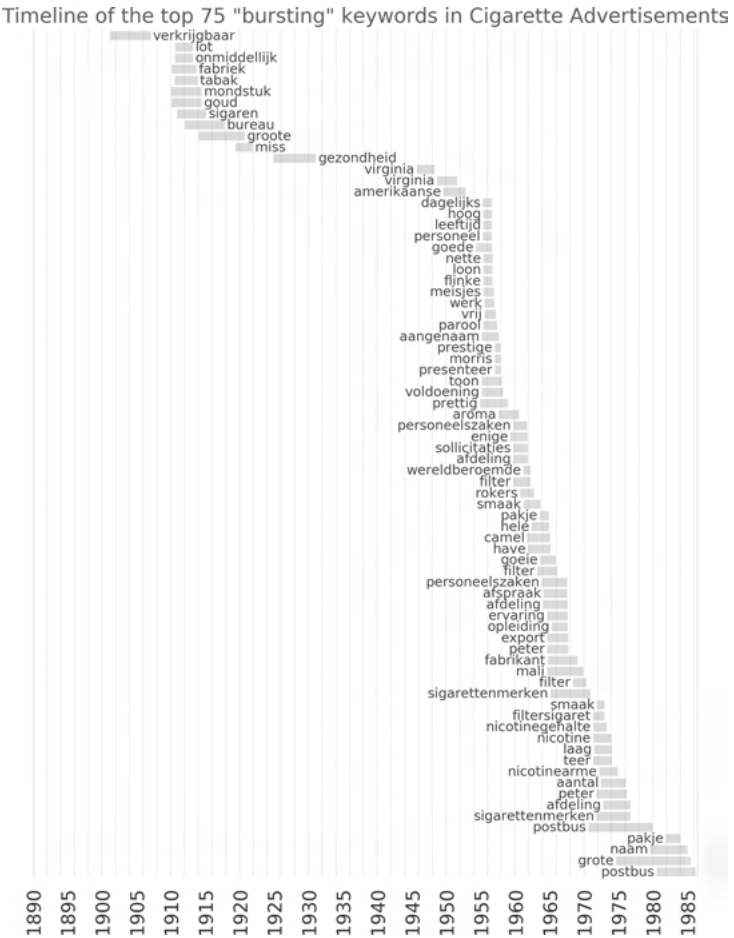


Fig. 10: Top 50 Bursty Nouns and Adjectives in Cigarette advertisements, 1890–1990.

3.3.2 Detecting Distinctive Features

A different line of questioning might focus on the differences in vocabulary between cigarettes with varying product nationalities. For example, were American cigarettes more commonly associated with a mild taste, or were Egyptian cigarettes presented as artisanal? One method to gauge the differences between texts is through the use of a machine learning classifier. These classifiers learn which textual features are predictive of a certain class of texts. For example, which words are indicative of either American cigarettes or Virginia cigarettes.

This method assumes that language use between the two remains separated over time. In other words, if features are first introduced in Virginia cigarettes and then co-opted by American cigarettes, it might be more challenging to use this feature to distinguish between the two. For this reason, we can also train classifiers for separate periods, given that we have enough data. If the classifiers fails to separate the two corpora, we can infer that the differences between language use were not very large. For each of these periods, we can also investigate the most informative features, *i.e.* the features that the algorithms rely the most on to make the classification.

We only select nouns, adjectives, adverbs, and verbs in the text.[28] Next, we label the corpus as either American or British, based on whether the ads contain references to the United States or to Virginia and the United Kingdom. After removing explicit references to these nationalities, we train a Naive Bayes Classifier on the data.[29] Based on the occurrence of particular words, the classifier learns to predict whether a text is an advertisement for American or British cigarettes. For the period 1920–1980, in which both types of cigarettes were represented, the classifier can distinguish between advertisements for American and Virginia cigarettes with an accuracy of 0.88, and respective F_1-scores of .92 and 0.79.

Subsequently, we can examine the most informative features for the classifier to label an ad as either American or Virginia. Noteworthy results include brand names such as 'Lucky Strike', 'Roxy', and 'Camel' for American cigarettes, and 'Derby' and 'Chief Whip' for Virginia cigarettes. Moreover, words such as 'cork' (*kurk*), 'mouth piece' (*mondstuk*), 'purity' (*zuiverheid*), and 'health' (*gezondheid*) were predictive for Virginia cigarettes, while 'connaisseurs' (*liefhebbers*), 'packages' (*pakjes*), 'filter', and 'smoking pleasure' (*rookgenot*) are all predictive of American cigarettes.

There was one company primarily responsible for connecting these features to Virginia cigarettes: 'Ardath'. This British cigarette manufacturer boasted that the purity of its Virginia tobacco led to a better tasting and healthier cigarette. This link was particularly strong in advertisements for the brand Chief Whip, which Ardath described as the "zenith of purity" (*toppunt van puurheid*). Together with the company Wills, Ardath distinguished the taste of Virginia cigarettes from that of American cigarettes. Ardath and Wills both denounced saucing and blending – two key features of the American cigarettes – and distanced themselves from American cigarettes in doing so. In 1925, Wills claimed

28 For this, we rely on Spacy. https://spacy.io.

29 Without removing these explicit references, it would be very easy for the classifier to determine in which category a text would fall. We use the NLTK implementation of the Naive Bayes Classifier (https://www.nltk.org/_modules/nltk/classify/naivebayes.html).

that its Virginia cigarettes consisted of 100% pure Virginia tobacco, without the addition of Greek, Indonesian, or Turkish tobaccos. In advertisements for Chief Whip, a doctor claimed that the cigarette was "absolutely pure and free of all surrogates and sauces." Advertisers presented Chief Whip as a pure and unprocessed cigarette.

The features used by the classifier point to three distinctive material aspects of the American cigarette: its length, its filter tip, and its packaging. Three additions that American cigarette producers introduced in Europe after the Second World War. First, in the 1950s, the longer, king size cigarette was introduced to Dutch consumers, after enjoying great success in the United States. Advertisers linked the increase in length to the United States to help familiarize the Dutch smoker with the long cigarette. In a 1955 advertisement, the brand So Long referred to the United States in their explanation of a king size cigarette: "In America, a cigarette longer than 85mm is called King Size."[30] Advertisers used the link with the United States to help acquaint Dutch consumers with longer cigarettes.

The second significant change was the introduction of filter cigarettes. Amid growing health concerns in the United States, American cigarette manufacturers introduced the purportedly healthier filter cigarettes in the early fifties. In the context of American filter cigarettes, purity did not denote the unblended nature of cigarettes, but the purifying effects of filters. Advertisements claimed that a lengthy filter would lead to "more and purer smoking."[31] Filter cigarettes became hugely popular, and by the 1970s, almost ninety percent of the cigarette market consisted of filter cigarettes.[32]

Finally, technological developments in the United States changed the look and feel of cigarette packaging. These innovations offered advertisers new ways to link the product to the United States. American companies were the first to package cigarettes mechanically in plastic-wrapped cardboard boxes. Until then, consumers bought cigarettes per piece, and stored them in less practical tin boxes. The new method of packaging not only referred to the United States because of its origin, but it also carried cultural connotations that resonated with American culture. Advertisers described the packaging of the American cigarette as flat, practical, modern, famous, or fancy.

30 "So Long advertisement," *Het Vrije Volk*, December 13, 1955.

31 "Sir Richard advertisement," *De Telegraaf*, January 5, 1962.

32 Allan Brandt (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books, p. 244.

The information gained through this method can be used for closer examinations of particular characteristics, as demonstrated above. A more in-depth analysis falls outside of the scope of this chapter.[33]

4 Conclusion

Advertisements are a rich and varied source for the study of public discourse. This chapter showcases how we can apply computational methods to metadata information to learn more about advertising in the past. The ability to distinguish between different types of advertisements can help to streamline further inquiry. Moreover, based on a case study on product nationalities in cigarette advertisements, we showed how text mining can help to better understand the historical trajectories between products and particular nationalities. Additionally, we demonstrated how machine learning can help to identify trending topics and distinctive words in advertising discourse.

There is always an interplay between prior knowledge, the choice and implementation of the computational methods, and the interpretation of results. Therefore, it is important that we are explicit about the assumptions that we introduce into our computational modelling. This prevents us from massaging the data or optimizing the method to merely find what we want to find. As Joshua Epstein succinctly shows, models can help us with more than prediction and can help us with the data collection, explanation of data, and the surfacing of dynamics in data.[34] While the methods introduced in this article can certainly be applied to different domains, they always need to be fine-tuned to the data and specific question. Moreover, relying on a combination of methods, including ones that show uncertainty in results, can further improve the transparency and the value of computational methods in historical research. While the text quality is often far from perfect in digitized newspapers, there is still much we can extract from this abundance of source material using computational methods. With the increasing amount of digitized newspaper collections in different countries, we can also envision expanding our efforts to study advertisements from a transnational perspective.

33 See Melvin Wevers (2017). "Consuming America: The United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890–1990." Ph.D. Dissertation. Utrecht: Utrecht University, for such an analysis.

34 Joshua M. Epstein (Oct. 2008). "Why Model?" In: *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

Bibliography

Arnold, Taylor and Lauren Tilton (Mar. 2019). "Distant viewing: analyzing large visual corpora." In: *Digital Scholarship in the Humanities*. DOI: 10.1093/digitalsh/fqz013. eprint: https://academic.oup.com/dsh/advance-article-pdf/doi/10.1093/digitalsh/fqz013/28082598/fqz013.pdf. URL:https://doi.org/10.1093/digitalsh/fqz013.

Åskegaard, Søren and Güliz Ger (1998). "Product-Country Images: Towards a Contextualized Approach." In: *European advances in consumer research* 3.1, pp. 50–58.

Barman, Raphaël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan (Jan. 2021). "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & Digital Humanities* HistoInformatics. URL:https://jdmdh.episciences.org/7097.

Bilgin, Aysenur, Laura Hollink, Jacco van Ossenbruggen, Erik Tjong Kim Sang, Kim Smeenk, Frank Harbers, and Marcel Broersma (Oct. 2018). "Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history." In: *IEEE 14th International Conference on eScience, e-Science 2018*, pp. 486–496. DOI: 10.1109/eScience.2018.00137.

Brandt, Allan (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. New York: Basic Books.

Chung, Joon Son, Relja Arandjelović, Giles Bergel, Alexandra Franklin, and Andrew Zisser- man (2015). "Re-Presentations of Art Collections." In: *Computer Vision – ECCV 2014 Workshops*. Ed. by Lourdes Agapito, Michael M. Bronstein, and Carsten Rother. Cham: Springer International Publishing, pp. 85–100.

Daems, Joke, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen (June 2019). "'Workers of the World'? A Digital Approach to Classify the Interna- tional Scope of Belgian Socialist Newspapers, 1885–1940." In: *Journal of European Periodical Studies* 4.1, pp. 99–114. DOI: 10.21825/jeps.v4i1.10187.

Epstein, Joshua M. (Oct. 2008). "Why Model?" In: *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

Fox, Stephen R. (1997). *The Mirror Makers: A History of American Advertising and Its Creators*. English. Urbana: University of Illinois Press.

Gienow-Hecht, Jessica C. E. (Jan. 2006). "Always Blame the Americans: Anti-Americanism in Europe in the Twentieth Century." In: *The American Historical Review* 111.4, pp. 1067–1091.

Hill, Mark J. and Simon Hengchen (Dec. 2019). "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." In: *Digital Scholarship in the Humanities* 34.4, pp. 825–843.

Hoffmann, Dietrich, Ilse Hoffmann, and Karam El-Bayoumy (2001). "The Less Harmful Cigarette: A Controversial Issue. A Tribute to Ernst L. Wynder." In: *Chemical research in toxicology* 14.7, pp. 767–790.

Hull, Gordon (Mar. 2016). "Cultural Branding, Geographic Source Indicators and Com- modification." In: *Theory, Culture & Society* 33.2, pp. 125–145.

Kleinberg, Jon (2002). "Bursty and Hierarchical Structure in Streams." In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 25.

Kroes, Rob (2006). "European Anti-Americanism: What's New?" In: *Journal of American History* 93.2, pp. 417–432.

Lears, T. Jackson (1994). *Fables of Abundance: A Cultural History of Advertising in America*. New York: Basic Books.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld (2020). "The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from16 Million Historic Newspaper Pages in Chronicling America." In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, pp. 3055–3062. DOI: 10.1145/3340531.3412767. URL:https://doi.org/10.1145/33405313412767.

Liebl, Bernhard and Manuel Burghardt (2020). "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, pp. 351–373. url:http://ceur-ws.org/Vol-2723/long20.pdf.

Marchand, Roland (1985). *Advertising the American Dream: Making Way for Modernity, 1920–1940*. Berkeley: University of California Press.

Marshall, Margaret (1995). *Contesting Cultural Rhetorics: Public Discourse and Education, 1890–1900*. Ann Arbor: University of Michigan Press.

Menapace, Luisa, Gregory Colson, Carola Grebitus, and Maria Facendola (2011). "Consumers' Preferences for Geographical Origin Labels: Evidence from the Canadian Olive Oil Market." In: *European Review of Agricultural Economics* 38.2, pp. 193–212.

Moretti, Franco (2013). *Distant Reading*. London: Verso.

Postman, Neil (Dec. 2005). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. London: Penguin.

Schudson, Michael (1982). *The Power of News*. Cambridge: Harvard University Press.

Shechter, Relli (2006). *Smoking, Culture and Economy in the Middle East: The Egyptian Tobacco Market 1850–2000*. London: Tauris.

Thakor, Mrugank (1996). "Brand Origin: Conceptualization and Review." In: *Journal of Consumer Marketing* 13.3, pp. 27–42.

van Eijnatten, J. and Ruben Ros (2019). "The Eurocentric Fallacy: A Digital-HistoricalApproach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990)." In: *International Journal for History, Culture and Modernity* 7. doi: https://doi.org/10.18352/hcm.580

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*. Malta: SCITEPRESS, pp. 484–496.

Vree, Frank van (1989). *De Nederlandse pers en Duitsland 1930–1939: een studie over de vorming van de publieke opinie*. Groningen: Historische Uitgeverij.

Wevers, Melvin (2017). "Consuming America: The United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890–1990." Ph.D. Dissertation. Utrecht: Utrecht University.

Wevers, Melvin and Thomas Smits (Apr. 2020). "The Visual Digital Turn: Using Neural Networks to Study Historical Images." In: *Digital Scholarship in the Humanities* 35.1, pp. 194–207. doi: 10.1093/llc/fqy085.

Petri Paju, Heli Rantala, Hannu Salmi

Towards an Ontology and Epistemology of Text Reuse

Cycles of Information Flows in Finnish Newspapers and Journals, 1771–1920

Abstract: The article explores the ontological and epistemological ramifications of text reuse, drawing on a digitized corpus of newspapers and journals from the National Library of Finland and covering the time span of 149 years from 1771 to 1920. The article examines three types of reuse cycles, rapid, slow and mid-range repetition. The argument is that text reuse has ontological ramifications on how the processes of a media network are conceived. With ontology we mean that the study of history always includes conceptualizations, either implicit or explicit, of which kinds of entities and things, as well as forms of being, there were in the past. Text reuse offers a perspective for the analysis of these "forms of being." In the epistemological part of the study, the article studies the aspects that influence and may bias the results, focusing on the material conditions of digitization process, the problems of metadata, and the possible methodological nationalism of drawing on nationally siloed corpora.

Keywords: text reuse detection, newspaper history, media history, ontology, epistemology

1 Introduction

Text reuse detection is an expanding field of research. It offers possibilities for exploring large text corpora and the circulation of texts and text passages. It has been successfully employed, for example, in the study of how authors have quoted previous literary works, such as ancient classics.[1] It has also proved to be an efficient tool in understanding how newspapers share each other's contents.[2]

1 M. Büchler, G. Crane, M. Moritz, and A. Babeu. 2012. "Increasing recall for text re-use in historical documents to support research in the humanities." In: (Proceedings) Second International Conference on Theory and Practice of Digital Libraries, vol 7489, pp. 95–100. DOI: 10.1007/978-3-642-33290-6_11.

2 D. A. Smith, R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: Proceedings of the Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-

https://doi.org/10.1515/9783110729214-012

In this article, we explore the ontological and epistemological ramifications of text reuse. The article draws on the research project *Computational History and the Transformation of Public Discourse in Finland* (Academy of Finland, 2016–2019), in which we studied text reuse in the Finnish press.[3] Our work is based on a digitized corpus of newspapers and journals from the collection of the National Library of Finland, covering a time span of 149 years, from 1771 to 1920. The advantage of the Finnish corpus is that it is complete, including in principle all published titles and their issues up to the year 1920. The corpus consists of all kinds of serial and periodical publications which in principle could reprint texts from each other. By drawing on both newspapers and journals we wanted to emphasize a comprehensive view on periodical press. From the perspective of Finnish history, the period is also essential: until 1809, Finland was part of Sweden, and after the War of 1808–1809, it came under Russian rule. Finland became an independent state in 1917. Throughout the timespan covered in this study, periodicals in Finland were published both in Swedish and Finnish.

Although the reprinting of particular texts in a range of different locations can be regarded as an old and well-acknowledged practice, a systematic examination of this phenomenon has not been possible until the digitization of the press. Our primary research material derives from the digitized and OCR'd corpus, including newspapers and journals, published by the National Library of Finland, in sum 5 million pages. For this project, we developed our own solution for text reuse detection, which we called *text-reuse-BLAST*. It was based on NCBI BLAST (National Center for Biotechnology Information Basic Local Alignment Search Tool), originally developed for detecting similarities in biomedical sequences. As a software originally created for comparing and aligning DNA and protein sequences, text

2013.pdf. H. Rantala, A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, and F. Ginter. 2019. "Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma." Historiallinen Aikakauskirja 1: 53–67. H. Salmi, A. Nivala, H. Rantala, R. Sippola, A. Vesanto, and F. Ginter. 2018. "Återanvändningen av text i den finska tidningspressen 1771–1853." Historisk tidskrift för Finland 1: 46–76. H. Salmi., P. Paju, H. Rantala, A. Nivala, A. Vesanto, and F. Ginter. 2020. "The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective". Historical Methods: A Journal of Quantitative and Interdisciplinary History, https://doi.org/10.1080/01615440.2020.1803166.

3 The project was a large consortium, including researchers from the Universities of Turku and Helsinki and the National Library of Finland. The text reuse study was realized as a cooperation between historians (PI Hannu Salmi) and data scientists (PI Tapio Salakoski) at the University of Turku. The group included Filip Ginter, Asko Nivala, Petri Paju, Heli Rantala, Hannu Salmi, Reetta Sippola, Tapio Salakoski and Aleksi Vesanto. The software development was done by Aleksi Vesanto under the supervision of Filip Ginter.

reuse detection with BLAST is character-based – not word-based as in many previous text reuse detection solutions – and it requires preprocessing of the material so that the letters are encoded into the alphabet of 23 amino acids. BLAST was originally designed to be highly applicable to material that includes a substantial amount of noise. The tolerance of noise was significant in processing the Finnish corpus of newspapers and journals, which includes many OCR errors.[4] In the end, text-reuse-BLAST proved to be effective in recognizing textual similarity. From the whole corpus, we found 61 million hits or occurrences of similarity, which formed 13.8 million clusters of text reuse. These clusters are simply chains of similar passages of text that were initiated between 1771 and 1920 in Finland. At the outset, we set the minimal threshold for similarity to 300 characters.

This article analyzes the different cycles of text reuse within this material, especially their ontological and epistemological aspects. In an earlier work, we identified and studied long-term text reuse[5] and presented the main findings of the project in two publications.[6] Cases of long-term reuse were those in which texts were reprinted after a gap of 50 years or more. Some repetition chains were very long and slow, lasting over 140 years. In contrast to this, there was also very short-term, rapid, viral circulation of texts. Virality was manifested when the same text spread across the network within a few days or weeks. The duality between slow and quick repetition is challenged by the significant amount of what we call mid-range repetition, characterized by texts that were copied infrequently within 5, to 10, or even 20, years. Our argument is that text reuse has ontological ramifications on how we conceive the past processes of a

4 On this methodology, see A. Vesanto, A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter. 2017. "Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910." In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf. An updated discussion has been published in Salmi *et al.* 2021. For source code, see https://github.com/avjves/textreuse-blast.

5 H., Salmi, H. Rantala, A. Vesanto, and F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

6 Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179. Salmi, H., P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter. 2021. "The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective". Historical Methods: A Journal of Quantitative and Interdisciplinary History 54 (1): 14–28, https://doi.org/10.1080/01615440.2020.1803166.

media network. In information science, "ontology" refers to a model and an abstraction based on certain concepts, and often to a data structure. Our point of departure emphasizes a complementary, philosophical approach. With ontology we mean that the study of history always includes conceptualizations, either implicit or explicit, of which kinds of entities and things, as well as forms of being, there were in the past. Text reuse offers a perspective for the analysis of these "forms of being." Simultaneously, it is important to concentrate on the epistemological aspects of text reuse: our conditions of knowing about those "forms of being."

This article examines the three cycles of text reuse, rapid, slow and mid-range repetition, as entities that shed light on newspaper activity in the analyzed time frame. Thereafter, we focus on the epistemological aspects of text reuse and concentrate on three perspectives, on the material conditions of both newspaper publishing and digitization process, the problems of metadata, and the possible methodological nationalism of drawing on nationally siloed corpora.

2 Towards an Ontology of Reuse

In our analysis, we were able to detect 13.8 million text reuse clusters.[7] Since the Finnish OCR corpus is unsegmented, the process of text reuse detection was made on all newspaper content types, including news but also advertisements, anecdotes, obituaries, timetables, announcements, and any other items. These reuse clusters were not necessarily cases of conscious reuse from paper to paper. In this kind of a computational analysis, it was not possible to identify external sources, like letters sent to several editorial offices at the same time. For us, clusters were simply chains of similar passages. These chains were different in nature: in the present corpus, the shortest clusters included only two or three hits, but there were also long chains of repetition. These clusters represent flows of information from the late eighteenth century to the early twentieth century, and thus offer the possibility to explore the ontology of newspaper printing.

During the project we gradually separated the mass of reuse clusters into three main categories which broadly correspond and describe different cycles of reprinting. These categories help us understand and analyze the vast number of clusters. We first concentrated on rapid, short-term circulation and slow, long-term circulation. Within the rapid circulation, special attention was paid to exploring and measuring viral repetition. Finally, we included mid-range circulation to cover the

7 Text reuse data can be explored via a search interface at http://comhis.fi/clusters.

wide variety of reprinting between rapid and slow circulation. Table 1 presents key features of these three cycle types, forming our ontological considerations.

Importantly, these cycles of reprinting are not all mutually exclusive. A reprinted text could first spread very rapidly, "go viral", and then get reprinted again after a considerable time although this was untypical.

Table. 1: Distinctive features of the three reprinting cycle types.

Reprinting cycles	Rapid repetition (including viral circulation)	Mid-range repetition	Long-term repetition
Timespan	within a year; in viral cases: days or weeks	several years, ranging from 1 to 49 years	50 years or more
Share of all clusters	85,29 % (11,768,371 clusters)	14,67 % (2,023,615 clusters)	0,04 % (5,888 clusters)
Movement	synchronic, geographic (fast, incl. viral cases)	diachronic (both regularly and sporadically repeated texts)	diachronic (slow)
Contents, typically	wide range, from ads to news (incl. boilerplate)	wide range, from announcements to religious stories (incl. boilerplate)	news, literary works

2.1 Rapid Circulation of Information

A distinct category of text reuse is the rapid circulation of information. In the time period studied, this was mainly horizontal: news traveled in the geographical space within a short period of time. In the Viral Texts project, Smith, Cordell, and Maddock Dillon analyzed nineteenth-century American newspapers and divided the reprinted texts into fast and slow ones. In their fast set, the median lag time was under one year.[8] In our findings, most of the chains were also realized within twelve months, which comprised 85 percent of all reuse clusters, and these can be seen as fast repetition.[9]

8 Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf, 93.

9 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019.

We can define texts that spread within twelve months as being "fast," but not all of these will be viral texts. In present-day discourse, the word "viral" is regularly used in connection to social media, in describing the rapid sharing of media content: rapidity is its essential feature. The roots of "virality" are obviously in microbiology and medicine, and the concept refers to the capacity of viruses to replicate in host cells and cause diseases in an epidemic curve. In public discourse, "viral" suggests that something is "like a virus" or "spreads like a virus."[10] Although the concept has a contemporary undertone, it can also be applied to historical material: in the nineteenth century, media capacity grew exponentially. In Finland, there were so many publications all around the country towards the end of the century that it created a resonance base for viral information flows. We decided to measure this capacity by paying attention to printing locations around the country and the number of newspaper titles.

To be able to filter the results, we defined a *virality score*, reflecting the diffusion of the clusters and measuring both rapidity and capacity (in relation to the volume of the press and to the geographical coverage of the cluster). This score is calculated by multiplying the number of different publication titles within a cluster and the number of unique printing locations by the inverse of the elapsed time. Here, "elapsed time" means the length of the cluster in days. The virality score helped us to capture how many titles the information spread into and how many places it geographically reached, and then penalized the value the slower the process was.[11] After this, the values of clusters were normalized into the range 0–100. The most viral text proved to be a paid advertisement by the Finnish tobacco industry against American cigarettes, published between 1 and 31 March 1916 in 45 different newspapers or journals in 26 different locations, 75 times altogether.[12]

The use of a virality score in our study was an experiment in measuring virality in a historical setting. It can be further refined into a tool that helps in understanding the rhythms of information flows and how they saturated the prevailing media system. Of course, this can be done only within the limits of

---

Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf, 394–404.

10 Oxford Dictionary of English. 2015. 3rd online edn. Oxford: Oxford University Press. DOI: 10.1093/acref/9780199571123.001.0001. On virality, see H. Salmi, 2020. *What is Digital History?* Cambridge: Polity, 22–25.

11 The virality score is discussed more in detail in Salmi *et al.* 2021. For the code, see https://github.com/avjves/cluster-viral-score.

12 Cluster no. 11592519, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=11592519, accessed June 27, 2020.

the OCR corpus, which means that the possible external sources of information could not be automatically detected. The score, however, helped in the qualitative analysis of results. In the case of the most viral texts, like the tobacco ad, it might be questionable whether paid advertisements, the impetus of which came outside of the media sphere, can be regarded as viral cases at all, but on the other hand this was illuminating from the perspective of an ontology of text reuse. In our material, only 81 clusters had a virality score higher than 50. These cases also include journalistic content, but almost all of them seem to have had sources outside newspaper editorial offices. It is obvious that media capacity enabled the viral spread of news, and this shift was understood by contemporaries who took advantage of the efficiency of the printing press.

An alternative way of approaching virality in the study of nineteenth century press could be to concentrate on reachability, or cluster spread. This would mean emphasizing geographical coverage and scope of the movements of reprinted texts within a chosen window of time (say, one year). In the project we thought about this option too but chose to concentrate on rapidity of reprinting. It might however be that reachability could better reflect the conditions of the printed press before the late nineteenth century, and in further studies it might be especially promising when studying border-crossing flows of news. Exploring the geographical spread of clusters as complementary to rapidity could improve our understanding of virality as a historically changing phenomenon.

2.2 Long-Term Cycles of Reuse

One of the valuable characteristics of the Finnish digital newspaper corpus is its long timespan. The corpus offers a view into Finnish newspaper publishing in its full scale from the very first publications of the 1770s to the era of wide national coverage of the press in the 1920s. In the beginning of our project, we did not really have any hypotheses on the temporal scale of text reuse. In this respect, it was a surprise that the longest chains of reuse covered over 140 years, almost the whole timespan of the corpus. We have analyzed this feature of the press, the so-called long-term reuse, in detail elsewhere.[13]

13 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf. For an article in Finnish on the results, see H. Rantala, H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen

The discovery of long-term reuse – 50 years or more – reminds us of different temporalities that have been present in the press. In addition to news material and editorial texts, the papers have comprised a great variety of repetitive content, such as advertisements and announcements. This is not a Finnish phenomenon as such, but we are not aware of other studies that have focused on the practices of slow publishing or on the reuse of texts within as long a time span as a century.

In the case of long cycles of reuse, the time lag between the first appearance and republication can be counted in several decades. Sometimes, although rarely, the span of reuse is even one hundred years or more. This means that the papers have benefited from material from the earlier publications and republished both extracts and full stories or news from the old newspapers. Thus, past publications have been used as an archive of possible material for contemporary newspapers. With its historical depth, the press participated in constructing a sense of community. Referring to Pierre Nora's notion *lieux de mémoire*, we have earlier suggested that the press could be understood as a site of memory.[14] On the whole, this type of text recycling is not a dominant form of reuse in the Finnish press but it is worth attention. If rapid reuse was often horizontal movements in geographical space, long-term reuse was vertical or diachronic: information that traveled in time. This does not mean that long-term reuse would realize without geographical spread, but time seems to be its dominant property: this includes cases where the same item has been republished by the same paper later on, again and again, and cases where the geographical spread is not characterized by the saturation of the capacity of the press but evolved rather randomly in time.

We discovered that, within these long-lasting cases of text reuse, there is variation in the actual cycles of republishing. While some texts had been reused only once or twice during a very long period of time, there have also been cases in which the old text was activated several times during different decades. One

ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179.

14 For further details, Rantala, H., A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, F. Ginter. 2019. "Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma." Historiallinen Aikakauskirja 1: 53–67; Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf. See also P. Nora. 1997. "General introduction: Between memory and history." In (P. Nora, ed.) The Realms of Memory: Rethinking the French part, Vol. 1: Conflicts and divisions. Columbia University Press: New York, 1–21.

clear reason for this kind of cyclic reuse is anniversaries of certain culturally and/or politically loaded events. For example, in 1863 the gathering of the Finnish Estates was a major event since the Emperor of Russia had not summoned the Diet in over 50 years. Many Finnish newspapers published the opening speech of the Tsar at the Diet meeting. During the following decades, the same speech was republished every now and then by several papers. Then, in 1913, 50 years after the historical Diet meeting, the words of the Emperor were republished by over 40 Finnish newspapers to honor the anniversary.[15]

Apart from the above-mentioned remembrance of different national anniversaries or other important dates, newspapers and journals recycled a variety of old material. Among the clusters of long-term reuse, there are several anonymous stories or anecdotes, as well as old news clippings, which have probably been reprinted for the sake of amusement and curiosity. Furthermore, we have found examples in which the reprinting of old material had a clear connection to the contemporary culture, for example, of the political life of the country; for censorship reasons, it was sometimes impossible to describe the current state of affairs, but contemporary concerns could be implied by reprinting old content.[16]

On the whole, the slow cycles of reuse represent only a very small amount of texts of the corpora in question (see Table 1). This feature of the press is nevertheless interesting and offers the possibility to rethink those temporal scales in which the newspapers have operated. The existence of publishing cycles that have covered a century or more demonstrate the manifold functions of the press. Along with the topical functions, the press has many other scopes, including those that strengthen our ability to remember the past.

2.3 Mid-Range Text Circulation

In addition to rapid circulation and long-term use of texts, text reuse detection also revealed clusters that escaped both categories, or did not completely fit into either of them. We call this gray area *mid-range repetition*. The mid-range cycles comprise a significant share of total reuse cases. Their amount can be measured in different ways: it can range from anything over a year to those just under 50 years, thus covering a variety of reuse cycles.

15 Cluster no 6216311, http://comhis.fi/clusters/?f%5Bcluster_id%5D%5B%5D=6216311, accessed June 27, 2020.

16 Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. "The Long-Term Reuse of Text in the Finnish Press, 1771–1920." In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

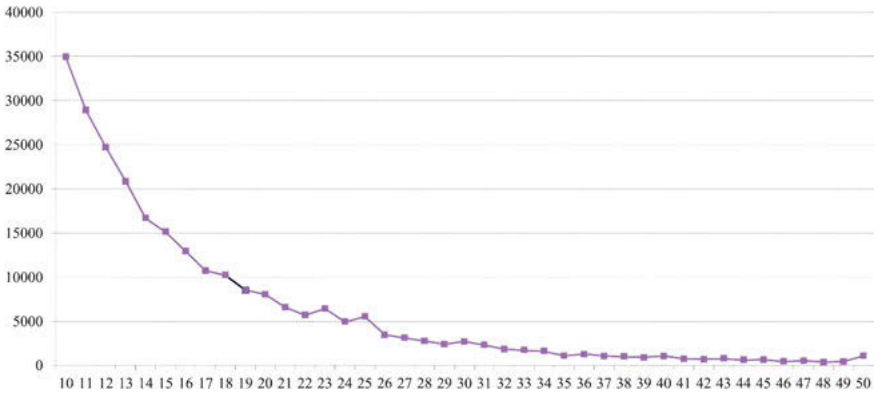


Fig. 1: The number of clusters in relation to their span (10 to 50 years). Source: comhis.fi, accessed September 8, 2022.

Instead of reuse cycles lasting a span of 1–49 years, we began to examine these cycles with simpler time frames, basically asking which kinds of texts were copied infrequently within 5 to 10, or even 20 years. To illustrate the distribution of clusters of different timespans, Fig. 1 shows clusters whose spans range from 10 to 50 years.

In our earlier publications, we only paid attention to fast and slow repetition, in a similar way to Smith, Cordell and Maddock Dillon.[17] However, after considering the results further, it seems clear that the dichotomy between fast and slow gives too limited a view on the types of text reuse. What we call 'mid-range text reuse' represents a significant, albeit often overlooked, practice of newspaper production. Mid-range repetition is mainly about undramatic content. It is easily forgotten especially compared to the texts encountered in long-lasting circulation chains. Studying mid-range circulation may help us to highlight the characteristics of other text circulation cycles. It may also be a fruitful entry point to further nuance the understanding of the forms of text reuse in the nineteenth century.

Typical cases of mid-range text reuse were official announcements using standard sentences annually and even across decades, and other announcements, such as for a research scholarship, published regularly. There were also

17 Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. "Infectious texts: Modeling text reuse in nineteenth-century newspapers." In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf, 93.

commercial, textual advertisements published repeatedly over several years either in one location or in several regions. This may also include texts that were reprinted by the same newspaper again and again. The virality score penalizes texts that were not printed by several papers, but these texts still have a significant role in newspaper publishing.

Religious texts and biblical stories were typically circulated in newspapers, and their long-lasting messages made them highly reusable. Their movements from paper to paper took years and even decades. Other short stories and educational texts are also among those that newspapers occasionally republished or borrowed from each other. For instance, one article encouraging farmers to use reindeer lichen and certain natural plants for fodder, or to add lichen to fodder when their crop of hay had failed, was circulated in newspapers at least 83 times around Finland during eight years from 1894 onwards.[18]

Further, the detection process has helped to recognize certain types of texts, or printing conventions, in which exact phrases were repeated. These were, for instance, strings of obituaries of different persons, spanning many years and repeating phrases (such as "died believing in the Savior"). These were then identified as similar and collected in clusters of text reuse. Since the detection process was set to find similarities that are over 300 characters, this means that the phrases were accompanied by other phrases. These can be regarded as boilerplate text, which we aimed to exclude by ignoring similarities under 300 characters. This did not completely succeed, however, since template phrases formed a sequel which was interpreted by BLAST as a larger textual whole. On the other hand, these texts were closely related to the development of newspapers as a media platform since, through these text templates, media visibility was sold and offered for various purposes. The templates were often not reused long-term, meaning decades or fifty years; they were shorter, but in many cases, long-lasting chains of repetition that tell of the changes in public writing patterns and also of how Finnish organizations learned to interact with and utilize the press effectively.

18 P. Paju. 2019. "Jäkälän paluu: Jäkälävalistus ja tekstien uudelleenkäyttö historiallisen tutkimusteeman jäsentäjänä. (Return of the Lichen. Lichen education and outlining a historical research topic by studying text reuse.)" Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41942380.

3 Epistemological Ramifications

Ontology and epistemology are always intertwined in research. Our ability to know about the past, about past entities and "forms of being," is conditioned and framed by our retrospective position, by the available sources, and by the methods selected for research. These conditions, we believe, do not prevent us from making conclusions on the ontological premises, but it is necessary to ponder how these conditions direct our perspective and influence our findings. The following discussion on the epistemological ramifications of text reuse detection is based on three aspects that need to be articulated. These are the materiality of the digital, the problem of metadata, and the question of methodological nationalism that must be considered when national and regional corpora are used.

3.1 Materiality of the Digital

When studying digitized periodicals as sources, the material conditions that framed the development of the press have to be kept in mind. This involves the materiality of newspaper publishing itself, and its many changes, in the nineteenth century.[19] Previous research has explored the development of text layout, including pagination, font size and number of columns as well as on material proportions, such as paper size and quality over this period.[20] We also studied the relationship of text reuse patterns with the growing number of characters per page and larger page sizes in the late nineteenth century, which shows that text reuse grew more moderately than would be indicated by absolute numbers of clusters in our results.[21] The absolute number of reuse clusters grew rapidly towards the end of the research period, which reflects the fact that the volume of the press increased accordingly. Paper size enlarged, especially

19 On previous discussion on source criticism, see for example, R. Abel. 2013. "The Pleasures and Perils of Big Data in Digitized Newspapers.". Film History, 25.1–2: 1–10. Milligan I. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." The Canadian Historical Review, 94.4: 540–69. M. Koolen, J. van Gorp, J. van Ossenbruggen. 2019. "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice." Digital Scholarship in the Humanities, 34.2: 368–85.

20 J. Marjanen, V. Vaara, A. Kanner, H. Roivainen, E. Mäkelä, L. Lahti, and M. Tolonen. 2019. "A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917." Journal of European Periodical Studies, 4.1: 54–77.

21 See the figure in Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. "Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920." Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179, 60.

in the 1880s and 1890s, while at the same time font size became smaller. This allowed even more information to be printed by the press. It must be noted, however, that the smaller the font size became, the more difficult it has been for the OCR to recognize the text. Therefore, the OCR accuracy is lower in the 1880s and 1890s than in the 1850s and 1860s, and likewise, there is more OCR noise towards the end of the period. From an epistemological perspective this is fascinating, since – although text reuse detection finds exponentially more results from the period compared to previous decades – it may well be that the real amount of actual reprinting cases is even higher than what could be retrieved with this selected method.

Further to this, the digitization process of newspapers is also impacted by material aspects that influence the results. In many countries, the preservation of old and fragile newspapers started with microfilming projects, since microfilm surrogates were regarded as a stable means for archival preservation.[22] The use of microfilm would also reduce the use of actual newspapers, which again would help the originals to be preserved. When digitization of Finnish newspapers started in the 1990s, it was done mostly on the basis of these microfilms, which was a technically inexpensive and practical solution. This allowed the National Library of Finland to proceed efficiently on the project. The first online collection was opened in 2001, and today all issues published prior to 1920 have been digitized and opened for researchers. The essential epistemological feature lies in the fact that optical character recognition was done on the images taken from the microfilm, not from the original newspaper.[23] If one compares the Finnish development to Sweden, for instance, many historical newspapers there have only been digitized for the first time in 2020. Because the National Library of Finland acted early on in large-scale OCR processing, the present collection contains a lot of noise and random, erroneous characters created by the OCR software in the

22 See, for example, A. Prescott. 2018. "Searching for Dr Johnson: The digitisation of the Burney newspaper collection." In (S. Gøril Brandtzæg, P. Goring and C. Watson, eds.) Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, Brill: Leiden, 49–71, 57.

23 On the history of the collection, see Bremer-Laamanen, M. 2006. "Connecting to the past – newspaper digitisation in the Nordic countries." Journal of Digital Asset Management, 2(3–4): 168–171. See also M. H. Beals, and E. Bell, with contributions by R. Cordell, P. Fyfe, I.G. Russell, T. Hauswedell, C. Neudecker, J. Nyhan, M. Oiva, S. Padó, M. Peña Pimentel, L. Rose, H. Salmi, M. Terras, and L. Viola. 2020. The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges. Loughborough. DOI: 10.6084/m9.figshare.11560059. https://www.digitisednewspapers.net/.

1990s and the early 2000s.[24] Overcoming this problem of OCR inaccuracy, in fact, motivated our project to study text reuse in the first place, and to choose an approach that is tolerant to noise. According to Aleksi Vesanto, text-reuse-BLAST can recognize similarity even if 40 percent of the characters are wrong.[25] Our chosen method thus enabled the project to circumvent the problem and also had the advantage of analyzing a corpus that is complete in relation to the real volume of newspaper publishing.

In these ways, the quality of OCR in the Finnish digital corpus still carries features of its production history, which again conditions the ways in which we can know about text reuse, and the past in general. It has to be added that this production history is an open process, and there are ongoing development actions to improve OCR accuracy; re-digitization is even being considered with selected newspapers.

3.2 The Problem of Metadata

Another aspect of our epistemological assessment deals with the metadata that is available for the OCR corpus of Finnish newspapers. The quality of the metadata used is in general very good, with exact timestamps and the names of publications. There are rare cases of metadata mistakes, but they do not impact the results in this large-scale exploration. There are other features, however, that do influence the investigation.

Most importantly, the Finnish OCR corpus is not segmented, which means that the metadata does not include information on how the texts are divided into different forms of content, like editorials, news items, advertisements, obituaries, and so on. The OCR corpus includes newspaper and journal issues as separate folders that include pages as XML files. This arrangement has two sides: on the one hand, all forms of newspaper publishing are mixed together and cannot be automatically separated from each other. On the other hand, page breaks form ruptures within the material in a way that influences the analysis. If there were segmentation, it would be easier to connect, for example, broken-up news stories,

24 On the problem of OCR noise, see J. Jarlbrink, and P. Snickars. 2017. "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive." *Journal of Documentation*, 73(6), 1228–43.

25 Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter. 2017. "Applying BLAST to Text Reuse Detection in Finnish News-papers and Journals, 1771–1910." In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

the beginning of which is on a different page from the end. In newspapers, the continuation is not necessarily on the next page. Our reuse clusters could perhaps be employed in developing categorization of the contents and in enhancing the metadata, but that is a future prospect.

The fact that the OCR corpus consists of pages influences the estimation of the extent of text reuse. Our text-reuse-BLAST identifies similar passages of text, but page breaks confuse the results. If, for example, an article has been printed twice, and in the first instance it is divided into two pages but in the second instance it is on one page, the algorithm finds two reuse clusters instead of one, since it treats the beginning and the end as different passages. Otherwise, text reuse detection has no upper limit: it identifies similarity as long as it remains similar, but page breaks interrupt the process of recognition. In our epistemological consideration, the previous point on the substantial OCR noise prevents the algorithm of finding all reuse cases, but here the trend is the opposite: page breaks have produced more text reuse clusters than would have been the case if the problem of page breaks had been overcome. This is a problem of course, but it does not undermine the usefulness of our reuse cluster database. It does mean, however, that the absolute numbers of clusters have to be discussed with reservations.

Let us return to the question of the boilerplate, which was mentioned in the earlier discussion on mid-range reuse. Text-reuse-BLAST does not require lower or upper limits for the recognition of similarity, but we set a lower limit of 300 characters to avoid too much boilerplate content. OCR noise also makes it problematic to quantify very short reuse cases. This also relates to the question of metadata. The metadata of the corpus includes information on images on the pages, but many graphic aspects of the paper have been ignored. This leads to the situation that in the OCR corpus smaller text passages might be connected together, although in the actual newspaper they were separated with a visual element. Thus, separate boilerplate texts might aggregate so that their sum exceeds 300 characters, and in the end mix the results. There are advantages too, since this makes boilerplate elements more perceptible in the overall interpretation of text reuse, which is historically fair, considering how much repetitive content there actually was in the newspapers of the past.

In addition to these points, another aspect of metadata must be mentioned. The OCR corpus of the National Library of Finland includes, as already noted, both newspapers and journals. We included all periodical press in the analysis to make it more comprehensive. There lies a problem, however, since some of the journals did not have an exact date of publication, at least to the day. In these cases, the timestamp is not as accurate as in the case of newspaper issues. This means that the time spans of the clusters might be

inaccurate in those cases where one of the publications of the reuse chain was a journal without an exact date. While this creates a problem, it seems that the participation of journals into reuse chains was limited, so this issue does not harm the results of the project much.

3.3 Methodological Nationalism and Digital History

In the last perspective in our epistemological assessment, we would like to draw attention to a more general question, the tendency of researchers to see nation-states as units of analysis, which has been called methodological nationalism.[26] Although the digitization of newspapers, and the creation of digital newspaper corpora have been a welcome and much-appreciated development for historians, digital newspaper history includes challenges that are embedded in the very processes of large digitization projects. As a rule, newspaper corpora are nationally, sometimes regionally, siloed collections provided by an operator/actor responsible for the collection and preservation of the national heritage.[27] As such, digital collections do not differ from other national collections available in the museums and libraries. The easy accessibility of digital corpora, however, might obscure the fact that from the perspective of transnational history, digital newspaper corpora are often biased by design.

For example, in the case of the Finnish history of print culture, one cannot understand the developments in the field without a wider context: Finland only became an independent state in 1917, after having been a part of the Swedish Kingdom until 1809, and a Grand Duchy within the Russian Empire between 1809 and 1917. The early decades of what is regularly defined as the Finnish newspaper history actually also belong to Swedish press history, since Finland was a province of the realm of Sweden. The Finnish National Bibliography regards the *Tidningar Utgifne Af et Sällskap i Åbo*, published in Turku, present-day Finland, in 1771 as the first Finnish newspaper. It was published in Swedish, as most Finnish newspapers were until the mid-nineteenth century. The present Finnish territory was detached from Sweden in 1809, but the

26 Wimmer, A., Schiller, N. G. 2003. "Methodological Nationalism, the Social Sciences, and the Study of Migration: An Essay in Historical Epistemology." The International Migration Review. 37 (2): 576–610. https://www.jstor.org/stable/30037750.

27 On the siloed nature of newspaper collections, see for example M. Ehrmann, E. Bunout, and M. Düring, 2017. Historical Newspaper User Interfaces: A Review. http://library.ifla.org/2578/1/085-ehrmann-en.pdf.

Swedish press stayed as an important reference point and source of information for Finnish newspaper editors. One clear reason for this was the common language: until the second half of the nineteenth century, Swedish was also a dominant language in Finnish newspaper publishing. When literacy among the Finnish-speaking population increased towards the end of the century, the volume of the Finnish-language press also grew in a rising curve.

In Finland, and certainly everywhere, newspapers had an active role in sharing information across borders. An epistemological problem lies in the issue that it is difficult to examine the role of transnational news flows with nationally siloed corpora, which tend to undermine these cross-border news flows. News traveled across state borders, and before the establishment of news agencies or electronic telegraph lines, other papers were often important sources of information, particularly for foreign news. For the historian interested in information flows, it is therefore useful to draw from several newspaper corpora and to try to find ways of combining them. In our project, we analyzed several cases of text reuse, where the chain of circulation actually started outside the Finnish corpus.[28] This can be done by combining qualitative close reading methods with computational tools. If one only draws on a nationally limited or restricted collection, there is a danger of strengthening the methodological nationalism that is inherent in the building of national cultural heritage collections. This bears not only on the issue of sources but also methods. Our text reuse detection method could not identify similarity across language borders, meaning that, for example, a news item published first in Swedish and then in Finnish could not be automatically clustered together. We identified many of these through close reading of clusters, but this method of detection could not be done on a corpus level. Machine translations are not a solution either, since OCR noise makes this more than challenging. This is an avenue for future research in newspaper corpora, as there is an urgent need to combine national and regional collections and find ways of identifying information flows across linguistic borders.

28 See, for example, Salmi, H., A. Nivala, H. Rantala, R. Sippola, A. Vesanto, F. Ginter. 2018. "Återanvändningen av text i den finska tidningspressen 1771–1853." Historisk tidskrift för Finland 1: 46–76.

4 Conclusion

This article has been an effort to articulate ontological and epistemological issues in text reuse detection, and contribute to the wider heuristic discussion on digitized newspaper collections. In the research project *Computational History and the Transformation of Public Discourse in Finland*, we did not start from specific research topics or preconditions, but from the newspaper corpus itself. As a collaboration between historians and data scientists, we developed a special method, text-reuse-BLAST, which proved to be highly productive in aligning similar passages in the OCR corpus. We produced a new dataset that included all reuse clusters and provided for the possibility of further exploring the movements and routes of information. Thus, the project has enriched the original digital collection by providing new entry points to the history of Finnish media. Internationally, the Finnish case is illuminating since newspaper publishing was in its state of emergence throughout the nineteenth century. Up to the 1850s, there were only very few printing locations, but thereafter there was rapid development. By 1920, the periodical press was already a network that covered the whole country, including a distinct division of labor between newspapers on a local level and nationally. Our method can also be applied elsewhere in the effort to understand how newspaper publishing developed over a long period of time. The ontological approach introduced in this article could similarly be useful in other national and international settings. We are continuing this research in the project *Information Flows over the Baltic Sea*, where we combine the Swedish-language newspaper collections from Finland and Sweden to understand transnational news flows.[29]

In our project we found out that text reuse characterized the whole period under study and was not only a phenomenon of the rise of the press in the late nineteenth century. From an ontological perspective, it is essential that news travelled both synchronically and diachronically, in geographical space but also in time. The ontology of reuse can be further developed in the future, for example, by concentrating on the "gray area" we described as mid-range repetition. These reuse cases can shed more light on the changes in the publishing practices and thus help to understand how newspapers and journals were culturally positioned and re-positioned in Finland from the late eighteenth century to the early twentieth century. By revealing how much the content of the newspapers

29 *Information Flows across the Baltic Sea: Swedish-language press as a cultural mediator, 1771–1918* (The Society of Swedish Literature in Finland, 2020–2023), https://blogit.utu.fi/informationsfloden/.

was shared, text reuse detection can, for instance, be used in highlighting what, if anything, made the individual papers stand out among the press. Reprinting patterns can further inform us what was *not* reprinted, and prompt us to ask why this was the case. These omissions of reprinting could reflect emerging political or other divisions in the press.

In the epistemological part of the study, we aimed to track down aspects that influence, and may bias, the results found with computational tools, drawing on existing corpora. The more general aim of these considerations has been to try to shift the discussion on a broader level to issues that are relevant for any project on historical newspapers in the effort to try to understand how they circulated information and, especially, how we, and under which conditions, can know about the past.

To conclude, there is an existential question embedded in our study and its results: How do databases produced in fixed-term research projects survive in the long run? For such datasets as the clusters of text reuse to remain usable, they need maintenance and preferably some improvements over time, all of which might be difficult to sustain in a research environment based mostly on external funding. In similar undertakings, this is a vital perspective to conceive already in the outset. Projects in computational history need long-term research infrastructures both nationally and internationally, in order to secure the sustainability of the field in the future.

Database

Vesanto, A., F. Ginter, H. Salmi, A. Nivala, R. Sippola, H. Rantala, and P. Paju. 2018. Text Reuse in Finnish Newspapers and Journals, 1771–1920, http://comhis.fi/clusters.

Bibliography

Abel, R. 2013. "The Pleasures and Perils of Big Data in Digitized Newspapers.". Film History, 25.1–2:1–10.

Beals, M. H. and E. Bell, with contributions by R. Cordell, P. Fyfe, I.G. Russell, T. Hauswedell, C. Neudecker, J. Nyhan, M. Oiva, S. Padó, M. Peña Pimentel, L. Rose, H. Salmi, M. Terras, L. Viola. 2020. The Atlas of Digitized Newspapers and Metadata: Reports from Oceanic Exchanges. Loughborough. DOI: 10.6084/m9.figshare.11560059. https://www.digitisednewspapers.net/.

Bremer-Laamanen, M. 2006. "Connecting to the past – newspaper digitization in the Nordic countries." Journal of Digital Asset Management, 2(3–4): 168–171.

Büchler, M., G. Crane, M. Moritz, A. Babeu. 2012. “Increasing recall for text re-use in historical documents to support research in the humanities.” In: (Proceedings) Second International Conference on Theory and Practice of Digital Libraries, vol 7489, pp. 95–100. Doi: 10.1007/978-3-642-33290-6_11.

Ehrmann, Maud, Estelle Bunout, and Marten Düring. “Historical Newspaper User Interfaces: A Review.” In *Proceedings of the 85th IFLA General Conference and Assembly*, 1–26. Athens, Greece: IFLA Library, 2019. https://doi.org/10.5281/zenodo.3404155.

Jarlbrink, J. and P. Snickars. 2017. “Cultural heritage as digital noise: nineteenth century newspapers in the digital archive.” Journal of Documentation, 73(6),1228–43.

Koolen, M., van Gorp, J., van Ossenbruggen, J. 2019. “Toward a Model for Digital Tool Criticism: Reflection as Integrative Prac-tice.” Digital Scholarship in the Humanities, 34.2: 368–85.

Marjanen, J., V. Vaara, A. Kanner, H. Roivainen, E. Mäkelä, L. Lahti, and M. Tolonen. 2019. “A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917.” Journal of European Periodical Studies, 4.1: 54–77.

Milligan, I. 2013. “Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010.” The Canadian Historical Review, 94:4: 540–69.

Nora, P. 1997. “General introduction: Between memory and history.” In (P. Nora, ed) The Realms of Memory: Rethinking the French part, Vol. 1: Conflicts and divisions. Columbia University Press: New York,1–21.

Oxford Dictionary of English. 2015. 3rd online edn. Oxford: Oxford University Press. DOI: 10.1093/acref/9780199571123.001.0001.

Paju, P. 2019. ”Jäkälän paluu: Jäkälävalistus ja tekstien uudelleenkäyttö historiallisen tutkimusteeman jäsentäjänä. (Return of the Lichen. Lichen education and outlining a historical research topic by studying text reuse.)” Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41942380.

Prescott, A. 2018. “Searching for Dr Johnson: The digitisation of the Burney newspaper collection.” In (S. Gøril Brandtzæg, P. Goring and C. Watson, eds.) Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, edited by. Brill: Leiden, 49–71.

Rantala H., H. Salmi, A. Vesanto, F. Ginter. 2019. ”Tekstien pitkä elämä: Ajassa liikkuvat tekstit suomalaisessa sanomalehdistössä 1771–1920.” Ennen ja nyt (history journal online) 2/2019, https://research.utu.fi/converis/portal/Publication/41858179.

Rantala, H., A. Nivala, H. Salmi, P. Paju, R. Sippola, A. Vesanto, F. Ginter. 2019. “Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma.” Historiallinen Aikakauskirja 1: 53–67.

Salmi, H. 2020. What is Digital History? Cambridge: Polity.

Salmi, H., A. Nivala, H. Rantala, R. Sippola, A. Vesanto, F. Ginter. 2018. “Återanvändningen av text i den finska tidningspressen 1771–1853.” Historisk tidskrift för Finland 1: 46–76.

Salmi, H., H. Rantala, A. Vesanto, F. Ginter. 2019. “The Long-Term Reuse of Text in the Finnish Press, 1771–1920.” In: (Proceedings) 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark 6–8 March 2019: 394–404, http://ceur-ws.org/Vol-2364/36_paper.pdf.

Salmi, H., P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter. 2021. “The Reuse of Texts in Finnish Newspapers and Journals, 1771–1920: A Digital Humanities Perspective”. Historical Methods: A Journal of Quantitative and Interdisciplinary History 54 (1): 14–28, https://doi.org/10.1080/01615440.2020.1803166.

Smith, D.A., R. Cordell, E. Maddock Dillon. 2013. “Infectious texts: Modeling text reuse in nineteenth-century newspapers.” In: (Proceedings) Workshop on Big Humanities. IEEE Computer Society Press 2013, 86–94, http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf.

Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, F. Ginter. 2017. “Applying BLAST to Text Reuse Detection in Finnish News-papers and Journals, 1771–1910.” In: (Proceedings) 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017 (Linköping 2017), http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.

Wimmer, A., Schiller, N. G. 2003. “Methodological Nationalism, the Social Sciences, and the Study of Migration: An Essay in Historical Epistemology.” The International Migration Review. 37 (2): 576–610. https://www.jstor.org/stable/30037750.

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

Estelle Bunout

Contextualising Queries: Guidance for Research using Current Collections of Digitised Newspapers

Abstract: This chapter presents a snapshot of current research practices using digitised newspapers, with a particular focus on the strategies of contextualisation of the queries conducted in these collections. The question of the contextualisation seems to be a common preoccupation for researchers in the humanities, while remaining difficult to define, without a core understanding of how the digitised newspapers collections have been produced and published. Here, we propose some pointers for practical contextualisation of queries in these collections: how to recreate the context of the digitised historical source by understanding how collections are built, what happens when they query them. Beyond the keyword search, other dimensions of information made available by the digitisation, can inspire researchers to create another type of context: a content-based one. The paper reviews selected practices of such content metadata created in currently available collections.

Keywords: digital source criticism, digitised newspapers, interface critique

1 Introduction

Summarising the uses of historical digitised newspapers is as vain as trying to summarise their contents. However, whatever the use, some challenges remain common and above all, the structural elements of this source remain the same, independently of the research question. While the digitisation changes how one collects and reads individual articles, mainly via queries, it can also help bring forward these structuring elements. Indeed, the attraction of this source relies on its frequency and diversity of contents, which enable both longitudinal and

Acknowledgments: This chapter is the result of a workshop organized as part of the *impresso* project in January 2020 at the University of Luxembourg. We thank the participants and the *impresso* team members for the rich exchanges, Sally Chambers for raising awareness on the conception of newspapers collections as data as well as Maud Ehrmann, Marten Düring and Risto Turunen for their insightful feedback on this text. The project `*impresso* - Media Monitoring of the Past' has been supported by the Swiss National Science Foundation under grant CR-SII5_173719.

https://doi.org/10.1515/9783110729214-013

targeted, as well as local and national queries. Newspapers are used by many researchers, either as a central or complementary source, or as a substitution for missing archives on actors that left few traces of their own (and who are beyond the reach of oral history), in order to tackle research questions ranging from analyses of language, discourses, and iconography in the fields of political, cultural, technical, social and of course, media history. The size remains the most challenging dimension, and almost inevitably (and fortunately) implies that such collections are kept by institutions with the goal of preserving cultural heritage, be it the analogue or digitised version of newspaper collections. At the same time, the institutional frame for the preservation and publication of digitised newspaper collections makes the existence of platforms to query and interact with these collections possible, with the aforementioned allure of the keyword-based exploration of newspapers – the "bottom-up" approach,[1] the unexpected associations, and the existence of polysemy for a given term. This heuristic potential is often underlined for the digitised newspapers collections, followed typically by the concern of the interpretability of the results hits and the contextualisation of each identified item. To answer this questions, one needs to look at several layers of context: the institutional and technical context in which a collection has been prepared and published and of course, the historical context of the source. The collections of digitised newspapers, i.e., the holdings of a particular institution, have been prepared and published in various formats. For instance, the collection of Austrian newspapers has been made available on a dedicated portal (ANNO) by the Austrian national library.[2] It can also be accessed via another portal where users can visualise and interact, via graphs, with the metadata of this collection (ONBLabs).[3] This platform hosts output of research, that has been conducted using the Austrian digitised newspapers collection.[4] Finally, the collection has also been shared within a transnational research project, NewsEye,[5] where it has been enriched with, among other things, named entities.

1 Erik Koenen (2018). "Digitale Perspektiven in der Kommunikations- und Mediengeschichte." In: *Publizistik* 63.4, pp. 535–556. URL: https://doi.org/10.1007/s11616-018-0459-4.

2 https://anno.onb.ac.at.

3 https://labs.onb.ac.at/en/topic/historic-newspapers.

4 See Monika Kovarova-Simecek (2022). "Kulturgeschichte der Popularisierung von Börsennachrichten in Wien (1771–1914). Eine historische Analyse unter Anwendung von ANNO/ONB und ONBLabs." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History inthe Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter.

5 https://www.newseye.eu.

This example illustrates another type of versatility of use for digitised newspaper collections: on top of the news articles, obituaries, wedding announcements, fiction, and advertisements that are published in conservative, liberal, socialist, and business-oriented newspapers, a wealth of tools and portals open these collections up for researchers to explore them and create their own research corpus, namely to select items to discuss their research question. The massive efforts made to digitise newspaper collections that have been conducted in past decades have triggered fruitful discussions among historians, ranging from awe in the hope of accessing the "big unread", to worry concerning the risks of the new biases introduced by digitisation, such as a lack of standards for the process, issues surrounding the representation of the titles being digitised, diverging quality in metadata, and Optical Character Recognition (OCR).[6] In parallel, the past few decades have witnessed an intensive use of these collections, while adapting to these constraints. In summary, transposed onto digitised newspapers, the question of building an apparatus for digital source criticism[7] can be broken down into three dimensions: the interface and query criticism; the appropriation of digitised documentation (policies, processes, and outputs); and the reflection on the "mechanisation of the heuristic proces",[8] meaning the translation of research questions into interaction with the source material.

We propose hereunder to materialise these general questions of source criticism into a practical questionnaire that can help researchers situate their research corpus. This paper is not so much a wish list of what the digitised newspapers collections should offer, but rather a snapshot of what has been done and a digest of the lessons learned from past decades of research. We rely on a selective review of the research produced in the field, as well as an expert workshop that was organised in January 2020, in the context of the *impresso* project,[9] where a dozen of researchers gathered to share their experiences and visions on their research using digitised newspapers. Using this experience, we look at the strategies researchers have deployed to contextualise their findings and make this rich source

6 Erik Koenen (2018). "Digitale Perspektiven in der Kommunikations- und Mediengeschichte." In: *Publizistik* 63.4, pp. 535–556. URL: https://doi.org/10.1007/s11616-018-0459-4.

7 Andreas Fickers, Dana Mustata, and Anne-KatrinWeber (2019). "The Rise of Television." In: *The Handbook of European Communication History*. John Wiley & Sons, Ltd, pp. 239–255. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119161783.ch13.

8 Huub Wijfjes (2017). "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." In: *Tijdschrift voor Mediageschiedenis* 20.1, pp. 4–24. URL: http://www.tmgonline.nl/index.php/tmg/article/view/277.

9 https://impresso-project.ch.

"amenable" to their research questions.[10] We now look at how digitisation has impacted these research practices, before looking at what these collections contain in order to finally understand how the collection tools, and their enrichment, can support an academic use of these collections.

2 "Quietly Central" Changes of Research Practices by the Digitisation of Historical Newspapers

The amalgamation between collections and interface has sparked calls for caution,[11] pointing at bias in selection, the decontextualization of queries, OCR mistakes, etc. Several papers have discussed the strategies for performing the most efficient query, and cautioned against invisible changes in a database over time.[12] Others have taken a broader approach, looking at all interactions,[13]

10 Ryan Cordell (2016). "What Has the Digital Meant to American Periodicals Scholarship?" In: *American Periodicals* 26.1, pp. 2–7. URL: http://www.jstor.org/stable/44630657.

11 See Sarah Oberbichler and Eva Pfanzelter (2022). "Tracing Discourses in Digital Newspaper Collections - A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter; Mark J. Hill (May 2016). "Invisible interpretations: reflections on the digital humanities and intellectual history." In: *Global Intellectual History* 1.2, pp. 130–150. URL: https://doi.org/10.1080/23801883.2017.1304162.

12 Hieke Huistra and Bram Mellink (2016). "Phrasing history: Selecting sources in digital repositories." In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49.4, pp. 220–229. ISSN: 0161-5440. URL: https://doi.org/10.1080/01615440. 2016.1205964; Charles Upchurch (2012). "Full-Text Databases and Historical Research: Cautionary Results from a Ten-Year Study." In: *Journal of Social History* 46.1, pp. 89–105. URL: https://academic.oup.com/jsh/article/46/1/89/922483.

13 Sanna Kumpulainen and Elina Late (2022). "Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities." In: *Journal of the Association for Information Science and Technology* n/a. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24608; Elina Late and Sanna Kumpulainen (Jan. 1, 2021). "Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources." In: *Journal of Documentation* 78.7, pp. 106–124. URL: https://doi.org/10.1108/JD-04-2021-0078.

however the question of historians' practice – to what extent reflection on this source is integrated into their research and analysis – too often remains marginal. Indeed, the question is often less about use,[14] and more about access.

To summarise, we can see two main types of use: to locate articles on a theme, event, or person, and the analysis of a corpus of articles (circulation, distribution, weight etc.). In both cases, the question arises of the efficiency of the search tools, the quality of the digitisation, and the representativeness of the collections. The observation of Underwood on the "quietly central" change, addressing the implicit but omnipresent practice of querying "electronic databases", can be transposed onto the digitised newspapers and their use by historians.[15] Following Cordell, when looking at what the "digital meant to American Periodicals Scholarship",[16] we propose to tackle the challenges linked to the digitisation of historical newspapers and to collect the responses created by both the institutions who author these collections and the researchers who have worked with them in recent decades. Many of these challenges may sound trivial or rehashed. Meanwhile the past decade has witnessed a wealth of collected experience and lessons learned in dealing with these challenges, that are scattered across publications. To cite the issues listed by Cordell: the invisibility of the search algorithm that selects from an unknown pool of elements; the "archival gaps" hidden behind long lists of query results, that give a "false idea of completeness"; and the "sheer" quantity of results, potentially creating the illusion of importance for a phenomenon, or of correlations between terms. In other words, the simple task of typing a keyword into the search box of a particular collection interface triggers a series of algorithmic actions, and relies on a series of preceding decisions, all of which are invisible, yet central to any research based on these results.

These changes in research practice were echoed during the aforementioned expert workshop. Working with digitised newspapers has given a new dimension to the enjoyment and the "flaneur" attitude of researchers, Abel has led to

14 Ian Milligan (2013). "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." In: *The Canadian Historical Review* 94.4, pp. 540–569. URL: https://muse.jhu.edu/article/527016.

15 Ted Underwood (2014). "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." In: *Representations* 127.1, pp. 64–72. ISSN: 0734-6018. URL: http://www.jstor.org/stable/10.1525/rep.2014.127.1.64.

16 Ryan Cordell (2016). "What Has the Digital Meant to American Periodicals Scholarship?" In: *American Periodicals* 26.1, pp. 2–7. URL: http://www.jstor.org/stable/44630657.

engagement with new layers of information, and has radically changed how research corpora are created.[17] From the shared experience of the participants to the workshop, we could derive new outlooks on the digitised newspapers, new strategies to handle the impact of the mediation of the search engines and interfaces on the article collection and finally, new challenges that emerge from the multiplication of platforms and the need to contextualise individual articles collected on these interfaces, presented below in Tab. 1.

Tab. 1: Current practices and strategies of humanist use of digitised newspapers collections.

Researchers postures	Research approaches
Outlook on digitised newspapers collections	Rationalism: do the most with what is there, what has been digitised, what tools are available and are robust, for a given content analysis, making the most of the available quality of digitised newspapers. Pleasure: explore the diversity of contents, unexpected associations spurred by a keyword search, rediscover forgotten expression, words.
Strategies for navigating collections	Query archaeology: keep track of the query steps. Metadata excavating: treating the metadata as part of the source material, which requires a dedicated source criticism, contextualisation, and reconstruction of their creation. Looking from the outside in: deduce how a digitised newspaper collection has been constructed – what are the search engine settings, what is the coverage of a particular collection, what is the quality of the digitisation, what are the action levers to interact with the digitised newspapers?
From collections to research corpus creation	Silo breaking: connecting collections, media types, languages. Publishing the research corpus: reference it, as well as the tool used. Contextualised exploration: rebuild the contextual information that a reader (automatically) received. Layout analysis: what did the newspapers look like, how did they change over time? What was the materiality of the newspapers, and how can we keep track of this in the digitised context? Exploration, searching for similar objects to a given item.

17 Richard Abel (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1–2, pp. 1–10. ISSN: 0892-2160. DOI: 10.2979/filmhistory.25.1–2.1.

This rather abstract description of the changes brought about by digitisation might be better understood with a practical example: research on the media representation of Europe illustrates the versatility of this source material, containing textual and visual elements. A keyword search made it possible to look for mentions of Europe in the press, and already represents a fundamental change in the constitution of a research corpus. Looking at how research is conducted in journals that cover a similar topic, the first step of each use case was to define what Europe was (relations between Germany and France for instance, as Grunewald and Bock did[18]). The keyword search brings the research practices quietly and naturally much closer to the *Begriffsgeschichte* or conceptual history.[19] Changing the ways to collect historical media sources impacts, almost mechanically, the observations made about them. As has been shown by Greiner, the presence of Europe in selected German, British and US newspapers during the first half of the 20th century was dominated by comparisons with the US, questions of infrastructure and modernisation, and the enmity with the Soviet Union, and only very marginally linked to the various types of political projects that were mushrooming in this period.[20]

This contrasts, in his view, with the historiography of these movements, inherently zooming in on their media presence, while his research offered perspective on the true proportion of this presence. The keyword search underlying the constitution of a research corpus does not erase the questions of definition, but rather shifts to the issue of "query criticism": what the relevant keywords are, and how to select the relevant use in the case of polysemy. Bergamini and Mourlon-Druol created a corpus of press articles from major press titles from several European countries, this time covering the media presence of the European Union. Here, the challenge lay (alongside the varying accessibility of these titles) in differentiating the various European institutions (e.g. the Council of Europe) and connecting the relevant declination of the EU institutional forms (the

18 Michel Grunewald and Hans Manfred Bock (1997). *Le discours européen dans lesrevues allemandes = Der Europadiskurs in den deutschen Zeitschriften: (1918–1933)*. Convergences 3 3. Bern Berlin [etc.: P. Lang. 405 pp.

19 Elaine Zosa et al. (Dec. 18, 2020). "The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections." In: *Journal of Data Mining& Digital Humanities* HistoInformatics. URL: https://jdmdh.episciences.org/6728/pdf.

20 Florian Greiner (June 30, 2014). *Wege nach Europa: Deutungen eines imaginierten Kontinents in deutschen, britischen und amerikanischen Printmedien, 1914–1945*. Wallstein Verlag. 520 pp.

European Council, the EU agencies etc.).[21] They adopted strategies of weighting individual articles with hits to reduce the presence of articles that mention the currency "euro", for instance, and used topic modelling to cluster the articles and differentiate the uses of the terms. One interesting challenge they faced was the overlap of digitised and digitally created materials for the period of the early 2000s, and to harmonise the varying structures of metadata.

Media presence not only equates with articles covering "Europe" in its polymorphic manifestations: in a more cultural history flavoured approach, Eijnatten worked on weather reports in the Dutch press and their changing representation of Europe.[22] This research required a vastly different sets of skills and the preprocessing of the digitised newspaper collection. To collect this type of content, they needed to identify the tables and articles containing it, which requires a combination of visual identification and content-based structuring of the collection.

This brief excursion at the crossroads of European and media historiography exemplifies the changes and continuity of research in the context of the digitisation: the change of scale can bring a change of focus; however, some core scholarly practices remain crucial to the analysis of digitised newspapers contents.

As Fickers had already formulated in 2012, source criticism remains a defining task of scholarly historical research and must be transposed in the context of digitised sources, which implies interdisciplinary collaboration. Applying this to the specific case of digitised newspapers also creates a dependency on infrastructure, digitisation policies and documentation, and interface design decisions. Making digitised newspapers "amenable" to historical research implies bridging complex tools, applied to large collections, with nonetheless rather "straightforward" final uses, such as reading, counting and preparing corpus statistics. This can be managed by digital artefacts of historical sources, as researchers tendentially search for a series of articles or visual elements,[23] such as advertisements. The legitimacy of digitisation seems not to be the point of discussion anymore, but rather the question of contextualisation, both historical

21 Enrico Bergamini and Emmanuel Mourlon-Druol (2021). *Talking about Europe: exploring 70 years of news archives | Bruegel*. URL: https://www.bruegel.org/2021/03/talkingabout-europe-exploring-70-years-of-news-archives/- .

22 Joris Van Eijnatten (2019). "Something about the Weather. Using Digital Methods to Mine Geographical Conceptions of Europe in Twentieth-Century Dutch Newspapers." In: *BMGN - Low Countries Historical Review* 134.1, pp. 28–61. URL: http://www.bmgn-lchr.nl/articles/10.18352/bmgn-lchr.10655/.

23 Trevor Owens and Thomas Padilla (2020). "Digital sources and digital archives: historical evidence in the digital age." In: *International Journal of Digital Humanities*. URL: https://doi.org/10.1007/s42803-020-00028-7.

and technical.[24] Many experiences have been gathered in both aspects, and this is what we collect in the following.

To unpack this question, we first look at the typical content of a digitised newspaper collection and the diversity of collection types that have emerged over the past few decades. Based on this sketch of digitised newspapers collections, we'll move to identifying interesting practices of documentation and the publication of the datasets that the institutions that hold them have produced, and more specifically how institutions and researchers have made use of the catalogue metadata to support a first layer of contextualisation of the queries. Finally, we look at noteworthy initiatives to make use of the content of these collections to create data-driven context.

3 Digitised Newspapers Collections: Navigating through Consistencies and Idiosyncrasies

What are digitised newspapers collections? In order to better understand what we do when we search in digitised newspaper collections, it helps to understand how these collections are structured. Generally speaking, a digitised newspaper collection starts with a complete analogue collection of one or several newspaper titles. The institution holding these dedicates some budget to the digitisation, which is often conducted outside, by a specialised company, which uses tools and software tailored to the task. The digitisation consists of scanning the pages, creating images and metadata for each page. These pages are then run through OCR software, which produces two types of information: a transcription of the textual content of each page, a tracking of the position of each letter or word on the page. This information is then stored in separate files, following a chosen standard.

24 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. DOI: http://doi.org/10.18146/2213-0969.2012.jethc004; Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen (June 2019). "Toward a model for digital tool criticism: Reflection as integrative practice." In: Digital Scholarship in the Humanities 34.2, pp. 368–385. URL: https://academic.oup.com/dsh/article/34/2/368/5127711; Andreas Fickers (2020). "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?" In: *Zeithistorische Forschungen* 17.1, pp. 157–168.

The trend has been to follow the standards set by the Library of Congress: the mets and alto format. To give a gross simplification, the mets standard functions like a kind of table of contents for all the ALTO files, which contains the textual and "spatial" distribution of the text for each page. Later, these files are stored in a database at the institution that ordered this digitisation, and an index is created of the textual content. This index keeps track of which words (or combinations of words) appear on which page (or article) of which issue of which title. When one searches the collection via the search box of the institutional interface, this is the basis for it. Once the resulting list has been produced and a given document (a page or an article) has been selected, the viewer usually displays a facsimile (a scan of the page) with the transcribed text, either within the facsimile or side by side.

From this very simple description, we can already anticipate that each step contains a possibility of variation in their implementation. This means that even if one is aware of the general logic of a digitisation campaign for newspapers, much of the practical knowledge needs to be deliberately collected.[25] The complexity and diversity of this process has been exposed by the work conducted in the context of the Oceanic Exchange project, in the Atlas of Digitised Newspapers and Metadata.[26] In this report (and database), the authors mapped and "translated" the contents of the metadata, describing each of the elements in each collection connected by the project.

As historical materials, the newspapers produced and stored are not so homogeneous when looking at the individual issues. This makes an overview of the content of the collection potentially difficult. Without such detailed analytical work, there are some descriptions published alongside the collections that can be helpful to understand what the collection of digitised newspapers is.

To sum-up, digitised newspapers collections contain images, OCR output and metadata in varying degrees of granularity. A simple way to have a better understanding of what constitutes these collections is to consult batches published by collections holders, feeding into the approaches to collections as

25 Tessa Hauswedell et al. (June 1, 2020). "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archivesof historical newspapers." In: *Archival Science* 20.2, pp. 139–165. URL: https://doi.org/10.1007/s10502-020-09332-1.

26 Melody Beals and Emily Bell (2020). *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough. DOI: 10.6084/m9.figshare.11560059.v2.

data.[27] For instance, the National Library of Luxembourg offers several packages of digitised newspapers for researchers to download.[28]

Digitised newspapers collection beyond mainstream titles and institutional silos. Many of the crucial steps in the preparation and publication of digitised newspaper collections depend therefore on policy choices of the hosting library.[29] Given the availability of analogue collections of newspapers, sufficient funding, and the priority given to digitising their holdings, libraries have been progressively scanning and extracting the textual and image contents, storing them in databases, and sharing access with their users via search portals.

However, even in the absence of dedicated publications on the history of their digitisation campaign, some institutions have provided information on the financing of sub-collections. For instance, Scriptorium[30] has added in the search interface information on each sub-collection: the financial source of the digitization; a short history of the title(s) contained in the sub-collection; and the list of the missing issues. This information, however, when also used as a filter for the search interface, does not translate into further metadata, which would apply to all sub-collections.

As an example of thematic guidance through the institutional collection, Gallica[31] displays topical and geographical collections, which give an impression of the diversity that the label "digitized newspapers" covers: from local and national newspapers to the trench newspapers of WWI, newspapers printed in a colonial context, and so on. These groupings are, however, not useable as filters when conducting a query in the search page (or only indirectly, via "collection ou thème", where one has to use the original holder of a given title or title type

27 Rachel Wittmann et al. (Dec. 16, 2019). "From Digital Library to Open Datasets." In: *Information Technology and Libraries* 38.4, pp. 49–61. URL: https://ejournals.bc.edu/index.php/ital/article/view/11101.

28 https://data.bnl.lu/data/historical-newspapers/.

29 See Tessa Hauswedell et al. (June 1, 2020). "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers." In: *Archival Science* 20.2, pp. 139–165. URL: https://doi.org/10.1007/s10502-020-09332-1. Another central issue is the ownership of the contents, see the contrast between TROVE and BL as discussed by Thomas Smits (2016). "Making the News National: Using Digitized Newspapers to Study the Distribution of the Queen's Speech by W. H. Smith & Son, 1846–1858." In: *Victorian Periodicals Review* 49.4, pp. 598–625. URL: http://www.jstor.org/stable/26166579.

30 The digital library of the Bibliothèque Cantonale Universitaire de Lausanne, https://scriptorium.bcu-lausanne.ch.

31 From the French National Library, https://gallica.bnf.fr/html/und/presse-et-revues/presse-et-revues?mode=desktop.

to select the one that is relevant). The anchoring of these collections under the umbrella of national libraries may obscure their linguistic, historical and geographical diversity: for instance, Gallica contains titles published in colonial contexts, or by French diaspora in Turkey or China.

Another example of historical legacies is the collection published by the Digitale Forum Mittel- und Osteuropa (DiFMOE), which contains titles from the German minorities in Central and Eastern Europe. To raise awareness of this legacy, the interface of the Dutch Royal Library's newspaper collection has included "distribution area" as a filter, to enable the distinction between colonial, national and regional titles. Collections have been aggregated beyond state borders in international or federal structures, such as Chronicling America or the Europeana Newspapers.[32] A core difference between these two types of aggregation lies in the level of coordination at the stage of the digitisation. In the European context, Europeana holds duplicates of collections that were produced and designed in each country, whereas in the US context, the design of digitisation was more centralised, and thus harmonised.

Mitigating OCR misidentifications. What are the key elements to be aware of when performing queries in these interfaces? One recurrent concern is the impact of OCR quality on the resulting list. Some institutions offer some transparency on the OCR regime under which given parts of their collections have been digitised: the Royal Dutch Library, in its periodical portal Delpher, offers an option to filter by OCR generation, helping users become aware of varying levels of quality.

However, it might not simply be the quality of recognition of individual letters that is the most determining factor. Moreux tackled this question by analysing the queries performed on Europeana,[33] as well as dissecting their performance.[34] In summary, the longer a keyword is, but above all the less popular it is, the less chance it has of being properly recognised and indexed. Indeed, one has to take into account that it is not the raw recognition of letters that is the base of content queries, but the content that has been indexed in the database. The OCR process

32 Benjamin Charles Germain Lee et al. (2020). "The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america." In: *arXiv*, arXiv–2005; Nuno Freire et al. (2019). "Opening Digitized Newspapers Corpora: Europeana's Full-Text Data Interoperability Case." In*: 2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany.

33 https://www.europeana.eu/en/collections/topic/18-newspapers.

34 Jean-Philippe Moreux (Aug. 2016). "Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment." In: *IFLA News Media Section, Lexington, August 2016, At Lexington, USA*. Lexington, United States: IFLA. URL: https://hal-bnf.archives-ouvertes.fr/hal-01389455 (visited on 07/13/2021).

can include the use of dictionaries to correct or disambiguate misidentified words. If a word is not part of the dictionary used, then it might not be corrected if misidentified. Also, for most of the collections, the masthead containing the title of the newspaper is excluded from the search space. This selection of the source material where the query is conducted reduces the search space but often results hits remain dispatched on many pages in a interface, making this assessment difficult. However, in spite of missing documentation, researchers need develop strategies to assess the quality of OCR in different titles or periods, by using a few test-keywords, of different lengths or frequency and focus on segments of time or titles in the collection they are using (see the examples in the list below).

Exploring digitised newspapers beyond keyword search. Some libraries have given access to their collections not only via a traditional search interface, but also via APIs or prepared samples of the raw data-set. Going a step further, some collections were prepared for linguistic analysis, such as DiaCollo,[35] where one can focus on "the company a word keeps"[36] and search for collocations with a high level of precision. However, most of the time, the central feature offered by a newspaper interface is a keyword search and the possibility of filtering the resulting list, based on bibliographic metadata, such as newspaper title, date, etc. In recent years, libraries and dedicated research projects, such as the *impresso* project, have produced content-based filters or annotations, enabling a different exploration of the digitised newspaper collections. For instance, the filters based on topic modelling can sometimes be used as (imperfect) proxies for rubrics, such as radio programs, sports sections or front-news, which gives immediate insights on the distribution of the results across titles and contexts. Often, however, these features are stored in separate interfaces and the wealth of interactions and information produced remains in the shadow of the main tool: the keyword search.

Beyond the scope of documentation: the case of censorship and political orientation. Some information on the content of a corpus cannot be found in the history of its collection or the documentation of its digitisation, or requires a highly detailed domain knowledge, in order to be transposed as "data". Two simple examples, often given by historians are the question of censorship and the political orientation of a given title. On the latter issue, the problem seems to be twofold: first, the boundaries between political orientations can fluctuate, and this does not necessarily cover the entire content of a given newspaper. Some newspapers are connected to a political party as their organ and it would appear simple to

35 https://clarin-d.de/de/kollokationsanalyse-in-diachroner-perspektive.

36 John Rupert Firth (1957). *A synopsis of linguistic theory, 1930–1955.* URL: https://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf.

attach this information as metadata or, for instance, in the case of a continued publication under Nazi occupation, as the BNL did for the *Luxemburger Wort*,[37] one could imagine transforming this asterisk within the short note on the title into a metadata filter. But the question remains of how useful such metadata would be in filtering the results of a query: what would be the added value, and what kind of chain of biases would such a filter bring and overwrite in certain traditional fields of research? However, it is mainly the second part of the issue that blocks the implementation of such a filter, namely the need to add time stamps to this information, which requires a lot of manual work and would often remain highly debatable, or would simply not be possible for all the titles of a collection by the institution holding the collection.

The issue of censorship is subject to similar challenges: where to systematically collect the information on which version of the issue has been stored in analogue and subsequently digitised? One could imagine adding metadata that informs on a particular law being in place at a given time, focusing on a specific aspect, but again, this would require significant manual labour; it could also be argued that is the responsibility of the researcher to be aware of this historical context. This is especially true since censorship is also a complex topic in itself, and could include softer official "guidance", as was the practice in the GDR.[38]

An interesting example has been given by Giullian, who relates the case of a targeted search for articles that the researcher knew had been published but could not be found in the digitised collection of the Soviet *Pravda*.[39] This could inspire strategies to prompt researchers working with potentially sensitive titles and periods to make some comparative work between the analogue and digitised collections.

Finally, although it may seem obvious to seasoned researchers, a feature of newspapers is their "transparency" concerning themselves (Vella), they should not be the primary source to find out the technical, legal and economical constraints or the daily organisation of the work of a newspaper.

37 https://eluxemburgensia.lu/periodicals/luxwort.

38 Anke Fiedler and Michael Meyen (Oct. 2, 2015). "The steering of the press in the socialist states of Eastern Europe: the German Democratic Republic (GDR) as a case study." In: *Cold War History* 15.4, pp. 449–470. URL: https://doi.org/10.1080/14682745.2015.1028531.

39 Jon C. Giullian (2013). ""Seans chernoi magii na Taganke": The Hunt for Master and Margarita in the Pravda Digital Archive." In: *Slavic & East European Information Resources* 14.2–3, pp. 102–126. URL: https://doi.org/10.1080/15228886.2013.813374.

To sum up, the answer to the question "what to do with a million newspaper pages"[40] has transformed into a need to better understand what digitised newspaper collections are before being able to critically assess the significance of the findings and find the most relevant to interact with these rich collections. Some elements are (relatively) consistent across collections (digitisation steps, access via keyword search) while others are vary within one collection or between collections (OCR quality, metadata as research filters, historical context of the newspapers production).

Thankfully, libraries and digitised newspaper collection holders in general have produced a wide range of tools and (scattered) documentation to raise awareness among their users, particularly researchers. It remains the researchers' responsibility to assess the source material in the context of ease of access, to find strategies to manage massive and multiple corpora, and to search for answers to the core questions concerning the availability of digitised collections.

Here are a few questions that could help researchers assess the collection they are using:

Collection description

- What titles and what time span are covered?
- What is the context of creation of these digitised (sub)collections?
- What regional/national diversity is represented in the collection?
- How many languages are present in the titles?
- Does the collection specialise in newspapers or contain other documents, archives?

Digitisation context and output

- What is the institutional frame of the digitisation (public/private institution, cultural heritage institutions, such as galleries, libraries, archives, and museums (GLAM)/project)?
- Is the digitisation ongoing? Are the titles digitised entirely or for limited portions?
- Is the content searchable with keywords (OCR)?
- Is there another portal that published this collection and is there differently processed?
- Are there corpus statistics available (numbers of tokens, pages, issues)?

40 Robert B. Allen, Weizhong Zhu, and Robert Sieczkiewicz (2010). "What to Do With a Million Pages of Digitized Historical Newspapers?" In: URL: https://www.ideals.illinois.edu/handle/2142/14932.

Search settings and user interactions

- Can I store the resulting hits? Can I add labels to selected items?
- Can I export the resulting hits? In bulk/individually? Which format?
- What are the effects of lemmatisation and indexation of the digitised content on the search results? Do the queries "apples" and "apple" have the same results?
- Is the search conducted at page level or article level? Can I search for two consecutive words (e.g. Spanish flue) within one article, or do they appear on one page/issue?
- Can I test the "vulnerability" of the chosen keywords? Is it prone to OCR misidentification - and what options are there in the search page to mitigate this?
- How "fuzzy" are the search settings? Are all the results exact matches? In the case of multilingual collections, are the keywords translated into several languages?
- Can I limit my search to a date range, a type of item (articles, advertisement, obituaries, tables ...)?
- Can I search for images (adverts, maps, photographs, cartoons)?

Access: What type of access is made possible with the digitised newspapers collection?

- Can I access via an interface and/or via an API?
- Are there ready-made data-sets to download, and what do they contain (images, text, METS/ALTO files)?
- Are there other portals of the same institution that offer different interactions than the general interface (via "labs" for data visualisation, use of NLP outputs such as named entities)?

4 Mining the Digitised Newspapers: Tools to Contextualise the Queries and Corpus Creation

While it is nearly impossible (and not necessarily interesting) to reconstruct the context in which an individual article has been produced, adding structural information to a collection, such as indications of rubrics, can help to assess the content of a query or a research corpus. In other words, how can researchers identify in which section of a newspaper a particular or a list of articles has been published? Trying to answer his implies integrating into the use of digitised newspapers collection, the research outcome produced by media history,

on the emergence and dissemination of given newspapers rubrics[41] but also, in the digital context, looking at corpus statistics of a given collection. Doing so, one can try and check whether the signals observed are specific to the research corpus or an artefact of the collection or period, title.

The context offered by digitisation. Digitisation has created a trend towards the decontextualised *use*[42] of individual articles: going quickly from a keyword to a results list does indeed "hide" the context of the issue where the article was published. This fact, multiplied by the hundreds or thousands of matches, makes it more difficult to reconstruct the context than when one collects the relevant articles by leafing through issues or scrolling microfilms.

As has been discussed by Walma, a query for "morphine" appears in a variety of contexts, even in a search environment that offers many filtering options such as Delpher, where while querying, one may see how many hits are found in documents identified as "articles" vs. "advertisements".[43] For this research, Walma managed to manually classify 1020 documents into categories to see whether "morphine" appeared in the news, a "feuilleton", or in science rubrics. From there (and using other tools), the author could analyse the specificity of discourses on the use of morphine produced for literary or reporting purposes, and how they interacted with each other. This intermediary step, inserting thematic research into the more general media history, is part of any research using this material. Though it is often achievable for researchers to manually label their collection of articles based on article types, it raises the question of how to create indicators informed by media history to create content-based context for the queries.

The transcription of newspapers has opened them up to the generation of content-based metadata, which can be used to create another type of context to each article (also presented in part II of this volume – see the chapters by Langlais, Wevers, and Paju). The content-based metadata are information, indications based on tools that mine the content for this source, at the article or page level. This means that they were not manually added and therefore are not

41 Marie-Ève Thérenty and Sylvain Venayre (Sept. 30, 2021). *Le monde à la une. Une histoire de la presse par ses rubriques*. Illustrated édition. Paris: Anamosa. 368 pp.

42 The digitisation does not prevent from leafing through a collection and conduct manual inspection as done with analogue sources. The digital format makes simply the keyword-based exploration more attractive.

43 L. W. B. Walma (2015). "Filtering the "News": Uncovering Morphine's Multiple Meanings on Delpher's Dutch Newspapers and the Need to Distinguish More Article Types." In: TS: *Tijdschrift voor Tijdschriftstudies*. URL: http://dspace.library.uu.nl/handle/1874/324205.

determined by historical context but by algorithms. For instance, one could imagine using the methods developed by PC Langlais to define and detect the genre of articles (or rubrics) to detect political orientation. This would mean that the content of individual articles would be labelled based on vocabulary or named entities or other content-type that would be the source of the metadata, rather than a label added in the context of the cataloguing of a title. This type of metadata does not exist (yet?) and would need to be subjected to a discussion on the reliability of the tool used to detect such content types. This hypothetical example aims to simply underline the significant difference between what follows and what has been presented just before. It is also a different approach from applying tools to mine a selected part of a given collection. Here, again, this type of analysis becomes meaningful when applied to the entire collection.

The central question here is not so much what tools exist and how they could be applied to a collection, but rather how these layers of information created via content-based metadata can inform the researchers about the context of the collection, or more precisely, what contextual information can be useful when browsing and querying newspaper collections. We now look at what has already been produced and used, before briefly discuss why creating such information can help make a significantly more interesting use of digitised newspaper collections.

Contextualisation inspired by media history: patterns, seasonality and genres. Once there are computer-readable materials produced by the digitisation process, with a minimal structure given by the catalogue metadata as listed earlier, digitised newspaper collections open themselves up to a number of computations of their content, the simplest being word counts. This method was fruitfully used for research purposes by Turunen.[44]

The output of such computations is not always simple to analyse or even to implement given the size of the collections, and requires a similarly critical statistical review. At the same time, these simple tasks, of counting words and looking for when they appear the most, can give interesting insights into a collection (and create grounds for comparison between collections).[45] In this case, the statistical measures inform not only a research question but highlight structural elements of the collection that impact every research question, using this source. Using the British and US digitised newspaper collections

44 Risto Turunen (Nov. 10, 2021). *Shades of Red*. Työväen historian ja perinteen tutkimuksen seura. URL: https://helda.helsinki.fi/handle/10138/336197.

45 Fabon Dzogang et al. (2016). "Discovering Periodic Patterns in Historical News." In: *PLOS ONE* 11.11. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0165736, As shown by.

available at the time, they looked at the most frequent words appearing on a given month over the years, and could highlight, for instance the "periodic seasonal return of certain infectious diseases", the celebration of national (with the word "fireworks") or religious (with the words "Santa Claus") holidays, as well as crop seasons. This type of work brings out the repetitive, ritual contents of the newspapers and when applied to an entire collection, can produce an interesting element to this evasive "context", here being partially defined by the seasonality of the press content.

Another type of pattern that digitised newspaper lend themselves to is the recognition of rubrics and other types of segmentation within the newspapers. Attempting to mimic what "humans can recognise",[46] Barman et al. worked with visual and content features to train the recognition of certain types of rubrics, such as the "feuilleton", to create more precise and targeted segment identification of newspapers.[47] Certain elements of newspapers, such as tables, are more difficult for such tools to differentiate (for instance between train tables and stock exchange tables), but offer an interesting approach to enrich and potentially extract targeted contents. With such an identification, linked to a collection, one could imagine using this research output as filters, but also to create entirely new types of data-sets to conduct further research on, for instance, the "last minute" reprints of press agencies.

This type of identification has been conducted with other tools by Langlais and is available for the feuilletons of certain French titles,[48] or for very particular items such as obituaries, by the BNL. These examples remain, however, the output of research and are not integrated into any interface. The integration of such types of metadata as filters or as tools for the visual exploration of a collection of digitised newspapers remains complex, as the parameters (or Natural Language Processing models) set for the production of such corpus statistics can greatly impact the results, making such information potentially difficult to

46 Chiel van den Akker (Jan. 2, 2018). "What are patterns in the humanities?" In: *Interdisciplinary Science Reviews* 43.1, pp. 74–86. URL: https://doi.org/10.1080/03080188.2017.1296265.

47 Raphaël Barman et al. (2021). "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & Digital Humanities*. DOI: 10.5281/zenodo.4065271.

48 See the Generothèque, http://www.numapresse.org/generotheque/items/show/1 and Pierre-Carl Langlais (2022). "Classified News - Revisiting the history of newspaper genre with supervised models." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter.

assess for users.[49] Nonetheless, the potential opened up by such information when connected to the collection itself in our view constitute a true Eldorado, with all its promises and pitfalls.

5 Outlook: Connecting Collections as the Current Frontier of the Eldorado

The current status of the use of digitised newspapers can be described as either strikingly simple, not to say primitive, with the popularisation of the practice of keyword search, or extremely (not to say disproportionately) complex, once we try to contextualise the collected material (via a keyword search). How can we adjust research practices and reach an acceptable communication of the findings, without reproducing the specification of the digitisation, search engine, copyrights etc.? And how can we navigate the collections that our research question leads us to, with their specific content and the varying access options to connect them?

The issue of scale of the digitised newspaper collections plays a central role: the same tools can be applied to portions or to the entire collection, but the output is completely different. For instance, topic modelling has been applied in many cases to analyse the corpora generated by a keyword query[50] or to an entire collection (such as in the context of the *impresso* project). In the first type of use, the goal is to test the hypothesis or detect the specificity of the research corpus in contrast to a test corpus, whereas in the second, the goal is to give indications on the content of the articles in the context of a query.

Two trends in scholarly and institutional practices may help us to manage these challenges. For the scholarly part, we can observe the increasing trend to publish data papers, where authors can specifically explain how they discovered their research corpus. For the institutional part, code blocks are increasingly published, that can be "plugged" into a specific collection to better interact with

49 See the discussion around the implementation of TM-based filters in *impresso* - in the FAQs, or the implementation of NER for Dutch/French/Austrian newspapers, https://lab.kb.nl/dataset/europeana-newspapers-ner.

50 Quintus Van Galen and Bob Nicholson (2018). "In Search of America." In: *Digital Journalism* 6.9, pp. 1165–1185. URL: https://doi.org/10.1080/21670811.2018.1512879; Maarten van den Bos and Hermione Giffard (2016). "Mining Public Discourse for Emerging Dutch Nationalism." In: *Digital Humanities Quarterly* 010.3.

it, without downloading and processing the entire collection on our own laptops. An interesting example is the Australian collection TROVE, which has made notebooks of code available to query and analyse its contents.[51] These notebooks lower the threshold of using the TROVE collection as data for scholars with reasonable understanding of programming.For more digitally literate humanists scholars, a more indirect connection has been proposed for the British Library (BL) as well as for the GDR Press collection.[52] In the latter case, a script to extract articles from the GDR Press collection has been published, enabling further text mining on this collection, whereas in the first case, Yann Ryan pairs the presentation of R-code to be applied to the BL collection with the indication of the opportunities that such a workflow opens, such as generating maps with the help of the geographical metadata attached to newspapers titles.

Ultimately, the issue of contextualisation remains at the cross-road of the evolution of scholarly practices and the resources made available by the collection holders. In this regard, efforts on both side multiply.

Bibliography

Abel, Richard (2013). "The Pleasures and Perils of Big Data in Digitized Newspapers." In: *Film History* 25.1–2, pp. 1–10. DOI: 10.2979/filmhistory.25.1-2.1.

Akker, Chiel van den (Jan. 2, 2018). "What are patterns in the humanities?" In: *Interdisciplinary Science Reviews* 43.1, pp. 74–86. URL:https://doi.org/10.1080/03080188.2017.1296265.

Allen, Robert B., Weizhong Zhu, and Robert Sieczkiewicz (2010). "What to Do With a Million Pages of Digitized Historical Newspapers?" In: URL:https://www.ideals.illinois.edu/handle/2142/14932.

Barman, Raphaël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan (2021). "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers." In: *Journal of Data Mining & Digital Humanities*. DOI: 10.5281/zenodo.4065271.

Beals, Melody and Emily Bell (2020). *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough. DOI: 10.6084/m9.figshare.11560059.v2.

Bergamini, Enrico and Emmanuel Mourlon-Druol (2021). *Talking about Europe: exploring 70 years of news archives | Bruegel*. URL:https://www.bruegel.org/2021/03/talkingabout-europe-exploring-70-years-of-news-archives/.

51 Tim Sherratt (n.d.). *Trove newspapers - GLAM Workbench*. URL: https://glamworkbench.net/trove-newspapers/.

52 Yann Ryan (2020). *A Short Guide to Historical Newspaper Data, Using R*. URL: https://bookdown.org/yann_ryan/r-for-newspaper-data/; "Public Projects / ddrpresse-gettext" (2021). URL: https://gitlab.zeitgeschichte-digital.de/public-projects/ddrpresse-gettext.

Bos, Maarten van den and Hermione Giffard (2016). “Mining Public Discourse for Emerging Dutch Nationalism.” In: *Digital Humanities Quarterly* 010.3.

Cordell, Ryan (2016). “What Has the Digital Meant to American Periodicals Scholarship?” In: *American Periodicals* 26.1, pp. 2–7. URL:http://www.jstor.org/stable/44630657.

Dzogang, Fabon, Thomas Lansdall-Welfare, FindMyPast Newspaper Team, and Nello Cristianini (2016). “Discovering Periodic Patterns in Historical News.” In: *PLOS ONE* 11.11. DOI: 10.1371/journal.pone.0165736.

Eijnatten, Joris Van (2019). “Something about the Weather. Using Digital Methods to Mine Geographical Conceptions of Europe in Twentieth-Century Dutch Newspapers.” In: *BMGN - Low Countries Historical Review* 134.1, pp. 28–61. URL:http://www.bmgn-lchr.nl/articles/10.18352/bmgn-lchr.10655/.

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. DOI: http://doi.org/10.18146/2213-0969.2012.jethc004.

Fickers, Andreas (2020). “Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?” In: *Zeithistorische Forschungen* 17.1, pp. 157–168.

Fickers, Andreas, Dana Mustata, and Anne-Katrin Weber (2019). “The Rise of Television.” In: *The Handbook of European Communication History*. John Wiley & Sons, Ltd, pp. 239–255. URL:https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119161783.ch13.

Fiedler, Anke and Michael Meyen (Oct. 2, 2015). “The steering of the press in the socialist states of Eastern Europe: the German Democratic Republic (GDR) as a case study.” In: *Cold War History* 15.4, pp. 449–470. URL:https://doi.org/10.1080/14682745.2015.1028531.

Firth, John Rupert (1957). *A synopsis of linguistic theory, 1930–1955*. URL:https://cs.brown.edu/courses/csci2952d/readings/lecture1-firth.pdf.

Freire, Nuno, Antoine Isaac, Twan Goosen, Daan Broeder, Hugo Manguinhas, and Valentine Charles (2019). “Opening Digitized Newspapers Corpora: Europeana’s Full-Text Data Interoperability Case.” In: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany.

Galen, Quintus Van and Bob Nicholson (2018). “In Search of America.” In: *Digital Journalism* 6.9, pp. 1165–1185. URL:https://doi.org/10.1080/21670811.2018.1512879.

Germain Lee, Benjamin Charles, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S Weld (2020). “The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america.” In: *arXiv*, arXiv–2005.

Giullian, Jon C. (2013). ““Seans chernoi magii na Taganke”: The Hunt for Master and Margarita in the Pravda Digital Archive.” In: *Slavic & East European Information Resources* 14.2–3, pp. 102–126. URL:https://doi.org/10.1080/15228886.2013.813374.

Greiner, Florian (June 30, 2014). *Wege nach Europa: Deutungen eines imaginierten Kontinents in deutschen, britischen und amerikanischen Printmedien, 1914–1945*. Wallstein Verlag. 520 pp.

Grunewald, Michel and Hans Manfred Bock (1997). *Le discours européen dans les revues allemandes = Der Europadiskurs in den deutschen Zeitschriften: (1918–1933)*. Convergences 3 3. Bern Berlin [etc.: P. Lang. 405 pp.

Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras, and Emily Bell (June 1, 2020). “Of global reach yet of situated contexts: an examination of the implicit and explicit

selection criteria that shape digital archives of historical newspapers." In: *Archival Science* 20.2, pp. 139–165. URL:https://doi.org/10.1007/s10502-020-09332-1.

Hill, Mark J. (May 2016). "Invisible interpretations: reflections on the digital humanities and intellectual history." In: *Global Intellectual History* 1.2, pp. 130–150. URL:https://doi.org/10.1080/23801883.2017.1304162.

Huistra, Hieke and Bram Mellink (2016). "Phrasing history: Selecting sources in digital repositories." In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49.4, pp. 220–229. URL:https://doi.org/10.1080/01615440.2016.1205964.

Koenen, Erik (2018). "Digitale Perspektiven in der Kommunikations- und Mediengeschichte." In: *Publizistik* 63.4, pp. 535–556. URL:https://doi.org/10.1007/s11616-018-0459-4.

Koolen, Marijn, Jasmijn van Gorp, and Jacco van Ossenbruggen (June 2019). "Toward a model for digital tool criticism: Reflection as integrative practice." In: *Digital Scholarship in the Humanities* 34.2, pp. 368–385. URL:https://academic.oup.com/dsh/article/34/2/368/5127711.

Kovarova-Simecek, Monika (2022). "Kulturgeschichte der Popularisierung von Börsennachrichten in Wien (1771-1914). Eine historische Analyse unter Anwendung von ANNO/ONB und ONBLabs." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter.

Kumpulainen, Sanna and Elina Late (2022). "Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities." In: *Journal of the Association for Information Science and Technology* n/a. URL:http://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24608.

Langlais, Pierre-Carl (2022). "Classified News - Revisiting the history of newspaper genre with supervised models." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter.

Late, Elina and Sanna Kumpulainen (Jan. 1, 2021). "Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources." In: *Journal of Documentation* 78.7, pp. 106–124. ISSN: 0022-0418. URL:https://doi.org/10.1108/JD-04-2021-0078.

Milligan, Ian (2013). "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." In: *The Canadian Historical Review* 94.4, pp. 540–569. URL:https://muse.jhu.edu/article/527016.

Moreux, Jean-Philippe (Aug. 2016). "Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment." In: *IFLA News Media Section, Lexington*, *August 2016*, *At Lexington, USA*. Lexington, United States: IFLA. URL:https://hal-bnf.archives-ouvertes.fr/hal-01389455 (visited on 07/ 13/2021).

Oberbichler, Sarah and Eva Pfanzelter (2022). "Tracing Discourses in Digital Newspaper Collections - A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage." In: *Digitized newspapers - A New Eldorado for historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitisation*. Ed. by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert. De Gruyter.

Owens, Trevor and Thomas Padilla (2020). "Digital sources and digital archives: historical evidence in the digital age." In: *International Journal of Digital Humanities*. URL:https://doi.org/10.1007/s42803-020-00028-7.

"Public Projects / ddrpresse-gettext" (2021). URL:https://gitlab.zeitgeschichte-digital.de/public-projects/ddrpresse-gettext.

Ryan, Yann (2020). *A Short Guide to Historical Newspaper Data, Using R*. URL: https://bookdown.org/yann_ryan/r-for-newspaper-data/.

Sherratt, Tim (n.d.). *Trove newspapers - GLAM Workbench*. URL:https://glam-workbench.net/trove-newspapers/.

Smits, Thomas (2016). "Making the News National: Using Digitized Newspapers to Study the Distribution of the Queen's Speech by W. H. Smith & Son, 1846–1858." In: *Victorian Periodicals Review* 49.4, pp. 598–625. URL:http://www.jstor.org/stable/26166579.

Thérenty, Marie-Ève and Sylvain Venayre (Sept. 30, 2021). *Le monde à la une. Une histoire de la presse par ses rubriques*. Illustrated édition. Paris: Anamosa. 368 pp.

Turunen, Risto (Nov. 10, 2021). *Shades of Red*. Työväen historian ja perinteen tutkimuksen seura. URL:https://helda.helsinki.fi/handle/10138/336197.

Underwood, Ted (2014). "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." In: *Representations* 127.1, pp. 64–72. URL:http://www.jstor.org/stable/10.1525/rep.2014.127.1.64.

Upchurch, Charles (2012). "Full-Text Databases and Historical Research: Cautionary Results from a Ten-Year Study." In: *Journal of Social History* 46.1, pp. 89–105. URL: https://academic.oup.com/jsh/article/46/1/89/922483.

Vella, Stephen (Sept. 3, 2008). "Newspapers." In: *Reading Primary Sources: The Interpretation of Texts from Nineteenth and Twentieth Century History*. Ed. by Miriam Dobson and Benjamin Ziemann. 1st edition. London; New York: Routledge, pp. 192–208.

Walma, L. W. B. (2015). "Filtering the "News": Uncovering Morphine's Multiple Meanings on Delpher's Dutch Newspapers and the Need to Distinguish More Article Types." In: *TS: Tijdschrift voor Tijdschriftstudies*. URL:http://dspace.library.uu.nl/handle/1874/324205.

Wijfjes, Huub (2017). "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." In: *Tijdschrift voor Mediageschiedenis* 20.1, pp. 4–24. URL:http://www.tmgonline.nl/index.php/tmg/article/view/277.

Wittmann, Rachel, Anna Neatrour, Rebekah Cummings, and Jeremy Myntti (Dec. 16, 2019). "From Digital Library to Open Datasets." In: *Information Technology and Libraries* 38.4, pp. 49–61. URL:https://ejournals.bc.edu/index.php/ital/article/view/11101.

Zosa, Elaine, Lidia Pivovarova, Jussi Kurunmäki, and Jani Marjanen (Dec. 18, 2020). "The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections." In: *Journal of Data Mining & Digital Humanities* HistoInformatics. URL:https://jdmdh.episciences.org/6728/pdf.

Monika Kovarova-Simecek

Kulturgeschichte der Popularisierung von Börsennachrichten in Wien (1771–1914)

Eine historische Analyse unter Anwendung von ANNO/ONB und ONBLabs

Abstract: The article deals with the question of how digital archives (ANNO/ONB) together with digital methods and semantic technologies provided by ONBLabs (Text Mining, Plotly, IIIF-API) can be used in historical research. It illustrates their application with the example of a research project investigating the popularisation of financial news in Vienna from the stock exchange establishment in 1771 until its temporary closure in 1914. The advantage of working with digitised newspaper archives and methods lies, in particular, in the possibility to outline the development of financial news over significantly longer periods and to identify potential correlations with economic, political and social contexts. Based on a keyword search across the digital holdings of ANNO and a record of 4,028,974 hits in up to 174 newspapers, the project examined (1) when financial terms and concepts first appeared in the Viennese press and how they evolved over the upcoming decades, (2) which factors might have influenced the development of financial news in the Viennese press (e.g. singular events such as the foundation of the Vienna Stock Exchange, changes of media policy such the abolition of censorship, growing interest of the population in the Viennese stock market, or a combination of multiple aspects), (3) whether financial news in terms of volume and intensity was cyclical or counter-cyclical about specific developments such as the railway boom or the founding vertigo in the 19th century, (4) in which contexts (e.g. politics, sports, culture, satire, etc.) financial news were published and which social discourses could be observed in the specific contexts, (5) whether and in which periods financial newspapers emerged, and (6) which newspapers can be considered particularly relevant to explaining the genesis of the financial journalism in Austria. The paper also discusses the limitations of digital newspaper portals such as the deterministic character of the portal architecture, given research logic, accessibility of metadata, semantic interoperability, and the quality of OCR, as well as their impact on the research process and results. The paper concludes with the implications of the use of digital archives on the historical communication research and gives an outlook on the integration of further features.

Keywords: digitised newspapers, cultural history, stock exchange history

https://doi.org/10.1515/9783110729214-014

1 Einleitung

Das nachfolgend skizzierte Forschungsprojekt ist Teil einer kommunikationshistorischen Untersuchung zur Kulturgeschichte der Popularisierung von Börsennachrichten in Wien von der Gründung der Wiener Börse 1771 bis zu ihrer vorübergehenden Schließung 1914. Diese Untersuchung verfolgt einerseits das Ziel, die Genese des Wiener Finanzjournalismus anhand von publizistischen Leistungen wie Zeitungen und Formaten retrospektiv zu beschreiben, andererseits aber auch seine Entwicklung im Lichte der politischen, ökonomischen und gesellschaftlichen Rahmenbedingungen zu analysieren und seine gesellschaftlichen Auswirkungen zu rekonstruieren.[1] Dazu werden mehrere methodische Zugänge miteinander kombiniert, wobei die Anwendung digitaler Zeitungsarchive (ANNO – AustriaN Newspapers Online) zusammen mit digitalen und semantischen Technologien der ONBLabs (Text Mining, Plotly, IIIF) in der Erkenntnisgenerierung ganz wesentlich waren.[2] Dieser Beitrag ist der Frage gewidmet, wie digitale Archive und Methoden im Rahmen historischer Kommunikations- und Presseforschung im Allgemeinen und ANNO sowie ONBLabs im Besonderen eingesetzt werden können. Dies wird am Beispiel eines konkreten Forschungsprojektes veranschaulicht, in dessen Rahmen untersucht wird, (1) welche börsenrelevanten Begriffe und Konzepte erstmals Eingang in die Wiener Presse fanden und wie sich ihre Präsenz im Laufe der Zeit veränderte, (2) welche Zeitungen in der Genese des Wiener Finanzjournalismus als besonders relevant erachtet werden können, (3) ob und in welchen Zeiträumen es eigene Finanz- oder Börsenzeitungen gab, (4) ob dies mit singulären Ereignissen wie z. B. der Gründung und Reorganisation der Wiener Börse, dem Staatsbankrott oder dem Fall und der Wiedereinführung der Zensur korreliert, (5) ob die Börsenberichterstattung gemessen am Umfang und der Intensität zyklisch oder antizyklisch zu bestimmten Entwicklungen wie dem Eisenbahnboom oder dem Gründungsschwindel im 19. Jahrhundert verlief und (6) in welchen Kontexten (z. B. Politik, Sport, Kultur, Mode, Satire etc.) börsenspezifische Inhalte thematisiert wurden und welche gesellschaftlichen Diskurse sich hier abzeichneten.[3]

---

1 Die publizistische historische Forschung wird hier als Wirkungsforschung im Sinne von Lerg (1992, 78) aufgefasst.

2 Zur Beschreibung der genannten Portale siehe Kap. 3.1 sowie 3.5.

3 Diese Forschungsfragen, welche Gegenstand eines Dissertationsprojektes sind, werden in diesem Beitrag nicht allumfassend beantwortet. Sie sollen jedoch verdeutlichen, mit welchen Fragenstellungen Forscherinnen und Forscher digitale Zeitungsarchive adressieren können.

2 Anwendung digitaler Archive und Methoden in der historischen Kommunikationsforschung

Über die Vorteile, qualitative und quantitative Ansätze zu einem hybriden Forschungsdesign zu integrieren, herrscht in den historischen Wissenschaften vor dem Hintergrund der Digitalisierung ein breiter Konsens (u. a. Haber, 2011; Koller, 2016; Schmale, 2010, 2013a, 2013b; Wettlaufer, 2016; Zaagsma, 2013). Die auf Basis digitalisierter historischer Quellen mögliche quantitative Analyse des Materials über längere Zeiträume und in größerem Umfang kann Muster, Strukturen und Zusammenhängen offenlegen, welche es einerseits ermöglichen, Phänomene gesamtheitlich und in größeren politischen, ökonomischen, gesellschaftlichen und kulturellen Kontexten zu betrachten, und welche andererseits Orientierung für die weitere Arbeit der Forscherinnen und Forscher schaffen. Dadurch können qualitative Zugänge auf eine breitere empirische Basis gestellt, Forschungserkenntnisse vertieft und die Interpretationsmöglichkeiten erweitert werden (Johnson und Onwuegbuzie 2004:21–22; Koenen et al. 2018:8)

Die Idee des komplementären *distant and close reading* (Moretti 2016) wird im Kontext der historischen Kommunikationsforschung durch zwei Entwicklungen der letzten Jahrzehnte gestützt, welche auch die Bedeutung digitaler Zeitungsarchive unterstreichen. Der *digital turn* hat „the creative use of online archives and a willingness to imagine a new kind of research" befördert (Nicholson 2013:63). Das kommt wiederum kulturgeschichtlichen Zugängen zugute, die im Sinne des *cultural turn* die gestaltende Macht von Sprache und Kultur in den Fokus der Forschung rücken, wodurch gerade Zeitungen als Kulturspeicher und Diskursspiegel an Bedeutung gewinnen (Koenen 2018:550).

Die Integration digitaler Zugänge in der medien- und kommunikationshistorischen Forschung kann grundsätzlich auf der materiellen und der methodischen Ebene vollzogen werden. Zum einen können Forscherinnen und Forscher auf digital vorliegende kommunikations- und pressehistorische Quellen zugreifen und zum anderen können die als Digitalisate verfügbaren Textkorpora quer über Zeiträume und Quellenbestände computergestützt analysiert werden (Koenen et al. 2018: 9; Stöber 2016: 304). Damit können die kognitiven und forschungsökonomischen Grenzen einer quantitativ angelegten kommunikationshistorischen Forschung (Schmale 2010:96; Wilke 1987:49) oder auch das von Marc Bloch beklagte Missverhältnis von Analyse und Synthese überwunden (Bloch 1994:155), und der von Max Weber bereits 1924 in „Soziologie des Zeitungswesens" formulierte Wunsch „zu messen [...], wie sich denn der Inhalt der Zeitungen in quantitativer Hinsicht verschoben hat [...]" (Weber 1988:441) mit einer größeren Wahrscheinlichkeit realisiert werden.

3 Anwendung von ANNO und ONBLabs in der historischen Analyse der Popularisierung der Börsennachrichten in Wien 1771–1914

In dem vorliegenden Forschungsprojekt, das die Popularisierung der Börsennachrichten in Wien zwischen 1771 und 1914 entlang der oben skizzierten Forschungsfragen untersucht, kamen ANNO[4], das digitale Zeitungs- und Zeitschriftenarchiv der Österreichischen Nationalbibliothek (ONB) sowie ausgewählte Anwendungen von ONBLabs[5] bei der Quellensuche, -auswahl und -auswertung zum Einsatz. Um die Forschungsfragen zu beantworten, wurden (1) ANNO nach börsenjournalistisch relevanten Schlüsselbegriffen durchsucht, (2) die im HTML-Format auf der ANNO-Website verfügbaren Metadaten der Suchbegriffe (Titel, Ausgabendatum und Anzahl der Treffer) mittels Web Scraping (webscraper.io) in einer strukturierten, statistisch auswertbaren Form extrahiert (CSV-Format), (3) die Daten in weiterer Folge mit zusätzlichen Metadaten wie Kontext und Zielgruppe angereichert, (4) die Daten in Form interaktiver Diagramme mit Python und Plotly (plot.ly) visualisiert und schließlich (5) Zeitungssammlungen, auf denen die Visualisierungen basieren, unter Verwendung der IIIF-API (iiif.io) erstellt.[6] Um die Möglichkeiten, aber auch die Grenzen dieser Anwendungen in den Kontexten historischer Kommunikations- und Presseforschung zu verstehen, werden ANNO und ONBLabs nachfolgend v. a. im Hinblick auf ihre forschungsrelevanten Aspekte beschrieben.

3.1 Austrian Newspapers Online

Die Österreichische Nationalbibliothek (ONB) startete im Jahr 2003 eine Digitalisierungsinitiative, im Zuge deren historische Zeitungs- und Zeitschriftenbestände in Digitalisate transformiert werden (ANNO/Österreichische Nationalbibliothek 2003). Das Ergebnis liegt heute als ANNO – AustriaN Newspapers Online (http://anno.onb.ac.at/anno) vor. Auf dem Portal sind die digitalisierten Bestände kostenfrei und zeitlich uneingeschränkt zugänglich und auch im TXT- und PDF-Format verfügbar.

4 http://anno.onb.ac.at Abfrage am 19.2.2019.

5 https://labs.onb.ac.at Abfrage am 19.2.2019.

6 Im Zentrum dieses Beitrags steht der quantitative Zugang. Dieser stellt allerdings nur einen Teil eines Methodenkomplexes dar, der quantitative und qualitative Zugänge im Sinne des Mixed Methods Ansatzes kombiniert.

Ein zentrales Anliegen von ANNO ist eine möglichst umfassende und vollständige Erfassung des historischen Pressematerials. Das wird durch das Zusammentragen der Bestände aus unterschiedlichen Archiven und Bibliotheken ermöglicht.[7] Der Anteil an vollständig erschlossenen Beständen liegt bei ca. 22% (271 von 1.258 Titeln), wobei die Bestände laufend ergänzt und komplettiert werden.[8] Die Volltextsuche deckt mehr als 90% des Bestandes des digitalisierten Materials ab, wobei hier schwerpunktmäßig die Jahrgänge 1689–1949 erfasst sind. Das entspricht aktuell ca. 20 Millionen Seiten, die OCR-basiert[9] auf Wortebene durchsuchbar sind (ANNO/Österreichische Nationalbibliothek 2019).

3.2 Datenabfrage und -erhebung

Dieses digitale Repository wurde auch in dem vorliegenden Forschungsprojekt als primäre Ressource herangezogen. Dazu wurden im ersten Schritt via Volltextsuche, die auf den Methoden des Text Mining basiert, theorie- und empiriegeleitet jene Begriffe erhoben, die die Börsenberichterstattung in dem Untersuchungszeitraum prägten. Die Volltextsuche ergab für den Begriff *Börse** als Ergebnis 226.976 Ausgaben[10]. Hinzu kommen noch weitere börsengängige Begriffe, welche Hinweise auf Börsenberichterstattung liefern können und daher mit Hilfe der Volltextsuche mit den folgenden Resultaten (Anzahl der Ausgaben) abgefragt wurden: „*Wiener Börse*“ (116.075)[11], *Aktie** (164.647)[12], *Obligation** (145.528)[13] und *Pfandbrief** (113.594)[14]. Aus der Analyse dieser Daten kann beispielsweise abgeleitet werden, ob Tages- und Wochenzeitungen über börsenbezogenen Themen regelmäßig (täglich oder wöchentlich) oder nur gelegentlich berichteten. Ersteres kann

7 Eine komplette Erschließung der Bestände kann nicht immer gewährleistet werden, z. B. wenn die physischen Bestände nicht mehr existieren. Das ist mitunter auch dem Umstand geschuldet, dass „diese Druckgattung von Bibliothekaren, Bibliophilen und Bibliographen nicht dieselbe Zuwendung und Wertschätzung wie handschriftliche Dokumente, Inkunabeln und Frühdrucke erf[u]hren“ (Duchkowitsch 1980, 56).

8 Stand gemäß Auskunft der ONB vom 19.4.2019.

9 Optical Character Recognition.

10 Abfrage am 19.2.2019. Um ein möglichst umfassendes Bild zu gewinnen, wurde die Suchabfrage mit weiteren Suchoptionen (Wildcards, Phasensuche und Abstandsuche) präzisiert. Das soll auch hier via (*) verdeutlich werden. Andere Suchmodi würden zu abweichenden Ergebnissen führen.

11 Abfrage am 19.2.2019.

12 Abfrage am 4.3.2019.

13 Abfrage am 4.3.2019.

14 Abfrage am 4.3.2019.

wiederum als ein Hinweis für eine Börsenrubrik und eine hohe Bedeutung der Börsennachrichten für die Leserschaft der Zeitung gedeutet werden. Eine erste Durchsicht der Ergebnisse zeigt allerdings, dass die Anzahl der Treffer im Text von Ausgabe zu Ausgabe mitunter stark schwankt.[15] Das bedeutet, dass es unabhängig von der Periodizität mehr und weniger intensive Phasen der Börsenberichterstattung gegeben haben mag, weshalb die Erhebung der Treffer im Text als unerlässlich erscheint.

Um diese Daten in eine strukturierte, statistisch auswertbare, und somit für weitere Analyse verwertbare Form zu überführen, wurden die mit dem Suchbegriff verbundenen Metadaten (Titel, Datum der Ausgabe und Anzahl der Treffer), die auf der Website im HTML-Format vorliegen, mit Hilfe von Web Scraping[16] maschinell extrahiert. Im Rahmen dieses Forschungsprojekts wurde das Google Chrome Web Scraper Plugin (www.webscraper.io) verwendet (siehe Abb. 1). Die extrahierten Daten werden ins CSV-Format übergeführt und stehen als solche zum Download zur Verfügung. Für die weitere statistische Analyse ist eine Aufbereitung der Daten als Excel-Spreadsheet erforderlich.

3.3 Ergebnisse der Volltextsuche

Die keyword-basierte ANNO-Abfrage entlang der oben skizzierten Suchparameter ergibt umfassendes Datenmaterial mit insgesamt 4.028.912 Treffern (siehe Tab. 1). Eine genauere Betrachtung der Daten zeigt, wie sich die einzelnen Suchbegriffe hinsichtlich der Treffer im Text, der Anzahl der Ausgaben sowie der Anzahl der Titel verteilen (Kapitel 3.3.1), wie sich die Suchergebnisse auf Trefferebene auf die einzelnen Zeitungen und Zeitschriften verteilen und welche Titel im Hinblick auf die Börsenberichterstattung als besonders relevant erscheinen (Kapitel 3.3.2), und wie sich die Präsenz der Suchbegriffe in der Presse titelunabhängig im Zeitverlauf entwickelte (Kapitel 3.3.3). Diese Resultate dienen zum einen als Orientierungslandkarte für weitere, tiefergehende Analysen, lassen aber bereits größere Zusam-

15 So ergibt die Suchabfrage nach dem Begriff *Börse** in der *Neuen Freien Presse* am 20.11.1873 32 Treffer, am 10.4.1901 nur 4 Treffer.

16 Als Web Scraping wird ein Verfahren bezeichnet, mit dessen Hilfe Daten aus Websites ausgehoben und in einer maschinenlesbaren Form aufbereitet werden. Dabei wird zunächst jede zu extrahierende Website automatisiert heruntergeladen. Welche Websites heruntergeladen werden sollen, wird über die URL-Parameter (sog. Query-String) übermittelt. Im nächsten Schritt wird der Ort der Daten auf der Website festgelegt, die extrahiert werden sollen. Für die eigentliche Extraktion der Daten stehen unterschiedliche Möglichkeiten zur Verfügung.

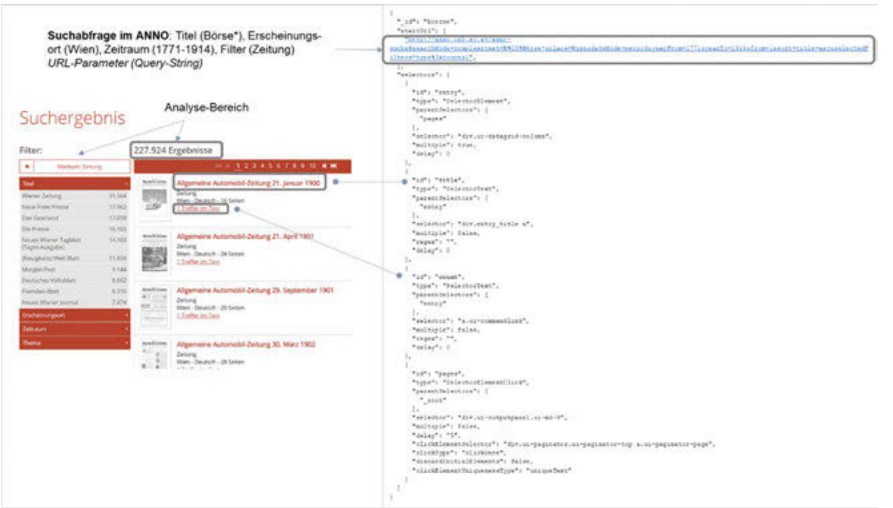


Abb. 1: Metadatenset im JSON-Format für die ANNO-Suche am Beispiel des Begriffs Börse*, http://anno.onb.ac.at, Abfrage am 19.2.2019.

menhänge erkennen und erlauben Annahmen hinsichtlich mancher Determinanten der Börsenberichterstattung.

3.3.1 Suchresultate nach Anzahl der Treffer, Ausgaben und Titel

Die Volltextsuche ergab zum Zeitpunkt der Erhebung[17] auf der Treffer- und Ausgabenebene unabhängig von den einzelnen Medientiteln folgende Ergebnisse (siehe Tab. 1). Eine erste deskriptiv-statistische Betrachtung zeigt, dass die Suchabfrage bei den Begriffen *Börse** und *Aktie** zu den höchsten Suchergebnissen führt. Auch die Normierung der Daten mit einer durchschnittlichen Anzahl der Treffer pro Ausgabe bestätigt dieses Bild. Mit sieben Treffern pro Ausgabe sind die Suchbegriffe *Börse** und *Aktie** häufiger zu finden als die anderen börsenspezifischen Termini.

17 Diese erfolgte am 19.2.2019 für die Suchbegriffe *Börse** und *„Wiener Börse“* und am 14.3.2019 für die Begriffe *Aktie**, *Obligation** und *Pfandbrief**. Zumal ANNO historische Zeitungen und Zeitschriften kontinuierlich digitalisiert und verfügbar macht, werden diese Ergebnisse von einer Abfrage, die zu einem anderen Zeitpunkt durchgeführt wurde, zwangsläufig abweichen.

Tab. 1: Ergebnisse der keyword-basierten Suche nach Anzahl der Treffer, Ausgaben und Titel.

Suchbegriff	Anzahl der Treffer im Text	Anzahl der Ausgaben	Anzahl der Title
Börse*	1.685.401	226.976	174
"Wiener Börse"	235.597	116.074	123
Aktie*	1.158.312	164.647	171
Obligation*	594.101	145.470	158
Pfandbrief*	355.501	113.591	145

3.3.2 Suchergebnisse nach Medientiteln

Neben der Feststellung der Häufigkeit börsengängiger Begriffe in der Presse ermöglicht die Analyse auch einen Einblick in die Verteilung der Suchergebnisse auf die einzelnen Zeitungen (siehe Tab. 2). Dabei zeigt sich, dass quer über alle Suchbegriffe rund 90% der Suchergebnisse auf 20 Titel entfallen, während die übrigen 10% sich auf mehr als 100 Blätter verteilen. Dies lässt erste Schlüsse auf die Relevanz einzelner Zeitungen in der Entwicklung des Finanzjournalismus zu. Da die Erscheinungsdauern und -frequenzen der Top 20 Zeitungen teils deutlich variieren (einzelne decken den gesamten Untersuchungszeitraum ab wie die *Wiener Zeitung*, andere sind nur sehr kurz erschienen wie *Der Reporter*) und manche Titel zum Erhebungszeitungpunkt nicht vollumfänglich als Digitalisate verfügbar waren, ist es erforderlich, die Ergebnisse zu normalisieren. Dazu wurde zusätzlich die Anzahl der Treffer in Verhältnis zu der Anzahl der ausgewerteten Ausgaben gesetzt (siehe Anzahl der Treffer pro Ausgabe in Tab. 2). Diese Betrachtung lässt darüber hinaus Schlüsse auf die durchschnittliche Intensität der Börsenberichterstattung in den jeweiligen Blättern zu, welche zum Teil zu einer anderen Reihung führt als die Anzahl der Treffer gesamt.

Dabei zeigt sich quer über alle Begriffe, dass die politische Tagespresse, die sogenannte Großpresse, im Zusammenhang mit Börsennachrichten eine zentrale Stellung einnahm. Bezogen auf die gesamte Anzahl der Treffer und die Intensität[18] dominieren die *Neue Freie Presse* (1864–1914), die *Wiener Zei-*

18 Die Intensität wird als die durchschnittliche Anzahl der Treffer pro Ausgabe pro Jahr definiert.

Tab. 2: Suchresultate für den Begriff *Börse** nach Zeitungstiteln in Wien 1771–1914.

Suchbegriff: 'Börse*'

Rang	Titel	Erscheinungs-zeitraum	Erhebungs-zeitraum	Auswertungs-zeitraum in Jahren	Anzahl der Ausgaben gesamt	Geschlossener Bestand J/N	Anzahl der Treffer gesamt	%	% kum.	Anzahl der Treffer pro Ausgabe	Rang
(1)	Neue Freie Presse	1864-1939	1864-1914	51	17 962	ja	325 003	19%	19%	18,1	(2)
(2)	Wiener Zeitung	1703-heute	1771-1914	144	31 564	ja	218 491	13%	32%	6,9	(11)
(3)	Die Presse	1848-1896	1848-1896	49	16 165	ja	176 384	10%	43%	10,9	(7)
(4)	Neues Wiener Tagblatt (Tages-Ausgabe)	1867-1945	1867-1914	43	14 164	nein	167 151	10%	53%	11,8	(5)
(5)	(Neuigkeits) Welt Blatt	1874-1943	1874-1914	41	11 959	ja	102 900	6%	59%	8,6	(9)
(6)	Das Vaterland	1860-1911	1860-1911	52	17 059	ja	81 634	5%	64%	4,8	(18)
(7)	Neues Wiener Journal	1893-1939	1893-1914	22	7 474	ja	59 617	4%	67%	8,0	(10)
(8)	Wiener Allgemeine Zeitung	1880-1934	1880-1914	10	3 416	nein	53 818	3%	70%	15,8	(4)
(9)	Deutsches Volksblatt	1889-1922	1889-1914	26	8 662	ja	50 541	3%	73%	5,8	(13)
(10)	Morgen-Post	1854-1886	1854-1886	33	9 144	ja	49 427	3%	79%	5,4	(15)
(11)	Fremden-Blatt	1847-1919	1847-1914	33	8 370	nein	48 527	3%	76%	5,8	(14)
(12)	Neues Fremden-Blatt	1865-1876	1865-1876	12	8 370	ja	33 434	2%	81%	4,0	(20)
(13)	Reichspost	1893-1938	1893-1914	20	5 702	nein	30 402	2%	83%	5,3	(17)
(14)	Wiener Handelsblatt	1865-1877	1865-1877	12	2 399	nein	26 509	2%	84%	11,1	(6)
(15)	Gemeinde-Zeitung	1862-1890	1862-1890	16	3 524	nein	18 827	1%	86%	5,3	(16)
(16)	Deutsche Zeitung	1871-1907	1872-1874	3	771	nein	18 470	1%	87%	24,0	(1)
(17)	Illustrirtes Wiener Extrablatt	1872-1928	1872-1914	6	1 938	nein	16 961	1%	88%	8,8	(8)
(18)	Wiener Sonn- und Montags-Zeitung	1863-1936	1863-1914	47	2 723	nein	16 240	1%	89%	6,0	(12)
(19)	Der Reporter	1870-1875	1870-1875	4	865	nein	14 804	1%	90%	17,1	(3)
(20)	Wiener Landwirtschaftliche Zeitung	1868-1943	1868-1914	38	2 632	nein	12 544	1%	90%	4,8	(19)
	Sonstige						163 717	10%	100%		
	Gesamtergebnis						**1 685 401**	**100%**			

tung (1771–1914), die *Presse* (1848–1896), das *Neue Wiener Tagblatt* (1867–1914), das *Vaterland* (1860–1911), das *Fremden-Blatt* (1847–1914) sowie *(Neuigkeits) Welt Blatt* (1874–1914).[19] Bis auf das *(Neuigkeits) Welt Blatt* wurden alle anderen Blätter vor dem Börsenkrach 1873 gegründet. Inwieweit diese Zeitungen die Börsenberichterstattung im Zuge der Börseneuphorie intensivierten und zu dieser möglicherweise beitrugen, kann anhand der kombinierten Frequenz- und Intensitätsanalyse einzelner Zeitungen veranschaulicht und diskutiert werden (siehe Abb. 2).

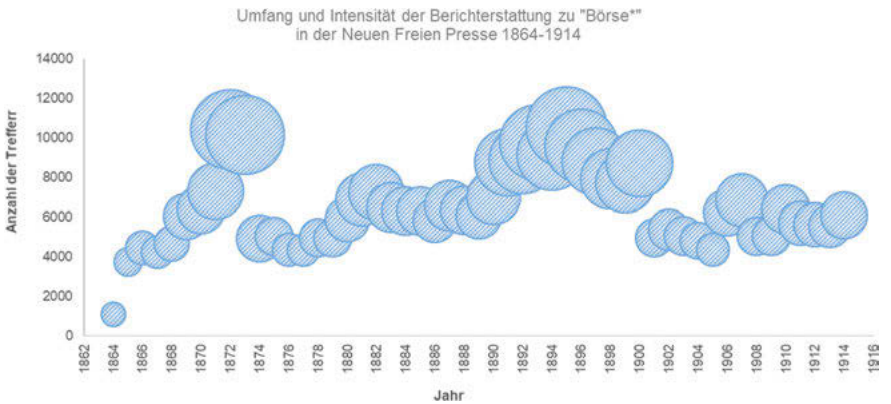


Abb. 2: Frequenz- und Intensitätsanalyse zum Begriff *Börse** in der *Neuen Freien Presse* 1864–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

Angesichts der Leserkreise der 20 Zeitungen kann angenommen werden, dass Börsennachrichten so gut wie alle Bevölkerungsschichten erreichen konnten. Zu der Leserschaft der *Neuen Freien Presse* zählte vor allem das Großbürgertum, das liberale Bürgertum sowie die in der Donaumonarchie lebenden assimilierten Juden. Das mittlere und kleinere Bürgertum sowie Beamte wurden durch das *Neue Wiener Tagblatt* erreicht. Diese beiden Zeitungen waren bisherigen Erkenntnissen zufolge börsenneutral bis börsenfreundlich eingestellt und wurden insbesondere von Ärzten, Advokaten, Gutsbesitzern, Fabrikanten, Professoren und Schriftstellern gelesen (Weiss 1958; Haider und Haller 1998, 94). Das zumindest auf den ersten Blick

19 Die Zeitangaben beziehen sich nicht auf die Erscheinungsdauer der jeweiligen Zeitung, sondern auf den hier relevanten Betrachtungszeitraum. Der Erscheinungszeitraum der *Wiener Zeitung* erstreckt sich von 1703 (bis 1780 unter dem Titel Wienerisches Diarium) bis heute. Das *Neue Wiener Tagblatt* erschien bis 1945, das *Fremdenblatt* bis 1919 und das *(Neuigkeits) Welt* Blatt bis 1943.

tendenziell börsenfeindliche *Vaterland* galt hingegen als das Organ der föderalistisch-feudalen Aristokratie (Veith 1937; Bruckmüller 2001, 279). Gemäß der ANNO-basierten Analyse gab es in Wien nur wenige fachspezifische Börsen- bzw. Finanzzeitungen.[20] Unter den wichtigsten 20 Blättern rangieren nur zwei, das *Wiener Handelsblatt* (1865–1877) und *Der Reporter* (1870–1875). Die Existenz beider Blätter währte mit zwölf bzw. fünf Jahren allerdings nicht lange.

3.3.3 Suchresultate im Zeitverlauf von 1771 bis 1914

Die Betrachtung der Suchergebnisse im Zeitverlauf offenbart zum einen, wann einzelne börsenspezifische Begriffe erstmals Eingang in die Presse fanden und zum anderen, welche Entwicklung die Präsenz dieser Begriffe in der Presse in dem Untersuchungszeitraum nahm. Nicht zuletzt lässt die politische, wirtschaftliche und gesellschaftliche Kontextualisierung dieser Entwicklung erste Schlüsse auf mögliche konstitutive Einflussfaktoren zu. So wird ersichtlich, dass mit der Gründung der Börse 1771 nicht zwangsläufig auch ihre Medialisierung vonstattenging, und dass börsenrelevante Begriffe in der Wiener Presse erst ab den 1820er-Jahren zwar häufiger, aber noch immer nur im geringen Ausmaß vorkamen. Auffallend ist bei allen fünf Begriffen die starke Konzentration der Treffer in der zweiten Hälfte des Untersuchungszeitraums (siehe Abb. 3 und 4). Anteilsmäßig macht die Anzahl der Treffer zu *Börse** bis 1848 nur 2% (34.585 von 1.685.401) aus, die restlichen 98% verteilen sich sehr unterschiedlich auf die Zeitspanne zwischen 1849 und 1914. Gleichzeitig zeigt sich, dass mit der vorübergehenden Lockerung der Zensur 1848 und dem darauffolgenden Anstieg neuer Zeitungen und Zeitschriften im Raum Wien (Paupié 1966) Börsennachrichten nicht im gleichen Ausmaß Eingang in die Presse fanden und die Entwicklung der Börsenberichterstattung von den 1840er-Jahren bis zum Wirtschaftsboom in den späten 1860er-Jahren eine stetig gemäßigte war. Ein erster moderater Anstieg wird bei allen fünf

20 Hier muss allerdings kritisch angemerkt werden, dass ANNO den zwischen 1850 und 1914 in Wien existierenden Bestand an Finanz- und Börsenzeitungen derzeit noch nicht repräsentativ abbildet. Anhand von statistisch-historischen Studien zur Wiener Presse (Winckler 1875, Richter 1888, Zenker 1893), Zeitungsannoncen und städtischen Registern konnten für diesen Zeitraum insgesamt 27 Finanz- bzw. Börsenzeitungen festgestellt werden. Von diesen waren zum Zeitpunkt der Untersuchung nur die zwei genannten im ANNO für die Volltextsuche erschlossen. Die *Wiener Geschäftszeitung* (1865–1873) war zwar im ANNO als Digitalisat verfügbar, jedoch nicht für die Texterkennung erschlossen. Der Großteil der damals existierenden Finanz- und Börsenzeitungen fand demnach keinen Eingang in die quantitative Auswertung (zum Repräsentativitätsproblem von digitalen Zeitungsarchiven siehe auch Kapitel 3.4.1).

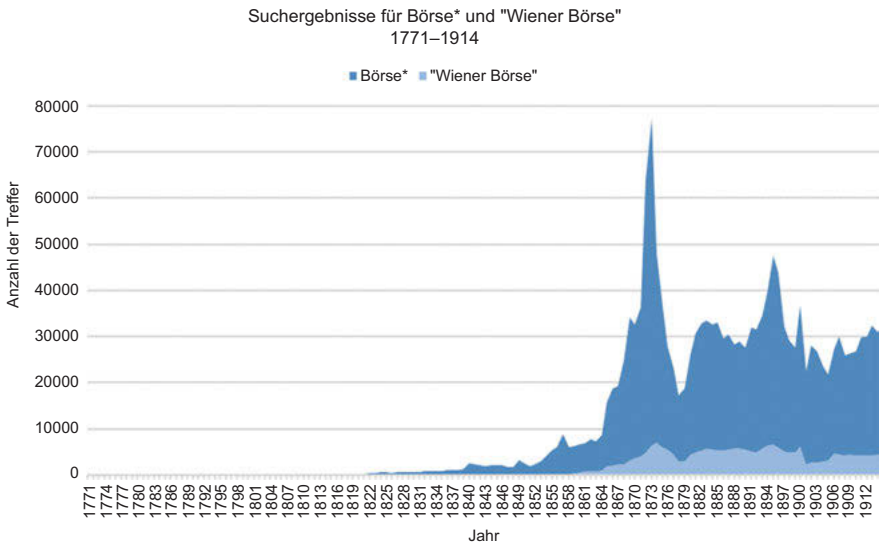


Abb. 3: Suchergebnisse für die Begriffe *Börse** und „*Wiener Börse*“ in der Wiener Presse 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

Suchbegriffen in den 1850er-Jahren sichtbar, bevor ab Mitte der 1860er-Jahre – getragen von positiver Konjunkturentwicklung, guten Erntejahren und dem Ausbau des Eisenbahnnetzes (Sandgruber 1995, 243), eine massive Zunahme börsenbezogener Themen in der Presse zu beobachten ist. Diese Entwicklung erreichte 1873 mit dem Wiener Börsenkrach ihren Höhepunkt. Auch wenn sich unmittelbar nachher wieder ein Rückgang der Medialisierung der Börse abzeichnet, blieb das Niveau in den Folgejahren insgesamt deutlich höher als vor der Krise der 1870er-Jahre.

Als Erklärungsansätze für den oben skizzierten und in Abb. 3 einsehbaren Verlauf kommen mehrere, sich vermutlich gegenseitig beeinflussende Faktoren in Betracht: (1) das steigende ökonomische Bewusstsein und das Interesse der Bevölkerung am wirtschaftlichen Geschehen zu partizipieren, wofür Information und Aufklärung unabdingbar waren, (2) der insbesondere durch die Märzrevolution 1848 eingeleitete medienpolitische Wandel (Lunzer 1992; Paupié 1960), (3) die von der Bevölkerung geforderte und letztlich gesetzlich verankerte Meinungs- und Pressefreiheit sowie die Anschaffung des Zeitungsstempels und der Inseratssteuer (Paupié 1960) und nicht zuletzt (4) der wirtschaftliche Aufschwung der 1850er- und der späten 1860er-Jahre, der auch das Interesse „des kleinen Mannes“ am Börsengeschehen weckte und ein an Finanzinformationen interessiertes Publikum hervorbrachte (Reich 1947; Veith 1937).

Ein ähnlicher Verlauf zeigt sich auch bei den an der Wiener Börse vordergründig gehandelten Finanzinstrumenten (siehe Abb. 4), der auch die bisherigen Erkenntnisse zur Diffusion von Finanzinstrumenten in Österreich-Ungarn (Komlos 1983; März 1983) widerspiegelt. Kurse der Staatspapiere (Obligationen) fanden bereits sehr früh Eingang in die Wiener Presse. Maria-Theresia bediente sich sehr bewusst der Presse, um die Bevölkerung für den Erwerb von Obligationen für die Zwecke der Staatsfinanzierung zu sensibilisieren (Franc 1952:22). Die Obligationen genossen ähnlich wie Pfandbriefe nicht die gleiche Popularität wie Aktien, aber auch sie waren in der Gründerzeit zwischen 1867 und 1873 deutlich stärker in der Presse thematisiert. Bei Aktien ist eine verzögerte, aber umso intensivere Entwicklung zu beobachten und auch hier korrespondiert die Präsenz der Aktie an der Wiener Börse mit jener in der Presse. Die ersten Aktien an der Wiener Börse (Aktien der Österreichischen Nationalbank) wurden 1818 emittiert und blieben bis in die 1840er-Jahre die einzigen (Baltzarek 1973, 44), bevor im Zuge des Eisenbahnbooms in den 1840er- und 1850er-Jahren weitere hinzukamen. Während der Wirtschaftskrise zu Beginn der 1860er-Jahre scheinen Aktien ihre Popularität vorübergehend eingebüßt zu haben, was sich in der Gründerzeit allerdings rasch änderte. Der steile Anstieg der Treffer zu *Aktie** vor 1873 und der genauso steile Fall nach dem Börsenkrach zeigen, wie sehr Aktien die Berichterstattung prägten. Das niedrige Niveau hielt bis zur Jahrhundertwende an. Erst nachher gewannen Aktien erneut deutlich an Präsenz, während Obligationen und Pfandbriefe weiterhin auf einem deutlich niedrigeren, jedoch sehr konstanten Niveau in der Presse vertreten waren.

3.4 Implikationen der Nutzung von ANNO in der pressehistorischen Forschung

Die Anwendung von ANNO eröffnet Forscherinnen und Forschern neue Perspektiven. Die Möglichkeit der Auswertung umfassender pressehistorischer Textkorpora über längere Zeiträume hinweg erlaubt es, sich dem Material mit neuen Forschungsfragen anzunähern, die ansonsten unbeantwortet bleiben müssten. Gleichzeitig geht die Forschung mit Hilfe pressehistorischer digitaler Portale, so auch von ANNO, mit Limitationen einher, deren sich Forscherinnen und Forscher bei der Konzeption und der Durchführung ihrer Projekte bewusst werden und die auch bei der Interpretation der Forschungsergebnisse Berücksichtigung finden sollten. Das betrifft im Wesentlichen die Repräsentativität und die Architektur des Portals, sprachlich-syntaktische und semantische Aspekte sowie die Qualität der Texterkennung.

Abb. 4: Suchergebnisse für *Aktie**, *Obligation** und *Pfandbrief** in der Wiener Presse 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

3.4.1 Repräsentativität der im ANNO erfassten Digitalisate

Digitale Portale sind Orte einer konstruierten Repräsentativität. Forschen in digitalen Umgebungen lenkt die Aufmerksamkeit zwangsläufig auf jene Quellen, die als Digitalisate vorhanden sind. Dadurch laufen Forscherinnen und Forscher Gefahr, pressehistorische Publikationen einer Epoche allein aufgrund der Tatsache, dass sie digital zugängig sind, als besonders bedeutsam wahrzunehmen, während nicht digitalisierte Quellen aus dem Blickfeld geraten. Der Fokus auf digitalisierte Quellen beim gleichzeitigen Ausblenden nicht digitalisierten Quellenmaterials verzerrt aber womöglich die Wahrnehmung der historischen Realität (Bingham 2010, 229). Obzwar der Bestand von ANNO bereits sehr groß ist, umfasst das Repository (noch) nicht alle Zeitungen bzw. Zeitungsausgaben. Auch in börsenjournalistischen Kontexten fehlen einige Periodika, vor allem Fachzeitungen und Fachzeitschriften wie z. B. die *Realzeitung* (1770–1786), die *Wiener Handlungszeitung* (1783–1785) oder der *Oesterreichische Oekonomist* (1869–1873), die in der Genese des österreichischen Wirtschafts- und Finanzjournalismus womöglich aber eine prägende Rolle spielten. Diese thematische Lücke ist insofern nicht überraschend, als dass Finanzjournalismus in der historischen Kommunikations- und Presseforschung bislang verhältnismäßig wenig Beachtung fand, die Digitalisie-

rungsstrategie von ANNO aber intensiv beforschte Themen priorisiert (ANNO/Österreichische Nationalbibliothek 2003). Die Bewusstmachung dieses Zusammenhangs hilft, die Einschätzung darüber zu treffen, inwieweit ANNO die für den eigenen Forschungsgegenstand relevanten Pressebestände abdecken könnte, und inwieweit davon auszugehen ist, dass relevante Bestände fehlen. Dies verdeutlich, dass die Nutzung digitaler Zeitungsarchive die qualitative Auseinandersetzung mit bisheriger Forschung nicht obsolet macht, sondern diese dem quantitativen Zugang vorangehen muss.

3.4.2 Einfluss der Portalarchitektur auf den Forschungsprozess

Wie jedes digitale Archiv verfügt auch ANNO über eine spezifische Portalarchitektur mit eigenen Logiken des Selektierens, Indizierens und Systematisierens des historischen Pressematerials wie von Fickers (2016) beschrieben. Die Architektur mit vordefinierten Suchpfaden und Bewertungskriterien hilft auf der einen Seite, determiniert auf der anderen Seite aber auch die an das Portal gestellten Fragestellungen und somit die Forschungsperspektiven. Erschwerend kommt hinzu, dass die Metadaten und Algorithmen, auf denen die Filter- oder Klassifizierungskriterien basieren, nicht immer nachvollziehbar sind und sich manche Gliederungskriterien, bei ANNO z. B. das Kriterium *Relevanz*, dadurch den Forscherinnen und Forschern nicht erschließen. Die Einschränkung bei der Ausgabe der Suchresultate auf Top 10, z. B. bei der Reihung der Ergebnisse nach Titeln, lenkt die Aufmerksamkeit auf die überrepräsentierten Medien, wobei ihr Größenverhältnis zu den restlichen Resultaten verborgen bleibt.

Ein weiteres, im Hinblick auf die textstatistische Auswertung auftretendes Problem stellt der Ausgabemodus von ANNO dar. Die Summe der Ergebnisse bezieht sich auf die Anzahl der Zeitungsausgaben und nicht auf die Summe der Treffer im Text (siehe Abb. 5). Die Trefferanzahl ist jedoch für die meisten statistischen Textanalysen die wichtigste Erhebungsebene und insbesondere dann von Relevanz, wenn die Präsenz spezifischer Begriffe über längere Zeiträume erfasst und analysiert werden soll und die Erscheinung dieser Begriffe von Titel zu Titel und von Ausgabe zu Ausgabe stark schwankt. In dem Fall hat die Anzahl der Ausgaben nur eine unzureichende Aussagekraft und führt zu einer verzerrten Wahrnehmung der medialen Darstellung spezifischer Phänomene. Die Extraktion der Daten auf Trefferebene ist bei einem ähnlich gelagerten Forschungsinteresse daher jedenfalls anzuraten.

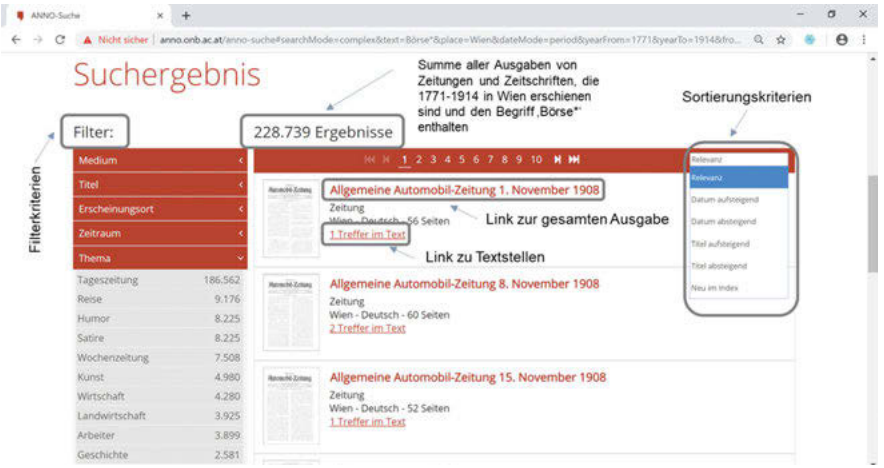


Abb. 5: Ausgabe der Suchresultate im ANNO, http://anno.onb.ac.at, Abfrage am 19.2.2019.

3.4.3 Sprachlich-syntaktische und semantische Aspekte

Weitere Einschränkungen betreffen die sprachlich-syntaktische und semantische Ebene. Diese Limitationen sind weitestgehend vom Portal unabhängig, und insofern in der Arbeit mit jedem pressehistorischen Repository mitzubedenken. Hier kommt vor allem der Wandel der Schreibweise und der Terminologie im Zeitverlauf zum Tragen. In Finanzkontexten ist es z. B. die orthographische Veränderung des Begriffs *Aktie*. Bis ins 19. Jahrhundert hinein wurde vorwiegend die Schreibweise mit c (Actie) bevorzugt. Es kann allerdings kein spezifischer Zeitraum ausgemacht werden, wann diese Schreibweise durch jene mit k (Aktie) ersetzt wurde. Die Schreibweise war weniger vom Zeitraum, sondern vielmehr vom jeweiligen Titel abhängig. Diesem Problem kann im ANNO durch die Nutzung von Wildcards (A?tie) begegnet werden. Ein ähnliches Phänomen lässt sich auch am Begriff *Cours* bzw. *Kurs* (im Sinne eines Wertpapierkurses) beobachten. Hier hat sich im Laufe des 19. Jh. die neue Schreibweise *Kurs* durchgesetzt. Bei der keyword-basierten Suche führt die Eingabe von *Cours* im *Österreichischen Beobachter* zu lediglich 61 Ergebnissen (Ausgaben), die Eingabe von *Kurs* in derselben Zeitung hingegen zu 4.010 Ergebnissen (Ausgaben).[21] Ähnlich wie beim Begriff *Actie* bzw. *Aktie* hat sich die neue Schreibweise nicht plötzlich durchgesetzt, vielmehr vollzog sich der sprachliche Wandel eher schleichend, denn die 61 Ausga-

21 Abfrage im ANNO am 6.6.2020.

ben mit der alten Schreibweise verteilen sich auf die Dauer von 1811 bis 1847. An diesem Beispiel zeigt sich die Gefahr der Fehleinschätzung der Begriffsrelevanz beim Erschließen eines Themas und verdeutlicht die Wichtigkeit der Kenntnis zeitgenössischer Ausdrucksweise. Anhand des ersten Ergebnisses würde der *Österreichische Beobachter* möglicherweise als eine in der Genese des österreichischen Finanzjournalismus irrelevante Zeitung eingestuft werden, wohingegen das zweite Ergebnis die Bedeutung der Zeitung als sehr hoch einstufen lassen würde.

Darüber hinaus ist die keyword-basierte Suche bedeutungs- und kontextunabhängig. Das geht mit der Problematik einher, dass die Suchergebnisse gegebenenfalls auch jene Begriffe inkludieren, die in dem spezifischen Untersuchungskontext irrelevant sind. Anhand des Begriffs *Börse* kann die Problematik gut veranschaulicht werden. Im Rahmen der vorliegenden Untersuchung wird *Börse* grundsätzlich als Handelsplatz für Wertpapiere definiert. In medialen Diskursen kann *Börse* sinngemäß allerdings auch als Institution, Gebäude oder Standort thematisiert werden. Der Begriff *Börse* wird aber auch als Ausdruck für Brieftasche oder Portemonnaie verwendet, was einen Teil der Suchergebnisse für den Forschungskontext unbrauchbar macht. Das ist z. B. dann der Fall, wenn Zeitungen über gestohlene Börsen im Sinne eines Brieftaschendiebstahls berichteten[22], oder wenn sich in Frauenzeitschriften Anleitungen zum Sticken einer Börse finden.[23] Selbst wenn der Kontext ähnlich beschaffen ist und Börse in journalistischen Beiträgen als Markt thematisiert wird, besteht die Möglichkeit, dass in dem konkreten Beitrag der Markt für landwirtschaftliche Produkte und nicht Wertpapiere gemeint war.[24] In Ermangelung der Möglichkeit semantischer Suche können Forscherinnen und Forscher die Suche in ANNO mit Booleschen Operatoren, Wildcards, Phrasensuche oder Abstandssuche präzisieren und so die Treffsicherheit erhöhen.

Neben den orthografischen Veränderungen war bei der Datenerhebung auch der Wandel von Finanzterminologie zu bedenken. Am Beispiel des *Österreichischen Beobachters* können die Veränderungen in der Börsenberichterstattung bei den beiden für Schuldverschreibungen verwendeten Begriffen *Obligation* und *An-*

22 Siehe z. B. *Morgen-Post*, Ausgabe vom 14.9.1864.

23 Siehe *Allgemeine Frauen-Zeitung*, Jg. 1892, Nr. 8.

24 Siehe *Wiener Zeitung*, Ausgabe vom 11.1.1900. Die Presseberichterstattung über die „Börse für landwirthschaftliche Produkte" findet sich nicht nur in facheinschlägigen Blättern, sondern auch in allen größeren Tageszeitungen wie der *Wiener Zeitung*, der *Neuen Freien Presse*, der *Presse* etc. Eine Volltextabfrage für ‚Börse für landwirthschaftliche Produkte' ergab 4.699 Treffer in 4.499 Ausgaben, das entspricht 0,0279% des Gesamtergebnisses für den Suchbegriff *Börse**, welche sich allerdings nur auf den Zeitraum von 15 Jahren zwischen 1890 und 1904 beschränken und somit das Gesamtbild nicht wesentlich verändern.

leihe veranschaulicht werden. Der Begriff Anleihe setzte sich gegenüber dem Obligationsbegriff nach 1820 nachhaltig durch, obwohl beide Begriffe nach wie vor in der Finanz- und Börsenberichterstattung repräsentiert waren (siehe Abb. 6). Auch der Begriff *Staatspapiere* kam während der gesamten Erscheinungsdauer des *Österreichischen Beobachters* vor, im Vergleich zu *Anleihen* und *Obligationen* allerdings in einem deutlich niedrigeren Ausmaß.[25]

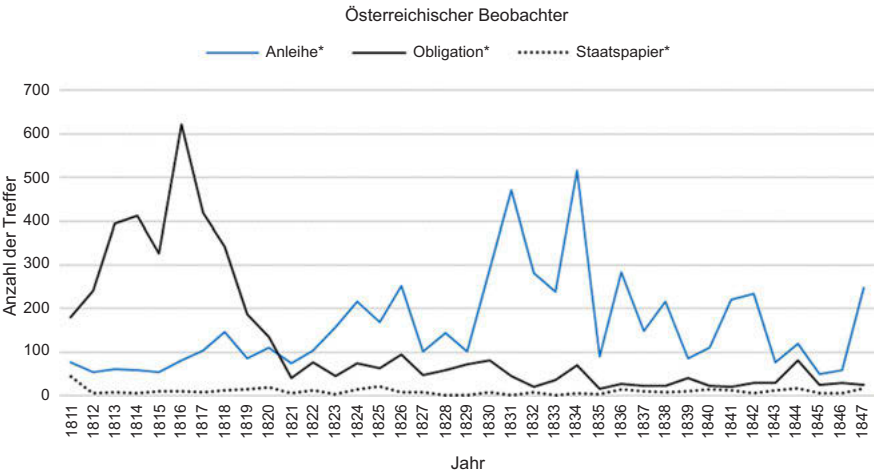


Abb. 6: Begriffswandel in der Finanzberichterstattung am Beispiel des *Österreichischen Beobachters* 1811–1847, https://labs.onb.ac.at, Abfrage am 19.2.2019.

Die Abkehr von der fachspezifischen Ausdrucksweise kann als ein Versuch gedeutet werden, die Bedürfnisse des lokalen Lesepublikums zu bedienen, zu dem nicht (mehr) ausschließlich Kaufleute, Händler und mit dem Finanzwesen vertraute Personen zählten, sondern zunehmend auch breite Bevölkerungsschichten, die für den Erwerb von Staatsanleihen animiert werden sollten. Die mangelnde Kenntnis der Finanz- und Börsenbegriffe sowie ökonomischer Zusammenhänge, welche in weiten Teilen der Laienpublika vermutet werden darf, stellte in diesem Zusammenhang ein Hindernis dar. Die zumindest teilweise Abschaffung der sprachlichen Barriere zielte möglicherweise auf ein besseres Lese- und Interpretationsverständnis von Börsennachrichten im Sinne von Hall (1973) ab.

25 Zum Erhebungszeitpunkt am 15.2.2020 ergab die Suche im ANNO für den Suchbegriff *Anleihe** 5.899 Treffer in insgesamt 2.954 Ausgaben und für den Begriff *Staatspapier** 387 Treffer in insgesamt 325 Ausgaben. Die Suchergebnisse für den Begriff *Obligation** wurden am 4.3.2019 erhoben und ergaben 4.487 Treffer in insgesamt 2.383 Ausgaben.

3.4.4 Texterkennung

Die Volltextsuche im ANNO basiert auf maschineller Erschließung der Quellenbestände mittels Optical Character Recognition (OCR). Dabei kann es auch bei ANNO-Beständen in manchen Texten zu einer hohen Fehlerdichte kommen (Resch 2018, 22). Dies kann mehrere Ursachen haben wie uneinheitliche Schrift, ungleiche Zeichen- und Wortabstände sowie Text- und Silbentrennung, schlechte Druckqualität oder Unschärfe des Bildes, was insbesondere die Differenzierung von ähnlichen Buchstaben wie z. B. des „Schaft-s“ und „f“ der Frakturschrift schwierig macht (Gupta, Jacobson, und Garcia 2007).

Auch in den OCR-erschlossenen Jahrgängen treten vereinzelt Lücken aufgrund von Problemen mit der Texterkennung auf. Als Beispiel kann die *Morgen-Post* angeführt werden. Die Zeitung ist zwar als Digitalisat über ihren gesamten Erscheinungszeitraum von 33 Jahren (1854–1886) verfügbar, die Volltextsuche für den Begriff *Börse** blieb in den Jahren 1865–1868 allerdings ergebnislos. Gleichzeitig war die Trefferquote vor und nach diesem Zeitraum unverändert hoch. Angesichts des konkreten Zeitraums, der von einer allmählichen wirtschaftlichen Erholung geprägt war, schien es auch unwahrscheinlich, dass das Blatt auf die Börsenberichterstattung gänzlich verzichtet hätte.[26] Eine genauere Betrachtung der Ausgaben zeigte, dass Börsennachrichten auch über diese Jahre auf einem unverändert hohen Niveau zum Inhalt der Zeitung gehörten. Mit einer leicht adaptierten Volltextsuche unter Anwendung von Booleschen Operatoren (B?rse statt Börse) konnte die Lücke letztlich geschlossen werden.

3.5 Visual Analytics mit Tools der ONBLabs

Für die Nutzung der mittlerweile sehr großen Datenbestände von ANNO im Rahmen von Forschungsvorhaben sind digitale Technologien erforderlich, welche insbesondere in der Datenanalyse und Datenanreicherung unterstützen und höhere Transparenz gewährleisten. Diese werden von ONBLabs (https://labs.onb.ac.at), einer relativ neuen Initiative der Österreichischen Nationalbibliothek, in Form unterschiedlicher Tools bereitgestellt. So können Forscherinnen und Forscher beispielsweise mit der IIIF-API und der Schnittstelle SACHA (Simple Access to Cultural Heritage Assets) eigene Sammlungen errichten, die auf ihren spezifischen Anforderungen und Interessen basieren, mit Python/Plotly interaktive Grafiken erstellen oder die offenen Metadaten mit SPARQL abfragen. ONBLabs

26 Stand April 2019.

stellen auch digitale Datensätze bereit, welche für eigene Zwecke weiterverwendet werden können. Diese Sammlungen können freigegeben, geklont, in andere Tools exportiert und heruntergeladen werden.[27] Im Rahmen des vorliegenden Forschungsprojektes wurden die zuvor aus den ANNO-Beständen gewonnenen und angereicherten Daten mit Hilfe interaktiver Grafiken analysiert, welchen eigenen IIIF-Kollektionen zugrunde liegen.

In dem vorliegenden Forschungsprojekt lag der Schwerpunkt auf der Visualisierung im Sinne einer explorativen Analyse und im Zuge dessen der Generierung entsprechender IIIF-Kollektionen.[28] Informationsvisualisierung[29] hat als eine „method for seeing the unseen" (McCormick, DeFanti, und Brown 1987) bei der Analyse insbesondere sehr großer Datenmengen gegenüber einer statischen, tabellarischen Darstellung oder der Narration mittels Text den Vorteil, multiple Informationen simultan darstellen zu können (Krempel 2005, 25) und Muster sichtbar zu machen, die sonst kaum an die Oberfläche hervordringen würden (Graham et al. 2016, 70–72). Die mit Hilfe der ONBLabs-Tools (Python, Plotly) erstellten interaktiven Graphiken erfüllen – über die Präsentations- und Kommunikationsaufgabe hinaus – vor allem die explorativ-analytische Funktion (Rehbein 2017, 332), wobei sie einerseits im Sinne vom komplementären *distant and close reading* (Moretti 2016) übergeordnete Strukturen sichtbar machen, andererseits aber auch detaillierte Einblicke in zeitliche oder sachliche Sequenzen erlauben.

3.5.1 Interaktive visuelle Darstellung der Medientitel

Für die Auswahl der Daten, die in Form von interaktiven Grafiken aufbereitet wurden, dienten die Ergebnisse der text-statistischen Analyse als Ausgangspunkt. Im ersten Schritt wurden jene zehn Zeitungen in den Fokus genommen, die bei dem Begriff *Börse** die höchste Frequenz und Intensität aufweisen. Diese können – nach Identifikation und Ausschluss kontextfremder Beiträge (siehe Kap. 3.4.3) – als für die historische Rekonstruktion der Genese des österreichischen Finanzjournalismus in weiterer Folge als besonders forschungsrelevant betrachtet werden (siehe Abb. 7).

27 Daten, die über ONBLabs bereitgestellt werden, werden unter CC0 lizenziert.

28 Das Projekt kann unter https://labs.onb.ac.at/en/topic/financial-news/ eingesehen werden (letzter Zugriff am 8.6.2020).

29 Definitorisch stellen Visual Analytics gegenüber der Informationsvisualisierung eine funktionale Erweiterung dar, indem die Informationsvisualisierung keinen Selbstzweck darstellt, sondern in den Forschungsprozess eingebunden ist (Rehbein 2017, 330).

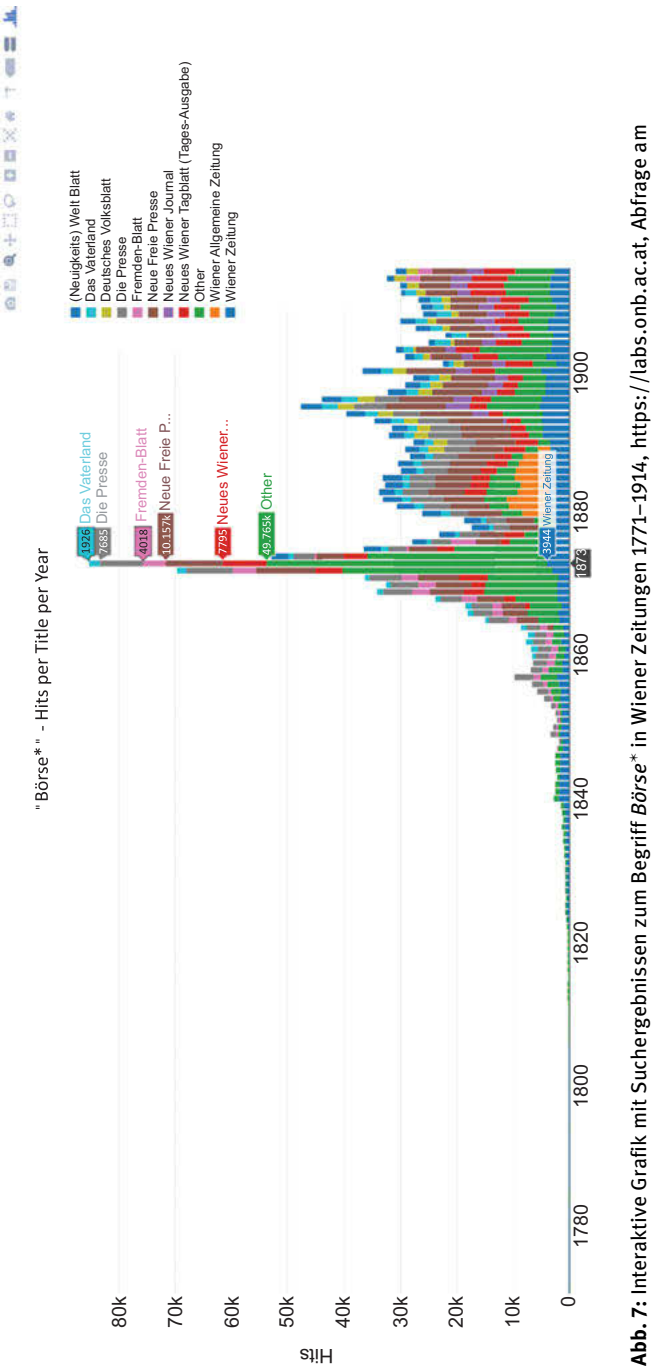


Abb. 7: Interaktive Grafik mit Suchergebnissen zum Begriff *Börse** in Wiener Zeitungen 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

Diese Darstellung zeigt die Veränderung der medialen Präsenz des Börsenbegriffs im Zeitverlauf und hilft damit, besonders berichtsintensive Zeitperioden und mögliche Einflussfaktoren in der Entwicklung der Berichterstattung zu identifizieren sowie Lücken im Datenbestand zu erkennen. Der Vorteil gegenüber einer rein statischen Darstellung liegt vor allem in der Möglichkeit, bestimmte Zeitungen und Zeiträume ein- und auszublenden und auf diese Weise schnell detaillierte Einsichten zu gewinnen (siehe Abb. 8).

Die oberen Graphiken basieren auf der Auswertung der Häufigkeit des Suchbegriffs *Börse** in den genannten Zeitungen. Diese Darstellung sagt jedoch wenig über die Intensität der Berichterstattung von bzw. über die Börse aus, zumal die einzelnen Periodika sich in Bezug auf die Erscheinungsdauer (diese umfasst z. B. bei der *Wiener Zeitung* 144 Jahre, bei der *Wiener Allgemeinen Zeitung* im Vergleich nur 10 Jahre) und die Periodizität unterscheiden. In Folge führt die Reihung der Zeitungen nach den beiden Kriterien, Frequenz und Intensität, teils zu unterschiedlichen Ergebnissen. Abbildung 9 visualisiert daher als Ergänzung zu den obigen Darstellungen die Intensität, in der der Suchbegriff *Börse** in den Top 10 Wiener Zeitungen vorkommt. Dabei wird die Intensität als die durchschnittliche Anzahl der Treffer im Text pro Ausgabe pro Jahr definiert.[30]

Diese Darstellung macht die Relevanz der Top 10 Zeitungen im Vergleich zu der insgesamt großen Anzahl der verbleibenden, ausgewerteten Zeitungen, die hier unter *Other* zusammengefasst werden, deutlich. Die Intensität ist hier trotz der kumulierten Betrachtung von 164 Blättern auffallend niedriger als in allen anderen Medien. Ebenso zeigt sich in Abb. 9 die langsam, aber kontinuierlich wachsende Intensität der Börsenberichterstattung in der *Wiener Zeitung* und die gesteigerte Intensität der Börsenberichterstattung in der Gründerzeit. Auch in dieser Grafik ist es möglich, einzelne Titel und Zeitabschnitte ein- und auszublenden und so schnell zu Detailinformationen zu gelangen.

3.5.2 Interaktive visuelle Darstellung der Publika

Der bisherigen Forschung zufolge stellen die Publika in ihrem Umfang und ihrer Zusammensetzung für das Aufkommen und die Entwicklung des Finanzjournalismus eine konstitutive und prägende Bedingung dar (u. a. Langenohl und Wetzel 2014; Radu 2017). Daher war ein Teil der Untersuchung der Frage gewidmet, welche Wiener Bevölkerungsgruppen als Leserkreise spezifischer Zeitungen und Zeitschriften mit Börsennachrichten bzw. börsenbezogenen The-

30 Damit wird auch die schwankende Anzahl der Ausgaben von Jahr zu Jahr normalisiert.

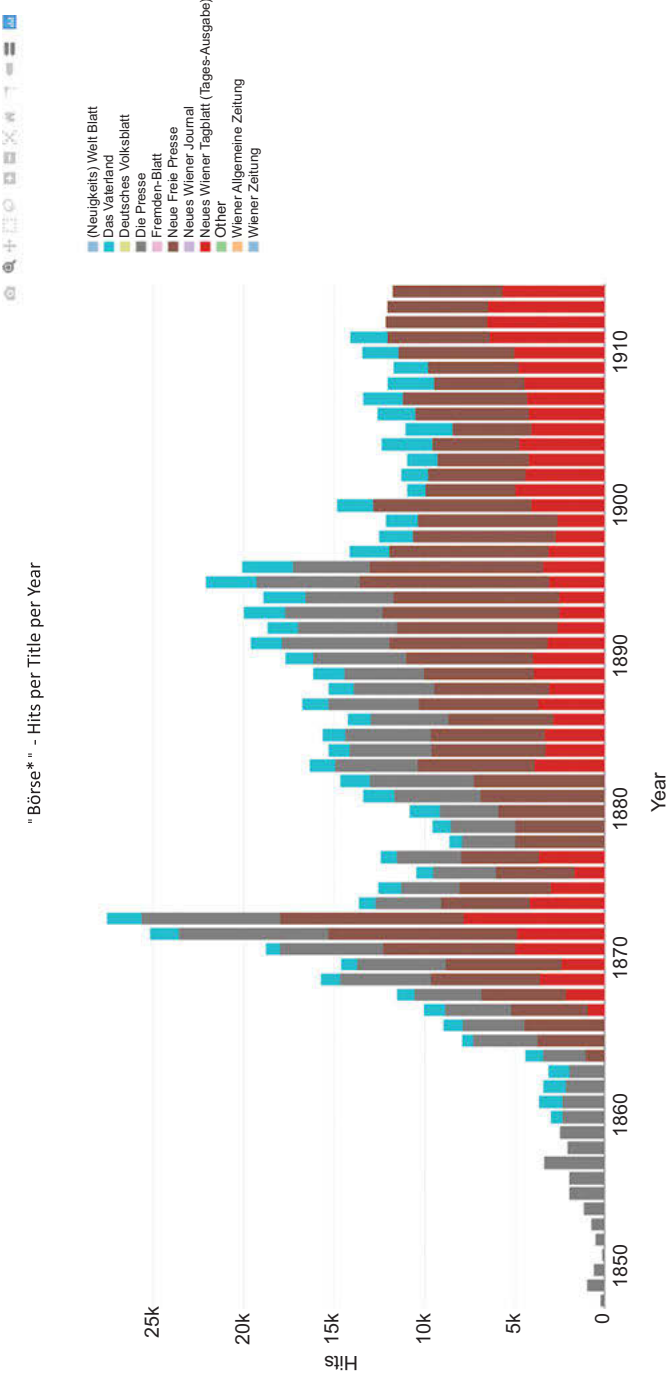


Abb. 8: Detailansicht der interaktiven Graphik zum Begriff *Börse** in Wiener Zeitungen 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

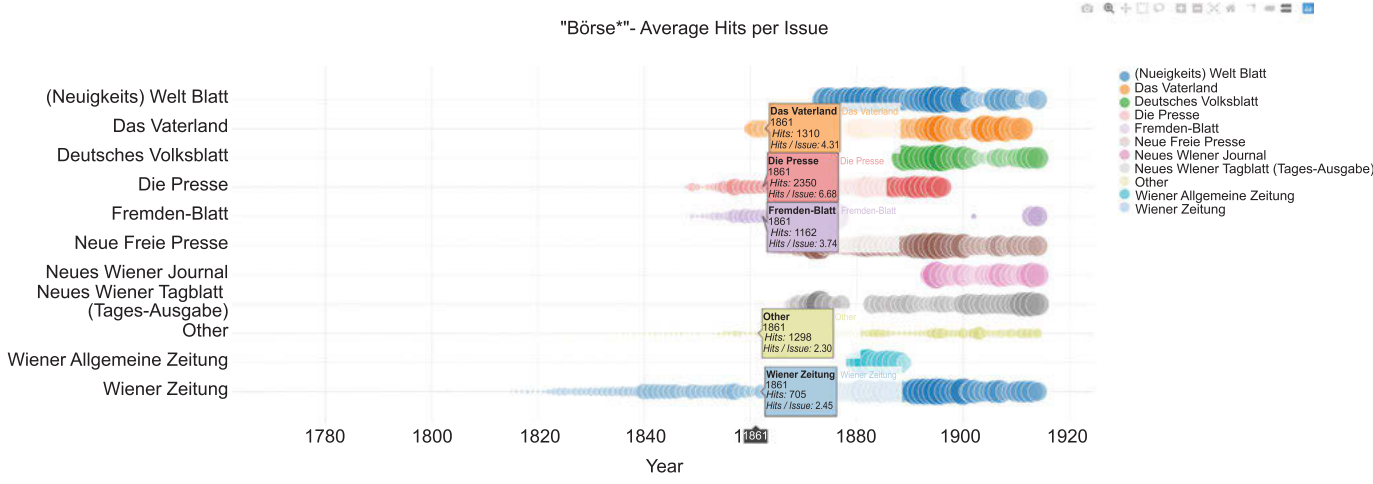


Abb. 9: Intensität der Berichterstattung zum Suchbegriff *Börse** in der Wiener Presse 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

men in Berührung kamen, und ob zwischen den Publika in Bezug auf einzelne Zeitungen und Zeiträumen Unterschiede festgemacht werden können.

Derzeit lassen die Gliederungs- und Filterkriterien von ANNO kaum oder nur sehr eingeschränkt Schlüsse auf spezifische Rezipientenkreise zu. Zwar können die Ergebnisse u. a. nach dem Kriterium *Thema* gefiltert werden, über das z. B. *Arbeiter* als eine eigene Kategorie angeführt werden, jedoch weisen die Kategorien insgesamt weder eine durchgängige Systematik auf noch sind die Zuordnungslogiken für Forscherinnen und Forscher transparent und nachvollziehbar.[31] Daher wurden in dem vorliegenden Forschungsprojekt die extrahierten Daten (Titel, Datum, Anzahl der Treffer) um zwei weitere Kriterien, Kontext und Zielgruppe, manuell angereichert. Die zeitungsspezifischen Kontexte wurden auf Basis der inhaltlichen Ausrichtung der jeweiligen Blätter beschrieben.[32] Aufbauend auf dieser Kategorisierung wurden unter Berücksichtigung zeitgenössischer Quellen (v. a. Winckler 1875; Richter 1888; Zenker 1893) und der Erkenntnisse bisheriger Forschung zu zeitungsspezifischen Publika (u. a. Veith 1937; Reich 1947; Weiss 1958; Haider und Haller 1998) sowie der Sichtung vorhandener Pränumerationsanzeigen der jeweiligen Blätter die Leserkreise definiert und damit die potenziell erreichbaren Schichten der Wiener Bevölkerung identifiziert.

Wie die explorative Analyse der 20 wichtigsten Zeitungen bereits gezeigt hat (siehe Tab. 2), war die politische Tagespresse, die sogenannte Großpresse, die bedeutendste Mediengattung bei der Verbreitung von Börsennachrichten wie Börsenkursen und Kommentaren zum Börsengeschehen, aber auch von sonstigen investorenrelevanten Informationen: Ankündigungen von Generalversammlungen und Bekanntgaben der dort gefassten Beschlüsse, Kommentare zur Entwicklung einzelner Unternehmen, Informationen über Neubesetzungen von Präsidien und Verwaltungsräten oder auch der Lob und die Kritik an deren Tätigkeit und schließlich zahlreiche Anzeigen, die für Aktienerwerb und die Partizipation an der Börsenspekulation warben. Diese Zeitungen weisen im Vergleich auch die höchsten Auflagenzahlen auf. Allein die *Neue Freie Presse*, in

31 Beim Filterkriterium *Thema* sind die Gliederungskategorien Periodizität (Tageszeitungen, Wochenzeitungen), Kontexte/Interessen (Reise, Wirtschaft, Satire, Humor, Landwirtschaft, Kunst, Sport) sowie Zielgruppen (Arbeiter), wobei Satire und Humor stets zu gleichen Ergebnissen führen und weitere Überschneidungen und Mehrfachzuordnungen von ANNO nicht transparent offengelegt werden.

32 Die induktiv hergeleiteten Kontextkategorien sind Politik, Wirtschaft, Land- und Forstwirtschaft, Recht, Religion, Medizin, Militär, Kunst, Humor/Satire, Haushalt und Mode, Jugend, Pädagogik, Technik, Sport und Berufsfachzeitschrift. Da die politische Tagespresse aufgrund der inhaltlichen Vielfalt nicht nur einem Thema zuordenbar ist, wurde noch die Kategorie *Politik, Wirtschaft, Kunst* hinzugefügt.

der auch die Börsenberichterstattung von ihrer Gründung im Jahr 1864 an eine große Rolle spielte, zählte mit einer Auflage von 25.000 Exemplaren in den 1870er-Jahren, die bis auf 55.000 Exemplare um die Jahrhundertwende anwuchs, zu den auflagenstärkten Zeitungen Europas (Walter 1994:51). Die gewichtige Rolle der Tagespresse wird in Abb. 10 (grüne Säulen) augenscheinlich. Die Tagespresse adressierte viele Gesellschaftsschichten und sprach damit ein breites Publikum an, wenn auch einzelne Blätter tendenziell von liberalen (Bürgertum) oder konservativen (Aristokratie) Kreisen präferiert wurden.[33] Aber auch bei Humor und Satire sowie bei Sport- und Freizeitzeitschriften wurde aufgrund bisheriger Forschung (u. a. Haas 1982) von der Rezeption durch breite Publika ausgegangen.

Neben der breiten Öffentlichkeit konnten auf Basis facheinschlägiger Zeitungen und Zeitschriften (Special-Interest-Magazine) sowie Berufsfachzeitschriften, in denen auch Börsenthemen adressiert wurden, spezifische Leserkreise als Zielgruppen der Börsenberichterstattung beschrieben werden.[34] Die Detailansicht der Grafik zeigt, dass diese spezifischen Gruppen vor und nach dem Wiener Börsenkrach 1873 sowie im Zuge des Börsenhypes und der Finanzkrise 1895 intensiv mit Börsenthemen konfrontiert wurden, was diese Bevölkerungsgruppen in den Augenmerk der weiteren Untersuchungen rückt. Neben den breiten Bevölkerungsschichten, welche mit der Tagespresse, Humor- und Freizeitzeitschriften erreicht wurden, zählen in der Gründerzeit vor allem die Handel- und Gewerbetreibenden zu den Adressaten von Börsennachrichten, genauso wie Ärzte, Soldaten und Juristen. Interessant ist auch die Erkenntnis, dass kurz vor, aber besonders nach der Jahrhundertwende, das Thema Börse in einem auffallend hohen Ausmaß auch Eingang in Arbeiterzeitungen fand (siehe Abb. 11).

Jene Zeitungsausgaben, auf denen die interaktiven Graphiken basieren, können als IIIF-Kollektionen direkt über die jeweilige Grafik abgerufen werden. Diese Option stützt die Idee des komplementären *distant and close reading* (Moretti 2016) und ermöglicht den schnellen Zugriff auf das pressehistorische Material. Allerdings muss hier kritisch angemerkt werden, dass die derzeitigen Nutzungsmöglichkeiten mit Limitationen einhergehen. Zum einen ist die Anwendung der IIIF-API aus urheberrechtlichen Gründen zeitlich begrenzt, wes-

33 Diese Unterscheidung wird hier nicht getroffen, wäre aber als ein Aspekt der weiterführenden Forschung anzudenken.

34 Über die facheinschlägigen Zeitungen und Zeitschriften wurden Frauen, Kunst- und Kulturinteressierte, Land- und Forstwirte, Ärzte, Soldaten, Juristen, Erzieherinnen und Erzieher, Handel- und Gewerbetreibende sowie Jugendliche als Kategorien ausgemacht. Die Berufsfachzeitschriften richteten sich an Beamte, Buchhändler, Drogisten, Drucker, Feuerwehrmänner, Friseure, Hausarbeiterinnen, Hausbesitzer, Hebammen, Techniker und im Transportwesen Beschäftigte.

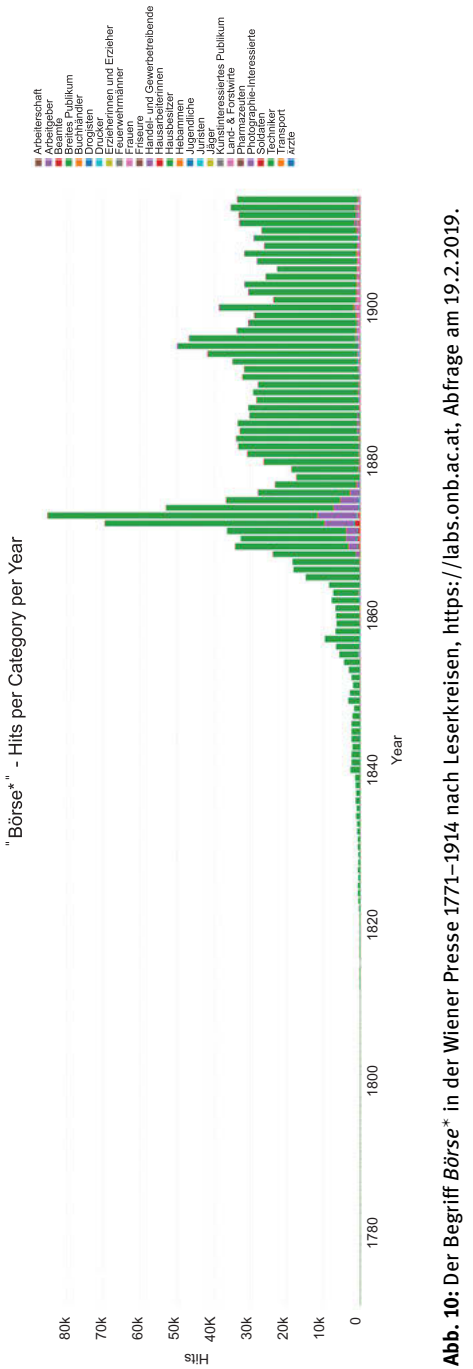


Abb. 10: Der Begriff *Börse** in der Wiener Presse 1771–1914 nach Leserkreisen, https://labs.onb.ac.at, Abfrage am 19.2.2019.

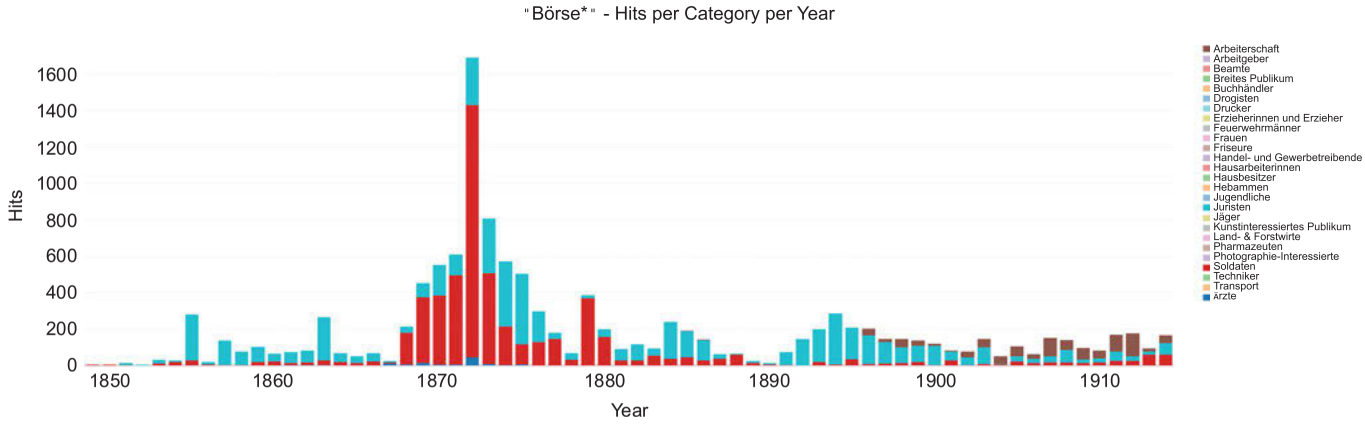


Abb. 11: Ausgewählte Leserkreise von Börsennachrichten in der Wiener Presse 1771–1914, https://labs.onb.ac.at, Abfrage am 19.2.2019.

halb die den Graphiken hinterlegten IIIF-Kollektionen nur Zeitungsausgaben bis zum Jahr 1878 umfassen. Spätere Ausgaben müssen direkt im ANNO-Archiv eingesehen werden. Ein weiterer Nachteil der aktuell verfügbaren Aufbereitung besteht darin, dass in den als IIIF-Kollektion verfügbaren Ausgaben die Hervorhebung von Suchbegriffen verloren geht, was die schnelle Identifikation der relevanten Stellen erschwert und erst recht die Sichtung der Digitalisate im ANNO-Portal notwendig macht. Zumindest der zweite Aspekt kann bei einer Fortsetzung des Projekts – das hier vorliegende Forschungsprojekt war eines der ersten in Kooperation mit den ONBLabs – adressiert und für Forschungszwecke weiter ausgebaut werden.

4 Conclusio und Ausblick

Das Forschen mit Hilfe digitaler Zeitungsarchive und Methoden eröffnet für kommunikations- und pressehistorische Forschung neue Forschungsperspektiven und Erkenntnismöglichkeiten, verändert aber auch die Forschungslogiken und -prozesse. Zu den wesentlichen Begleiterscheinungen des *digital turn* (Nicholson 2013), der auf den Fundamenten des *cultural turn* aufsetzen kann, zählt der Umstand, dass historische Zeitungen und Zeitschriften noch stärker und intensiver als Quelle historischer Forschung Beachtung finden und sich mit zunehmender Wahrscheinlichkeit auch in anderen Forschungskontexten als wertvolle Ressource erschließen lassen. Die aufgrund von zeitlichen und kognitiven Beschränkungen bisher in der historischen Kommunikations- und Presseforschung dominierenden qualitativen Forschungsdesigns können durch die Möglichkeit, umfassende Text- und Pressekorpora quantitativ auszuwerten, auf eine breite empirische Basis im Sinne von *culturomics* (Michel et al. 2011) gestellt werden. Die in diesen Datenlandschaften erkennbaren Muster und Strukturen sowie die abwechselnde und sich ergänzende Entfernung vom und Annäherung an das pressehistorische Material im Sinne von Morettis (2016) *distant and close reading* lassen Schlussfolgerungen über besonders relevante Zeiträume und Medien zu und offenbaren mögliche Zusammenhänge, denen in weiterführender Forschung nachgegangen werden kann.

Gleichzeitig müssen Forschungsprozesse neu gedacht und die damit verbundenen Limitationen berücksichtigt werden. Die Näherung an die Quellen erfolgt in der Anwendung digitaler Archive und Methoden nicht top-down (von Titel zum Wort), sondern bottom-up (vom Wort zum Titel), wobei die Analyse des Materials der gegebenen Portalarchitektur folgt, die dadurch unweigerlich den Forschungsprozess beeinflusst (Bingham 2010; Koenen 2018, 541; Resch 2018, 24). Da Forscherinnen und Forscher in Ermangelung von Programmierschnittstellen

die Portalarchitektur (z. B. Gliederungs- und Ordnungskriterien) vielfach nicht verändern können und als User vorgegebenen Strukturen folgen müssen (Brake 2012), sollten die der Portalarchitektur immanenten Recherchelogiken kritisch reflektiert und bei der Interpretation der Ergebnisse berücksichtigt werden. Mit den Anwendungen der ONBLabs wird diese Problematik, welche auch bei ANNO anzutreffen ist, von der ONB gezielt adressiert und im Sinne der Forschung versucht zu lösen.

Für die vorliegende Untersuchung erwies sich die Nutzung von ANNO und die Anwendung von ONBLabs trotz der skizzierten Limitationen für die Erkenntnisgewinnung als überaus hilfreich. So zeigte sich, dass die Gründung der Wiener Börse im Jahre 1771 als alleinige konstitutive Bedingung für die Medialisierung und Popularisierung von Börsennachrichten nicht ausreichend war und die Entfaltung von Finanzjournalismus weiterer Faktoren wie pressepolitischer Maßnahmen und technologischer Entwicklungen bedurfte. Gestützt auf eine breite empirische Basis konnte aufgezeigt werden, dass die Börsenberichterstattung insbesondere ab den 1850er-Jahren vor dem Hintergrund des Eisenbahnbooms, der pressepolitischen Lockerungen und der Einführung der Telegrafie an Bedeutung gewann und ihre Intensität mit finanzwirtschaftlichen und konjunkturellen Entwicklungen korreliert. Diese erreichte in der Gründerzeit ihren Höhepunkt und mit dem Wiener Börsenkrach 1873 vorerst ihr Ende. In der Krise 1895 findet dieses Phänomen eine erneute Bestätigung. Dabei waren weniger finanzwirtschaftliche Fachzeitungen und -zeitschriften (z. B. Börsenzeitungen) von tragender Bedeutung als vielmehr die politische Tagespresse, mit deren Hilfe breite Bevölkerungsschichten erreicht werden konnten. Die Börse trat aber auch in anderen medialen Kontexten auf wie z. B. in humoristisch-satirischen Blättern, Sport- und Freizeitzeitschriften, Kunst- und Kulturzeitschriften, Kirchenzeitungen u. v. m. ebenso wie in Berufsfachzeitschriften, über welche spezifische Leserkreise angesprochen wurden.

Diese Aspekte zu beleuchten wäre ohne die Anwendung eines digitalen Archivs und digitaler Methoden kaum möglich gewesen. Nur so konnte der gewählte Betrachtungszeitraum (1771–1914) in diesem Umfang unter die Lupe genommen werden. Wie jedoch am Beispiel des dargelegten Forschungsprojektes auch mehrfach ersichtlich wurde, ersetzt die maschinell-gestützte, quantitative Analyse pressehistorischer Korpora nicht die Notwendigkeit einer qualitativ-heuristischen Herangehensweise (Lauer 2013; Nicholson 2013, 69), denn schlussendlich geht es darum, die extrahierten Informationen kontextbezogen zu interpretieren. Das setzt historisches Kontextwissen (Wettlaufer 2016) genauso voraus wie die Sichtung noch nicht digitalisierten Quellenmaterials (Bingham 2010, 229–30; Nicholson 2013, 67).

Dieser Beitrag soll die Möglichkeiten der digitalen Datensätze, Tools und Dienste der Österreichischen Nationalbibliothek aufzeigen, die im Rahmen kommunikations- und pressehistorischer Forschung Anwendung finden können. Das hier skizzierte Forschungsprojekt soll auch ermutigen, kollaborative und interdisziplinäre Ansätze bei der Generierung, Anreicherung und Wiederverwendung historischer Datensätze zu verfolgen. In diesem Projekt hat sich der Austausch zwischen Archivarinnen und Archivaren, Historikerinnen und Historikern sowie IT-Expertinnen und -Experten als sehr fruchtbar erwiesen und hat verdeutlich, dass digitale Zeitungsarchive auf den Austausch mit Forscherinnen und Forschern nicht verzichten sollten. Wenn digitale historische Zeitungs- und Zeitschriftenarchive gezielt (auch) der historischen Forschung dienen und zur Weiterentwicklung von *digital history* sinnvoll beitragen sollen, wird die Expertise von Historikerinnen und Historikern vor allem bei der Gestaltung von Portalarchitekturen, Definition von Klassifizierungen, Annotationen sowie der Integration von Knowledge Graphs im Sinne der semantischen Suche nicht nur sinnvoll, sondern unerlässlich sein.[35] Eine verstärkte Entwicklung in diese Richtung wäre jedenfalls zu begrüßen, um die bereits bestehenden technologischen Potenziale nicht nur in der kommunikations- und pressehistorischen Forschung zur Entfaltung zu bringen.

Bibliographie

ANNO/Österreichische Nationalbibliothek. 2003. ANNO – AustriaN Newspapers Online. A digitisation initiative of the Austrian National Library. Vortrag bei der Postkonferenz „Newspapers and the Press in Central and Eastern Europe: Access and Preservation“ bei der 69. IFLA Konferenz.

ANNO/Österreichische Nationalbibliothek. 2019. ANNO-Suche. Volltextsuche in ausgewählten Zeitungen.

Baltzarek, Franz. 1973. Die Geschichte der Wiener Börse: Öffentliche Finanzen und privates Kapital im Spiegel einer österreichischen Wirtschaftsinstitution. Wien.

Bingham, A. 2010. The Digitization of Newspaper Archives: Opportunities and Challenges for Historians. Twentieth Century British History vol. 21 (2) S. 225–231.

Bloch, Marc. 1994. Für eine vergleichende Geschichtsbetrachtung europäischer Gesellschaften. In: Alles Gewordene hat Geschichte: die Schule der ANNALES in ihren Texten; 1929–1992, Reclam-Bibliothek, herausgegeben von M. Middell und S. Sammler. Leipzig, S. 121–164.

Brake, Laurel. 2012. Half Full and Half Empty. Journal of Victorian Culture 17 (2), S. 222–229.

Bruckmüller, Ernst. 2001. Sozialgeschichte Österreichs. 2. Auflage, Wien.

35 Zu Annotationen siehe z. B. Rapp (2017).

Duchkowitsch, Wolfgang. 1980. Zeitung und Bibliothek. Der Stand der Erschließung österreichischer Zeitungen des 17. und 18. Jahrhunderts und Vorstellungen für den Soll-Zustand. In: Das historische und wertvolle Buchgut in der Bibliotheksverwaltung, herausgegeben von O. Mazal und E. Irblich. Wien, S. 55–61.

Fickers, Andreas. 2016. Digitale Metaquellen und doppelte Reflexivität. https://www.hsozkult.de/debate/id/diskussionen-2954 (letzter Zugriff, 17.3.2019).

Franc, Lucia. 1952. Die Wiener Realzeitung: ein Beitrag zur Publizistik der theresianisch-josefinischen Epoche. Dissertation, Universität Wien.

Graham, Shawn, Ian Milligan, und Scott Weingart. 2016. Exploring big historical data: the historian's macroscope. London.

Gupta, Maya R., Nathaniel P. Jacobson, und Eric K. Garcia. 2007. OCR Binarization and Image Pre-Processing for Searching Historical Documents. Pattern Recognition 40 (2), S. 389–397.

Haas, Hannes. 1982. Die politische und gesellschaftliche Satire der Wiener humoristisch-satirischen Blätter vom Zusammenbruch der Monarchie bis zum Justizpalastbrand (1918–1927). Dissertation, Universität Wien.

Haber, Peter. 2011. Digital past: Geschichtswissenschaft im digitalen Zeitalter. München: Oldenbourg Verlag.

Haider, Hans, und Haller, Günther (Hrsg.). 1998. 150 Jahre „Die Presse": ein Stück Österreich [237. Sonderausstellung des Historischen Museums der Stadt Wien in Zusammenarbeit mit der Zeitung „Die Presse", 16. Mai bis 30. August 1998], Sonderausstellung des Historischen Museums der Stadt Wien. Wien.

Hall, Stuart. 1973. „Encoding and Decoding in the television discourse".

Johnson, R. Burke, und Anthony J. Onwuegbuzie. 2004. Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational Researcher 33 (7), S. 14–26.

Koenen, Erik. 2018. Digitale Perspektiven in der Kommunikations- und Mediengeschichte: Erkenntnispotentiale und Forschungsszenarien für die historische Presseforschung. Publizistik 63 (4), S. 535–56.

Koenen, Erik, Christian Schwarzenegger, Lisa Bolz, Peter Gentzel, Leif Kramp, Christian Pentzold, und Christina Sanko. 2018. Historische Kommunikations- und Medienforschung im digitalen Zeitalter. medien & zeit 33 (2), S. 4–17.

Koller, Guido. 2016. Geschichte digital: historische Welten neu vermessen. 1. Auflage, Stuttgart.

Komlos, John. 1983. The Diffusion of Financial Technology into the Habsburg Monarchy Toward the End of the Nineteenth Century. In: Economic Development in the Habsburg Monarchy in the Nineteenth Century, her ausgegeben von J. Komlos. New York, S. 137–163.

Krempel, Lothar. 2005. Visualisierung komplexer Strukturen: Grundlagen der Darstellung mehrdimensionaler Netzwerke. Frankfurt am Main.

Langenohl, Andreas, und Dietmar J. Wetzel. 2014. Finanzmarktpublika: Moralität, Krisen und Teilhabe in der ökonomischen Moderne. 1. Auflage, Wiesbaden.

Lauer, Gerhard. 2013. Die digitale Vermessung der Kultur, Geisteswissenschaften als Digital Humanities. In: Big Data: das neue Versprechen der Allwissenheit, edition unseld Sonderdruck, herausgegeben von H. Geiselberger. Berlin, S. 99–116.

Lerg, Winfried. 1992. Programmgeschichte als Forschungsauftrag. In: Medien- und Kommunikationsgeschichte: ein Textbuch zur Einführung, Studienbücher zur Publizistik-

und Kommunikationswissenschaft, herausgegeben von M. Bobrowsky, W. Duchkowitsch, und H. Haas. Wien, S. 78–87.
Lunzer, Marianne. 1992. Parteien und Parteienpresse im wirtschaftlichen und gesellschaftlichen Wandel des 19. Jahrhunderts. In: Medien- und Kommunikationsgeschichte: ein Textbuch zur Einführung, Studienbücher zur Publizistik- und Kommunikationswissenschaft, herausgegeben von M. Bobrowsky, W. Duchkowitsch, und H. Haas. Wien, S. 105–115.
März, Eduard. 1983. The Austrian Crédit Mobilier in a Time of Transition. In: Economic Development in the Habsburg Monarchy in the Nineteenth Century, herausgegeben von J. Komlos. New York, S. 117–135.
McCormick, Bruce, Thomas DeFanti, und Maxine Brown. 1987. Visualization in Scientific Computing. Computer Graphics 21, S. 1–14.
Michel, J.B., et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, S. 176–182.
Moretti, Franco. 2016. Distant reading. Konstanz.
Nicholson, Bob. 2013. „THE DIGITAL TURN: Exploring the Methodological Possibilities of Digital Newspaper Archives". Media History 19 (1), S. 59–73.
Paupié, Kurt. 1960. Handbuch der österreichischen Pressegeschichte : 1848–1959. Band 1. Wien.
Paupié, Kurt. 1966. Handbuch der österreichischen Pressegeschichte : 1848–1959. Band 2. Wien.
Radu, Robert. 2017. Auguren des Geldes: eine Kulturgeschichte des Finanzjournalismus in Deutschland 1850–1914. Göttingen, Bristol.
Rapp, Andrea. 2017. Manuelle und automatische Annotation. In: Digital Humanities: eine Einführung, herausgegeben von F. Jannidis, H. Kohle, und M. Rehbein. Stuttgart, S. 253–267.
Rehbein, Malte. 2017. Informationsvisualisierung. In: Digital Humanities: eine Einführung, herausgegeben von F. Jannidis, H. Kohle, und M. Rehbein. Stuttgart, S. 328–342.
Reich, Josef. 1947. Die Wiener Presse und der Wiener Börsenkrach 1873 im wechselseitigen Förderungsprozess. Dissertation, Universität Wien.
Resch, Claudia. 2018. ‚Zeitungs Lust und Nutz' im digitalen Zeitalter. Partizipative Ansätze zur Erschließung historischer Ausgaben der Wiener Zeitung. medien & zeit 33 (2), S. 20–31.
Richter, Heinz Moritz. 1888. Die Wiener Presse. In: Wien 1848–1888. Denkschrift zum 2. December 1888. Band 2, herausgegeben von Gemeinderath der Stadt Wien, S. 407–468.
Sandgruber, Roman. 1995. Ökonomie und Politik: österreichische Wirtschaftsgeschichte vom Mittelalter bis zur Gegenwart. Wien.
Schmale, Wolfgang. 2010. Digitale Geschichtswissenschaft. Wien.
Schmale, Wolfgang. 2013a. Digital Humanities – Einleitung: Begriff, Definition, Probleme. Historische Mitteilungen 26, S. 86–93.
Schmale, Wolfgang. 2013b. Digitale Vernunft. Historische Mitteilungen 26, S. 94–100.
Stöber, Rudolf. 2016. Historische Methoden in der Kommunikationswissenschaft. Die Standards einer Triangulation. In: Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft, herausgegeben von S. Averbeck-Lietz und M. Meyen. Wiesbaden, S. 303–318.
Veith, Emil. 1937. „Die Krise von 1873 und die Wiener Presse". Dissertation, Universität Wien.
Walter, Edith. 1994. Österreichische Tageszeitungen der Jahrhundertwende: ideologischer Anspruch und ökonomische Erfordernisse. Wien.

Weber, Max. 1988. Soziologie des Zeitungswesens. In: Gesammelte Aufsätze zur Soziologie und Sozialpolitik, herausgegeben von M. Weber. Tübingen, S. 434–441.
Weiss, Johann. 1958. Die Wirtschaft als gestaltender Faktor in der österreichischen Pressegeschichte von der Einführung des Buchdrucks in Wien (1482) bis zur Revolution des Jahres 1848. Dissertation, Universität Wien, Wien.
Wettlaufer, Jörg. 2016. Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern. Zeitschrift für digitale Geisteswissenschaften.
Wilke, Jürgen. 1987. Quantitative Methoden in der Kommunikationsgeschichte. In: Wege zur Kommunikationsgeschichte, herausgegeben von M. Bobrowsky und W. R. Langenbucher. München, S. 49–57.
Winckler, Johann. 1875. Die periodische Presse Österreichs: eine historisch-statistische Studie.
Zaagsma, Gerben. 2013. On Digital History. BMGN – Low Countries Historical Review 128(4), S. 3.
Zenker, Ernst Viktor. 1893. Geschichte der Wiener Journalistik während des Jahres 1848. Wien, Leipzig.

Malorie Guilbaud Perez

Analyser un processus mémoriel au travers des archives de presse numérisées et physiques

l'exemple du Triangle Fire

Abstract: On March, 25th 1911, the three top floors of the Asch building, close to Washington Square in downtown Manhattan New-York City, were stormed by an unprecedent fire. This tragic event – 146 lost their lives – was told in detail on front and inside pages of numerous newspapers of international, national and even local coverage. This large interest gave birth to specific narratives, groundworks of a complex commemorative process. Eventually, the tragedy embraced the American National History. This article explores how difficult it is to build a consistent collection of historical articles related to the event, despites the enormous amount of articles theoretically available. Searching for the most efficient key words and the most effective database, browsing several digital newspapers collections, it seeks to understand what the characteristics of these first narratives were and how did they flow from New York to the domestic territory and abroad.

Keywords: digitised newspapers, labor history, query criticism

Le 25 mars 1911, au sein du Lower East Side de Manhattan, survient l'incendie de l'atelier de fabrication textile new-yorkais[1], la « Triangle Factory », installée aux trois derniers étages du Asch building, et qui, parmi les cinq cents employés de la compagnie, provoqua 146 victimes, quasi exclusivement de jeunes femmes migrantes, originaires d'Europe de l'Est ou du Sud. Cet accident industriel[2] engendra des réactions multiples dont notamment une forte mobilisation des employés new-yorkais de l'industrie textile massivement présents à la suite du drame lors de cérémonies collectives comme à l'occasion des procès des propriétaires de la Triangle, Max Blanck et Isaac Harris. Preuve de l'actualité de cet évènement, cette tragédie a été récemment rappelée par l'ancienne candidate à l'investiture démo-

1 Green, Nancy L., *Ready-to-wear and ready-to-work: a century of industry and immigrants in Paris and New York*, Duke University Press, Durham (1997).

2 Le Roux, Thomas (dir.), *Risques industriels : Savoirs, régulations, politiques d'assistance, fin XVII^e^-début XX^e^ siècle*, Presses universitaires de Rennes, 2016.

https://doi.org/10.1515/9783110729214-015

crate pour l'élection présidentielle de 2020, Elisabeth Warren[3], un exemple militant choisi par la sénatrice du Massachusetts pour dénoncer le « sweatshop system »[4] d'hier à aujourd'hui. L'objectif global de notre étude auquel participe cet article est ainsi d'étudier les processus à l'œuvre dans l'inscription de ce drame dans la mémoire nationale américaine, comment l'accident industriel, le fait divers, a dépassé ce statut initial pour devenir un événement plus complexe.

Au début du XX^{e} siècle, la presse écrite est un média de masse, la source d'information comme de formation de l'opinion la plus importante dans l'ensemble des territoires industrialisés[5]. La dynamique des publications est particulièrement importante aux Etats-Unis[6], accompagnant le tournant industriel et urbain de la nation américaine. Tout au long du siècle précédant l'événement, nombreux titres de presse ont pu acquérir de plus en plus d'autonomie, voire d'indépendance vis-à-vis des pouvoirs politiques[7], relayant les différents courants de réformes qui traversent alors la société américaine. Ils sont également devenus des objets de consommation de plus en plus diffusés, de plus en plus populaires[8], entre les mains d'entrepreneurs aux activités économiques diversifiées[9], mais conçus par des journalistes se professionnalisant. La presse américaine, anglophone ou non, même en se limitant ici à son contenu textuel[10], a donc été l'un des socles

3 Kaplan, Thomas, « Elizabeth Warren, at Washington Sq. Park Rally, Promises to Take On Corruption », *New York Times*, 16 septembre 2019; « She spoke near the site of the Triangle Factory Fire of 1911, which killed 146 garment workers, most of them women. The fire spurred a push to improve workplace safety, which Ms. Warren harnessed as a parallel for the far-reaching change she wants to pursue as president », érigeant la tragédie de la Triangle comme le pire du sweatshop system.

4 Barraud de Lagerie, P., « The wages of sweat: a social history perspective on the fight against sweatshops », *Sociologie du Travail*, volume 55, supplément 1, novembre 2013, pages e1–e23.

5 Kalifa, Dominique, Philippe Régnier, Marie-Ève Thérenty, Alain Vaillant (dir.), *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIX^{e} siècle*, Paris, Nouveau Monde éditions, 2011.

6 Kaplan, Richard L., « From Partisanship to Professionalism: The Transformation of the Daily Press », in: Carl F. Kaestle/Janice A. Radway (edd.), *A History of the hook in America, Volume Four. Print in Motion: The Expansion of Publishing and Reading in the United States, 1880–1940*, Chapel Hill UNC Press, 2009.

7 Risley, Ford, « Politics and Partisanship », in: David Sloan/Lisa Mullikin Parcell (edd.), *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002.

8 Exemple de la « penny press » comme le New York Sun de Benjamin Day, voir Whitby, Gary et David Sloan, « The purposes of the press: a reinterpretation of American journalism history », in: *Annual meeting of the association for education in journalism*, 1981.

9 Exemple de Pulitzer voir Blevens, Fred, « Publishers », in: David Sloan/Lisa Mullikin Parcell (edd.), *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002.

10 L'analyse de nombreux dessins de presse illustrant ces articles fait l'objet d'une étude séparée non encore publiée.

du processus de fabrication de l'incendie comme un événement. La circulation de celui-ci a pu ainsi être vecteur d'informations, autant que de mémoires[11] comme de protestations[12]. Ce média est ainsi tout autant témoin de l'événement, que reflet, imparfait, de celui-ci, porteur d'une vision, d'une compréhension, d'une sémantique spécifique, lesquels évoluent selon les époques, les lieux, les communautés ou les titres de presse considérés[13]. Une presse donc qui ne fait pas que témoigner, mais dont les choix d'édition sont révélateurs d'enjeux politiques et sociétaux qui traversent une nation tout entière et l'interrogent sur les valeurs qui la fondent.

L'enjeu de la présente étude est de parvenir à retracer, en partie grâce à la presse, la circulation diachronique et synchronique de l'évènement, à identifier des continuités et des ruptures. Nous commençons donc par construire un corpus de documents pertinents, accessibles via une consultation in situ et/ou un accès en ligne, ce dernier étant de plus en plus facilité au travers de la multiplication d'initiatives de numérisation des publications des plus diverses[14]. Comment choisir et interroger les bases de données disponibles pour constituer un corpus adéquat aux objectifs de la recherche qui puisse répondre à nos hypothèses ? Ainsi, après avoir présenté les enjeux de la constitution du corpus initial de la recherche centré sur un titre de presse, nous aborderons l'importance autant que les difficultés rencontrées à composer des corpus miroirs pertinents permettant d'approfondir l'analyse initiale, avant d'évoquer les exploitations possibles de ces sources numérisées.

1 Constitution du corpus initial

Le choix initial a été de se focaliser sur un titre de presse new-yorkais, accessible en ligne et toujours publié depuis 1911, le *New York Times*. A la date de l'incendie, le *Times* est l'un des huit quotidiens new-yorkais en circulation. Il

11 Lenoble, Benoît, « Célébrations », in : *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIX^e^ siècle*, Paris, Nouveau Monde éditions, 2011.

12 Wrona, Adeline, « La presse en son miroir, dénonciations », in : *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIX^e^ siècle*, Paris, Nouveau Monde éditions, 2011.

13 Bouchet, Thomas, « Évènement, opinion et politique », in : *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIX^e^ siècle*, Paris, Nouveau Monde éditions, 2011.

14 Acland, Charles R. et Eric Hoyt, *The Arclight guidebook to media history and the digital humanities*, Sussex, Reframe Books, 2016.

n'est pas encore le plus important, mais a déjà imposé sa différence[15] avec les deux journaux les plus vendus, *The World* acheté par Joseph Pulitzer en 1893, et *The Journal* détenu W. R. Hearst. Sa ligne éditoriale renouvelée par son acquéreur en 1896, Adolph Osch, s'incarne au travers du refus du sensationnel, « yellow journalism », à l'inverse d'autres tirages[16]. Valorisant une presse de qualité, le *New York Times* vise donc un lectorat plutôt aisé et cultivé, mais ne perd pas pour autant la volonté de s'imposer dans le paysage médiatique newyorkais, ce qui conduit son propriétaire à faire le choix de vendre le numéro à un penny seulement. En 1911, la dynamique du *Times* est positive, toujours plus d'exemplaires sont vendus (de 9.000 en 1896 à 352.000 en 1921[17]) et son lectorat s'agrandit, tandis que ses concurrents directs stagnent ou déclinent. Ces choix permettent au *New York Times* d'acquérir une reconnaissance, « the respectable New York Times », à l'échelle nationale et internationale : « It represents the best rather than the average of American journalism »[18]. Malgré les difficultés rencontrées par la presse écrite traditionnelle, il reste encore aujourd'hui l'un des journaux les plus importants.

Le *New York Times* a développé assez tôt des outils numériques avec plus ou moins de succès : éphémère *New York Pulse* entre 1983 et 1986, *Continuous News Desk* en 1991, site internet dès 1996. Depuis, le journal n'a cessé de diversifier son offre en ligne notamment pour résister à la multiplication des numériques gratuits. En 2014, le *New York times* lance pour ses abonnés la « TimesMachine »[19], un lecteur de microfilm virtuel, qui rend progressivement accessible les articles du journal depuis sa création. Une première vague a permis de numériser les articles publiés entre 1851 et 1980 (11 millions d'articles, 2,5 millions de pages), une seconde a mis à disposition ceux publiés entre 1981 et 2002 (8.035 éditions,

15 Royot, Daniel et Susan Ruel, *Médias, société et culture aux Etats-Unis,* Orphys, Ploton, 1996, p. 16–17.

16 Folkerts, Jean et Dwight Teeter, « Chapitre 9: Mass Markets and Mass culture », in: *Voices of a Nation : A History of Mass Media in the US,* MacMillan, 1994.

17 Davis, Elmer, «History of the New York Times 1851–1921 », New York Times, 26 septembre 1921, p. 12.

18 Schudson, Michael, *Discovering the news, a social history of American Newspapers,* New York, Basic Books, 1978.

19 Cotler, Jane et Evan Sandhaus, « How to build a TimesMachine », *New York Times,* 01 février 2016.

1,4 million d'articles, 1,6 million de pages), les articles postérieurs au 31 décembre 2002 étant accessibles via le moteur de recherche du site internet[20].

Nous avons choisi d'utiliser TimesMachine, car c'est un outil relativement performant qui d'une part n'isole pas les articles ciblés, mais les replace dans la publication originelle et qui d'autre part permet d'exporter les articles à différents formats pour les intégrer à un gestionnaire de références[21]. Cet outil a donc permis la constitution d'un premier corpus d'étude, restreint à un titre de presse new-yorkais, mais qui permet l'analyse de la circulation de l'incendie de la fabrique sur le siècle écoulé après son déroulement. La recherche d'articles traitant de l'événement a été réalisée sur l'ensemble des pages de la publication sur une période allant du 25 mars 1911 au 31 décembre 2002[22]. Nous recherchons alors tous les types d'articles liés directement ou indirectement à l'évènement, que celui-ci soit le sujet de l'article ou bien une référence évoquée au sein du texte. Sur ce point la numérisation des journaux et le lecteur virtuel permettent de ne plus se restreindre à une recherche par index thématique, qui référence les articles en fonction des thèmes principaux qui y sont abordés, mais permet une recherche par mots-clés offrant une plus grande exhaustivité. Il a donc fallu d'abord déterminer quelle combinaison de mots clés utiliser afin de borner correctement la recherche sans qu'elle soit trop discriminante[23]. Pour y parvenir, nous avons déterminé quels étaient les mots les plus évidents qui permettaient de définir l'événement et nous les avons testés séparément sur le moteur de recherche. Les résultats bruts présentés par ordre décroissant dans le tableau ci-dessous nous ont permis d'écarter les moins discriminants (le mois de mars) comme ceux qui l'étaient trop.

Cette première recherche sur la base de données nous conduit à ne pas retenir les termes les plus discriminants comme les noms propres (« Asch » building ou « Blanck » l'un des propriétaires de la compagnie, dont les noms sont parfois mal orthographiés et transformés en « ash » ou « blank ») ou des noms trop spécifiques comme « shirtwaist » souvent abrégé en « waist » dans la plupart des articles. Comme le montre le Tab. 2 ci-dessous, nous avons ensuite

20 Ainsi même avec une unité de publication l'accès aux différents articles du New-York Times n'est pas entièrement libre d'accès, restreint pour une partie de la période chronologique considérée aux seuls abonnés.

21 Pour cette étude nous utilisons Zotero qui est un logiciel de gestion de références libre, gratuit et open source.

22 Borne chronologique imposée par le lecteur virtuel, les articles plus récents sont accessibles directement via le moteur de recherche du site internet du journal.

23 Waller, Gregory A., « Search and re-search: digital print archives and the history of multisited cinema », in: Charles R. Acland/Eric Hoyt, *The Arclight guidebook to media history and the digital humanities*, Sussex, Reframe Books, 2016.

Tab. 1: Présentation des résultats synthétiques d'une recherche par mots-clés simples sur TimesMachine (1911–2002), résultats bruts (décembre 2019), en nombre d'articles référencés.

Mot-clé possible pour définir l'événement	Nombre d'articles contenant le mot-clé	Mot-clé possible pour définir l'événement	Nombre d'articles contenant le mot-clé
MARCH	67.970.054	WAIST	3.362.180
FIRE	41.833.002	TRIANGLE	2.996.169
FACTORY	16.585.043	COMPANY	2.166.923
HARRIS	12 960.934	SHIRTWAIST	175.870
ASH	6.362.450	ASCH	106.538
1911	6.208 163	BLANCK	29.129
BLANK	4.793.726		

testé les associations de mots afin de restreindre le nombre d'articles potentiels correspondant réellement à notre enquête. Les combinaisons les plus adéquates semblent être « fire + factory + triangle » ou « fire + factory + waist », des combinaisons de trois mots d'usage courant, pouvant définir au mieux l'évènement, mais aussi être utilisés dans des contextes très différents. Se pose alors la question de l'adjonction d'un quatrième mot-clé qui permettrait de restreindre davantage les résultats obtenus.

Tab. 2: Présentation des résultats synthétiques d'une recherche par mots-clés sur TimesMachine (1911–2002) et sur le site du New York Times en ligne (2003–2019), résultats bruts (décembre 2019), en nombre d'articles référencés.

Période	Fire + Factory + Waist	Fire + Factory + Triangle	Pertinence de l'ajout qu'un quatrième mot-clé ?						
			+ 1911	+ company	+ shirtwaist	+ waist	+ harris	+ blanck	+ asch
1911–2002	**8.140**	**1.279**	**287**	**664**	**230**	**321**	**182**	**497**	**117**
2003–2019	**7.101**	**814**	**230**	**410**	**219**	**254**	**131**	**403**	**66**

Il apparaît d'après les résultats obtenus que la combinaison la plus pertinente soit « fire + factory + triangle », à laquelle l'adjonction d'un quatrième mot serait trop restrictive. Nous avons donc testé cette hypothèse et consulté l'ensemble des articles proposé par TimesMachine pour vérifier leur contenu et évaluer la

pertinence de la sélection effectuée par le moteur de recherche. Comme le montre le Tab. 3 ci-dessous, le nombre d'articles correspondants réellement aux objectifs de la recherche pour la période 1911–2002 s'établit finalement à 376[24]. Ce qui signifie que sur les 1.279 articles sélectionnés par TimesMachine, seul un tiers traite réellement de l'incendie de la Triangle Factory. L'ajout d'un quatrième mot clé n'améliore pas réellement la pertinence de la recherche. Le ciblage de celle-ci n'est réellement meilleur que dans un seul cas, celui de l'ajout du mot « shirtwaist ». La pertinence s'établit à 60%, cependant cette recherche ne rend accessible que 136 articles alors que nous en avons dénombré 376 concrètement en lien avec notre thème d'étude. Dans cinq cas sur sept, l'ajout d'un quatrième mot-clé restreint trop les résultats obtenus.

Tab. 3: Présentation évaluant la fiabilité de la recherche par mot clé sur TimesMachine par nombre d'articles référencés (décembre 2019).

Période entre 1911 et 2002	**Fire + Factory + Triangle**	**Avec l'ajout d'un quatrième mot-clé**						
		+ 1911	**+ company**	**+ shirtwaist**	**+ waist**	**+ harris**	**+ blanck**	**+ asch**
Résultats bruts	1.279	287	664	230	321	182	497	117
Résultats filtrés	376	95	152	136	94	37	41	35
% réellement utilisable	**30%**	**33%**	**23%**	**60%**	**30%**	**20%**	**8%**	**30%**

Tous les articles obtenus par cette recherche ne sont donc pas pertinents, car bien que contenant les mots-clés, ils ne traitent pas de l'événement recherché. La pertinence de la recherche s'établit donc autour des 30%. Celle-ci est moindre quand des noms propres sont utilisés comme mots-clés (les noms des propriétaires Harris et Blanck). Un écart entre les résultats et les articles effectivement utiles, car les trois mots-clés retenus sont des mots utilisés couramment que l'on retrouve dans des contextes très différents notamment des articles liés aux deux conflits mondiaux, au conflit vietnamien, ou encore des faits divers et des informations culturelles. Pour ces derniers aspects, le manque de fiabilité est en partie dû au processus de segmentation réalisé au sein des journaux numérisés. La recherche

24 Pour la période 2003–2019 il est de 89.

par mots-clés s'effectue aussi au sein d'espaces non segmentés en articles,par exemple des pages présentant essentiellement des listes : listes nominatives[25] ou encore listes d'activités culturelles possibles, de livres disponibles, de spectacles du moment (critiques littéraires, programmes télévisuels, actualités du Princeton's Triangle Club, attractions, sorties « spare times ou leisure times »). Ces résultats non pertinents viennent gonfler le nombre d'articles potentiels renvoyés par le moteur de recherche.

D'autre part, on observe également que la fiabilité de la recherche par mots-clés est plus forte, quasi excellente, sur les 36 mois qui suivent la tragédie. Par contre, elle s'étiole plus on s'éloigne de la date de l'évènement. Un facteur permet en partie d'expliquer ce constat, c'est la forme des articles du Times. Ceux-ci sont, en effet, de plus en plus fournis au fil des années pouvant s'étendre sur une, deux, trois voire quatre pages en fonction des thématiques, des « dossiers » traités, donc la possibilité de trouver la combinaison des trois mots-clés est plus forte, l'interface ne limitant pas les résultats à une seule page mais proposant l'intégralité de l'article même si celui-ci est filé sur plusieurs pages. Néanmoins, l'usage du lecteur virtuel proposé par le New York Times s'est révélé à la fois très accessible quant à son utilisation et de qualité quant aux documents obtenus. Les articles ne sont pas isolés, mais proposés à chaque fois en lien avec l'exemplaire complet du journal au sein duquel ils ont été publiés. La qualité de numérisation des pages est bonne et très homogène, la segmentation des articles assez performante, ce qui permet de constituer[26] un corpus de qualité pouvant être traité numériquement après océrisation.

2 Constitution des corpus miroirs

À l'issue de l'exploitation des ressources du *New York Times* nous avons donc un corpus de 435 articles (corpus 1) directement lié à l'événement ou faisant référence explicitement à celui-ci, publiés à New York entre 1911 et 2019. Or, pour pouvoir appréhender la circulation de l'événement dans l'espace et dans le temps, l'analyse de ce seul titre de presse n'est pas suffisante même si sa diffusion ne se limite pas à New York. Nous avons besoin d'autres collections numérisées accessibles, des publications locales, nationales et internationales, proposant une

25 Par exemple les listes des nécessiteux « neediest » new-yorkais qui appellent à la générosité des lecteurs du *Times* en présentant précisément le cas de chaque famille.

26 Chaque article a pu être téléchargé aux différents formats proposés (.txt ou .pdf) et intégré à un gestionnaire de références, ici Zotero.

durée de parution suffisamment étendue qui puisse soutenir la comparaison avec le *New York Times*, afin d'identifier et d'analyser les mécanismes synchroniques et diachroniques de présentation de l'événement, de circulation au sein de la société américaine et au-delà. C'est ce que nous avons appelé des corpus miroirs, des corpus qui nous permettent d'interroger les récits construits autour de l'événement dans un support singulier.

Les titres de presse sont particulièrement nombreux sur le territoire américain. La publication de journaux aux États-Unis a atteint son maximum en 1909 avec 2.600 titres en circulation, pour un tirage moyen de 13.531 exemplaires (un tirage moyen plus élevé pour les éditions du dimanche). En parallèle, le lectorat abonné augmente passant en 1870 de 11,5% de la population éduquée à 45% en 1930. Pour l'année 1991, on ne compte plus que 1.610 quotidiens en circulation, représentant un tirage cumulé de 60 millions d'exemplaires par jour[27].

Pour avoir un accès, libre cette fois, à des collections numérisées, nous nous sommes tournés vers le site internet de la bibliothèque du congrès. En effet, le travail de numérisation des collections de journaux et périodiques américains fait l'objet d'un programme spécifique : le National Digital Newspaper Program (NDNP). Y sont associés le National Endowment for the Humanities (NEH) et la Bibliothèque du Congrès (Library of Congress), le premier assurant le financement des projets de numérisation soumis par différentes organisations (institutions, bibliothèques, associations, universités ...) et des États fédérés, le second fixant les cadres chronologiques des archives à numériser. Les données ainsi numérisées sont centralisées ensuite au sein d'un portail d'accès numérique, Chronicling America[28], qui permet d'effectuer une recherche par mots-clés au sein des pages des différents supports transmis à la bibliothèque du congrès[29].

Le portail Chronicling America rend donc a priori facilement accessible une collection numérisée de plus de 3.000 journaux[30], sur une période allant de 1789 à 1963. Même si les collections ne couvrent pas l'ensemble de la chronologie souhaitée, elles peuvent permettre en théorie l'analyse de la circulation

27 Folkerts, Jean et Dwight Teeter, « Chapitre 9: Mass Markets and Mass culture » in *Voices of a Nation : A History of Mass Media in the US*, MacMillan, 1994.

28 https://chroniclingamerica.loc.gov.

29 Une dynamique de numérisation observable des deux côtés de l'Atlantique, voir par exemple Rygiel, Philippe, « Nouvelles frontières de l'historien », *Historien à l'âge numérique*, Villeurbanne, Presses de l'Enssib (nouvelle édition en ligne), 2017.

30 Un des aspects donc de cet âge de l'abondance précisé par Rosenzweig, Roy, « Scarcity or abundance ? Preserving the past in a digital era », *American Historical Review*, 108–3, juin 2003, p. 735–762.

temporelle et spatiale de l'événement sur une période de cinquante ans. Cependant, comme le montre le Tab. 4, il nous apparaît rapidement que la collection proposée en ligne présente un déséquilibre géographique. En effet, le programme lancé en 2005 a été alimenté de manière disparate par différentes organisations (sociétés historiques, universités ...) ayant répondu à l'appel à projet: certaines ont soumis régulièrement des demandes de financement et ont rendu accessible une part conséquente de leurs journaux historiques (c'est le cas du Mississippi ou de l'Ohio), alors que d'autres états ne sont pas encore représentés dans la liste en 2019 (le Wyoming, le Rhode Island).

Tab. 4: Vue synthétique de la répartition par état des sources numérisées présentes dans Chronicling America par nombre de sources accessibles (Library of Congress, décembre 2019).

Les états les plus présents dans les collections de la LC			**Les états les moins présents dans les collections numérisées de la LC**		
	Nombre de publications accessibles	**Nombre de financements reçus depuis 2005**		**Nombre de publications accessibles**	**Nombre de financements reçus depuis 2005**
Mississippi	254	3	Wyoming	0	1 (2019)
Ohio	160	5	Virgin Islands	0	1 (2019)
Tennessee	126	3	Rhode Island	0	1 (2019)
South Carolina	125	3	Massachusetts	2	0
Louisiana	122	3	Puerto Rico	4	4
North Carolina	108	4	New Jersey	10	2
Delaware	106	3	Maine	13	2

De plus, un déséquilibre temporel apparaît également. L'étude de la composition de la collection fait apparaître que les journaux numérisés sont en nombre important pour une période précédant notre événement. Le cœur de la collection est ainsi composé de journaux du XIX^{e} siècle, qui pour un grand nombre ne sont plus édités en 1911, comme le montre le Tab. 5 ci-dessous.

Très peu de titres couvrent en théorie une période adéquate pour notre recherche, d'autant plus que les collections par titre ne sont pas complètes. Les numérisations par titre ne sont pas exhaustives et couvrent une période bien plus courte. Par exemple, le quotidien *Dermott News* publié en Arkansas entre 1910 à 1977 n'est accessible via la bibliothèque du congrès que pour les numéros couvrant la période de 1913 à 1919. Ces carences dans les collections numérisées

Tab. 5: Nombre de publications numérisées disponibles sur Chronicling America correspondant à notre période d'étude (décembre 2019).

Nombre total de publications sur l'ensemble du territoire	Dont publications encore éditées entre 1911 et 1927	Dont publications encore éditées entre 1928 et 1945	Dont publications encore éditées entre 1946 et 1962	Dont publications encore éditées après 1963	Dont publications éditées en continu entre 1911 et au moins 1963	Publications éditées entre 1911 et au moins 1963 entièrement disponibles
3.157	1.144	560	399	288	210	0

sont inhérentes aux modalités de constitution de la base de données qui a été réalisée par strate chronologique : les projets soumis pour le premier cycle devaient ainsi proposer des numérisations de publications éditées entre 1900 et 1910, le cycle suivant s'est focalisé sur la période 1880–1910, puis 1880–1922, puis 1860–1922, puis 1836–1922, pour enfin recouvrir la période 1960–1963 pour le dernier cycle proposé en 2019. Ce qui conduit à un déséquilibre important des publications proposées.

La base de données de la bibliothèque du congrès ne pouvait donc que partiellement répondre à la constitution d'un corpus miroir sur le temps long, les possibilités étant majoritairement restreintes à la période 1911–1922, soit la diffusion de l'événement uniquement sur la décennie lui succédant et non sur le siècle ou le demi-siècle écoulé. Mais elle présente l'avantage d'avoir une emprise territoriale assez importante bien qu'inégale. Comme le montre le Tab. 6, l'exploitation de la base de données de Chronicling America nous donne ainsi accès en théorie à 14.537 pages de journaux comportant les mots-clés ciblés présentés dans le Tab. 1. Si l'on réduit ce nombre à la présence de ces mots clés uniquement en première page des publications, on tombe à 1.262 pages théoriquement pertinentes. Une fois étudiées individuellement nous obtenons un corpus effectif de 307 unes de presse (corpus 2) mettant réellement en avant le Triangle Fire, publiées entre mars 1911 et novembre 1915, par une diversité de 124 titres de presse différents.

Ce premier corpus complémentaire nous permet d'envisager la circulation de l'événement entre 1911 et 1915 seulement, mais au sein d'une majorité d'États du territoire américain. Il n'est cependant pas suffisant pour répondre aux objectifs de notre recherche sur le temps long d'un siècle. Nous avons donc essayé de compléter le corpus obtenu par les ressources numérisées mises à disposition gratuitement par la Bibliothèque du Congrès en utilisant le moteur de recherche de

Tab. 6: Nombre de pages (toutes, seulement les « unes », et celles évaluées comme pertinentes référencées par le moteur de recherche du portail Chronicling America pour notre recherche par mot-clé « fire + factory + triangle », publiées entre 1911 et 1963 (date limite imposée) pour chaque état ayant participé avant l'année 2019 au projet Chronicling America (Library of Congress – décembre 2019).

Territoires proposant des publications entre 1911 et 1963	Potentiel ciblé par mots-clés		Unes « réelles »	nb titres de presse	Territoires proposant des publications entre 1911 et 1963	Potentiel ciblé par mots-clés		Unes « réelles »	nb titres de presse
	Toutes les pages	Unes seulement				Toutes les pages	Unes seulement		
CONNECTICUT	**1.006**	100	**30**	2	CALIFORNIA	**43**	5	**4**	2
NEW YORK	**813**	27	**23**	3	LOUISIANA	**136**	21	**4**	4
WASHINGTON	**190**	32	**17**	6	NORTH CAROLINA	**151**	29	**4**	4
D. OF COLUMBIA	**5.501**	124	**14**	3	OHIO	**124**	6	**4**	2
IOWA	**281**	42	**14**	3	TEXAS	**207**	12	**4**	1
OREGON	**241**	27	**13**	4	ALASKA	**81**	19	**3**	3
ARIZONA	**257**	32	**12**	4	ARKANSAS	**38**	6	**3**	2
VERMONT	**255**	45	**10**	4	DELAWARE	**135**	10	**3**	1
MICHIGAN	**154**	17	**9**	2	HAWAI	**101**	4	**3**	2
NEBRASKA	**235**	21	**9**	4	KENTUCKY	**75**	10	**3**	3
SOUTH CAROLINA	**91**	13	**9**	5	MINNESOTA	**153**	14	**3**	1
WEST VIRGINIA	**177**	13	**9**	5	IDAHO	**200**	21	**2**	2

NEW JERSEY	**309**	32	**8**	2	KANSAS	**248**	11	**2**	2
NORTH DAKOTA	**343**	28	**8**	4	PENNSYLVANIA	**306**	11	**2**	2
OKLAHOMA	**128**	21	**8**	4	COLORADO	**60**	3	**1**	1
SOUTH DAKOTA	**136**	15	**8**	4	GEORGIA	**14**	8	**1**	1
FLORIDA	**157**	25	**7**	3	MARYLAND	**137**	16	**1**	1
ILLINOIS	**155**	21	**7**	3	NEVADA	**66**	11	**1**	1
NEW MEXICO	**175**	17	**7**	3	TENNESSEE	**99**	9	**1**	1
UTAH	**178**	11	**7**	3	ALABAMA	**1**	0	**0**	0
INDIANA	**484**	43	**6**	2	MAINE	**45**	0	**0**	0
MISSOURI	**114**	13	**6**	3	MASSACHUSETTS	**0**	0	**0**	0
MONTANA	**132**	11	**6**	4	PIEDMONT	**0**	0	**0**	0
VIRGINIA	**221**	33	**6**	4	PUERTO RICO	**0**	0	**0**	0
MISSISSIPPI	**121**	8	**5**	4	WISCONSIN	**263**	22	**0**	0

la plateforme numérique ProQuest[31], largement présente au sein des universités américaines et accessible depuis leurs sites, comme depuis la Bibliothèque publique de New-York, afin d'évaluer si la circulation temporelle de l'événement était réalisable sur d'autres séries de publications que celles du *New York Times*.

Les collections réunies par cet agrégateur de base de données s'élèvent, d'après son interface, à plus de six milliards de pages numérisées, dont vingt millions de pages de journaux (donc proche des 16 millions proposés par la bibliothèque du Congrès), comprenant à la fois des pages numérisées pour les journaux historiques mais aussi des articles de publications récentes (de plus de trois mois à la date de la consultation). La recherche effectuée via ProQuest nous permet d'interroger simultanément 122 bases de données sur une plage de recherche démarrant au 25 mars 1911. En ne demandant que les résultats liés aux journaux et périodiques dont sont exclues les annonces et les publicités, le nombre d'articles à priori correspondant à l'événement est conséquent comme le montre le Tab. 7.

Tab. 7: Nombre de résultats (articles ou pages) renvoyés par le moteur ProQuest pour des requêtes combinant différents mots-clés, et avec un filtre temporel restreignant aux publications après le 25 mars 1911.

Mots-clés utilisés	Résultats bruts sans restriction	Résultats avec sources filtrées : quotidiens, journaux, dépêches, magazines, revues, périodiques historiques	Résultats avec type de document filtré : exclusion des publicités et des petites annonces
FACTORY + FIRE	**2.774.761**	**2.618.715**	**2.040.713**
TRIANGLE + FIRE	**555.692**	**458.060**	**391.284**
TRIANGLE + FACTORY	**331.437**	**283.297**	**192.735**
TRIANGLE + FACTORY + FIRE	**136.974**	**105.603**	**70.842**
TRIANGLE + FACTORY + FIRE + 1911	**24.188**	**11.649**	**9.192**
TRIANGLE + FACTORY + FIRE + WAIST	**11.890**	**5.237**	**3.747**

31 Utilisation de ce moteur de recherche payant accessible sur abonnement par le site de la bibliothèque publique de New York, puis par celui de l'Université de Cornell lors de la consultation d'archives physiques au sein du Kheel Center for Labor-Management documentation and archives, Catherwwod Library, Cornell University, New-York State.

Ce programme nous autorise donc une recherche simultanée sur une centaine de publications éditées entre 1911 et aujourd'hui, soit plus de 70.000 articles ou pages correspondant aux mots-clés déjà utilisés pour la requête du *New York Times*. Environ un tiers seulement de ces publications proposent des durées d'édition après 1911 et des disponibilités supérieures à celles de Chronicling America, le reste de la base de données étant constitué soit par des journaux historiques antérieurs à 1911 ou par des publications récentes accessibles uniquement pour les numéros des vingt ou trente dernières années comme le montre le Tab. 8 ci-dessous.

Si l'on affine encore ces résultats aux seuls titres dont la période de publication inclut l'année 1911 et se prolonge sur le siècle, nous obtenons 18 titres de presse répertoriés au sein du Tab. 9. Cela forme un corpus de 926 articles, pouvant soutenir la comparaison avec le *New York Times*. Ils peuvent nous permettre d'étudier d'une part la circulation de l'événement sur le siècle et d'autre part au sein d'une partie du territoire américain (11 États représentés). Nous pouvons aussi choisir de nous focaliser uniquement sur l'état ou la ville de New York afin de comparer les récits produits par les différents titres de presse. Nous pouvons aussi focaliser notre intérêt sur des publications communautaires, car trois notamment sont issues de la communauté juive.

Au final, la constitution de ces corpus miroirs est donc très contrainte par la difficulté d'accès autant que par l'indisponibilité de certaines sources numérisées. Le processus de numérisation de la presse écrite américaine est encore très largement inachevé. Les bases de données mettent en avant des quantités importantes de documents, mais cette apparente abondance est trompeuse à plus d'un titre. Les priorités données par les programmes qui financent le processus relèvent de logiques s'accordant parfois mal à celles des chercheurs. Les collections disponibles ne nous permettent donc d'étudier la circulation de l'événement au sein du territoire américain sur le siècle écoulé que de manière partielle. D'autre part, les publications accessibles sont majoritairement des publications anglophones d'information générale qu'il faut mettre en regard aussi de publications plus spécialisées comme les publications syndicales ou communautaires. Grâce à la consultation des archives du syndicat américain ILGWU (International Ladies Garment Workers Union), acteur important de la mémoire de l'événement, nous avons pu collecter au sein du Kheel Center de l'Université de Cornell 142 articles de presse spécialisés, notamment les articles du magazine *Justice*, édité par le syndicat ouvrier. Les corpus constitués sont restreints pour l'instant aux articles en langue anglaise. Ils sont donc à compléter par d'autres sources : des journaux communautaires dont les publications sont encore nombreuses sur le territoire américain au début du XX^{e} siècle ou encore des articles issus de la presse internationale non-anglophone dans le but d'élargir l'étude de la circulation de l'événement.

Tab. 8: Titres de journaux accessibles sur ProQuest avec une période d'édition étendue (jusque décembre 2019), par lieu de publication.

	VILLE	ETAT	TITRE de la publication	période disponible	
				début	fin
1	LOS ANGELES	CA	LOS ANGELES TIMES	1881	RECENT
2	HARTFORD	CO	HARTFORD COURANT	1764	RECENT
3	WASHINGTON	DC	WASHINGTON POST	1877	CURRENT
4	ATLANTA	GE	ATLANTA CONSTITUTION	1946	1984
5	CHICAGO	IL	CHICAGO TRIBUNE	1872	RECENT
6	PLAINFIELD	IL	PLAINFIELD COURIER NEWS	1894	1962
7	BALTIMORE	MD	SUN BALTIMORE	1837	1994
8	BOSTON	MA	BOSTON GLOBE	1872	2019
16	NYACK	NY	JOURNAL-NEWS	1889	1990
17	NEW YORK	NY	NEW YORK TIMES	1857	RECENT
18	NEW YORK	NY	NEW YORK TRIBUNE	1842	1962
19	LONG ISLAND	NY	NEWSDAY	1940	1991
20	ALLENTOWN	PE	MORNING CALL	1939	RECENT
21	PHILADELPHIA	PE	JEWISH EXPONENT	1897	present
22	PHILADELPHIA	PE	PHILADELPHIA INQUIRER	1860	2001
23	PHILADELPHIA	PE	PHILADELPHIA TRIBUNE	1912	2001

9	BOSTON	MA	CHRISTIAN SCIENCE MONITOR	1908	CURRENT	24	PITTSBURGH	PE	PITTSBURGH POST GAZETTE	1927	RECENT
10	BOSTON	MA	JEWISH ADVOCATE BOSTON	1909	2018	25	PITTSBURGH	PE	PITTSBURGH PRESS	1887	1992
11	ASBURY PARK	NJ	ASBURY PARK PRESS	1905	2011	26	AUSTIN	TE	AUSTIN STATESMAN	1923	1971
12	BINGHAMTON	NY	BINGHAMTON	1905	1960	27	BURLINGTON	VE	BURLINGTON FREE PRESS	1885	2007
13	NEW YORK	NY	NEW YORK AMSTERDAM NEWS	1922	recent	28		CANADA	GLOBE AND MAIL	1936	CURRENT
14	NEW YORK	NY	DAILY NEWS	1920	2009	29		CANADA	TORONTO DAILY STAR	1900	RECENT
15	ROCHESTER	NY	DEMOCRAT AND CHRONICLE	1884	RECENT	30		IRELAND	IRISH TIMES	1921	CURRENT
16	NEW YORK	NY	FORWARD	1897	RECENT	31		ISRAEL	JERUSALEM POST	1950	1988
17	ITHACA	NY	ITHACA JOURNAL	1914	2012	32		UK	GUARDIAN	1959	2003

Tab. 9: Présentation, par titre de publication, du nombre d'articles ou de pages numérisés accessibles sur ProQuest effectivement pertinents (décembre 2019).

VILLE	ETAT	TITRE de la publication	période disponible		POTENTIEL	nombre d'articles pertinents publiés à partir du 25 mars 1911
			début	fin		
LOS ANGELES	CA	LOS ANGELES TIMES	1881	RECENT	664	**70**
HARTFORD	CO	HARTFORD COURANT	1887	RECENT	531	**77**
WASHINGTON	DC	WASHINGTON POST	1877	CURRENT	1432	**65**
CHICAGO	IL	CHICAGO TRIBUNE	1872	RECENT	564	**61**
BALTIMORE	MD	SUN BALTIMORE	1837	RECENT	317	**44**
BOSTON	MA	DAILY GLOBE	1908	CURRENT	283	**40**
BOSTON	MA	CHRISTIAN SCIENCE MONITOR	1908	CURRENT	107	**31**
BOSTON	MA	JEWISH ADVOCATE BOSTON	1909	2018	65	**14**
ASBURY PARK	NJ	ASBURY PARK PRESS	1905	2011	2500	**51**
ROCHESTER	NY	DEMOCRAT AND CHRONICLE	1884	RECENT	4984	**76**
NEW YORK	NY	FORWARD	1994	RECENT	71	**71**
NEW YORK	NY	NEW YORK TRIBUNE	1842	1962	323	**160**
PHILADELPHIA	PE	JEWISH EXPONENT	1897	PRESENT	38	**24**
PHILADELPHIA	PE	PHILADELPHIA INQUIRER	1860	2001	10 326	**54**

Tab. 9 (suite)

VILLE	ETAT	TITRE de la publication	période disponible		POTENTIEL	nombre d'articles pertinents publiés à partir du 25 mars 1911
			début	fin		
PITTSBURGH	PE	PITTSBURGH PRESS	1887	1992	6 370	**47**
BURLINGTON	VE	BURLINGTON FREE PRESS	1885	2007	1051	**26**
	CANADA	TORONTO DAILY STAR	1900	RECENT	668	**15**

3 Les exploitations possibles

La constitution de corpus exploitables pour répondre aux objectifs de l'étude a donc été particulièrement complexe et reste encore inachevée à ce stade. Son exploitation comporte, elle aussi, de nombreuses difficultés, et ce, à plusieurs niveaux. D'une part, la qualité des documents proposés est d'une grande hétérogénéité. Les moteurs de recherche donnent accès à des images plus ou moins précises des articles en fonction du document source. Pour certaines collections, les documents proposés ne sont pas issus de prise de vue directe des pages de journaux, mais utilisent les images réalisées pour la conservation sur microfilm sans que celles-ci aient été améliorées. Or la faible qualité de l'image d'un article complexifie son exploitation. Les moteurs de recherche rendent accessibles les articles sous plusieurs formats, image ou texte, or pour beaucoup le texte issu du processus d'océrisation utilisé est d'assez mauvaise qualité[32], au point de compromettre le processus d'exploitation sémantique par exemple[33]. Très souvent, ni la date de réalisation de l'archive, ni la mention du processus utilisé ne sont précisés. Nous avons donc préféré collecter des images. Ce sont pour la majorité des images de pages entières au sein desquelles il faut

32 Mutuvi, Stephen, Antoine Doucet, Moses Odeo et Adam Jatowt, « Evaluating the Impact of OCR Errors on Topic Modeling », *Lecture Notes in Computer Science*, vol 11279, Springer, Cham, 2018, p. 3–14.

33 Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty et Antoine Doucet, « An Analysis of the Performance of Named Entity Recognition over OCRed Documents », in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, p. 333–334.

découper l'article cible afin de constituer un corpus uniquement basé sur des images d'articles à océriser[34] selon une méthodologie uniforme pour tous[35].

La collecte et la structure des corpus ont été organisées au sein de Zotero. Ce logiciel de gestion de références a été utilisé ici pour stocker et organiser les articles collectés sous différents formats : image, texte ou PDF. Au sein de cette bibliothèque virtuelle, chaque article possède une fiche d'identité propre permettant de préciser des informations concernant sa source, sa forme, son contenu, etc. À partir donc des articles et de leurs fiches informatives, il nous semble possible grâce à des protocoles spécifiques[36] dont nous maîtrisons les paramétrages, tout en prenant toutes les précautions nécessaires en matière d'exploitation de ces données[37], d'éprouver nos différentes hypothèses de travail[38]. Tout d'abord en observant la circulation, ou non, de l'événement dans l'espace à différentes échelles par rapport à son épicentre new yorkais. Une circulation dans le temps qui permet d'ancrer l'événement dans la mémoire collective nationale, quitte à produire un récit biaisé de l'événement. C'est également éprouver sa circulation au sein des différents groupes composant la société américaine, groupes ethniques, classes sociales, communautés urbaines et rurales. Les corpus de travail obtenus nous permettent également de mettre en évidence des évolutions, des spécificités ou des différences sémantiques pour qualifier l'événement[39] au sein de chaînes de valeurs identifiables. Les récits créés autour de l'événement répondent à des objectifs qui leur sont propres. Ils colorent l'événement au travers du vocabulaire employé, pour le décrire, pour définir ses acteurs, les victimes comme les survivants, pour élaborer ses origines ou déterminer ses conséquences. Ils lui donnent une résonance particulière. Ils nous permettent d'appréhender les multiples dimensions de l'événement

34 Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani et Jean-Philippe Moreux, « Impact of OCR errors on the use of digital libraries: towards a better access to information », in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, p. 1–4.

35 Une méthode de traitement des données développée par le laboratoire L3i de l'Université de La Rochelle avec lequel nous travaillons.

36 Dans ce cadre nous collaborons avec le laboratoire L3I de l'Université de La Rochelle, spécifiquement Cyril Suire et Antoine Doucet, investis dans le projet NewsEye : A Digital Investigator for Digital Newspapers (www.newseye.eu).

37 Carbou, Guillaume, « Analyser les textes à l'ère des humanités numériques. Quelques questions pour l'analyse statistique des données textuelles », in : *Les Cahiers du numérique*, vol. 13, no. 3, 2017, pp. 91–114.

38 Barats, Christine, Jean-Marc Leblanc et Pierre Fiala, « Chapitre 5 – Approches textométriques du web : corpus et outils », *Manuel d'analyse du web en Sciences Humaines et Sociales,* sous la direction de Christine Barats et Armand Colin, 2013, pp. 99–124.

39 Châteauraynaud, Francis et Josquin Debaz, « Prodiges et vertiges de la lexicométrie », *Socio-informatique et argumentation*, 23 décembre 2010.

attribuées par les titres de presse qui lui donnent ses caractéristiques propres, par association ou exclusion. C'est enfin retracer sa circulation spatiale dans un premier temps au sein du territoire américain en produisant une cartographie dynamique de la diffusion de la nouvelle au sein du territoire américain et au niveau international. Puis dans un second temps d'évaluer la persistance de la référence de l'événement au sein de publications non new-yorkaises.

Enfin, l'accès à des titres de presse numérisés nous a donc permis de ne pas utiliser la technique de l'échantillonnage pour l'étude de la presse. Néanmoins, la constitution d'un corpus d'étude opérationnel s'est révélée toute aussi difficile, en totale antinomie avec l'apparente abondance de la ressource vantée par les plateformes de mise à disposition. Si la numérisation des journaux autorise, depuis son ordinateur personnel, un accès a priori facile, ou du moins facilité, à un nombre toujours croissant de ressources physiquement disséminées dans plusieurs sites, l'accès concret à l'exemplaire numérisé du périodique sélectionné se dérobe encore souvent et, quand bien même nous parvenons à l'atteindre, son double numérique peut se révéler difficilement exploitable et susciter à son tour de nombreuses interrogations. Constituer un corpus de presse doit de plus prendre en compte des éléments qui dépassent la simple accessibilité numérique, notamment les contextes sociaux, économiques, politiques, culturels d'élaboration et de circulation des titres de presse. Chaque publication est spécifique, et ses caractéristiques propres doivent être prises en compte dans l'élaboration du corpus de recherche. Dans le cadre de cette recherche nous avons pu ainsi constater que la numérisation partielle des collections nécessite d'être complétée par la consultation d'archives physiques. Une recherche comme celle que nous avons entreprise ne peut donc être réalisée uniquement avec les archives numérisées[40].

Bibliographie

Acland Charles R., Hoyt Eric, *The Arclight guidebook to media history and the digital humanities*, Sussex, Reframe Books, 2016.

Barats Christine, LeBlanc Jean-Marc, et Fiala Pierre, « Chapitre 5 – Approches textométriques du web : corpus et outils », *Manuel d'analyse du web en Sciences Humaines et Sociales*, sous la direction de BARATS Christine, Armand Colin, 2013, pp. 99–124.

Barraud de Lagerie P., "The wages of sweat: a social history perspective on the fight against sweatshops", *Sociologie du Travail*, volume 55, supplément 1, novembre 2013, pp. e1-e23.

40 Wijfjes, Huub, « Digital Humanities and Media History. A Challenge for Historical Newspaper Research », Tijdschrift Voor Mediageschiedenis, vol. 20, no 1, juin 2017, p. 4–24 (www.tmgonline.nl).

Blevens Fred, « Publishers », in Sloan David, Mullikin Parcell Lisa, *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002.

Bouchet Thomas, « Evènement, opinion et politique », *in La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011.

Carbou, Guillaume. « Analyser les textes à l'ère des humanités numériques. Quelques questions pour l'analyse statistique des données textuelles », *Les Cahiers du numérique*, vol. 13, no. 3, 2017, pp. 91–114.

Châteauraynaud Francis, DEBAZ Josquin, « Prodiges et vertiges de la lexicométrie », *Socio-informatique et argumentation*, 23 décembre 2010.

Chiron, G., Doucet, A., Coustaty, M., Visani, M., & Moreux, J. P. (2017, June). Impact of OCR errors on the use of digital libraries: towards a better access to information. In 2017 *ACM/IEEE Joint Conference on Digital Libraries (JCDL)* pp. 1–4. IEEE.

Cotler Jane, Sandhaus Evan, « How to build a TimesMachine", *New York Times*, 01 février 2016.

Davis Elmer, "History of the New York Times 1851–1921", *New York Times*, 26 septembre 1921, p 12.

Folkerts Jean, Teeter Dwight, "chapitre 9: Mass Markets and Mass culture", in *Voices of a nation : a history of mass media in the US*, MacMillan, 1994.

Green Nancy L., *Ready-to-wear and ready-to-work: a century of industry and immigrants in Paris and New York*, Duke University Press, Durham, 1997.

Hamdi A., Jean-Caurant A., Sidere N., Coustaty M., and Doucet A., "An Analysis of the Performance of Named Entity Recognition over OCRed Documents", in *2019 ACM/IEEE Joint Conference on Digital Libraries* (JCDL), 2019, pp. 333–334.

Le Roux Thomas (dir.), *Risques industriels : Savoirs, régulations, politiques d'assistance, fin XVIIe-début XXe siècle*, Presses universitaires de Rennes, 2016.

Kalifa Dominique, Regnier Philippe, Therenty Marie-Ève, Vaillant Alain [dir. *], La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011.

Kaplan Richard L., "From Partisanship to Professionalism: The Transformation of the Daily Press," in Kaestle Carl F. and Radwav Janice A., *A History of the hook in America*, Volume *Four. Printin Motion: The Expansion of Publishing and Reading in the United* States, *1880–1940*, Chapel Hill UNC Press, 2009.

Lenoble Benoît, « Célébrations », in *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011.

Mutuvi Stephen, Doucet Antoine, Odeo Moses, Jatowt Adam, « Evaluating the Impact of OCR Errors on Topic Modeling", *Lecture Notes in Computer Science*, vol 11279, Springer, Cham, 2018, pp. 3–14.

Risley Ford, " Politics and Partisanship", in Sloan David, Mullikin Parcell Lisa, *American journalism: history, principles, practices*, Jefferson NC, Mc Farland, 2002.

Rosenzweig Roy, « Scarcity or abundance? Preserving the past in a digital era *», American Historical Review*, 108–3, juin 2003, pp. 735–762.

Royot Daniel, Ruel Susan, *Médias, société et culture aux Etats-Unis*, Orphys, Ploton, 1996, pp. 16–17.

Schudson Michael, *Discovering the news, a social history of American Newspapers*, New York, Basic Books, 1978.

Whitby Gary, Sloan David, “The purposes of the press: a reinterpretation of American journalism history”, in *Annual meeting of the association for education in journalism*, 1981.

Waller Gregory A., « Search and re-search: digital print archives and the history of multi-sited cinema”, in Wijfjes, Huub. « Digital Humanities and Media History. A Challenge for Historical Newspaper Research ». Tijdschrift Voor Mediageschiedenis, vol. 20, no 1, juin 2017, pp. 4–24. www.tmgonline.nl

Wrona Adeline, « La presse en son miroir, dénonciations », in *La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Nouveau Monde éditions, 2011.

Zoé Kergomard

A Source Like Any Other?

Including Digitised Newspapers in a "Hybrid" Research Project

Abstract: This contribution discusses the potentials and challenges of using digitised newspapers as a source in relation to other newspaper collections, digitised or not. The author argues that digitisation makes the pitfall of media-centrism even more visible and calls for an even more careful contextualisation of media sources. Instead of focusing too narrowly on a sample of seemingly "representative" digitised newspapers, digitisation may in fact invite us to multiply the types of sources and perspectives we include in our research. Dealing with such diverse sources, both digitized and non-digitized, requires that we highlight how we think about, construct, and analyse our corpus. Ultimately, digitisation can lead to a more exploratory and iterative research approach and thus an understanding of the research corpus as an evolving, interconnected, and reflexive collection of diverse sources.

Keywords: digitised newspapers, corpus creation, media history

Today, many historians engage with digital history as a mix of analogue and digital practices, not least because there are still many small-scale, individual research projects that are not formally embedded in digital history.[1] When looking at the still unexplored possibilities offered by digitised newspapers, dividing between "traditional", meaning "qualitative" historians interested in small-scale research on the one side, and digital, "big data" historians working with quantitative methods on the other seems increasingly artificial (Wijfjes, 2017, 6–8). Tim Hitchcock reframed the popular metaphor of the macroscope in digital humanities by suggesting to take into account large and small scales, as well as everything in between (Hitchcock, 2014). "Blended reading" has been suggested as a way to reconcile distant with close reading (Stulpe and Lemke, 2016). As Julia Laite recently argued, digitization also offers myriad possibilities for a "small history", yet "traditional" historians are left to wonder how exactly to do that (Laite). There is an agreement that digitization calls for more reflexivity and, in the context of digitised newspapers, for an attentive digital source criticism (Gibbs and Owens, 2013; Maurantonio, 2014; Rygiel, 2017). But if digital history cannot be

1 I would like to thank the editors and my colleagues at the German Historical Institute Paris and particularly Mareike König and Jürgen Finger for their thoughtful comments on this paper.

https://doi.org/10.1515/9783110729214-016

just a "complement" for traditional historians (Wijfjes, 2017, 4), how does it change their research processes? Concretely, how do we include digitised newspapers into a diverse and hybrid (meaning digitised and non-digitised) source corpus? Are digitised newspapers a source like any other? What implications does it have on the way we conceptualise and organise our research?

In this contribution, I discuss the possibilities of using digitised newspapers as sources among others for projects that may or may not focus primarily on the media. During my previous research on Swiss post-war politics, I began using collections of Swiss digitised newspapers and wondered which place I should give to those sources and, consequently, to the press as a collective actor in my analysis. I argue that digitization makes the pitfall of media-centrism even more visible and calls for an even more attentive contextualization. Taking my current research on the changing interpretations of electoral turnout in France, Germany and Switzerland after 1945 as a starting point, I discuss how digitised newspapers can be included in a diverse corpus – and the impact this may have on how we think of such a corpus. For this contribution, I focus on the part of my research regarding Switzerland, the country for which I have access to the largest digitised newspapers collections. I discuss the challenges I face when collecting and analysing my sources and the possibilities opened by including multiple perspectives – beyond newspapers.

1 Newspapers in Historical Research: A Source Like Any Other?

As an historian with no formal training in digital humanities, I did not reflect much on my first encounters with digitised newspapers at the beginning of the 2010s. Sure, I was fascinated by the rapid development of digitised collections and platforms in Switzerland and the new options they offered for research and teaching.[2] But my approach to digitised newspapers was oftentimes merely

2 While some newspapers (i.e. the *Neue Zürcher Zeitung*) launched their own digitization projects and restricted access to their subscribers, Swiss libraries and archives have been particularly active in launching digitization projects for large and small newspapers and developing platforms in the 2010s. Federalism shows here its advantages, in that it allowed smaller and larger institutions to engage in their own projects, and its drawbacks – even if they are connected and well referenced, the multiplicity of platforms makes it hard to keep up. For an overview see Digitalisierte Schweizer Zeitungen, E-Newspaper Archives, http://www.e-newspaperarchives.ch/?a=p&p=onotherplatforms. Accessed 18 June 2020. For a presentation and a review of two prominent

instrumental. With little effort, I could find information on a little-known personality or a political movement. Soon however, I became interested in the media *per se* in the course of my PhD research on the history of election campaigns in post-war Switzerland (Kergomard, 2020a). While I focused on the perspective of political parties, I also wanted to better understand how media coverage of election campaigns changed over time and how it affected campaigning in Switzerland, following the discussion of "mediatization" as one of the key transformations of politics throughout the 20th century (Swanson and Mancini, 2016; Bösch and Frei, 2006).

However, as noticed in other Western European countries (Vella, 2009, 193; Bingham, 2010, 225–26; Broersma and Harbers, 2007, 1153), the historiography of 20th century, and particularly post-war Switzerland, has largely dwelt little on the press, neither as an actor nor as a type of source worthy of a specific discussion. Recent studies in political or social history often quote newspapers merely for factual evidence or anecdotes.[3] As Fred Gibbs and Trevor Owens suggested, traditional narrative conventions in history might be at fault here: the emphasis on narration hinders us to interrupt the story we are telling in order to discuss how we came to the document in question and how we interpret it (Gibbs and Owens, 2013, 159). As a result, while many researchers have benefited from digitization to access newspaper sources, many questions about their research process are left unanswered. Did they have access to a digitised archive, to a microfilm or to the original newspaper? Did they target specific newspaper(s), and, if yes, over which period of time? Did they use thematic press clippings, compiled by archivists or sometimes by actors themselves? How did they select the articles they quote? As Ted Underwood noted, finding and selecting any kind of digitised sources requires "algorithms to explore a big dataset, and the search process may well have shaped my way of framing the subject, or my intuitions about the representativeness of sources. The scholarly consequences of search practices are difficult to assess, since scholars tend to suppress description of their own discovery process in published work" (Underwood, 2014, 65). Furthermore, by merely quoting articles as factual evidence, we fail to analyse the representations and discourse they convey, or even to relate the newspaper in question to its

projects, Impresso. Media Monitoring of the Past and Enewspaper Archives, (Ehrmann, *et al.*, 2020; Natale, 2019).

3 Not surprisingly, this is the case of (otherwise excellent) reference books on 20th century Switzerland, but also of recently published dissertations. Scholars merely adapt to the constraints of (analogue) book publication. In order to save space, I choose to publish a critical review of my sources and methodology on my scientific blog rather than in my book, see (Kergomard, 2020a).

production and reception context. As a result, we tend to neglect or make invisible the role of the press as a collective actor in the processes we are studying.

Whereas general historiography neglects to historicize the press for itself, this endeavour can be found in works rooted in media history. This segmentation of research preceded newspaper digitization (Bingham, 2010, 2). In recent years, press historians in Switzerland have asked about the long-term institutionalisation of the Swiss press as a professional field, not least through concentration processes (Meier and Häussler, 2010), and approached journalists as historical actors embedded in specific trajectories, networks and professional cultures (Clavien, 2017). In Swiss media studies and political science, there is also a tradition of large projects of content analysis building on (analogue and later digitised) newspaper databases with the objective of identifying long-term transformations in politics and society. In the 1990s, a major research project in media studies reconstructed the hierarchy of "media events" over several decades in order to follow political discussions over time (Imhof *et al.*). While this project also questioned the relationship of journalists to other types of actors, such as social movement activists (Imhof, 1996; Romano, 1998), its very methodology, not least because of its labour-intensive nature, tended to restrict the research perspective to a handful of (overwhelmingly German-speaking) newspapers, supposedly representative of the press field. Furthermore, this approach also ran the risk of taking "events" in press as a proxy for transformations in other fields, for instance, by equating editorial battles between party newspapers of the 1940s with tensions between political parties – although one of the functions of party newspapers was precisely to stage and dramatise political polarisation. Paradoxically, restricting the analysis to press sources while excluding other types of sources may lead to minimising the specificities of the press.

Long before large-scale digitization, both historians and media scholars have thus struggled with the issue of media-centrism (Schudson, 1997, 463–66; Hampton, 2013, 2–3) and with it, with the challenge of contextualising newspapers as media sources. It is tempting to treat newspapers as a constant entity in order to study changes happening in other fields, although their very materiality is constantly affected by changes specific to the media field. Quantitative studies assessing the mediatization of election campaigns in post-war Switzerland, for instance, have a hard time assessing whether their increased media coverage related to their growing importance for political actors or whether it (also) resulted from changes in the media itself – with the increased dramatisation and personalization of politics on television, but also the new narrative strategies of new party-independent newspapers and magazines from the 1960s onwards (Udris, 2013, 6; see also Udris et al., 2015; Kriesi, 2012). By increasingly covering campaigns, journalists also asserted their autonomy *vis-à-vis* politicians, who were in

parallel professionalising their political communication with the help of emerging experts in "public relations". In revealing the backstage of politics, they meant to show the public that they were not falling for the increasingly shining stories served by politicians and their consultants – a circular process in political communication described in many post-war democracies (Fink and Schudson, 2014; Riutort, 2020, 56–74).

Contextualising newspapers by historicizing the relationships between journalists and politicians is therefore always a complex endeavour, not least because we oftentimes struggle with the usual linear narrative of a profession becoming gradually autonomous and objective (Curran, 2009; Broersma and Harbers, 2019, 1154). In post-war Switzerland, the parallel decline of party newspapers and rise of independent newspapers in the first post-war decades are never as clear-cut as is usually assumed (Blum, 2005; Ladner, 2005; Donges, 2005). From one region to another and from one newspaper to another, change occurred in a myriad of ways, if only because whereas some newspapers survived with a completely different line, others tried to adapt but were ultimately closed or merged with other newspapers. Newspapers that remained could also take different ways to distance themselves from the party to which they stood close. Looking at publication and advertising policies is revealing in this regard: the reference Swiss-German newspaper *Neue Zürcher Zeitung*, for instance, was still circulating (unpaid) content on behalf of the Radical Party throughout the 1970s, slowly opening up to ads coming from other parties – but foremost from other right-wing parties. Yet this highly regarded newspaper engaged in a distanced commentary of Swiss and international politics, and could therefore be understood neither as a classical "party newspaper" nor as an "independent newspaper". Beyond older publications that kept track of these changes in real time (Hosang, 1974; Gruner, 1977, 226–34), we lack studies exploring these complex transformations in detail, although it is not only of interest for media scholars, but also essential background information for anyone exploring digitised newspapers. Especially when they distinguish articles from ads, digitised newspapers collections could be a tremendous basis to ask about structural newspaper transformations by looking at advertisement policies, as well as style, form and genres (Broersma, 2007; Broersma and Harbers, 2018; for an overview Wijfjes, 2017, 11). With or without digitization, working with newspaper articles thus requires to relate them to their production and reception context – just as "any other" source.

2 The Temptations of Digitization, or Media-Centrism 2.0

While digitization opens new possibilities to study newspapers for themselves, it also raises the issue of media-centrism in new ways. Not only is it tempting to develop projects based entirely on digitised newspapers (see the chapter in this volume by C-L. Gaillard), but collectively we create a new bias of its own if everyone focuses on the same sample of already digitised newspapers (Hobbs, 2013; Milligan, 2013; Wijfjes, 2017, 16–17). In my current research project, I have struggled with how much space to give to newspapers, and particularly, to digitised newspapers. I ask about the changing interpretations of electoral turnout in France, Germany and Switzerland after 1945, with the aim to uncover the changing political and social meanings of voting in a period which is often understood as the golden age and then the demise of electoral democracy (Conway, 2020; Müller, 2013). I draw inspiration from the concept of frame in media studies, meaning the "interpretative packages" that media coverage attaches to a specific issue (Gamson and Modigliani, 1989, see also D'Angelo, 2002). In my research, this means asking about the understandings and normative evaluations of voting, participation, and democracy that the press put forward when reporting over electoral participation. I am particularly interested in frames that present non-voting and its rise (from the 1960s onwards in Switzerland, from the 1980–1990s onwards in Germany and France) as a "social problem" requiring political action (Spector and Kitsuse, 2017). This additional step towards possible institutional reforms supposed to curb abstention (such as postal voting) can then be studied with the concept of agenda-setting, by asking if abstention was set at the institutional agenda (and for how long), whether possible solutions against it have been discussed, and whether political actors joined forces to enact and implement them.[4]

Clearly, my research questions would have been much more difficult to address in the world of analogue archives. With digitization, a straightforward approach could be to rely on digitised newspapers as a proxy of the public sphere. Thanks to Optical Character Recognition (OCR), a full-text keyword search would lead me to articles that would have been much more difficult to find otherwise. However, my transnational approach makes the above-mentioned bias all the more problematic. For Switzerland, I could access a wide range of digitised newspapers in all three languages, but missed out on several key Swiss-German newspapers that have not been digitised (such as the *Tages-Anzeiger* or the *Basler*

4 For recent discussions, see Hassenteufel, 2010; Princen, 2018.

Zeitung). There are much fewer French and German newspapers that have been digitised for the post-war era.[5] I would face a quantitative imbalance between the three countries that I could only solve by choosing one or two newspapers per country – with the tricky issue of pluralism and representativeness. As Andrew Hobbs has argued regarding the "deleterious dominance" of *The Times* in British historiography, there is a risk of mistaking prestige and centrality for representativeness, at the expense of other (particularly regional or local) newspapers (Hobbs, 2013). While Lev Manovich has advanced that with digitization "we no longer have to choose between data size and data depth" (Manovich, 2012, 466), in such cases we actually choose a restricted media pluralism when we limit ourselves to digitised newspapers – certainly at the expense of data scope.

Furthermore, by focusing only on newspapers, I would still have to be cautious not only about media-centrism, but also about diachronic comparability. Overemphasising the role of the media in defining and prioritising topics is already a well-known risk when studying agenda-setting. In this case particularly, focusing only on the interests of journalists might be particularly tricky. Discussing electoral participation and its decline as a symptom for a "democratic crisis" could very well belong to the dramatised story-telling of politics that many journalists adopted from the 1960s onwards. It might also have corresponded to their rediscovered role in Swiss democracy: contributing to citizenship education, strengthening the interest of the public for Swiss politics and thereby its inclination to vote.[6] Consequently, it would not be surprising if they increasingly remarked on electoral participation, particularly from the 1960s onwards, as it started to decline. Speaking about these topics in the media was also becoming detrimental for other (older or newer) professions seeking to legitimise their forms of expertise in the public sphere: communication experts, but also pollsters and pundits. As talk of a "crisis of representation" emerged in France in the 1980–1990s, political scientist Bernard Lacroix relativized this diagnosis and attributed it to the growing importance of these new experts in public debate. For Lacroix, lamenting about this crisis, the rise

5 The handful of digitalization projects after 1945 mostly come from the newspapers themselves, so that the articles are only accessible under a paywall (for instance *Le Monde, Frankfurter Allgemeine Zeitung*). See for Germany (Blome); for France, Bibliothèques numériques de presse francophone, Bibliothèque nationale de France, https://bnf.libguides.com/presse_medias/bibliotheques_numeriques. Accessed 19 June 2020.

6 The advent of television was detrimental in formalising this role. While press journalists did not have the same guidelines as in the public sector, they worked together with the public broadcasting actors to make the coverage of Swiss politics more attractive in the 1960s (Vallotton, 2006; Kergomard, 2020a).

of abstention and other forms of "problematic" electoral behaviour reinforced their legitimacy to decipher and interpret the moods of the "people" for journalists and political parties (Lacroix, 1994). While my aim is less to find out whether there actually was "a crisis" than to assess the circulation of such a diagnosis in the public sphere, Lacroix's provocative analysis points out the risk of looking only at a sort of media and expert bubble. Focusing only on a handful of digitised, oftentimes national and prestigious newspapers may lead to base one's analysis on a deforming magnifying glass for broader processes within and beyond the media – a problem much older than Twitter. Consequently, I have chosen to include not only digitised and non-digitized newspapers, but also other actors and sources in my research project. Instead of focusing on one type of actor and/or sources, I aim at exploring the distribution of discourse on electoral participation in the public sphere at large. Against the pitfalls of media-centrism, this approach allows me to approach framing and agenda-setting as multi-actor, contingent and open-ended processes (Hassenteufel, 2010, 56). My corpus is hence heterogeneous, including archives from political authorities and parties (both "analogue" and digitised – i.e. parliamentary debates), publications from experts (political scientists, pollsters) on the topic (digitised or not), and newspaper articles, whether non-digitized, digitised or born-digital for more recent decades.

3 Rethinking the Corpus?

The process of searching and collecting these diverse sources has led me to reflect on how I approached my research corpus. As mediaeval historians in particular have discussed, the notion of "corpus" has been increasingly used in history since the 1980s, but is rarely explicitly defined (Magnani, 2017). Its popularity has been partly supported by projects in other disciplines (particularly linguistics) working on large text corpora, oftentimes with quantitative methods, and with a stronger focus on the sources themselves than on the traditional interests of historians in source criticism and source contextualization. Digitization and the development of text mining methods have only sharpened the need to discuss this notion (Treffort, 2014). Now more than ever, historians are asked to reflect on the boundaries of their corpus, its homogeneity or heterogeneity and how they justify its representativeness, while other disciplines have a much more precise (and restrictive) understanding of what a corpus should be. For instance, the linguist Damon Mayaffre working on historical political speeches approaches a corpus as a "coherent and self-sufficient whole" through the connections existing between texts. Researchers should not distinguish between "within" and "outside" the

corpus: in order to analyse all elements with the same (ideally quantitative) methodology and reduce subjectivity, Mayaffre advocates including all contextual elements (i.e., biographical information) in the corpus at the beginning of a research process (Mayaffre, 2002, 5).

While historians share this interest for corpus as a formalisation of intertextuality (Treffort, 2014), asking to constitute one's corpus once and for all seems to go against the usual (but rarely formalised) archival research process, marked by the very limitations of archives: historians search and collect sources in an iterative manner, go back and forth between different possible trails, incidentally discover an interesting source where they had not initially looked … Does (and/or should) archival research remain an iterative process when it includes digitised sources? On the one hand, many aspects of digitised newspaper collection lead to a rather explorative search process. Digitised sources can be searched again and again, which helps to deal with the (diminishing but still unavoidable) limitations of digitised collections: missing pages, faulty article segmentation, and above all, the varying conditions of digitization and OCR processing between newspapers.[7] Precisely because of these issues, digitised sources may also evolve, which can require new search iterations – OCRization can be improved; automatic topic assignment can change. And even while topic modelling itself implies a closed corpus, historians are encouraged to use the discovered topics precisely to allow an explorative, open-ended research process (Underwood, 2014). On the other side, there can be an ambiguity in the status of the "article collections" that can now be directly exported out of some newspaper platforms (as in Switzerland *impresso*, and for general overview, see Ehrmann, *et al.*, 2019). This sophisticated feature helps scholars to quickly gather multiple articles and their metadata and easily export them for analysis. But when I started extracting my own collections, I realised it led me to think of them as a closed entity, which seemed at odds with the way I otherwise conducted my research. I was also at risk of idealising the coherence and exhaustiveness of such a collection and "freezing" it at a given time, instead of improving it and letting it grow with time.

This is where it helps me not to think of my digitised newspaper articles as a closed, "frozen", and ideally "representative" sub-corpus, but, rather, as just another part of my whole corpus understood as an evolving collection of diverse sources that gains its coherence through my reflecting on the iterative search process and through the connections I make between them. Of course, iteration has long been formalised as an integral part of qualitative research in social sciences

7 For a discussion of these boundaries and of current projects to reduce them, Broersma and Harbers, 2018, 1151; Ehrmann, *et al.*, 2020, 2.

(Lejeune, 2014), and many historians have also pointed to the advantages of going back and forth between various sources and, hence, points of views as a way to destabilise and decenter one's own assumptions and standpoint. This approach is for instance central in the "toolbox" of *histoire croisée* developed by Bénédicte Zimmermann and Michael Werner (Werner and Zimmermann, 2004). For my project, embracing (and constantly questioning) an iterative research process thus allows me to go back and forth between the different political and cultural contexts I am looking at.

Indeed, digitization might actually enhance the possibilities to engage in iterative research processes, since it facilitates exploration. As Lara Putnam puts it, "for the first time, historians can find without knowing where to look" (Putnam, 2016, 377). We are able to follow "faint trails of breadcrumbs" from digitised sources to physical archives (and vice-versa) with "substantially more chance of finding something" (Laite, 2020, 10). Tim Underwood advocates reflecting on the search process itself as a "hermeneutic spiral", since search strategies often encode "assumptions about the patterns we expect to find" (Underwood, 2014). Instead of seeing them only as limitations, the different biases of each search process could complement each other: just as "no single collection of volumes is perfectly representative of print culture, in practice, the best way to address questions of representativeness is often to pose the same question in multiple collections that have been selected in different ways" (Underwood, 2014). We may thereby jungle between traditional archival search, which is more constrained by fixed archival labels and categories (Laite, 2020), and the sort of "screwing around" made possible by digitization (Ramsay, 2014). This type of research process gains in scientificity by being well documented, constantly reflected and made explicit in our writing, which is why Owen and Gibbs suggest deemphasizing the importance of narrative when we write (Gibbs and Owens, 2013). While digitization certainly gives room to historical research projects based on closed, "frozen" corpora, it can therefore also be an incentive to rethink how we understand a corpus in the framework of a qualitative and iterative research process, namely as an evolving, interconnected and reflexive collection of diverse sources.

4 Digitization as an Incentive for Multi-Perspective, Iterative Research Processes

In order to explore the media coverage of electoral turnout in post-war Switzerland, I found that switching between different search processes and sources provides me with multiple points of entry that complemented each other. A first

door into Swiss politics of the late 20th century can be the *Année politique suisse*, a political yearly chronicle published by the Institute of Political Science at the University of Bern since 1966 and now available in a digitised form.[8] Every year, its collaborators chronicled political discussions and decision-making processes on the basis of press reviews. Going through the *Année politique suisse* enables me to identify topics that were discussed in relation to electoral turnout, such as postal voting. Clearly, such chronicles document topics and events that were deemed relevant to the chroniclers at the end of the year – in all likelihood because they remained discussed throughout the year or eventually led to political measures. I approach it foremost as a retrospective chronicle of successful agenda-setting. To go beyond the limited amount of newspaper articles that the *Année* sources, I also consult the press clipping at the Swiss Federal Archives on which it is based.[9]

Thematic press clippings allow a different entry: they were usually constituted on a real-time basis and are more likely to include topics that did not remain long in the media and/or did not make it to the institutional agenda. In the era of digitised newspapers, they will remain useful, since they allow a thematic entry into a wide range of newspapers, including those that have not been digitised. Institutions like the Swiss Economic Archives in Basel are therefore in the process of digitising their press clipping collections.[10] For archivists and scholars alike, digitization is a new incentive to reflect (retrospectively) on the constitution of these collections and hence on their inherent biases, parallel to the query criticism we exercise for digitised newspapers. Even if archivists have rarely explicated their selection criteria, they have usually kept the same topical categories over decades. At the Swiss Social Archives, for instance, the collection on "electoral participation" led me to a wide range of articles (including from larger and smaller newspapers linked to the Social Democratic Party) providing explanations and/or moral evaluations of non-voting. While the collection becomes larger in the 1970–80s, it would be hasty to conclude from there that the topic itself gained in importance – archivists could also have interpreted the topical category in a broader way than they used to.[11]

8 Institut für Politikwissenschaft an der Universität Bern; https://anneepolitique.swiss/.

9 Swiss Federal Archives, J2.300–01*, Institut für Politikwissenschaft der Universität Bern: Dokumentation zur schweizerischen Politik (1965-).

10 'Wirtschaftsdokumentation – Elektronische Zeitungsausschnittsammlung'. Schweizerisches Wirtschaftsarchiv, https://ub-easyweb.ub.unibas.ch/de/historische-bestaende/wirtschaftsdokumentation/. Accessed 19 June 2020.

11 Swiss Social Archives, Zeitungsartikel 37.0, Abstimmungen und Wahlen, Stimm- und Wahlbeteiligung, 1943–1993; 1993–2006.

Parallel to exploring the *Année politique suisse* and press clippings, I dig into digitised collections, starting with *impresso* and *E-newspaper-Archives.ch*. These search processes complement each other, sometimes in very practical ways: articles in press clippings are separated from their reading environment and digitised newspaper platforms can actually help to link them to illustrations or related articles of the same issue. This parallel search allows me to think critically about the biases of each option. Since I dropped the ambition to constitute a large corpus that would be fixed for the rest of my research project, I have decided to start with restrictive search criteria, which I can then gradually expand. On the platform *impresso* for instance, I begin by restricting my query to Swiss newspapers and choosing locations in Switzerland.

Keyword search has proved to be a double-edged sword. On the one hand, it allows to go much further than with the topical categories of press clipping. On the other hand, as David Deacon has noted, "keyword searching is best suited for identifying tangible 'things' (i.e. people, places, events and policies) rather than 'themes' (i.e. more abstract, subtler and multifaceted concepts)" (Deacon, 2017, 8).[12] While archivists could collect varied articles on the topic of "electoral participation" for their press clippings, using this phrase as a keyword on digitised platforms is too broad: I land on almost any short press dispatch written about elections and/or votes in parliaments. While I am not interested in the latter context, it is nevertheless semantically very close to my research topic, to the point that it would be too restrictive to exclude it from my search using Boolean operators. Using topic modelling, *Impresso* has identified semantic topics that were integrated as search facets. This option seems more interesting to explore topic distribution over an entire collection (Jacobi *et al.*, 2015; Ehrmann, *et al.*, 2020, 964), but it was not fruitful in my case, again because of this semantic proximity (i.e. articles about elections and votes in parliaments were assigned the same topic). I also tried to restrict my search to articles on the front page and/or displaying the search keyword in their title, which is a very useful feature of recent platforms. Yet, I found that incidental remarks on electoral participation were equally interesting and that, in any case, many articles centring on election participation had a completely different title.

Looking for the right balance between quantity (or "recall") and precision, I therefore experimented with more specific keywords directing the search to the context of mass elections and referenda, using fuzzy search options when possible to include possible OCR errors. In essence, while asking about "electoral

12 The first type of search is therefore also favoured by users, see the discussion of user behaviour studies and interface reviews in Ehrmann, *et al.*, 2019, 3–4.

participation" was too neutral, I could try to reverse the search and ask about "non-voting". For the French-speaking newspapers, the word "abstention" also gave too many results in a variety of contexts, again regarding votes in parliament but also in foreign policy. I then tried "abstentionnisme", a now dated term with a latent negative coloration of non-voting.[13] For the Swiss-German newspapers, I tried the older words "Wahlabstinenz" and "Stimmabstinenz" as well as the more neutral "Wahlenthaltung" and "Stimmenthaltung" (abstention at elections resp. referenda). While these keywords have helped me explore the discussion around non-voting, their moral connotation (or at least the implication that non-voting is a problem) is certainly a bias in itself, so that I should not assess the importance of this discussion in the Swiss post-war public debate based exclusively on these searches. Instead, I can study the different contexts in which non-voting was discussed and the various frames that were conveyed. Via the keyword search on *impresso*, I stumbled upon articles discussing electoral participation in contexts that I had not thought of: for instance, at national or cantonal holidays, politicians warned citizens of the "danger" it represented for Swiss democracy and insisted on voting as a civic duty.[14] Such serendipity effects help not only to "find without knowing where to look" (Putnam, 2016, 377) but also to avoid trapping the search in what I already know. Press clippings have led me to the traces of citizenship ceremonies ("Jungbürgerfeier"): designed to stress the importance of political responsibilities upon youngsters, they were regularly disrupted by activists of the 1968s and early 1980s movements and thus became a space where politicians and youngsters renegotiated the norms of citizenship.[15] I also look for specific concepts or diagnoses expressed parallel to non-voting, such as "political disenchantment" ("politische Verdrossenheit" in German). Because of their semantic precision, concepts can work well in a traditional keyword search and digitised newspapers are thus a goldmine to follow their emergence and circulation in the public sphere (Ronnes and Kessel, 2016).

Along my searches, query criticism both for digitised and non-digitized sources is an inherent part of my workflow and allows me to redirect my further research steps. I go back and forth between the selections operated by generations of archivists in press clippings, my own queries but also my discoveries in traditional archives. For instance, when looking for early statistical studies about non-voting, I

13 See in the French context Barbet, 2007.

14 For instance s.n. 'Ce fut samedi soir le traditionnel banquet de la restauration. M. Casaï dénonce les méfaits de l'abstentionnisme'. Journal de Genève, 14 January 1952, https://impresso-project.ch/app/issue/JDG-1952-01-14-a/page/JDG-1952-01-14-a-p0004/article/JDG-1952-01-14-a-i0054. Accessed on 23 May 2020.

15 Swiss Social Archives, Zeitungsartikel 14.3, Staatsbürgerliche Erziehung, 1945–1996.

noticed that many did not refer to "non-voting", but instead to "the non-voter" in their titles.[16] I then included the German term "Nichtwähler" to my queries,[17] which directly led me to articles trying to sketch out the profile of the "non-voter". It helped me identify a specific interest of journalists for establishing a profile of this peculiar citizen, which they quelled by discussing scientific studies.[18] Such connections between sources are, in the end, what "glues" my corpus together and are hence essential to document. I compile and enrich such information in a Zotero library with the aim of eventually making it available online. Thanks to the growing amount of metadata provided by platforms and the possibility to export them, I focus particularly on metadata categories such as author, named entities, article genre, and text reuse, since they also help me to better identify the specific discourse configurations in which electoral turnout is a topic. I can identify journalists writing frequently on the topic (particularly in editorials) as well as individuals or collective actors asked to intervene on it in the media. I then read the transcription of articles, correct the main OCR mistakes that hinder my understanding, all the while keeping the digitised collection open so that I can look at the facsimile and replace the article in the context of the newspaper. A last step is then to annotate my articles with keywords and comments and to connect them to related sources. This is already possible in reference management softwares like Zotero. CAQDAS (Computer-Assisted/Aided Qualitative Data Analysis Software) like Atlas.ti are a step above as they allow to annotate directly in the source material and help to organise a reflexive coding system.

For my analysis, quantification comes as an option among others to explore my sources. I cannot use it as a means of evidence, since my corpus is not designed to be representative of a larger entity and is not preprocessed for text mining (with the variety of sources I look at, that would be a project in itself). Even considering these flaws, quantification can yet be a way to ask new questions. The *impresso* platform has a useful n-gram option, which indicated rising occurrences for "abstentionnisme" from the 1960s onwards for all Swiss newspapers in question, and a steady decline from the 1980s onwards. But instead of concluding to a successful agenda-setting on the part of the media between the 1960s and the

16 One of the first German studies on the topic already used the phrase "Party of non-voters" as its title, Würzburger, Eugen. 'Die „Partei der Nichtwähler"'. Jahrbücher für Nationalökonomie und Statistik, vol. 33 (88), no. 3, 1907, pp. 381–89.

17 Other terms for "non-voter" such as Wahlabstinenzler, Stimmabstinenzler, abstentionniste were already included in my previous search thanks to fuzzy search options.

18 S.n. 'Zürichs schweigende Mehrheit wächst. Eine Analyse der Nichtwähler bestätigt fatale Tendenzen'. Die Tat, 20 January 1972, http://www.e-newspaperarchives.ch/?a=d&d=DTT19720120-01.2.1&srpos=5. Accessed 23 April 2020; Schwaar, Egon. 'Die Nichtwähler bei den Gemeinderatswahlen 1970'. Zürcher Statistische Nachrichten, vol. 48, no. 2, 1971, pp. 75–101.

1980s, mapping the actors at stake and their discursive contexts has revealed a more contrasted picture. While journalists had regularly commented upon the decline in turnout starting in the 1960s, it was oftentimes the same politicians who raised alarm on what they saw as a lack of "civic duty" (a recurrent frame) and the "danger" it caused for Swiss democracy.[19] Throughout the 1970s, federal and cantonal authorities commissioned experts and working groups with the task of explaining the decline in turnout and suggesting solutions to curb it. While their reports were unsure about what could be done, this explosion of interpretations on electoral participation changed the distribution of speech in the public sphere: "ordinary" citizens themselves, who had rarely been consulted on this topic, suggested their own takes in letters to the editors. They often insisted that they were dutiful voters, so that the discussion on non-voting in newspapers still mostly took place without non-voters.[20] But both they and some journalists started to criticise the "citizen bashing" and "crocodile tears", indulged by politicians, which might explain that the topic slowly started to lose attention in the public sphere.[21] This short-lived focalization on non-voting may have foremost crystallised the worries of political elites over the future of Swiss democracy in the age of television and new social movements (Kergomard, 2020c). Based on these first results, I can now deepen and expand my research to the other two countries in my study, in order to study the ways electoral participation was problematized in a transnational perspective. With these various entry points and tools, my analysis benefits from my going back and forth between the "frenetic fishing for information" and the "peaceful, monotonous" reading and annotating, as Claire-Lise Gaillard observed (Gaillard, 2018).

19 'Postulat Schalcher. Aktivierung der schweigenden Mehrheit'. Amtliches Bulletin der Bundesversammlung, vol. I, no. 11263, 20 March 1973, pp. 373–76.

20 Julliard, Horace. 'Monsieur le Rédacteur en chef ... L'affligeante médiocrité de notre personnel politique'. Journal de Genève, 29 October 1979, https://impresso-project.ch/app/issue/JDG-1979-10-29-a/page/JDG-1979-10-29-a-p0002/article/JDG-1979-10-29-a-i0008. Accessed 22 April 2020.

21 Diezi, Cécile. 'Regard. Le mutisme du citoyen'. L'Impartial, 19 January 1984, https://impresso-project.ch/app/issue/IMP-1984-01-19-a/page/IMP-1984-01-19-a-p0017/article/IMP-1984-01-19-a-i0171. Accessed 5 May 2020; Juilliard, Olivier. 'Monsieur le Rédacteur en chef ... Et si l'on punissait les abstentionnistes?' Journal de Genève, 29 October 1979, https://impresso-project.ch/app/issue/JDG-1979-10-29-a/page/JDG-1979-10-29-a-p0002/article/JDG-1979-10-29-a-i0009. Accessed 22 April 2020; Amstutz, Peter. "Wählerschelte statt Parteitag," Basler Zeitung, 5 September 1983. Swiss Social Archives, Zeitungsartikel 38.7, Bauern-, Gewerbe, und Bürgerpartei, 1934–1985.

5 Conclusion

What is the impact of including digitised newspapers in small-scale research projects? Digitization is certainly an incentive for more reflexivity that can be fruitful for all historians, regarding source criticism, but also search processes and source selection. As I discussed in this contribution, historians have struggled with the place to give to newspaper material long before digitization, as "any other" source, with specific production and reception contexts. Outside media history, they were often used for factual evidence without much contextualization, whereas media scholars reversely faced the pitfall of media-centrism. Digitization reinforces both problems, not least because it often leads us to focus on the same already digitised newspapers. Instead of focusing too narrowly on a sample of seemingly "representative" digitised newspapers, digitization can actually invite us to multiply the type of sources and perspectives we include in our research. But dealing with diverse, digitised and non-digitized sources asks that we clarify how we think of, construct and analyse our corpus. While the methodological reflection on digitised newspapers has until now mostly focused on large-scale (oftentimes quantitative) projects in media history and rather led to a "frozen" approach to corpora, digitization can also enhance a more explorative, iterative research approach and hence an understanding of corpus as an evolving, interconnected and reflexive collection of diverse sources. While these different approaches need to be clarified and reflected upon, they certainly complement each other, since all historical projects dealing with the media can shed light on its place as a collective actor mediating societal discussions and perceptions in historical processes (Schudson, 1997, 473).

Bibliography

Barbet, Denis. 'Quand les mots de l'abstention parlent des maux de la démocratie'. *Mots. Les langages du politique*, no. 83, 2007, pp. 53–67.

Bingham, Adrian. 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians'. *Twentieth Century British History*, vol. 21, no. 2, June 2010, pp. 225–31.

Blome, Astrid. 'Zeitungen'. *Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften*, edited by Busse, Laura, Wilfried Enderle, Rüdiger Hohls, Thomas Meyer, Jens Prellwitz and Annette Schuhmann, 2nd ed., 2018, 10.18452/19244.

Blum, Roger. 'Politischer Journalismus in der Schweiz'. *Politische Kommunikation in der Schweiz*, edited by Patrick Donges and Otfried Jarren, Haupt, 2005, pp. 115–30.

Bösch, Frank, and Norbert Frei, editors. *Medialisierung und Demokratie im 20. Jahrhundert*. Wallstein, 2006.

Broersma, M. J. *Form and Style in Journalism. European Newspapers and the Representation of News 1880–2005*. Peeters, 2007.

Broersma, Marcel, and Frank Harbers. 'Exploring Machine Learning to Study the Long-Term Transformation of News'. *Digital Journalism*, vol. 6, no. 9, Oct. 2018, pp. 1150–64.

Clavien, Alain. *La presse romande*. Antipodes, 2017.

Conway, Martin. *Western Europe's Democratic Age: 1945–1968*. Princeton University Press, 2020.

Curran, James P. 'Narratives of Media History Revisited'. *Narrating Media History*, edited by Michael Bailey, Routledge, 2009, pp. 1–21.

D'Angelo, Paul. 'News Framing as a Multiparadigmatic Research Program: a Response to Entman', in: Journal of Communication vol. 52, no. 4, Dec. 2002, pp. 870–888.

Da, Nan Z. 'The Digital Humanities Debacle'. *The Chronicle of Higher Education*, Mar. 2019. *The Chronicle of Higher Education*, https://www.chronictle.com/article/The-Digital-Humanities-Debacle/245986.

Deacon, David. 'Yesterday's Papers and Today's Technology: Digital Newspaper Archives and "Push Button" Content Analysis'. *European Journal of Communication*, vol. 22, no. 1, 2017, pp. 5–25.

Donges, Patrick. 'Politische Kommunikation in der Schweiz. Medialisierung eines "Sonderfalls"?' *Politische Kommunikation in der Schweiz*, edited by Patrick Donges and Otfried Jarren, Haupt, 2005, pp. 7–27.

Ehrmann, Maud, Estelle Bunout, and Marten Düring, 'Historical Newspaper User Interfaces: A Review'. *IFLA WLIC 2019*, 2019. *library.ifla.org*, http://library.ifla.org/2578/.

Ehrmann, Maud, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströber, and Raphaël Barman, 'Language Resources for Historical Newspapers: The Impresso Collection'. *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 958–68. *Association for Computational Linguistics*, https://www.aclweb.org/anthology/2020.lrec-1.121.

Fink, Katherine, and Michael Schudson. 'The Rise of Contextual Journalism, 1950s–2000s'. *Journalism*, vol. 15, no. 1, Jan. 2014, pp. 3–20.

Gaillard, Claire-Lise. 'Feuilleter la presse ancienne par Giga Octets'. *Le goût de l'archive à l'ère numérique*, 4 June 2018, http://www.gout-numerique.net/.

Gamson, William A., and Andre Modigliani. 'Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach'. *American Journal of Sociology*, vol. 95, no. 1, July 1989, pp. 1–37.

Gibbs, Fred, and Trevor Owens. 'The Hermeneutics of Data and Historical Writing'. *Writing History in the Digital Age*, edited by Kristen Nawrotzki and Jack Dougherty, University of Michigan Press: Digitalculturebooks, 2013, pp. 159–70.

Gruner, Erich. *Die Parteien in der Schweiz*. 2nd ed., Francke, 1977.

Hampton, Mark. 'Historical Approaches to Media Studies'. *The International Encyclopedia of Media Studies*, American Cancer Society, 2013, pp. 381–98.

Hassenteufel, Patrick. 'Les processus de mise sur agenda : sélection et construction des problèmes publics'. *Informations sociales*, vol. 157, no. 1, Feb. 2010, pp. 50–58.

Hitchcock, Tim. 'Historyonics: Big Data, Small Data and Meaning'. *Historyonics*, 9 Nov. 2014, http://historyonics.blogspot.com/2014/11/big-data-small-data-and-meaning_9.html.

Hobbs, Andrew. 'The Deleterious Dominance of The Times in Nineteenth-Century Scholarship'. *Journal of Victorian Culture*, vol. 18, no. 4, 2013, pp. 472–97.

Hosang, Balz Christian. *Parteien und Presse: die Beziehungen zwischen den politischen Parteien und der politischen Presse: ein Beitrag zum Problem der Meinungsbildung durch die politische Presse im Kanton Zuerich*. Haupt, 1974.

Imhof, Kurt. 'Eine Symbiose: Soziale Bewegungen and Medien'. *Politisches Raisonnement in der Informationsgesellschaft*, edited by Kurt Imhof and Peter Schulz, Seismo Verlag, 1996, pp. 165–86.

Imhof, Kurt, editors. *Zwischen Konflikt und Konkordanz. Analyse von Medienereignissen in der Schweiz der Vor- und Zwischenkriegszeit*. Seismo, 1993.

Institut für Politikwissenschaft an der Universität Bern. *Année Politique Suisse*. Institut für Politikwissenschaft, 1966.

Jacobi, Carina, Wouter Atteveldt, and Kasper Welbers, 'Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling'. *Digital Journalism*, vol. 4, Oct. 2015, pp. 1–18.

Kergomard, Zoé. 'L'histoire politique suisse est-elle ennuyeuse? Ou le potentiel encore inexploité des archives des partis politiques suisses'. *Histoire politique 2.0*, 30 Jan. 2020a, https://polsuisse.hypotheses.org/136.

Kergomard, Zoé. *Wahlen Ohne Kampf? Schweizer Parteien Auf Stimmenfang, 1947–1983*. Schwabe, 2020b.

Kergomard, Zoé. Knowledge on a Democratic "Silence": Conflicting Expertise on the Decline in Voter Turnout in Postwar Switzerland (1940s–1980s), in: *KNOW: A Journal on the Formation of Knowledge*, vol. 4, no. 2, 2020c, pp. 232–261.

Kriesi, Hanspeter. 'Personalization of National Election Campaigns'. *Party Politics*, vol. 18, no. 6, 2012, pp. 825–44.

Lacroix, Bernard. 'La "crise de la démocratie représentative en France". Eléments pour une discussion sociologique du problème'. *Scalpel*, vol. 1, 1994, pp. 6–29.

Ladner, Andreas. 'Die Parteien in der politischen Kommunikation. Mediendemokratie: Herausforderungen und Chancen für die politischen Parteien'. *Politische Kommunikation in der Schweiz*, edited by Patrick Donges and Otfried Jarren, Haupt, 2005, pp. 57–74.

Laite, Julia. 'The Emmet's Inch: Small History in a Digital Age'. *Journal of Social History*, vol. 53, no. 4, June 2020, pp. 963–89.

Lejeune, Christophe. *Manuel d'analyse qualitative : Analyser sans compter ni classer*. De Boeck, 2014.

Magnani, Eliana. 'Qu'est-ce qu'un corpus ?' *Les carnets de l'IRHT*, 2017, https://irht.hypotheses.org/3187.

Manovich, Lev. *Trending: The Promises and the Challenges of Big Social Data*. Edited by Matthew K. Gold, 2nd ed., University of Minnesota Press, 2012.

Maurantonio, Nicole. 'Archiving the Visual. The Promises and Pitfalls of Digital Newspapers'. *Media History*, vol. 20, no. 1, Jan. 2014, pp. 88–102.

Mayaffre, Damon. 'Les corpus réflexifs : entre architextualité et hypertextualité'. *Corpus*, no. 1, Nov. 2002.

Meier, Peter, and Thomas Häussler. *Zwischen Masse, Markt und Macht : Das Medienunternehmen Ringier im Wandel 1833–2009*. Chronos-Verlag, 2010.

Milligan, Ian. 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010'. *The Canadian Historical Review*, vol. 94, no. 4, Nov. 2013, pp. 540–69.

Müller, Jan-Werner. *Contesting Democracy: Political Ideas in Twentieth-Century Europe*. Reprint, Yale University Press, 2013.

Natale, Enrico. 'Compte rendu: e-newspaperarchives.ch'. *infoclio.ch*, 18 Feb. 2019, https://infoclio.ch/de/compte-rendu-e-newspaperarchivesch.

Princen, Sebastiaan. Agenda-Setting and Framing in Europe, in: Ongaro, Edoardo and Sandra Van Thiel (eds.): *The Palgrave Handbook of Public Administration and Management in Europe*. Palgrave, 2018, pp. 535–551.

Putnam, Lara. 'The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast'. *The American Historical Review*, vol. 121, no. 2, Apr. 2016, pp. 377–402.

Ramsay, Stephen. 'The Hermeneutics of Screwing Around; or What You Do with a Million Books'. *Pastplay: Teaching and Learning History with Technology*, edited by Kevin Kee, University of Michigan Press, 2014, pp. 111–20.

Riutort, Philippe. *Sociologie de la communication politique*. 3rd ed., La Découverte, 2020.

Romano, Gaetano. 'Die Überfremdungsbewegung als «neue soziale Bewegung». Zur Kommerzialisierung, Oralisierung und Personalisierung massenmedialer Kommunikation in den 60er Jahren'. *Dynamisierung und Umbau: Die Schweiz in den 60er und 70er Jahren*, edited by Mario König, Georg Kreis, Franziska Meister and Gaetano Romano, Chronos, 1998, pp. 143–59.

Romein, C. Annemieke, et al. 'State of the Field: Digital History'. *History*, vol. 105, no. 365, 2020, pp. 291–312.

Ronnes, Hanneke, and Tamara Van Kessel. 'Heritage (Erfgoed) in the Dutch Press: A History of Changing Meanings in an International Context'. *Contributions to the History of Concepts*, vol. 11, no. 2, Dec. 2016, pp. 1–23.

Rygiel, Philippe. *Historien à l'âge numérique*. Presses de l'Enssib, 2017.

Schudson, Michael. 'Toward a Troubleshooting Manual for Journalism History'. *Journalism & Mass Communication Quarterly*, vol. 74, no. 3, Sept. 1997, pp. 463–76.

Spector, Malcolm, and John I. Kitsuse. *Constructing Social Problems*. Routledge, 2017.

Stulpe, Alexander, and Matthias Lemke. 'Blended Reading'. *Text Mining in den Sozialwissenschaften*, edited by Matthias Lemke and Gregor Wiedemann, Springer VS, 2016, pp. 17–61.

Swanson, David L., and Paolo Mancini. *Politics, Media, and Modern Democracy: An International Study of Innovations in Electoral Campaigning and Their Consequences*. Greenwood Publishing Group, 1996.

Treffort, Cécile. 'Le corpus du chercheur, une quête de l'impossible? Quelques considérations introductives'. *Les Annales de Janua, actes des journées d'études*, vol. 2, no. Le corpus. Son contour, ses limites et sa cohérence, Apr. 2014, https://annalesdejanua.edel.univ-poitiers.fr/index.php?id=725.

Udris, Linard. *Schweizer Medien im Wahlkampf. Qualität der Medienberichterstattung vor den Eidgenössischen Wahlen 2011*. Schwabe, 2013.

Udris, Linards, Jens Lucht and Jörg Schneider, 'Contested Elections in Increasingly Commercialized Media. A Diachronic Analysis of Executive and Parliamentary Election News Coverage in Switzerland'. *Swiss Political Science Review*, vol. 21, no. 4, 2015, pp. 578–95.

Underwood, Ted. 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago'. *Representations*, vol. 127, Aug. 2014, pp. 64–72.

Vallotton, François. 'Anastasie ou cassandre? Le rôle de la radio-télévision dans la société helvétique'. *Radio und Fernsehen in der Schweiz: Geschichte der schweizerischen Radio- und Fernsehgesellschaft SRG 1958–1983*, edited by Theo Mäusli, Hier + Jetzt, 2006, pp. 37–82.

Vella, Stephen. 'Newspapers'. *Reading Primary Sources. The Interpretation of Texts from Nineteenth-and Twentieth-Century History*, edited by Miriam Dobson and Benjamin Ziemann, Routledge, 2009, pp. 192–208.

Werner, Michael, and Bénédicte Zimmermann. *De la comparaison à l'histoire croisée*. Editions du Seuil, 2004.

Wijfjes, Huub. 'Digital Humanities and Media History: A Challenge for Historical Newspaper Research'. *TMG Journal for Media History*, vol. 20, no. 1, June 2017, pp. 4–24.

Suzanna Krivulskaya

The Crimes of Preachers: Religion, Scandal, and the Trouble with Digitised Archives

Abstract: Newspaper digitization has opened up new opportunities for historical research, but it has also highlighted persistent methodological problems. Imprecisions of optical text recognition and shortcomings in the content of the sources themselves necessitate a reckoning with the promises and limitations of large digitized textual corpora, particularly when it comes to newspapers. Using a subset of U.S. newspapers from the late nineteenth and early twentieth centuries as a case study, this chapter considers the methodological problems inherent in working with digitized sources and recommends possible directions for navigating the problem of fragmentary evidence in both digital and analogue modes of historical analysis.

Keywords: digitised newspapers, U.S. freethinkers, elopement scandals

1 Introduction

The sound of an accidental tear of crumbling yellowed newspaper material is among the more unnerving side effects of archival research. This particular paper – whose crumbs populate the special collections desk long after the box has been returned to the stacks – is from Brooklyn, 1872, and not in great shape. "Not to worry," the panicked researcher remembers, "It has been digitized." A sigh of relief – though the exhale is made difficult by the archival dust particles inhaled by the researcher in the process of maneuvering the fragile artifact. Brown stains appear on white archival gloves with some frequency – as the researcher interacts with the rusting paper clips that the donor or archivist failed to remove prior to arranging the newspapers into acid-free archival boxes. Sometimes, an overlooked rusted staple might prick the finger, but as far as occupational hazards go, historians have it easy.

Technology has provided answers for the tears, the staples, and the occasional finger pricks; digitization may soon render archival visits that rely on newspaper research nearly obsolete. The crumbling records of the past now reside on server farms and are easily accessible, zoomable, and searchable with the click of a few buttons (and, usually, the paying of subscription costs). But while the problem of

https://doi.org/10.1515/9783110729214-017

access (and accidental damage by trembling hands) has largely been solved, the trouble with the archive – digitized or not – persists in multiple iterations. The issues are both technical and methodological: from the imprecision of text recognition to the unreliability of newspaper accounts as witnesses to historical events.

This chapter considers the promises and perils of working with digitized newspapers in the context of the Gilded Age (1870–1900) and Progressive Era (1900–1920) United States. It proceeds with an analysis of the usefulness of digitized sources at three different levels. First, using an edited collection of newspaper stories about religious leaders' alleged crimes, the study considers some of the challenges of both the digitization of the material and the reliability of the dataset. Next, zooming into a subset of preachers' crimes using sensational stories of ministerial elopements, the chapter analyzes the limits of newspapers as historical sources. Finally, examining just one of these elopement stories closely, the chapter proposes ways in which digitized newspapers can be useful starting points for research – as long as they are supplemented by careful cross-referencing and verification. In the end, digitization may exponentially enhance access to sources, but responsible scholarship will still need to rely on traditional methods of source selection, verification, and analysis – and not be too charmed by the seemingly straightforward accessibility of large digital corpora.

2 Crimes of Preachers

The 1870s were a tumultuous time for U.S. freethinkers. "Freethinker" is a term that came to loosely define people who rejected traditional religion but varied in their responses to religion's hold on American society: with some being defiantly anti-religious and others becoming involved in less conventional forms of religiosity.[1] Freethinkers, or "infidels" as they were sometimes called by their critics, faced profound persecution on account of their belief (or lack thereof). In 1873, a devout Christian moral crusader by the name of Anthony Comstock (1844–1915) convinced Congress to pass an act that would tighten federal surveillance of citizens' lives – freethinkers' lives among them. Although the so-called Comstock Act was primarily concerned with the suppression of "vice" (i.e., information and objects of sexual nature), freethinkers quickly understood that their freedoms were

1 Susan Jacoby, *Freethinkers: A History of American Secularism* (Princeton, N.J.: Holt Paperbacks, 2005), 20–21.

being threatened as well.[2] In response, they went on the offensive against the Christian architects of restrictive legislation. Pointing out hypocrisy in holy places had been the freethinkers' best weapon for some time, and they were now mobilized to deploy it strategically.

The man who took on the task of compiling the long list of ministerial misconduct was Myron Edward Billings (1837–1919), a decorated Civil War veteran, distinguished lawyer, and obsessive collector of sensational newspaper reports. Billings wrote books with titles like *Unholy Bible* and "many other liberal works."[3] His most famous publication by far was *Crimes of Preachers in the United States and Canada*, first printed in 1881 in the offices of D. M. Bennett (1818–1882), the founding editor of *The Truth Seeker* – the most influential freethought periodical of the century.

Bennett was also a prolific author of anti-religious literature. He had come out with a prototype of *Crimes of Preachers* three years earlier, with a book facetiously titled *The Champions of the Church: Their Crimes and Persecutions*.[4] The eleven-hundred-page volume was a substantive attempt at discrediting Christianity: beginning with the founder, Jesus of Nazareth himself. But Bennett also touched on the contemporaneous crisis among the clergy. The chapter "Sinful Clergymen" listed the alleged crimes of dozens of Christian ministers, all drawn from newspaper accounts.

Four years later, Billings would perfect what Bennett had begun. Billings explained that he had come across a list of ministerial crimes under the heading "Preachers' Pranks" in the *Cincinnati Commercial* and was astounded by both the volume and the variety of charges against the ministers. His astonishment soon "paled into insignificance," as he began to scan other newspapers for similar reports.[5] Having methodically compiled cases from eleven different papers (originating from the states of Massachusetts, New York, Ohio, Indiana, Illinois, Iowa, and Missouri), Billings came up with the first edition of *Crimes of Preachers*. Meticulously organized, Billings envisioned printing new editions annually – the steady supply of ministerial crimes, he believed, would provide enough material for regular updates. Billings also viewed the book as a collaborative project, asking

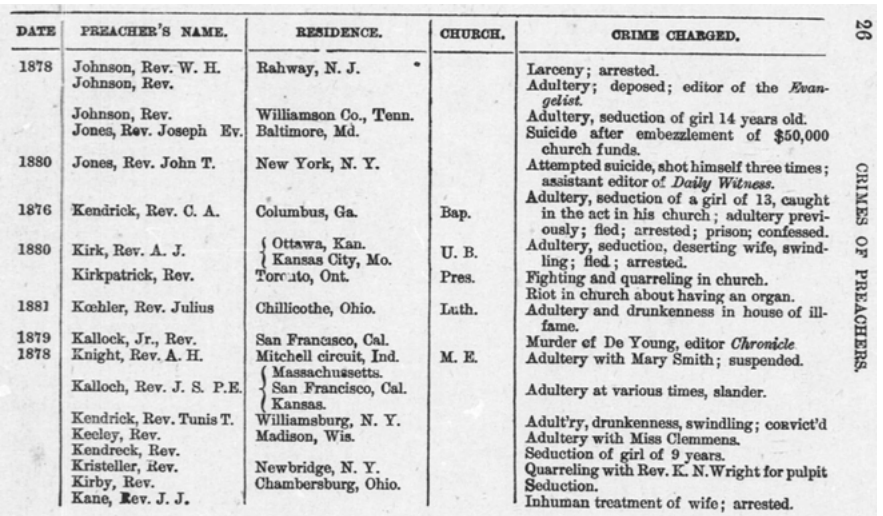
2 For examples of the long-standing connections between blasphemy laws and obscenity, see Leigh Eric Schmidt, *Village Atheists: How America's Unbelievers Made Their Way in a Godly Nation* (Princeton: Princeton University Press, 2016), 198, 312–313.

3 M. E. Billings, *Crimes of Preachers in the United States and Canada from May, 1876 to May, 1882*, 2nd ed. (New York: D. M. Bennett, 1882), title page.

4 D. M. Bennett, *The Champions of the Church: Their Crimes and Persecutions* (New York: D. M. Bennett, 1878).

5 Billings, *Crimes of Preachers*, 7.

readers to cut out newspaper clippings and forward them to his address in Waverly, Iowa. The audience responded. The second, expanded edition of *Crimes of Preachers*, published in 1882, contained more than twice the number of cases.



DATE	PREACHER'S NAME.	RESIDENCE.	CHURCH.	CRIME CHARGED.
1878	Johnson, Rev. W. H.	Rahway, N. J.		Larceny; arrested.
	Johnson, Rev.			Adultery; deposed; editor of the *Evangelist.*
	Johnson, Rev.	Williamson Co., Tenn.		Adultery, seduction of girl 14 years old.
	Jones, Rev. Joseph Ev.	Baltimore, Md.		Suicide after embezzlement of $50,000 church funds.
1880	Jones, Rev. John T.	New York, N. Y.		Attempted suicide, shot himself three times; assistant editor of *Daily Witness.*
1876	Kendrick, Rev. C. A.	Columbus, Ga.	Bap.	Adultery, seduction of a girl of 13, caught in the act in his church; adultery previously; fled; arrested; prison; confessed.
1880	Kirk, Rev. A. J.	Ottawa, Kan. / Kansas City, Mo.	U. B.	Adultery, seduction, deserting wife, swindling; fled; arrested.
	Kirkpatrick, Rev.	Toronto, Ont.	Pres.	Fighting and quarreling in church.
1881	Koehler, Rev. Julius	Chillicothe, Ohio.	Luth.	Riot in church about having an organ.
1879	Kallock, Jr., Rev.	San Francisco, Cal.		Adultery and drunkenness in house of ill-fame.
1878	Knight, Rev. A. H.	Mitchell circuit, Ind.	M. E.	Murder of De Young, editor *Chronicle*
	Kalloch, Rev. J. S. P.E.	Massachussetts. / San Francisco, Cal. / Kansas.		Adultery with Mary Smith; suspended.
	Kendrick, Rev. Tunis T.	Williamsburg, N. Y.		Adultery at various times, slander.
	Keeley, Rev.	Madison, Wis.		Adult'ry, drunkenness, swindling; convict'd
	Kendreck, Rev.			Adultery with Miss Clemmens.
	Kristeller, Rev.	Newbridge, N. Y.		Seduction of girl of 9 years.
	Kirby, Rev.	Chambersburg, Ohio.		Quarreling with Rev. K. N. Wright for pulpit
	Kane, Rev. J. J.			Seduction.
				Inhuman treatment of wife; arrested.

26 CRIMES OF PREACHERS.

Fig. 1: A page from the second edition of *Crimes of Preachers in the United States and Canada* (New York: D.M. Bennett) (1882).

For scholars of religion and sexuality, *Crimes of Preachers* at first presents itself as a treasure trove of deviance among clergy, especially given its digital availability in the public domain since 2008. But a cursory look at the source reveals some inherent problems – both methodological and technical. The tenth and final edition of *Crimes of Preachers* contains 140 pages of entries organized in horizontally printed columns as shown in Fig. 1 – the reader would have had to turn the book sideways to use it. The other 89 pages reverse the orientation, eschew columns, and list the entries in the date-name-place-denomination-accusation format.

First, the technical challenges. Digital history relies on datasets just as much as Billings relied on carefully arranged entries to produce his work. At first glance, the tabulated organization of the *Crimes of Preachers* lists appears as though it should aid the process of data extraction. This is not the case. For the first 140 pages, Optical Character Recognition (OCR) does not produce easily readable or well-organized text, given the formatting challenges of columned text perpendicularly arranged on book pages. Likewise, the quality of the scanned book renders some entries illegible. The inconsistent formatting of the remaining 89 pages adds to the problem: a number of entries are missing key data in

some of the categories, therefore rendering comma-separated values – which would have neatly separated the years, names, denominations, places, and accusations – obsolete. To successfully use the data in *Crimes of Preachers*, a mix of OCR and much more time-consuming, manual entry is essential.

Even with the data entered and processed, methodological challenges abound. Digitization and digital manipulation enable closer scrutiny of the original material and, in some cases, undermine its purported comprehensiveness. In both parts of the book, for example, key information is missing from at least 50% of the entries. Many entries list no date, place, or denomination. Of the 3,564 entries, 998 do not include dates, 395 do not include first names, 160 do not include locations, and 943 do not include the ministers' denominations. There are duplicates. Some entries provide only the last name of the alleged perpetrator, making it virtually impossible to compare the data in the book with the accusations printed in historical newspapers. Certain entries, e.g., Rev. J. S. Kalloch's record in Fig. 1, provide multiple locations for the same accused minister – making a potential geospatial representation of the data challenging. When considered together in a dataset format, the inconsistent details among the cases in *Crimes of Preachers* cast doubt on the original methodology of Billings and his informants. If Billings and his fellow freethinkers collected these stories from local newspapers and faithfully compiled them, would not each entry at the very least include the year in which the story appeared? Would not the majority of entries be richer and more consistent with the details of each alleged crime? In short: were Billings and company unreliable data collectors, or does the trouble with *Crimes of Preachers* extend much deeper – to the nineteenth-century newspaper source base itself?

3 Newspapers as Historical Sources

Newspapers are a tricky archive. In the United States, the first general interest newspapers did not appear until the 1830s. Prior to that, most newspapers were either partisan political vehicles or commercial advertisement venues. The 1833 penny press revolution made newspapers more affordable and entertaining, but the standard of objectivity never weighed too heavily on editors who embellished stories to sell copy. The rise of the Associated Press (AP) in 1846 was the first significant step toward universal reliability – or at least uniformity – in reporting. The AP gathered facts with minimal commentary and delivered them to newspapers by wire – thereby streamlining the process. Still, newspapers

continued to provide subjective commentary alongside objective reportage. Stories sold better than facts.

It was not until 1896, when Adolph S. Ochs (1858–1935) purchased the *New York Times*, that a standard of objectivity was articulated.[6] In his inaugural editorial for his newly acquired paper, Ochs wrote:

> It will be my earnest aim that The New-York Times give the news, all the news, in concise and attractive form, in language that is parliamentary in good society, and give it as early, if not earlier, than it can be learned through any other reliable medium; to give the news impartially, without fear or favor, regardless of any party, sect or interest involved; to make the columns of The New-York Times a forum for the consideration of all questions of public importance, and to that end to invite intelligent discussion from all shades of opinion.[7]

The problem of unreliability of newspapers as historical sources is perennial. The first U.S. historian to rely on newspapers as sources was John Bach McMaster (1852–1932) with his 1883 *History of the People of the United States*.[8] Historians have not stopped since, even as they have realized, regurgitated, and reframed arguments about newspapers' utility as historical sources. "While their contemporaneity is an important and valuable aspect of the news media that will recommend them to the historian," complained one critic in 1981, "not only do newspapers share all the weaknesses of human testimony [...], but they have quite a lot of others that are inherent in the medium itself."[9] The short timeline to publication, the dependence on subscriptions for profit, the partisanship, and the censorship are all legitimate reasons to doubt newspapers as the foremost authorities on historical events. Yet, as journalism historian Jerry W. Knudson put it in 1993, "history is concerned – or should be concerned – not only with what actually happened in any given time or place, but also with what people *thought* was happening, as revealed to them through the means of

6 For more on the history of U.S. journalism and the subject of objectivity in newspapers, see Frank Luther Mott, *American Journalism: A History of Newspapers in the United States through 250 Years, 1690–1940* (New York: The Macmillan Company, 1942); Michael Schudson, *Discovering The News: A Social History of American Newspapers* (New York: Basic Books, 1981); Hazel Dicken-Garcia, *Journalistic Standards in Nineteenth-Century America* (Madison: University of Wisconsin Press, 1989); Andrew Porwancher, "Objectivity's Prophet: Adolph S. Ochs and the *New York Times*, 1896–1935," *Journalism History* 36, no. 4 (October 1, 2011): 186–95.

7 Adolph S. Ochs, "Business Announcement," *New York Times*, 18 Aug. 1896, p. 4.

8 Jerry W. Knudson, "Late to the Feast: Newspapers as Historical Sources," *Perspectives on History*, 1993, https://www.historians.org/publications-and-directories/perspectives-on-history/october-1993/late-to-the-feast.

9 Joseph Baumgartner, "Newspapers as Historical Sources," *Philippine Quarterly of Culture and Society* 9, no. 3 (1981): 256–58.

mass communication, which may have conditioned their subsequent actions" (emphasis original).[10] In this sense, newspapers remain historians' central archive for studying what people thought happened in the past.

4 Tracking Gilded Age and Progressive Era Elopement Scandals

The Gilded Age and Progressive Era in the United States were defined by innovation and experimentation. This was also a time of unprecedented mobility, and many citizens seized the chance afforded to them to reinvent their lives – or to escape altogether and start new lives, elsewhere. Enterprising white men, in particular, were afforded the opportunity to travel to a new state or territory and claim a new life. A number of religious men seemed to follow this path by choosing to elope and start over. Married Protestant pastors and their (usually much younger) lovers were among the people who attempted to disappear and start a new life elsewhere in the nation that made that possibility increasingly more viable.

The history of press coverage of elopement scandals illustrates how newspapers can shed light on changing cultural preoccupations and concerns in different historical eras. Stories about ministers "mysteriously" eloping with women who were not their wives between 1870 and 1914 appear in *Crimes of Preachers* with surprising frequency – 180 cases, to be precise. Given that the cases were reportedly collected from newspaper reports, additional searches in digitized newspaper collections were performed in order to find more details about the cases – and to discover new cases, which were not listed in the book.

Along with the cases found in the freethinkers' chronicle, additional searches in digitized U.S. newspaper databases, such as *Chronicling America*, *ProQuest Historical Newspapers*, and *Newspapers.com,* reveal that altogether, at least 266 pastors abandoned their posts and families and took off in search for something better with other women between 1870 and 1914 (see Fig. 2).[11] Of the three databases used for this project, two are behind a paywall (*ProQuest Historical*

10 Knudson, "Late to the Feast: Newspapers as Historical Sources."

11 Suzanna Krivulskaya, "The Itinerant Passions of Protestant Pastors: Ministerial Elopement Scandals in the Gilded Age and Progressive Era Press," *The Journal of the Gilded Age and Progressive Era* 19, no. 1 (January 2020): 77–95. For an interactive map of the elopements, see Suzanna Krivulskaya, "Runaway Reverends," https://suzannakrivulskaya.shinyapps.io/ministerial-elopements/.

Newspapers and *Newspapers.com*), while *Chronicling America*, a joint project of the Library of Congress and the National Endowment for the Humanities, is free and open access. *Chronicling America* represents a diverse collection of newspapers which span the years 1789 to 1963 and come from forty-six states, Puerto Rico, and the District of Columbia. In the course of the project, participants from each locale were tasked with digitization of their microfilm holdings of newspapers collections that were most complete, diverse, and no longer in circulation (the latter was done in order to focus on representation and to decrease instances of duplication of some of the major newspapers available in other databases).[12] *ProQuest Historical Newspapers* is a paid subscription database which includes 80 newspaper titles and more than 100 million digitized pages, representing large regional newspapers such as *The New York Times* (1851–2016), *The St. Louis-Post Dispatch* (1874–2003) and *The Los Angeles Times* (1881–2011) in addition to some smaller regional papers, as well as international, Jewish, and Black newspaper collections.[13] *Newspapers.com* is the largest commercial database of digitized newspapers in the U.S. It contains over 620 million pages of digitized text representing more than nineteenth thousand newspaper titles from around the country, with publication dates ranging between 1690 and 2020.[14]

Across all three databases, a number of searches within a limited date range were performed for phrases like "minister eloped," "pastor eloped," "rev eloped," "reverend eloped," "pastor disappeared, "minister disappeared," "rev disappeared," "reverend disappeared" in addition to individual keywords in combination with required adjacent words or phrases. Results were manually selected and then stored in a Zotero database, which saved both newspaper metadata and the PDF of the article or of newspaper page(s) on which the story appeared.[15] All duplicates were cross-referenced and removed prior to compiling the final database of the 266 cases.[16]

12 "About Chronicling America," Library of Congress, https://chroniclingamerica.loc.gov/about/, accessed Nov. 14, 2020.

13 "ProQuest Historical Newspapers," ProQuest, https://about.proquest.com/products-services/pq-hist-news.html, accessed Nov. 14, 2020.

14 "About Newspapers.com," *Newspapers.com*, https://www.newspapers.com/about/, accessed Nov. 14, 2020.

15 "About," Zotero.org, https://www.zotero.org/about/, accessed Nov. 14, 2020; A curated version of the final database is available under the "Raw Data" tab of the digital project associated with this research. See Suzanna Krivulskaya, Matthew Sisk, and Daniel Johnson, "Ministerial Elopements," https://suzannakrivulskaya.shinyapps.io/ministerial-elopements/, accessed Feb. 24, 2020.

16 A curated version of the final database is available under the "Raw Data" tab of the digital project associated with this research. See Suzanna Krivulskaya, Matthew Sisk, and Daniel

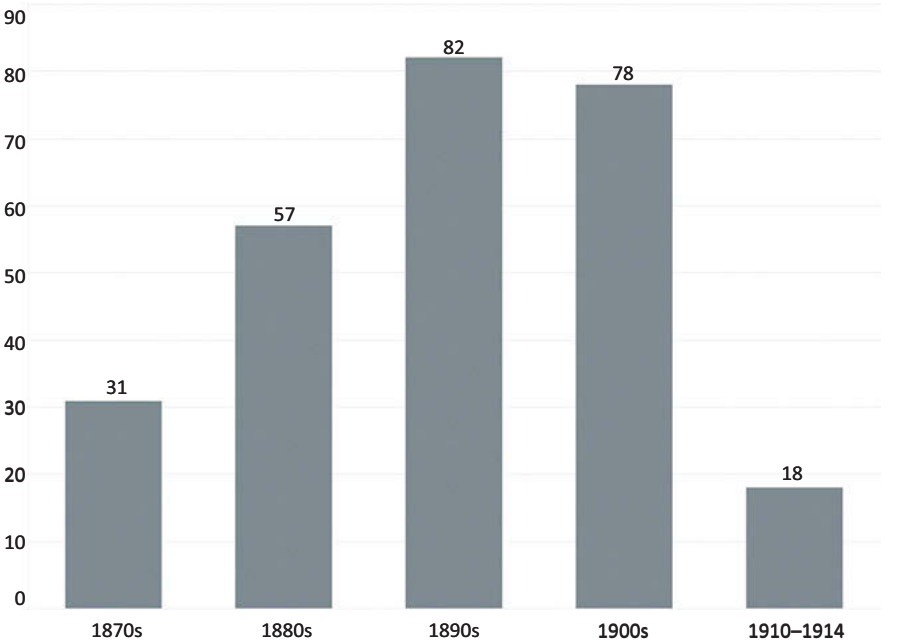


Fig. 2: Cumulative numbers of Protestant ministers' elopements by decade between 1870 and 1914. The last column's number represents the years 1910 to 1914 – and not the entire decade – to coincide with the last reissue of *Crimes of Preachers*.

The combination of the book and the newspaper articles is a serviceable starting point. Even if certain details are missing from some of the cases, their collective prominence in newspapers from this era points to a cultural preoccupation with the phenomenon of ministerial elopement. Still, just as with the larger dataset that appears in *Crimes of Preachers*, the problems with the elopement scandal subgenre found in digitized newspaper articles are multiple. They include issues of text recognition, stories being reprinted by secondary newspapers (when the original source does not survive), and the omission of key details for being able to fully track and understand every case. Consider, for example, the short article about the elopement of the Rev. John Plinket in 1887.

> An Elopement. – A Washington paper says: "A few days ago colored society was startled by the announcement that Cecelia Beckley, daughter of Rozier Beckley, a well-known Virginia politician, had eloped with Rev. John Plinket, of Virginia. The girl is a school

Johnson, "Ministerial Elopements," https://suzannakrivulskaya.shinyapps.io/ministerial-elopements/, accessed Feb. 24, 2021.

> teacher, and is very light, while the preacher is black as coal and by no means prepossessing in appearance. The color of the groom was too much for the mother, so she followed on to Fairfax county, where she captured her daughter and brought her back to Washington. It is said that when the newly-married preacher found that his unwilling mother-in-law had taken his bride home he nearly lost his religion."[17]

The story is a reprint from "a Washington paper," which, in turn, references its own ambiguous source by starting the last sentence with the words "It is said that" without identifying the source.[18] The principal actors are purportedly important people: a local politician and a minister. Yet the minister's only identifiable characteristics are being "black as coal" and "by no means prepossessing in appearance." The central conflict of the piece seems to be the darkness of the main characters' skin tone: with the minister's skin being "too much for the mother," whose daughter was "very light." The skin tone of the preacher apparently warranted a chase, and after the mother of the runaway bride-to-be followed the couple to Fairfax County, the daughter was returned to Washington.

Are readers in the early twenty-first century to believe that this elopement occurred? Or was this poorly fictionalized commentary on blackness – meant to present lighter skin color as more desirable? After all, the story takes place in the post-Reconstruction South, fraught with the smoldering vengeance of the defeated Confederacy. The last sentence, from an ambiguous source, makes a humorous observation about how the minister "nearly lost his religion" after losing his lover certainly constitutes editorializing, but how can historians ascertain with any degree of confidence whether the elopement occurred in the first place? Here again, Jerry W. Knudson's call to study what people believed happened – as opposed to what may actually have occurred – is useful. Whether the article tells historians anything about the specifics of Rev. John Plinket's individual scandal, it reveals quite a bit about white supremacy and the ways in which manufactured hierarchies of skin tone have continued to sustain it.

To address the bigger question of the usefulness and promise of digitized newspapers, Plinket's elopement story is only partially useful. It is brief, bare on details, and only the single article about the case survives (at least digitally and in the three newspaper databases consulted). It is helpful in confirming the entry in *Crimes of Preachers*, but it does not quite get at some of the other issues that digitization and historians' ability to consult multiple newspaper sources at once introduces. To examine issues with digitized sources further, let us fast-

17 "An Elopement," *Alexandria Gazette*, 21 Feb. 1887, p. 3, Chronicling America: Historic American Newspapers (Library of Congress), https://chroniclingamerica.loc.gov/lccn/sn85025007/1887-02-21/ed-1/seq-3/.

18 "An Elopement," *Alexandria Gazette*.

forward a few years and land in 1895 St. Louis, Missouri. The story of the married Presbyterian minister Rev. William J. Lee and his young lover who fled St. Louis at the end of the nineteenth century is rich with details and instructive for tracing the difficulties with using newspapers – digitized or not – as historical sources.

5 "Love Cannot Find Them"

The story begins with a misspelling. "Both Gone and Suffering Was Brought into a Benton Home," exclaimed a *St. Louis Post-Dispatch* headline on January 15, 1895 before proceeding to mix up the initials of the leading actor, "Rev. J. W. Lee, Pastor of Benton Presbyterian Church, the Cause."[19] The actual name of the alleged culprit was William J. Lee, and he was the recently resigned pastor of McCausland Avenue Presbyterian Church in St. Louis (Benton, where Lee and his family lived was a suburb of St. Louis). Lee was a well-known pastor with connections in high places. He was a respected gentleman who frequented social events and was active in the life of the city. It was also not unusual for Lee to go on overnight trips, but when he left home on New Year's Eve in 1894, his departure was both sudden and final – made all the more dramatic by the fact that Lee's rumored lover went missing as well.

The young May Ritchey (or Mae Riche, or May Ritchie, or Mae Ritchie, or Mae Richey – depending on the newspaper article) was the alleged accomplice and companion in Lee's escape. She was nineteen, a "pretty brunette," and William J. Lee's parishioner and next-door neighbor.[20] Lee was a fifty-six-year-old husband and father. To be fair, things in the Lee household had been deteriorating for some time. His wife Abbie suffered from a "nervous disorder" (whose symptoms sound like depression) and, with Lee's blessing, Mrs. Lee sought – and reportedly received – a cure from Christian Science in the late 1880s. But in the early years of the next decade, things began to unravel. The spouses constantly fought about money, and then there were rumors of Rev. Lee's affair with the young neighbor. Because of her condition, Mrs. Lee had been placed in a sanitarium in Pittsburgh, Pennsylvania, in the summer of

19 "Both Gone and Suffering Was Brought into a Benton Home," *St. Louis Post-Dispatch*, 15 Jan. 1895, p. 1.

20 The name, according to the 1880 U.S. census is, in fact, May Ritchey. For an example of a misspelling, see "Both Gone and Suffering Was Brought into a Benton Home," *St. Louis Post-Dispatch*.

1894. She was due to arrive back home in St. Louis just before Christmas, but when she got back to Missouri, "her heart failed her, and without entering the house or seeing her husband she took the first train and came" to stay with Lee's niece in a nearby town.[21] Meanwhile, Lee told his friends that he was going to California. May Ritchey informed her relatives that she was traveling to Alabama.

The first report of the disappearance – the one that confused William J. Lee's initials and misspelled May Ritchey's first name – included an extended quotation from Ritchey's mother. She had heard "the cruel gossip" about the alleged affair between the minister and her daughter but ignored it. "I am pained to confess," the mother said, "that I have come to the conclusion that his conduct has been unchristianlike [*sic*] and imprudent, and that he has cloaked his conduct from me and mine under the guise of falsehood."[22] The interview continued with a tone of profound sadness over the loss of innocence of the entire community.

Similar interviews would continue to appear in local newspapers in the months following the disappearance. Neighbors, relatives, church members, and distant relations of both Lee and Richey were asked to pitch in and provide some – any – evidence to help make sense of the case. While newspapers labeled all gossip "cruel" and insisted that no hard evidence had in fact been discovered to substantiate the allegations of elopement, they nonetheless embraced rumor as the next best thing to proof. Week after week, the story would be recapped, the pastor's conduct judged anew, and the sense of grief and disappointment recapitulated – but also savored, as the St. Louis newspapers continued to sell the story to a public apparently eager for more details of the scandalous. The mysterious disappearance of the duo would become so widely covered that reporters followed the case into the St. Louis Presbytery in order to see how the denomination would handle the case.

"A pretty girl and a gay minister of the gospel occupied the attention of the St. Louis Presbytery," reported the *St. Louis Post-Dispatch* in September of 1895, nine months after Lee's disappearance. Earlier that year, the Presbytery appointed an investigative committee, which "reported that it had been unable to secure anything but hearsay evidence against Rev. Mr. Lee and asked for further time." Lee was then issued an order to appear before the committee to answer to the charges of elopement. When, after ten days, Lee did not respond, he was given a second citation, which also produced no results. After all, the

21 "Rev. Lee's Wife: Visited Benton Heavily Veiled and Did Not See Her Husband," *St. Louis Post-Dispatch*, 17 Jan. 1895, p. 8.

22 "Both Gone and Suffering Was Brought into a Benton Home," *St. Louis Post-Dispatch*.

Fig. 3: Illustration and headline of the *St. Louis Post-Dispatch* story on Lee and Ritchey from August 1, 1897.

address to which the citations were being sent (Lee's Benton home) was the one place that was known not to contain the accused.[23]

By October, with Lee nowhere to be found, the St. Louis Presbytery was holding the absentee pastor's trial. Judge Selden P. Spencer was appointed by the Presbytery to defend Lee and to cross-examine the generally hostile witnesses. One witness, Lee's daughter Grace, described the frequent meetings between Lee and Ritchey that she had seen and reported having "charged her father with the improprieties which rumor had long urged against him."[24] Finally, on October 24, Lee was suspended from ministry in absentia after being found guilty of "undue familiarity with a female member of his flock."[25]

While the church reached a resolution, the newspapers did not. In August of 1897, two and a half years after the disappearance, the *St. Louis Post-Dispatch* dedicated the better portion of one of the pages of the Sunday edition to the story. "Love Cannot Find Them," the headline exclaimed, "Remarkable Disappearance of Rev. W. J. Lee and Miss Mae Richey: Has Death Hidden Them or Are They Concealed in the Multitude of a Great City?" (see Fig. 3).[26] Drawings of the alleged lovers adorned the two top corners of the page. An illustration of Lee's church occupied the middle. The lovers were nowhere to be found, the article concluded, but that did not mean that their story did not continue to produce interest and drive sales.

Seventeen different newspapers from Missouri, Kansas, Illinois, and Texas reprinted portions of the story with illustrations and a subheading designed to draw readers in: "Mysterious Disappearance of a Missouri Couple: Friends and

23 "Rev. W. J. Lee's Case: It Occupies the Attention of the St. Louis Presbytery," *St. Louis Post-Dispatch*, 19 Sep. 1895, p.6.

24 "Church Trial of Rev. Lee: His Daughter Tells of His Elopement with Mae Ritchie," *St. Louis Post-Dispatch*, 28 Oct. 1895, p. 3.

25 "Rev. W. J. Lee Found Guilty," *St. Louis Post-Dispatch*, 29 Oct. 1895, p. 3.

26 "Love Cannot Find Them," *St. Louis Post-Dispatch*, 1 Aug. 1897, p. 23.

Detectives Have Scoured the Country, but Whereabouts of the Couple Remain a Deep mystery – A Rural Church Scandal."[27] The story was making the rounds across the country because it received so much publicity in St. Louis, and while the details were scant, they were titillating enough to inspire incessant coverage.

The frustrating part of this story is that in spite of how many details the newspapers that covered it got wrong, they were right on one count: no one knows exactly what happened to Rev. William J. Lee and May Ritchey. The only piece of frustratingly unsatisfying evidence that can be verified is that by 1900, Lee's deserted wife Abbie had moved on. She and her daughter Grace moved uptown: to the Fountain Park area, about four and a half miles north-east of the Benton home they had shared with the runaway pastor Lee. In the 1900 U.S. census and then again in 1910 and 1920, Abbie's marital status is listed as "widowed." Abbie Foster King Lee died on August 22, 1924 at the age of eighty. Her daughter Grace never married and lived with her mother at least through 1920. Grace King Lee died in St. Louis in 1945. The Lee family tree ended there – unless, that is, Rev. Lee and May Ritchey successfully escaped their old lives and started new ones elsewhere, away from the scandal of St. Louis – or the historians' gaze.

6 Conclusion

Digitization has profoundly transformed the field of historical inquiry. Research that had been possible only by manually thumbing through fragile paper can now be done from the comfort of one's home – thousands of miles from the historical events in question or even their physical traces in carefully organized, categorized, and labeled archival iterations. But while the logistical problems with access have been overcome, the existential and theoretical issues of the archive – in digital or physical form – persist. As historian Carolyn Steedman puts it, the archive "cannot help with what is not actually there, with the dead who are not really present in the whispering galleries, with the past that does not, in fact live in the record office, but is rather, *gone* (that is its point; that is what the past is for); it cannot help with parchment that does not in fact speak" (emphasis original).[28] In the end, no degree of technological advancement can help salvage the quintessentially fragmentary nature of surviving evidence about the past.

27 See, for instance, "No Tidings of Them," *The Arlington Enterprise* (Arlington, Kansas), 10 Sep. 1897, p. 7.

28 Carolyn Steedman, *Dust: The Archive and Cultural History* (New Brunswick: Rutgers University Press, 2002), 81.

The same principle of fragmentary evidence applies to newspapers as sources of historical data – digitized or not. Optical character recognition makes it possible to track the Rev. William J. Lee story, but, unless the researcher knows to look for them, also obscures instances in which Lee's and May Ritchey's names were misspelled in dozens of articles – ones that would not come up in Boolean searches using the correct spellings. Nor will digitization correct the factual errors and embellished details of sensational stories; it will still require the trained eye of a trained professional to distinguish – or, at the very least, suggest – what details are likely to be true and what was spiced up for better newspaper sales. Aside from the ability to analyze large amounts of textual data, the gift that newspaper digitization has given to the historical profession lies, paradoxically, in its methodological shortcomings: digitization has revealed just how troublesome newspapers have always been as historical sources. Still, combined with a systematic approach of gathering and evaluating data – including a careful selection of keyword searches, addition of homonyms and spelling variations in advanced searches, OCR corrections, and cross-referencing with other sources – newspaper digitization has opened up a new world of possibilities for historical research. Even in the digital age, however, good historical scholarship will need to continue to rely on a combination of digital tools and traditional methods of investigation, validation, and interpretation.

Bibliography

"An Elopement." *Alexandria Gazette*. 21 Feb. 1887, https://chroniclingamerica.loc.gov/lccn/sn85025007/1887-02-21/ed-1/seq-3/.

Baumgartner, Joseph. "Newspapers as Historical Sources." *Philippine Quarterly of Culture and Society* 9, no. 3 (1981): 256–58.

Bennett, D. M. *The Champions of the Church: Their Crimes and Persecutions*. New York: D. M. Bennett, 1878.

Billings, M. E. *Crimes of Preachers in the United* States *and* Canada *from May, 1876 to May, 1882*, 2nd ed. New York: D. M. Bennett, 1882.

"Both Gone and Suffering Was Brought into a Benton Home." *St. Louis Post-Dispatch*. 15 Jan. 1895.

"Church Trial of Rev. Lee: His Daughter Tells of His Elopement with Mae Ritchie." *St. Louis Post-Dispatch*. 28 Oct. 1895.

Dicken-Garcia, Hazel. *Journalistic Standards in Nineteenth-Century America*. Madison: University of Wisconsin Press, 1989.

Jacoby, Susan. *Freethinkers: A History of American Secularism*. Princeton, N.J.: Holt Paperbacks, 2005.

Knudson, Jerry W. "Late to the Feast: Newspapers as Historical Sources." *Perspectives on History*. 1993. https://www.historians.org/publications-and-directories/perspectives-on-history/october-1993/late-to-the-feast.

Krivulskaya, Suzanna. "The Itinerant Passions of Protestant Pastors: Ministerial Elopement Scandals in the Gilded Age and Progressive Era Press." *The Journal of the Gilded Age and Progressive Era* 19, no. 1 (January 2020): 77–95.

"Love Cannot Find Them." *St. Louis Post-Dispatch*. 1 Aug. 1897a.

Mott, Frank Luther. *American Journalism: A History of Newspapers in the United States through 250 Years, 1690–1940*. New York: The Macmillan Company, 1942.

"No Tidings of Them." *The Arlington Enterprise* (Arlington, Kansas). 10 Sep. 1897.

Ochs, Adolph S. "Business Announcement." *New York Times*. 18 Aug. 1896.

Porwancher, Andrew. "Objectivity's Prophet: Adolph S. Ochs and the New York Times, 1896–1935." *Journalism History* 36, no. 4 (October 1, 2011): 186–95.

"Rev. W. J. Lee's Case: It Occupies the Attention of the St. Louis Presbytery." *St. Louis Post-Dispatch*. 19 Sep. 1895.

"Rev. W. J. Lee Found Guilty." *St. Louis Post-Dispatch*. 29 Oct. 1895.

"Rev. Lee's Wife: Visited Benton Heavily Veiled and Did Not See Her Husband." *St. Louis Post-Dispatch*. 17 Jan. 1895.

Schmidt, Leigh Eric. *Village Atheists: How America's Unbelievers Made Their Way in a Godly Nation*. Princeton: Princeton University Press, 2016.

Schudson, Michael. *Discovering The News: A Social History of American Newspapers*. New York: Basic Books, 1981.

Steedman, Carolyn. *Dust: The Archive and Cultural History*. New Brunswick: Rutgers University Press, 2002.

Tobias von Waldkirch

Korrespondentenberichte im *Journal de Genève* und ihre sprachlichen Muster

Das Beispiel der Kriegsberichterstattung im 19. Jahrhundert

Abstract: Correspondence reports in letter form represent an important genre for newspapers in the 19th century. Especially in times of war, international correspondent reports become important. In a diachronic comparison, changes in the textual structure of correspondent reports can be examined to see to what extent they are an indication of a change in journalistic culture.

The aim of this article is to analyse the correspondents' reports (French: *correspondances particulières*) during the Crimean War (CW: 1853–56) and the Franco-Prussian War (FPW: 1870–71). The focus is on the Swiss newspaper *Journal de Genève* (JdG). For both wars, there are historical factors that remain the same: Switzerland is uninvolved both times (and does not censor the press), the JdG sees itself as an institution of neutral observation and independent judgement during both wars. Against the background of these stable elements, the change in reporting per se is to be analysed.

Methodologically, the analysis is based on genre studies and corpus linguistics. Thanks to the Impresso platform, targeted article searches for the correspondents' reports published during the two wars are possible. In the following qualitative analysis, the reports are compared (in a holistic perspective) for features of their textual structure. This raises the question of the extent to which these features are shaped differently in each of the two wars. For example, the occurrence of the first-person singular (French: *je*), which refers to the reporting correspondent, is central in both wars. However, the use of *je* in the Crimean War appears to be partly different from the one in the Franco-Prussian War.

A corpus-based comparison of the occurrences of *je* over the two wars shows that the most frequent verb tenses that occur with *je* are the *présent* and the *passé composé* (CW: 44.1% / 32.7%; FPW: 41.9% / 30.1%, cf. Tab. 2 below). Thus, the perspectivation with the help of *je* appears to be related to a current event in both wars. Differences become apparent when comparing which occurrences of the first-person singular refer to the speaker and which ones are part of a quotation: In the Crimean War, most occurrences refer to the speaker; only 15.5% of the occurrences are *je*-forms from quotations. In contrast, 30.2% of the *je*-forms in the Franco-Prussian War are *je* of quotations (cf. bar graph 1 and 2 below).

A closer look shows that the frequent occurrence of *je* in quotations in the Franco-Prussian War is related to a paradigm shift in reporting. The printing of

https://doi.org/10.1515/9783110729214-018

speeches and interviews in correspondent reports is a new phenomenon; the interview represents a genre not yet present in the Crimean War. It is mainly high-ranking military officers who have their say. This change is connected to a new conception of reporting in the *Journal de Genève*: the editorial staff want to present the textual sources connected with an event as transparently as possible "*sous les yeux de ses lecteurs*", i.e., in front of the eyes of its readers (according to a programmatic editorial in October 1870). The interview has a special position in this context because it allows the correspondent to have a certain influence on its content by means of his choice of questions; at the same time, the way the correspondent obtains information can be presented as a transparent process.

The article also shows that a combination of qualitative and quantitative approaches proves fruitful; results from qualitative analyses can form the starting point for a quantitative analysis – and vice versa.

Keywords: digitised newspapers, media history, historical linguistics

1 Einleitung

Korrespondentenberichte in Briefform (französisch: *correspondances particulières*) stellen im 19. Jahrhundert eine zentrale Textsorte für Zeitungen dar (vgl. Vaillant 2014, Wauters 2012). Dabei spielen internationale Korrespondentenberichte eine zentrale Rolle: Sie informieren die Leserschaft über das internationale Geschehen und stellen so gleichsam ein „Fenster zur Welt" (Hillerich 2018: 9) dar. Gerade während Kriegen ist das Interesse an diesem Fenster besonders hoch und die damit verbundene Berichterstattung gewinnt an Bedeutung – was wiederum Veränderungen in der Kultur des Berichterstattens begünstigt und beschleunigt[1]. Im diachronen Vergleich lassen sich dabei textliche Strukturen von Korrespondentenberichten und deren Veränderungen daraufhin befragen, ob und inwiefern sie Ausdruck eines Wandels der journalistischen Darstellung von Welt – und somit eines Wandels journalistischer Kultur – darstellen[2].

Im Folgenden wird es darum gehen, dem Wandel in der Berichterstattung zu Kriegszeiten nachzuspüren, mit Fokus auf zwei bestimmte Konflikte, nämlich den Krimkrieg (1853–1856) und den Deutsch-Französischen Krieg (1870/1871). Gegenstand der Analyse sind die zu Kriegszeiten erscheinenden internationalen

1 Vgl. aus historischer Perspektive Baumgart (2010) zum Krimkrieg und Becker (2006) zum Deutsch-Französischen Krieg.
2 Vgl. hierzu Hanitzsch (2007).

Korrespondentenberichte des *Journal de Genève* (JdG)[3]. Diese Zeitung ist deshalb von besonderem Interesse, weil einige zentrale Faktoren während beider Konflikte identisch sind: So ist zum einen die Schweiz an keinem der beiden Kriege aktiv beteiligt und zensiert auch keine Presse-Erzeugnisse; zum anderen versteht sich das *Journal* sowohl in den 1850er- als auch zu Beginn der 1870er-Jahre in Bezug auf Auslandsnachrichten als eine Instanz der neutralen Beobachtung und der unabhängigen Beurteilung, was ihm in beiden Kriegen zu großer Aufmerksamkeit aus dem umliegenden Ausland gereicht[4]. Ferner bildet der hier gewählte Fokus auf Kriegsberichterstattung einen makrothematischen Hintergrund. Diese für beide Kriege stabil bleibenden Rahmenbedingungen bilden eine adäquate Voraussetzung, Wandel in der Berichterstattung zu betrachten und in Zusammenhang zu bringen mit dem sich stetig verändernden Pressewesen des 19. Jahrhunderts[5]. Methodologisch bilden Ansätze aus der Textsortenlinguistik und korpuslinguistische Zugänge die beiden Ausgangspunkte der Analyse, was in einem ersten Kapitel erläutert wird. Dabei gilt es auch aufzuzeigen, auf welche Weise die Impresso-Plattform eine solche Analyse überhaupt erst ermöglicht. Ergänzt wird das erste Kapitel mit einem Überblick über das *Journal de Genève* in der zweiten Hälfte des 19. Jahrhunderts. Das zweite Kapitel zeigt die Analysen der Korrespondentenberichte auf, mit einem Fokus auf die Stellung der Sprechinstanz *je* (1. Person Singular) und der Adressierung an das Lesepublikum durch *vous* (2. Person Plural) und der dementsprechend geschaffenen Kommunikationssituation. Dabei zeigt sich, dass insbesondere die Sprechinstanz *je* für beide zeitlichen Segmente ein Kernelement der textlichen Struktur darstellt und jeweils unterschiedlich ausgeprägt ist[6].

2 Methode(n)

2.1 Theoretischer Hintergrund und methodisches Vorgehen

Die Analysen erfolgen aus einer kulturlinguistischen Perspektive auf Textsorten[7]. Zu den möglichen Analysemethoden gehören – nebst schon länger etablierten

3 Vgl. https://impresso-project.ch/app/newspapers/JDG/metadata?sq= (10.10.2020).
4 Vgl. de Senarclens (1999, 30f. und 36f.), Chapuisat (1999 [1929], 55).
5 Vgl. für die Presse der französischsprachigen Schweiz Clavien (2017, 25–96).
6 Die nachfolgenden Analysen sind nicht als abgeschlossen zu werten, sondern stellen den gegenwärtigen Wissensstand im Rahmen meiner Dissertation zu journalistischen Kulturen im 19. Jahrhundert mit Fokus auf das JdG und die *Neue Zürcher Zeitung* dar.
7 Vgl. Linke (2003); Devitt (2004); Luginbühl (2014); Tardy/Swales (2014); Tienken (2015).

Methoden, die als 'klassische manuelle Analyse' apostrophiert werden können[8] – auch korpuslinguistische Zugänge[9]. „At their most basic level", so Tardy und Swales, seien Textsorten „formed in order to carry out actions and purposes" (Tardy/Swales 2014, 166). Entsprechend stellen Textsorten „kulturelle Artefakte" (Luginbühl 2014, 40) dar, die in einer bestimmten Kultur eingebettet sind und in einer „reciprocial relationship" (Devitt 2004, 27) Kultur sowohl mitkonstruieren als auch reproduzieren[10]. Wenn wir – ebenfalls 'at a most basic level' – Kultur als „a shared set of material contexts and learned behaviors, values, beliefs, and templates" (Devitt 2004, 25) betrachten, können Textsorten daraufhin befragt werden, welche Funktionen und Bedeutungen ihnen in einem bestimmten Handlungszusammenhang zukommen.

Texte, deren Lebenswelt vergangen ist, erscheinen uns als „aus ihrer ursprünglichen Situationsverankerung gelöst" (Adamzik 2016, 191); dasselbe trifft eo ipso auch auf Textsorten zu. Im Gegensatz zu uns bekannten Textsorten, deren Produktion und Rezeption weitgehend automatisiert ablaufen, ist es für historische Textsorten als „kulturell und historisch geprägte Phänomene" (Luginbühl 2014, 41) unerlässlich, dass man sich ihnen in einer Art Rekonstruktionsprozess mittels „Ergänzungs- und Schlussmechanismen" (Linke 2003, 46) annähert. Wichtig ist in dem Zusammenhang auch, wie Gruppen von Texten durch die Mitglieder einer Sprachgemeinschaft selbst benannt werden, denn „die Differenzierung von Texten [entspricht] einem ausgeprägten alltäglichen Bedürfnis" (Adamzik 2016, 327). Solche ethnokategorischen Bezeichnungen entbehren natürlich einer Systematik, geben aber oft Hinweise auf den Verwendungskontext bestimmter Arten von Texten. So wird in der Bezeichnung *correspondance particulière* nicht nur der postalische Übermittlungsweg expliziert, sondern durch das Attribut *particulière* auch der Umstand, dass die entsprechende Korrespondenz von einer Privatperson stammt. Hier wird sichtbar, dass sich allfällige Ansätze einer allmählich einsetzenden Professionalisierung des Korrespondentenwesens ethnokategorisch (noch) nicht niedergeschlagen haben[11].

Um Textsorten gleichsam aus ihrer Zeit heraus zu verstehen, bietet sich eine kontrastive Perspektive an; vorausgesetzt, dass eine kursorische diachrone Sichtung der fraglichen Texte nahelegt, dass diese als zusammengehörig betrachtet werden dürfen[12]. Ausschlaggebend dafür ist im JdG der Umstand, dass die Korre-

8 Vgl. die Überblicksdarstellung zu 'klassischen' Methoden in Tardy/Swales (2014, 168–175).
9 Vgl. Bubenhofer (2009) und Brommer (2018).
10 Vgl. Luginbühl (2014, 60f.).
11 Vgl. zur Professionalisierung im Journalismus im 19. Jahrhundert ausführlicher Van den Dungen (2008: Abschnitt 3 [ohne Seitenzahlen]).
12 Vgl. hierzu Luginbühl (2014, 92–98) sowie Tienken (2015, 471).

spondentenberichte sowohl während der 1850er- als auch der 1870er-Jahre jeweils mit *correspondance particulière*[13] betitelt und typographisch von anderen Texten abgehoben werden[14]. Die erwähnte kontrastive Perspektive lässt sich nun dergestalt etablieren, dass die fragliche Textsorte über zwei verschiedene Zeitabschnitte hin verglichen wird, um Merkmale zu eruieren, die in beiden zeitlichen Segmenten von Bedeutung sind. In den nachfolgenden Analysen werden der Krimkrieg und der Deutsch-Französische Krieg jeweils als ein Zeitsegment betrachtet. Als wichtiges Merkmal der Textstruktur erweist sich für diese beiden Zeitabschnitte beispielsweise die Kommunikationssituation mit einer Sprechinstanz in der ersten Person Singular, die sich direkt an ein Publikum wendet. Die im Vergleich über die beiden Zeitsegmente eruierten Merkmale bilden dann die Grundlage einer kontrastiven Detail-Analyse.

Ein solcherart angelegtes Vorgehen hat den Vorteil, dass sowohl Gemeinsamkeiten (d. h. das Vorhandensein der untersuchten Merkmale, so etwa die über *je* und *vous* etablierte Kommunikationssituation) als auch Unterschiede (d. h. deren spezifische Ausgestaltung pro Zeitabschnitt) Beachtung erfahren. Die dabei sichtbar gewordenen Unterschiede lassen sich im Spannungsfeld von Kontinuität und Wandel der Textsorte daraufhin befragen, inwiefern sie Ausdruck kultureller Ordnungen und allenfalls kulturellen Wandels sind[15] – im Falle von Zeitungstextsorten bezieht sich dies insbesondere auch auf journalistische Kultur[16]. Ziel dieses Vorgehens insgesamt ist es, der Voreingenommenheit der „historisch entfernten Rezipientenperspektive“ (Linke 2003, 46) soweit wie möglich entgegenzuwirken, denn die Merkmale werden auf eine solche Weise „induktiv und materialgeleitet“ (Adamzik 2016, 289) *am* Korpus herausgearbeitet und nicht *an* das Korpus herangetragen.

Es bietet sich an, dieses Vorgehen mittels korpuslinguistischer Methoden (vgl. unten 2.2) zu ergänzen. Auf diese Weise lässt sich beurteilen, inwiefern ‘manuelle’, qualitativ gemachte Beobachtungen über bestimmte Merkmale, verstanden als „Phänomene auf der Textoberfläche“ (Bubenhofer 2009, 52) auch aus einer quantitativen Perspektive als auffällig erscheinen. So soll der Frage nachgegangen werden, inwiefern bestimmte Merkmale, beispielsweise das Vorkommen der 1. Person Singular *je*, im Vergleich betrachtet für das ein oder andere Zeitsegment als musterhaft gelten können[17]. Solchermaßen sichtbar gewordene „Sprachgebrauchsmuster“ (Bubenhofer 2009, 42) erfordern wiederum eine qualitative Interpretation

13 Oder einer damals gängigen Abkürzung wie z. B. *corr. particulière*, vgl. unten 2.2.

14 Textanfang und -ende werden mittels feiner Querlinien markiert.

15 Vgl. Linke (2003, 47).

16 Vgl. ausführlich Hanitzsch (2007).

17 Vgl. auch Brommer (2018, 51–56).

und Klassifikation, bevor der Versuch unternommen wird, sie in einem größeren Kontext, d. h. als (möglichen) Ausdruck eines Wandels journalistischer Kultur zu bewerten. Untersucht man beispielsweise den Gebrauch der 1. Person Singular *je* mithilfe einer entsprechenden Lemmata-Suche über beide Zeitsegmente hinweg, so ist es unerlässlich, sämtliche entsprechende Okkurrenzen daraufhin zu überprüfen, ob sie der Sprechinstanz attribuiert werden können oder Teil eines Zitates sind (vgl. 3.2).

2.2 Textkorpus und korpuslinguistische Werkzeuge

Die Korrespondentenberichte, die für das beschriebene methodische Vorgehen von Interesse sind, lassen sich gezielt auf der Impresso-Plattform mit der Stichwortsuchfunktion ermitteln, da wie erwähnt in den fraglichen Zeitsegmenten die Korrespondentenberichte jeweils explizit mit *correspondance particulière* betitelt worden sind. Die so ermittelten Texte werden in einer Sammlung (engl. *collection*) zusammengefasst, wobei die Stichwortsuche, beschränkt auf die Artikeltitel, auch sämtliche damals gängigen Abkürzungen, die manchmal anstelle des ausgeschriebenen Begriffs Verwendung gefunden haben, umfassen muss, konkret: *corresp. partic.*; *corresp. part.*; *corr. particulière*. Die zwei zeitlichen Segmente, aus welchen sämtliche internationalen Korrespondentenberichte berücksichtigt worden sind, umfassen diejenigen Zeiträume zwischen den jeweils als Anfangs- bzw. Endpunkt der Kriege geltenden Daten[18]. Die beiden so erstellten Sammlungen lassen sich als CSV-Datei exportieren und nach Belieben weiterverwenden. Für den korpuslinguistischen Teil der Analyse wurden die Texte mit TreeTagger annotiert, d. h. für alle vorkommenden Wortformen die Informationen über Wortart und Grundform verfügbar gemacht. Anschließend erfolgte eine Bearbeitung der Daten durch das Programm Corpus Workbench, was gezielte Recherchen über die Plattform CQP-Web ermöglicht[19]. Genutzt wurden insbesondere die Lemmata-Suche sowie das Kollokationstool (vgl. 3.2). Das Teilkorpus 1 (TK1) für die Zeit des Krimkrieges umfasst dabei 768.488 Tokens; das Teilkorpus 2 (TK2) für den Deutsch-Französischen Krieg deren 491.149.

18 Krimkrieg: 4. Oktober 1853 (Kriegserklärung des Osmanischen Reiches) bis 30. März 1856 (Frieden von Paris), vgl. Baumgart (2010); Deutsch-Französischer Krieg: 19. Juli 1870 (Kriegserklärung des zweiten Kaiserreiches) bis 29. Januar 1871 (Waffenstillstand), vgl. Wawro (2009).

19 Die Annotierung und der Import auf CQP-Web erfolgten durch Klaus Rothenhäusler (Zürcher Hochschule für Angewandte Wissenschaften) und Noah Bubenhofer (Universität Zürich), denen ich meinen herzlichen Dank für ihre Hilfe aussprechen möchte. Vgl. zu den erwähnten Programmen vertiefend Bubenhofer (2006–2022).

2.3 Das *Journal de Genève* in den 1850er- bis 1870er-Jahren

Von einer Gruppe liberaler Politiker 1826 gegründet, ist das *Journal de Genève* seit 1850 eine Tageszeitung (mit sechs Ausgaben pro Woche); 1998 wird es mit der Lausanner Zeitung *Le Nouveau Quotidien* zum heute in Genf ansässigen Blatt *Le Temps* fusioniert. Die Zeit der 1850er- bis zum Beginn der 1870er-Jahre, in die der Krimkrieg und der Deutsch-Französische Krieg fallen, ist eine Zeit der Expansion und des Ausbaus für das JdG[20]. Dies geht einher mit mehreren Formatvergrößerungen zwischen 1856 und 1870; binnen knapp 15 Jahren wächst das Format um mehr als das Vierfache an[21]. Die Anzahl der Seiten pro Ausgabe bleibt über diesen Zeitraum stabil: drei Seiten für den redaktionellen Teil, die vierte und letzte enthält vorwiegend Anzeigen. Auch das Rubrikenrepertoire als Makrostruktur der Zeitung bleibt weitgehend identisch: Die erste Seite enthält unter dem Rubrikennamen *Confédération Suisse* die Inlandberichterstattung, gefolgt von der Auslandsberichterstattung unter *Étranger*, in welchen sich nach Herkunftsländern geordnet unter anderem die Korrespondentenberichte (*correspondances particulières*) finden[22]. Dem schließen sich die Rubrik *Faits divers* an sowie am Ende der dritten Seite die *Dépêches télégraphiques*; in der Zeit um 1870 zudem die Leserbriefe, *correspondances*. Das Feuilleton, von den ersten beiden Seiten jeweils das letzte Viertel beanspruchend, erscheint zu Kriegszeiten nur sehr sporadisch, denn oft muss – „en raison des événements qui mangeaient toute la place“ (Chapuisat 1999 [1929], 55)[23] – darauf verzichtet werden. Innerhalb dieser relativ stabil bleibenden Rubriken-Makrostruktur lassen sich vielgestaltige Phänomene des Wandels in der Berichterstattung ausmachen, wie sich dies nachfolgend am Beispiel der Korrespondentenberichte zu Kriegszeiten präsentiert.

20 Vgl. für einen historischen Überblick über das JdG: de Senarclens (1999, 17–39).

21 Vgl. für die Jahrzahlen der Formatvergrößerungen Chapuisat (1999 [1929], 50). Erscheint das JdG bis Ende 1855 noch in einem dreispaltigen Quartformat von ca. 20 x 25 cm, so wird es in den kommenden 15 Jahren viermal vergrößert und umfasst ab 1870 die Maße von ca. 41 x 53.5 cm bei einem sechsspaltigen Layout. Das Format wächst also um den Faktor 4.387.

22 Vgl. zu Zeitungsrubriken im 19. Jahrhundert von Waldkirch (2021).

23 Dt.: »wegen Ereignissen, die den gesamten Platz beanspruchen« (sämtliche deutschen Übersetzungen in Fußnoten: TvW).

3 Analyseresultate

3.1 Merkmale der *correspondances particulières*

Zunächst erfolgt eine Übersicht über qualitativ etablierte Merkmale, die sich für beide Zeitsegmente finden lassen. Anschließend werden deren zwei mit dem korpusbasierten Zugang vertieft: die Ausgestaltung der Kommunikationssituation anhand einer detaillierten Analyse der Pronomina-Okkurrenzen von *je* und *vous* sowie das (unterschiedliche) Vorkommen von *je*.

Für beide Zeitsegmente zeigt sich, dass die Textlänge der Korrespondentenberichte erheblich variiert, was in den 1850er-Jahren im kleinen Quartformat immer wieder dazu führt, dass eine Korrespondenz – oft ist es diejenige aus Paris[24] – bis zu einem Drittel des gesamten redaktionellen Teils einnimmt. Nebst Paris erscheinen regelmäßig Korrespondenzen aus London und Berlin, sporadisch auch aus Neapel und Turin; wichtigen Städten in der Entstehung des Königreichs Italien. Für die Zeit um 1870/1871 kommen zu den genannten Korrespondenzorten Brüssel, Rom, Wien und New York dazu, ferner Orte jenseits der Großstädte, in welchen während des Krieges Verhandlungen oder Vertragsunterzeichnungen stattfinden, so etwa Chaumont, Freiburg im Breisgau, St. Germain-en-Laye, Gersweiler (Saarbrücken), Lunéville und Versailles.

Sämtliche Korrespondenzen sind durch die Sprechinstanz *je* perspektivisch markiert und der Sprecher wendet sich mittels der zweiten Person Plural *vous* regelmäßig an ein Publikum. Dabei zeigt sich während beider Zeitsegmente, dass dort, wo ein Adressat mit *vous* expliziert wird, nicht zwingenderweise die Leserschaft gemeint ist, sondern die Redaktion, was in Formulierungen wie „je prie vos lecteurs de ne pas confondre [...]“[25] sichtbar wird. Der Leserschaft, die auf diese Weise – wenn überhaupt – nur indirekt erwähnt wird, kommt die Rolle eines nur implizit adressierten Publikums zu. Damit unterscheidet sich die Auslandsrubrik *Étranger*, wie die gesamte politische Berichterstattung überhaupt, von anderen Rubriken wie dem Feuilleton oder den *Faits divers*, in welchen im JdG,

24 de Senarclens (1999) weist für Paris in den 1850er-Jahren nur einen einzigen Korrespondenten aus (vgl. *ibid.*, S. 31); ein Blick in Zeitungsausgaben der 1850er-Jahren legt aber nahe, dass es in Paris mehrere Korrespondenten gegeben haben muss, wenn die Redaktion von »nos correspondants de Paris« spricht (vgl. z. B. im JdG vom 15.12.1854, 2). Da die Verwendung der 1. Person Singular allerdings in Korrespondenzen aus allen Städten gleichermaßen vorkommt, ist sie kaum als individuelles stilistisches Merkmal zu werten, sondern als damals üblicher Standard der Perspektivierung. Entsprechend ist die Autorenproblematik für die hiesige Fragestellung nur von geringem Interesse.

25 JdG vom 20.11.1855, dt.: „Ich bitte Ihre Leser, nicht... zu verwechseln“.

wie für französische Zeitungen jener Zeit üblich, ein „style conversationnel“ (Vaillant 2014, 2) vorherrscht, der sich in einer expliziten Adressierung an die Leserschaft manifestiert. Die Kommunikationssituation der Korrespondentenberichte lässt sich somit als Dialog zwischen dem schreibenden Korrespondenten und der Redaktion vor Publikum begreifen – und die Möglichkeiten einer solchermaßen konstruierten Kommunikationssituation werden in beiden Zeitsegmenten auf vielfältige Weise genutzt: Es finden sich metatextuelle Hinweise, zum Beispiel auf den Schreibprozess („au moment où je vous écris“[26]), Direktiven („je vous renvoie du reste à cet article“[27]), Bewertungen der Nachrichtenlage („je ne puis cependant vous garantir encore les détails“[28]) sowie Selbstinszenierungen („en correspondant impartial, je dois vous dire [...]“[29]). Die Briefform in der ersten Person bildet gleichsam den Rahmen, innerhalb dessen der Korrespondent als Sprechinstanz versucht, mit verschiedenen sprachlichen Mitteln das Gesagte als ‘erlebte Wirklichkeit’ darzustellen. In dieser so konstruierten Kommunikationssituation wird die Sprechinstanz *je* perspektivisch zum zentralen Moment der Informationsvermittlung. Dabei geschieht dies während der beiden Kriege teilweise unterschiedlich, wie der korpusbasierte Zugang nachfolgend aufzeigt.

3.2 Sprachliche Ausgestaltung der *correspondances particulières*: Krimkrieg und Deutsch-Französischer Krieg im Vergleich

Diese dialogisch angelegte Kommunikationssituation, etabliert durch die Personalpronomen *je* und *vous*, bildet nachfolgend den Ausgangspunkt für eine korpusbasierte Analyse. Pro Teilkorpus werden hierfür zwei Foci gesetzt: 1) das Pronomen *vous* als Kollokator von *je* und 2) das Pronomen *je* per se. Die Resultate werden anschließend qualitativ interpretiert und in einem größeren Kontext betrachtet.

Zunächst zu grammatischen Konstruktionen des Typs *je* + *vous* + [Verb]. Für beide Teilkorpora erscheint *vous* in der Kollokationsliste für den Suchbegriff *(je|j’)*[30] an erster Stelle[31], wobei Okkurrenzen aus Zitaten anschließend ma-

26 JdG vom 29.12.1870, dt.: „Jetzt, da ich Ihnen schreibe“.
27 JdG vom 18.01.1855, dt.: „im Übrigen verweise ich Sie auf diesen Artikel“.
28 JdG vom 04.08.1854, dt.: „Dennoch kann ich Ihnen die Details noch nicht garantieren“.
29 JdG vom 28.09.1870, dt.: „Als unparteiischer Korrespondent muss ich Ihnen sagen [...]“.
30 Beginnen Verbformen mit Vokal (z.B. *j’essaie*, ich versuche), wird *je* zu *j’* abgekürzt.
31 Bei einer Spannweite von fünf Tokens rechts und links.

nuell ausgezählt werden müssen. Letzteres geschieht an einem Sample von 25% zufällig ausgewählter Okkurrenzen:

Tab. 1: Okkurrenzen von *je + vous* + [Verb] in beiden Teilkorpora, Sprechinstanz vs. Zitate.

	Teilkorpus 1 Krimkrieg	**Teilkorpus 2 Deutsch-Frz. Krieg**
Total Okk. *je + vous* + [Verb]	1.220 Okk.	381 Okk.
Gesichtetes Sample (25%)	**305 Okk.**	**96 Okk.**
davon *je* Sprechinstanz	93.4% (=285 Okk.)	78.1% (=75 Okk.)
davon *je* Zitat	6.6% (=20 Okk.)	21.9% (=21 Okk.)

Das Personalpronomen *je* in der Konstruktion *je + vous* + [Verb] referiert im ersten Zeitsemgent (TK1) sehr viel öfter auf die Sprechinstanz *je*, nämlich in 93,4% der Fälle; mit 78.1% ist der entsprechende Wert im zweiten Zeitsegment (TK 2) klar tiefer (die Gründe dafür werden weiter unten erörtert). In der Auswertung derjenigen Okkurrenzen, in welchen *je* auf die Sprechinstanz referiert, werden die häufigsten Verben und Tempora, mit welchen sich das Korrespondenten-*je* an das Publikum wendet, sichtbar[32]:

Tab. 2: Okkurrenzen von *je + vous* + [Verb] in beiden Teilkorpora: häufigste Verben und Tempora.

		Teilkorpus 1 Krimkrieg	**Teilkorpus 2 Deutsch-Frz. Krieg**
1	Okkurrenzen gesichtet	285 Okk.	100 Okk.
2	*je + vous + [dire]*	29,2%	23,4%
3	*je + vous + [parler]*	9,3%	11,7%
4	je + vous + [Verb]: *présent*	44,1%	41,9%
5	je + vous + [Verb]: *passé composé*	32,7%	30,1%

Die Auswertung zeigt, dass in beiden Teilkorpora die Vollverben *dire* (Zeile 2: 29,2% resp. 23.4%) und *parler* (Zeile 3: 9.3% resp. 11.7%) in Kombination mit *je + vous* am häufigsten auftreten. Beide Verben dienen in der Regel dazu, Kohäsion

32 Für die Auswertung wurde das Teilkorpus 2, für welches bezüglich der Konstruktion *je + vous +* [Verb] nur 75 Okkurrenzen ermittelt werden konnten, die auf die Sprechinstanz verweisen (vgl. in Tab. 1), mit 25 zusätzlichen, zufällig ausgewählten Okkurrenzen erweitert.

herzustellen, indem auf bereits vermittelte Informationen verwiesen wird, typischerweise mit Formulierungen wie (1) „comme je vous l'ai dit"[33]; (2) „[x] dont je vous parle plus haut"[34]; (3) „[x] dont je vous parlais dans mes dernières lettres"[35]. Dabei kann auf eine Information [x] referiert werden, die sich in derselben Korrespondenz befindet, wie in (2), in früheren Korrespondenzen (3) oder aber es wird nicht explizit kenntlich gemacht, wo oder wann eine Information, auf die referiert wird, schon einmal genannt worden ist (1). Dieser letzte Fall tritt bedeutend öfter während des Krimkrieges auf als während des Deutsch-Französischen Krieges, was sich gleichsam auf Mikroebene als Ausdruck eines Wandels in der Rezeptionspraxis von Zeitungen interpretieren lässt: In der zweiten Hälfte des 19. Jahrhunderts geht die Tendenz von einer Ganzlektüre hin zu einer Selektivlektüre[36]. Wichtige Gründe hierfür sind der starke Anstieg der Textmenge (vgl. die unter 2.3 erwähnten Formatvergrößerungen) sowie ein Publikum, das im Laufe der Zeit anwächst und damit heterogenere Leseinteressen entwickelt[37]. Korrespondenten der 1850er-Jahre schreiben demzufolge tendenziell für ein Publikum, das jeden Tag die gesamte Zeitung durchliest, was in der Zeit um 1870/1871 nicht mehr in gleichem Maße gegeben ist. Entsprechend durften Korrespondenten in den 1850er-Jahren davon ausgehen, dass ein Hinweis wie oben in (1) durch das stets mitlesende Publikum richtig kontextualisiert würde – was bei der selektiver gewordenen Rezeptionspraxis um 1870/1871 kaum mehr gegeben war.

Untersucht man über beide Teilkorpora sämtliche auftretende Verben rechts von *je* bei einer Spannweite von 5 Tokens in Bezug auf Tempora und Modi (vgl. Tab. 2), treten das *présent* (Zeile 4: 44,1% resp. 41,9%) und das *passé composé* (Zeile 5: 32,7% resp. 30.1%) weitaus am häufigsten auf[38], wobei die große Mehrheit der *passé composé*-Formen solche betrifft, die auf semantischer Ebene als vollendete Gegenwart zu lesen sind, beispielsweise in einer Formulierung des Typs „je vous ai écrit ce matin de Nancy"[39]. Damit wird sichtbar, dass sich die Sprechinstanz in beiden Zeitsegmenten in erster Linie immer auf ein hic et nunc, ein aktuelles Geschehen bezieht.

Schaut man sich nur die Sprechinstanz *je* ohne Kollokator *vous* an, zeigt sich ein Unterschied in der Berichterstattung während der beiden Kriege. In Tab. 3 bilden sämtliche *je*-Okkurrenzen die Grundlage für einen quantitativen Vergleich:

33 JdG 10.09.1854, dt.: „Wie ich es Ihnen gesagt habe".
34 JdG 27.01.1854, dt.: „[x], wovon ich Ihnen weiter oben berichte".
35 JdG 26.07.1870, dt.: „[x], wovon ich Ihnen in meinen jüngsten Briefen berichtete".
36 Vgl. Püschel (1991, 438–441).
37 Vgl. Clavien (2017, 58–63); vgl. für deutschsprachige Zeitungen von Waldkirch (2021).
38 Mit einem Anteil von knapp 12% folgt in beiden Teilkorpora das *imparfait*.
39 JdG 18.08.1870, dt.: „ich habe Ihnen heute morgen aus Nancy geschrieben".

Tab. 3: *je*-Okkurrenzen in beiden Teilkorpora.

		Teilkorpus 1 Krimkrieg	Teilkorpus 2 Deutsch-Frz. Krieg
1	Tokens total	768.488	491.149
2	Lemma (je|j')	3.856 Okkurrenzen	1.772 Okkurrenzen
3	Pro Million Tokens	5.007 Okkurrenzen	3.502 Okkurrenzen

Der erste quantitative Vergleich zeigt mittels des Wertes pro Million Tokens (Zeile 3) eine Differenz in Bezug auf das Vorkommen der 1. Person Singular: 5.007 zu 3.502 Okkurrenzen. Um das Verhältnis von Sprechinstanz-*je* und Zitats-*je* zu ermitteln, wurde pro Teilkorpus wiederum eine Stichprobe von 25% zufällig ausgewählten Okkurrenzen gesichtet (vgl. Tab. 3, Zeile 2), also 962 Okkurrenzen für den Krimkrieg (25% von total 3.856); respektive 430 für den Deutsch-Französischen Krieg (25 % von total 1.772), was folgende Aufstellung ergibt:

Tab. 4: Okkurrenzen von *je* in Bezug auf Sprechinstanz und Zitate.

		Teilkorpus 1 Krimkrieg	Teilkorpus 2 Deutsch-Frz. Krieg
1	Gesichtete Stichprobe	962 Okkurrenzen	430 Okkurrenzen
2	davon *je* Sprechinstanz	84,5% (=813 Okk.)	69,8% (=300 Okk.)
3	davon *je* Zitat	15,5% (=149 Okk.)	30,2% (=130 Okk.)

Balkendiagramme veranschaulichen die Werte der Zeilen (2) und (3) aus Tab. 4, wobei das Größenverhältnis zwischen den Diagrammen (1) und (2) der Anzahl Tokens *je* pro Million Wörter entspricht (vgl. oben Tab. 3, Zeile 3); Diagramm (1) ist damit um den Faktor 1,43 größer als Diagramm (2).

Es zeigt sich somit, dass in den Korrespondentenberichten des Teilkorpus 1 die erste Person nicht nur generell eine höhere Frequenz aufweist als in Teilkorpus 2; auch verweisen die *je*-Okkurrenzen mit 84,5 % hier bedeutend öfter auf die Sprechinstanz als in Teilkorpus 2, dessen entsprechender Wert sich auf 69,8 % beläuft.

Schaut man die *je*-Okkurrenzen in ihrem Verwendungskontext näher an und legt dabei ein besonderes Augenmerk auf die Okkurrenzen in Zitaten, zeigt sich, dass die Zitats-*je* im Deutsch-Französischen Krieg als Ausdruck eines Paradigmenwechsels in der Berichterstattung des *Journal de Genève* gewertet werden können: Es fällt auf, dass in der Zeit des Deutsch-Französischen Krieges die Korrespondenzen regelmäßig Auszüge aus Reden oder vom Korrespondenten ge-

Diagramm 1: Sprechinstanz-*je* und Zitats-*je* im Krimkrieg (TK 1).

Diagramm 2: Sprechinstanz-*je* und Zitats-*je* im Deutsch-Französischen Krieg (TK 2).

führte Interviews enthalten, deren *je*-Okkurrenzen nicht auf die Sprechinstanz referieren. Das Interview stellt ein Genre dar, das im Krimkrieg noch gar nicht vorhanden ist. Dabei kommen vornehmlich hohe Militärs zu Wort, beispielsweise Patrice de Mac Mahon (Marschall von Frankreich) sowie auch Napoleon III. Damit beruhen Korrespondenzen nicht mehr ausschließlich auf der oben geschilderten dialogischen Kommunikationssituation, sondern werden durch das Interview als eine weitere mögliche Form der Informationsvermittlung hybridisiert. Dies verweist auf einer übergeordneten Ebene auf ein verändertes Verständnis der Berichterstattung selbst: Das JdG, im Deutsch-Französischen Krieg von der jeweils national geprägten (und zensierten) preußischen und französischen Presse angefeindet[40], sieht sich genötigt, seine journalistischen Werte in einem programmatischen Editorial zu explizieren[41]. Es will sich als „presse neutre" verstanden wissen, das Nachrichten über ein Ereignis nicht mehr nur vermittelt und allenfalls kommentiert, sondern die damit verbundenen Textquellen auch möglichst transparent wiedergibt; eine Zeitung sei dazu da, die mit einem Ereignis in Zusammenhang stehenden Texte, die „pièces du procès" auszubreiten – und zwar „sous les yeux de ses lecteurs"[42]. Dem Interview kommt dabei insofern eine Sonderstellung zu, als es dem Korrespondenten

40 Dabei allerdings mit sehr großem Interesse rezipiert, vor allem in Frankreich, vgl. de Senarclens (1999, 37).

41 Vgl. die Frontseite vom 20.10.1870.

42 Vgl. *ibid.*, dt.: „vor den Augen ihrer Leserschaft".

erlaubt, selbst ein solches 'pièce du procès' zu generieren und dabei mittels seiner Fragen einen gewissen Einfluss auf dessen Inhalt zu nehmen.

4 Conclusio

Die Analyseresultate zeigen Gemeinsamkeiten und Unterschiede zwischen den Korrespondentenberichten aus beiden Kriegen. Für beide Zeitsegmente ist die Perspektivierung über die Sprechinstanz in der 1. Person Singular *je* zentral; die jeweils damit einhergehenden hohen Frequenzen von Verbformen im *présent* und im *passé composé* zeigen, dass die Berichterstattung sich aktuellen Ereignissen widmet (vgl. 3.2). Allerdings zeigt sich auch, dass die 1. Person Singular während des Deutsch-Französischen Krieges leicht an Wichtigkeit abnimmt – wie dies genau zu beurteilen ist, wird weitere Forschung zeigen müssen. Dabei wird auch auf die Frage einzugehen sein, ob und inwiefern sich die Publikumsadressierung mittels *vous* in den beiden Kriegen unterscheidet und allenfalls auch als Ressource zu einer parasozialen Interaktion verwendet wird (vgl. 3.1).

Insgesamt wird ersichtlich, dass die Stellung der Sprechinstanz im Krimkrieg eine andere ist als im Deutsch-Französischen Krieg: Das Vorkommen von *je* in Zitaten ist während letzterem mit 30,2 % am Gesamtanteil der *je*-Okkurrenzen fast doppelt so hoch wie während des Krimkrieges (vgl. Diagramm 1 und 2). Erscheint im Krimkrieg somit der Korrespondent als die Instanz, aus deren Perspektive der Löwenanteil an Informationen präsentiert wird, gewinnt das Einfügen von Interviews in der Zeit um 1870/1871 an Bedeutung und mit ihm eine neue Form der journalistischen Transparenz; das Zustandekommen von Informationen wird damit gleichsam als transparent vor den Augen der Leserschaft inszeniert. Auch dieser Paradigmenwechsel wird in weiterer Forschung zu vertiefen sein.

In diesem Artikel ist versucht worden, zentrale Merkmale der Textsorte Korrespondentenbericht im 19. Jahrhundert über einen Vergleich zweier Zeitsegmente aus ihrer Entstehungszeit heraus zu eruieren. Dabei bietet die Möglichkeit, mit digitalisierten Zeitungen arbeiten zu können, große Vorteile – was aber abhängig vom jeweiligen Forschungsvorhaben immer auch kritisch hinterfragt werden muss[43], denn: *jede* Suche in einem digitalisierten Zeitungsbestand generiert Resultate. Was deren *Bedeutung* ist, sei es im kulturellen oder historischen Kontext, ist freilich eine andere Sache. Entsprechend zeigt sich, dass die qualitativ vorgenommene Interpretation und die kritische Kontextualisierung dieser Resultate

43 Hierzu ausführlich Kergomard in diesem Band.

zentral sind. Die Impresso-Plattform bietet dabei den großen Vorteil, nicht nur eine ganze Reihe von Zeitungen digital zugänglich zu machen, sondern diese auch dank einer übersichtlichen Benutzeroberfläche mit klar definierbaren und transparenten Suchfunktionen analysieren und untereinander vergleichen zu können. Waren diese zum Zeitpunkt der Arbeit an diesem Artikel naturgemäß noch nicht komplett ausgereift, so werden sie in Zukunft umso mehr in Anspruch genommen werden[44].

Bibliographie

Adamzik, Kirsten. 2016. Textlinguistik. Grundlagen, Kontroversen, Perspektiven. (2., völlig neu bearbeitete, aktualisierte und erweiterte Neuauflage). Berlin, Boston.

Baumgart, Winfried. 2010. Der Krimkrieg 1853–1856: Ein Überblick. In: Georg Maag, Wolfram Pyta, Martin Windisch (Hgg.), Der Krimkrieg als erster europäischer Medienkrieg. Berlin, S. 209–219.

Becker, Daniel. 2006. Deutschland im Krieg von 1870/71 oder die mediale Inszenierung der nationalen Einheit. In: Ute Daniel (Hg.), Augenzeugen. Kriegsberichterstattung vom 18. zum 21. Jahrhundert. Göttingen, S. 68–86.

Bubenhofer, Noah. 2006–2022. Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. [Elektronische Ressource: http://www.bubenhofer.com/korpuslinguistik/, abgerufen am 19.07.2020]

Bubenhofer, Noah. 2009. Sprachgebrauchsmuster – Korpuslinguistische Methoden als Methode der Diskurs- und Kulturanalyse (Sprache und Wissen, Band 4). Berlin, New York.

Brommer, Sarah. 2018. Sprachliche Muster. Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte. Berlin, Boston.

Chapuisat, Édouard (Hg.). 1999 (1929). Centenaire du Journal de Genève. Un siècle de vie genevoise. Genève.

Clavien, Alain. 2017. La presse romande. Lausanne.

Devitt, Amy J. 2004. Writing genres. Carbondale.

Hanitzsch, Thomas. 2007. Journalismuskultur: Zur Dimensionierung eines zentralen Konstrukts der kulturvergleichenden Journalismusforschung. Medien- und Kommunikationswissenschaft 3, S. 372–389.

Hillerich, Sonja. 2018. Deutsche Auslandskorrespondenten im 19. Jahrhundert. Die Entstehung einer transnationalen journalistischen Berufskultur. (Pariser Historische Studien, Bd. 110). Berlin.

Linke, Angelika. 2003. Sprachgeschichte – Gesellschaftsgeschichte – Kulturanalyse. In: Helmut Henne, Horst Sitta und Herbert Ernst Wiegand (Hgg.), Germanistische Linguistik: Konturen eines Faches. Tübingen, S. 25–65.

44 An dieser Stelle möchte ich allen Beteiligten des Impresso-Projektes meinen großen Dank für ihre Hilfe und Inspiration aussprechen.

Luginbühl, Martin. 2014. Medienkultur und Medienlinguistik. Komparative Textsortengeschichte(n) der amerikanischen „CBS Evening News“ und der „Schweizer Tagesschau“ (Sprache in Kommunikation und Medien, Band 4). Bern.

Miller, Carolyn R. 2015. Genre Change and Evolution. In: Natasha Artemeva, Aviva Freedman (Hgg.): Genre Studies around the Globe: Beyond the Three Traditions. Kentucky, S. 154–185.

Püschel, Daniel. 1991. Journalistische Textsorten im 19. Jahrhundert. In: Rainer Wimmer (Hg.), Das 19. Jahrhundert. Sprachgeschichtliche Wurzeln des heutigen Deutsch. Berlin, New York, S. 428–447.

Senarclens, Jean de (Hg.). 1999. Un journal témoin de son temps. Histoire illustrée du Journal de Genève 1826–1998. Genève.

Tardy, Christine M., Swales, John M. 2014. Genre analysis. In: Klaus Schneider, Anne Barron (Hgg.): Pragmatics of Discourse. Berlin, S. 165–187.

Tienken, Susanne. 2015. Muster – kulturanalytisch betrachtet. In: Christa Dürscheid, Jan Georg Schneider (Hgg.). Handbuch Satz, Äußerung, Schema. Berlin, Boston: de Gruyter, S. 464–484.

Vaillant, Alain. 2014. Le double jeu du journal, entre communication médiatique et correspondance privée. In: Guillaume Pinson (Hg.), La lettre et la presse: poétique de l’intime et culture médiatique. [http://www.medias19.org/index.php?id=341, update 27.5.2020]

Van den Dungen, Pierre. 2008. Écrivains du quotidien : journalistes et journalisme en France au XIX^{ème} siècle. Semen 25. [https://journals.openedition.org/semen/8108, update 1.1.2021]

von Waldkirch, Tobias. 2021. »daß auch für gediegenen Unterhaltungsstoff in erhöhtem Maße gesorgt ist« – Lesepublikum und Rubrikenrepertoire im 19. Jahrhundert am Beispiel der NZZ (1858, 1868/69, 1878). In: Susanne Tienken, Hartmut E.H. Lenk, Martin Luginbühl, Stefan Hauser (Hgg.). *Methoden kontrastiver Medienlinguistik*. Bern: Peter Lang, S. 259–273.

Wauters, Éric. 2012. Le procédé épistolaire dans la presse française de la Révolution à la Restauration. In: Guillaume Pinson (Hg.), La lettre et la presse: poétique de l’intime et culture médiatique. [http://www.medias19.org/index.php?id=327, update 1.6.2020]

Wawro, Geoffrey. 2009. The Franco-Prussian War. The German Conquest of France in 1870–1871. New York.

Fredrik Norén, Johan Jarlbrink, Alexandra Borg, Erik Edoff, Måns Magnusson

The Transformation of 'the Political' in Post-War Sweden

Abstract: This paper explores what was explicitly defined as 'political' during the post-war era, from 1945 to 1989, in two Swedish newspapers. Based on all extracted text blocks containing the term 'political', two research questions are examined: How has the use of the term "political" evolved over time? In which contexts was the concept inscribed, and how did these change over time? Inspired by conceptual history, the analysis is divided into three parts: an examination of 'political' through bigram extractions, contextual explorations using topic modeling, and a close reading of one particular topic over time, the topic labeled 'women'. The result shows an increased use of the term 'political' from the 1960s, with more things that were labeled as 'political'. The analysis reveals that the concept was broadened, but not entirely redefined.

Keywords: media history, conceptual history, topic modeling, post-war Sweden, newspapers, digital history

1 Introduction

30-year-old 'Sartre' is looking for love in July 1980. The man, using this signature in his ad in the Swedish newspaper *Aftonbladet*, writes that he is lonely and wants to meet a woman. He likes to play football, read, and listen to 'good music'. 'I have pretty strong opinions about most things political (socialist far to the left)'.[1] (*AB* 1980-07-13, all translations by the authors). The signature Sartre was not the only one declaring his political views in personal ads in the 1960s, 1970s and 1980s. Another unmarried man looking for a 25–30-year-old woman in 1975 writes: 'You should be politically aware with a socialist conviction. It is not all

1 All translations by the authors.

Acknowledgement: This research is part of the projects 'Welfare State Analytics: Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989' (project no. 2018-06063; westac.se/en), and 'Mining for Meaning: the Dynamics of Public Discourse on Migration' (project no. 2018-05170), both funded by The Swedish Research Council. The team would like to thank Miriam Hurtado Bodell for valuable assistance at the KBLab.

https://doi.org/10.1515/9783110729214-019

that counts, but it is important' (*AB* 1975-06-01). A woman in the north of Sweden, also writing in 1975, was hoping to meet a man, 40–45 years old: 'It's a plus if you are to the left politically' (*AB* 1975-08-03). A topic model on newspaper data from post-war Sweden captures many of these ads in one topic, increasing in weight in the late 1960s and peaking in 1979. This suggests that the 'political' had become a significant part of the private life of many Swedes – not just as an abstract dimension, but as a concrete notion. Moreover, in Swedish newspapers from the 1960s and 1970s, there were also bigrams such as 'political street theatre', 'political bachelor parties', 'political religion', 'political pea soup', 'political gender discrimination', and 'political celebrity journalism' (all are bigrams in Swedish).

Tracking the ways that the term 'political' (in Swedish 'politisk', 'politiska', and 'politiskt') was used over time in post-war Sweden makes it possible to examine discursive shifts. The purpose of this paper is to use computational approaches to study political trends in two Swedish newspapers, from 1945 to 1989, the morning paper *Dagens Nyheter* (*DN*, liberal) and the evening paper *Aftonbladet* (*AB*, liberal until 1956, then social democratic). How has the use of the term "political" evolved over time? In which contexts was the concept inscribed, and how did these change over time? The analysis is divided into three parts: an examination of 'political' through bigram extractions, contextual explorations using topic modeling and, finally, a close reading of one particular topic over time, the topic labeled 'women'.

1.1 The Press and the Political Landscape

The years 1945 to 1989 are commonly understood as the classical period of the Swedish welfare state, a time of both political stability and of political disruption. It was, for the most part, an era of governments led by social democrats launching a vast number of social reforms. The range of issues that the state took an interest in included almost every area of the Swedish society (Möller 2019).

In the 1950s, public debate in Sweden was characterized by political consensus: a social democratic dominance on the domestic arena, and sympathies towards liberal western democracies internationally. In a world divided by a communist east and a capitalist west, Swedes who argued for a 'third way' were seen as Soviet allies. In 1952, the influential editor of *DN*, Herbert Tingsten, declared 'the death of ideologies'. The acceptance for alternative political ideas, however, grew steadily from the early 1960s – first in the form of cultural relativism, later on as a radical critique of capitalism and established institutions (Frenander 1999). Furthermore, it is often argued that the new movements of the 1960s renegotiated and broadened the concept of the political (Östberg 2008).

Compared to continental Europe, Sweden's 1968 was less associated with student protests and more with internationally oriented solidarity and activist groups. Many of the democratization reforms fought for in Europe were already happening in Sweden (Östberg 2008). Regarding the impact of the 1960s' radicalism on Swedish society, Bjereld and Demker (2018) emphasize the teardown of traditional authorities: priests, teachers, adults, men, established high culture, and traditional party leaders. The ideological shift was followed by new legislation in many areas, such as tax laws making it worthwhile for married women to work, and a ban on physical punishment of children. A new leftist hegemony was established, and then to some degree overturned by neoliberalism in the 1980s. However, many vital ideas put forward by the left lived on – among them the critique of a powerful and controlling state (Frenander 1999; Boréus 1994).

News media and journalists in the 1960s and 1970s were part of, and contributed to, the general trend of questioning traditional authorities. Far from simply mirroring established politics, they started to raise critical questions and highlight social problems and political corruption. This new kind of journalism covered a broader range of issues. Everyday life, especially the daily work at factories and offices, became frequent topics in newspapers, radio and TV (Djerf-Pierre & Weibull 2001).

Until the mid-1960s, the arts, literature and culture sections in newspapers were dominated by essays and reviews of literature and art. If political ideas were debated, they usually followed established arguments of the editorial page. In the mid-1960s, however, the cultural editors became more independent. The two newspaper sections developed into two separate entities, sometimes arguing against each other, with the arts, literature and culture section often taking more radical positions. Gender equality and the situation within developing countries were among the issues debated. Literature and art continued to be covered, but reviewers started to discuss them in more political terms (Riegert & Roosvall 2017). These so-called left-wing tendencies in the Swedish metropolitan press were part of a general transformation, and at the fore were *AB* and *DN* (Gustafsson & Rydén 2010).

1.2 Theoretical Perspectives

'Politics' and 'political' are key concepts. Following the tradition of conceptual history (Ifversen 2011; Stråth 2013), key concepts are those that are contested and yet inescapable. Users might disagree about their meaning and how to use them, but they still need them. To define them, restrict, expand, or challenge their meaning and use, is part of a political struggle. The redefinition of critical

concepts is hence a way to redefine social reality, to interpret it and frame it, and perhaps legitimize goals and point out the necessity of certain actions. To conceptualize such a struggle as political is an example in itself. Defined differently, 'political' would have other meanings.

One way to conceptualize politics is to use spatial metaphors: a sphere, a field or a domain. Issues within the sphere are political. Those outside are not. The boundaries change over time, and non-political issues may become political, and vice-versa (Palonen 2006). Depending on the demarcations, the political sphere may include politics in a formal sense as well as political debate and extra-parliamentary movements. Politicization often refers to the process when boundaries are redrawn, and new issues and topics are introduced on the political agenda. Labeling something as 'political' is an essential part of the process. Palonen (2003: 182) states that 'by politicization, we can mark a phenomenon as political, as a *Spielzeitraum* for contingent action. "Politicization" thus refers to the act of *naming* something as political, including the controversies surrounding the acceptance of this naming' (Palonen 2003: 182). The expansion of what is defined as political indicates that new kinds of perspectives are introduced – and hence that new possibilities for action are constructed. The feminist redefinition of the personal as something political is one such example.

Building on a conceptual history approach is particularly useful in a computational context. From this perspective, political history is understood as a history of language use, and language use is an area where computational methods show their strength. Research questions are not translated into keyword searches, because research questions are already formulated as questions about keywords and how they are used. Yet, to trace things explicitly described as 'political' will not seize everything political. Issues and actors can be political even without the presence of the explicit term 'political'. In some cases, notions are implicit or absent because most readers will understand that the issues discussed are political. Our approach identifies many instances of labeling that are not part of an ongoing politicization or an expansion of the political sphere, for example, political parties, political elections, political debates. Still, with these limitations in mind, our analysis will provide an overview of emerging themes and of general trends of the 'political' in newspaper data, and how these relate to each other over time.

1.3 Methods and Sources

We have examined the 'political' in three kinds of contexts: in a micro context using bigram extraction, in a wider text block context using topic modeling, and in a publication context based on close reading, where topic co-occurrence

is related to genres and newspaper sections. With the extraction of bigrams, we analyze the words that follow the attribute word 'political' (in Swedish 'politisk', 'politiska' or 'politiskt').

In order to explore a wider context, we have turned to topic modeling using the Latent Dirichlet Allocation statistical model (LDA, Blei et al. 2003). The model offers an efficient way to study themes in a large corpus by assigning words in documents a probability value based on word co-occurrences. A topic can thus be understood as a discourse or theme occurring in the corpus (Blei et al. 2003; Blei et al. 2012). To estimate the parameters we used Gibbs sampling, one of the standard approaches to estimate topic models (Griffiths and Steyvers, 2004) as implemented in Mallet (McCallum, 2002). We used the last iteration of the Gibbs sampling algorithm to analyze the corpus.

The newspapers used have been digitized by the National Library of Sweden using the Optical Character Recognition (OCR) engine Abbyy and the segmentation tool Zissor. The public search interface provided by the library (tidningar.kb.se) consists of the lion's share of the newspapers published in Sweden from 1645 until today, around 27 million pages. All material published before 1905 is searchable in the public interface. Since we did our research at the Swedish National Library's lab for digital research,[2] we had access to copyright protected newspaper material 1945–1989. Because the newspaper data is not segmented into articles or sections, we had to base our analysis on a text block level. Text blocks are segments of visually cohesive text, usually a paragraph. The data was prepared in two steps: first, we extracted all text blocks that contained the word 'political' – 390 699 text blocks in total, accounting for about 27 million tokens, with an average of 69 tokens per text block. A recent evaluation of the page segmentation and of the OCR shows a high quality of our corpora compared to other historical newspaper materials, although this is far from perfect (Hurtado Bodell et al. 2020). Secondly, related to our topic modeling, we reduced the corpora further by discarding stop words (a commonly used list of Swedish stop words was complemented with a manually curated list), tokens that only occur five times or less, tokens containing only one character, and non-alphabetic characters.

2 'KBLab'. Text. Accessed 4 August 2021. https://www.kb.se/in-english/research-collaboration/kblab.html.

2 The Long Tail of Political Bigrams

As a first step to track the 'political' in the two newspapers, we examined bigrams – two adjacent words composed of the adjective 'political' and the noun it modifies (e.g. 'political party', 'political future', 'political women'). Conceptual history focuses on how specific terms are used, and this approach thus provides a first indication of how 'political' was utilized as a qualifier. Plotting the occurrences of 'political' bigrams reveals a significant increase from the late 1960s, a decrease from late the 1970s followed by a stabilization during the 1980s (Fig. 1).

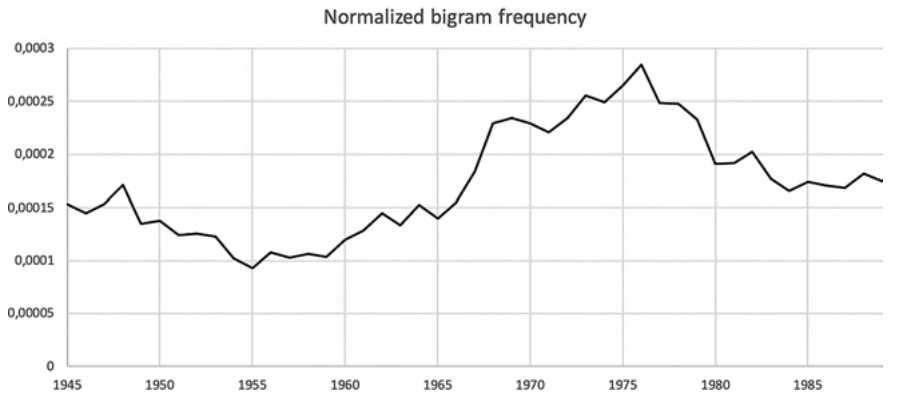


Fig. 1: The normalized frequency corresponds to the total number of 'political' bigrams per year divided by the total number of tokens in *Aftonbladet* and *Dagens Ny-heter* per year.

The bigram trend only tells us that the usage of the word 'political' as modifier increased. Plotting the frequency counts of all distinct bigrams (with 'political' as modifier), however, reveals a long tail of 'political' vocabulary, with a majority of bigrams only occurring once. The top lists of 'political' bigrams are relatively stable for each year and are dominated by themes or events that are traditionally described as political, such as 'political parties', 'political debates' and 'political prisoners' (see Tab. 1). This is true for at least the 100 most frequent bigrams.

The curve of bigram frequency counts – independently of year – drops rapidly and flattens out to a long tail. The graphs in Figs. 2 and 3 show the rank-frequency distribution of distinct 'political' bigrams counts for selected years and give a sense of how the proportion of the tail relates to a specific year. More specifically, these figures show the relation between 'political' bigram diversity (the frequency rank on the x-axis) and usage (the frequency count n on the y-axis). For example,

Tab. 1: The ten most frequent 'political' bigrams in nine different years from 1945 to 1989.

1945	1950	1955	1960	1965	1970	1975	1980	1985
ed [editor]	editor in chief	editor in chief	circles	(the) parties	prisoners	prisoners	prisoners	editor in chief
prisoners	editorial	circles	(the) parties	diary	asylum	(the) parties	(the) parties	(the) parties
circles	circles	(the) parties	(the) life	prisoners	(the) parties	parties	decision	parties
parties	party	parties	party	parties	reason	(the) prisoners	parties	asylum
(the) parties	committee	(the) situation	issues	party	parties	views	party	prisoners
activity	parties	issues	reason	reason	(the) debate	reason	reason	decision
editor	(the) parties	(the) debate	parties	issues	theater	party	asylum	party
(the) life	(the) situation	party	life	convers-ation	party	opponents	violence	(the) debate
life	refugees	reason	(the) debate	(the) debate	issues	refugees	commentator	reason
(the) situation	issues	(the) life	(the) situation	(the) life	(the) power	activity	issues	issues

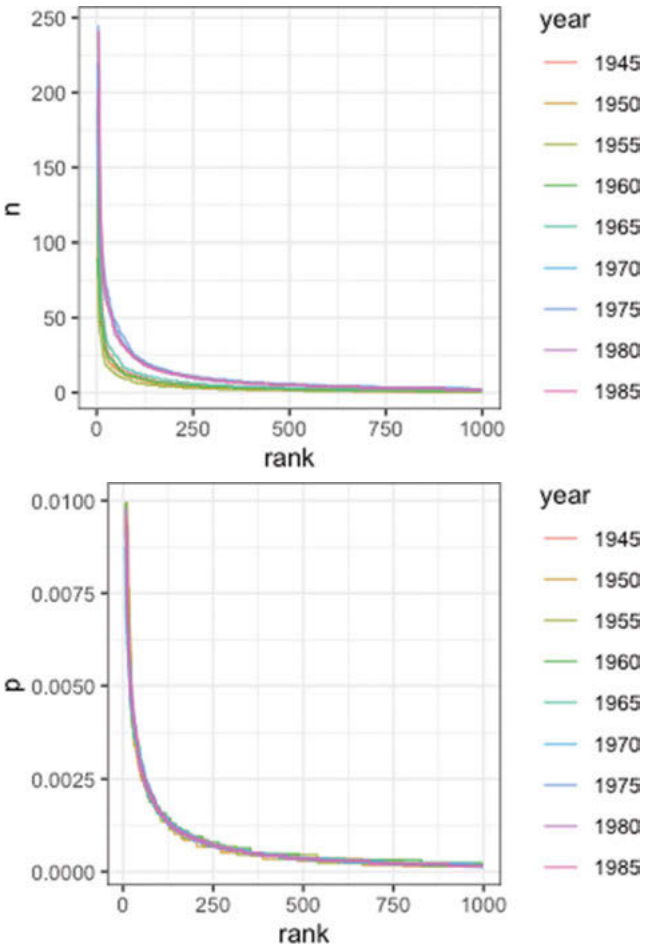


Fig. 2 (top) and 3 (bottom): Rank refers to the frequency rank of each bigram (i.e. the position of a bigram in the bigram frequency count distribution), *n* refers to the total frequency count of each unique bigram, and *p* to the (normalized) proportion of each bigram for each year, i.e. the frequency count of a given bigram divided by the total frequency counts of all bigrams. Hence, a higher trend line on the *x*-axis indicates increased diversity among the bigram for a specific year. The top and the end of the tail in each graph are cut.

taking the year 1970 of Tab. 1 above, the bigram 'political prisoners' would be positioned at rank 1 on the *x*-axis (left) with a high frequency count on the y-axis, resulting in a top-left point on the curve. The longer the tail on the *x*-axis, the more diverse set of 'political' bigrams. Here, we observe that an increase of 'political' bigrams counts (*n*) does cause an increase in the variety of bigrams (rank

distribution). However, this is mainly a function of the increased usage of the word 'political', as it is made clear by the normalized (p) curve in Fig. 3. Hence, when the term 'political' is used more often, it is also used to modify a more diverse set of activities and phenomena qualified as 'political'.

If we compare how many distinct bigrams it takes to reach 50 percent of the total bigram occurrences for each year (Fig. 4), however, it is possible to measure the lexical diversity of 'political' bigrams', i.e. to get a sense of how the usage of the term 'political' becomes more disparate or more concentrated. Hence, as the graph in Fig. 4 displays, 'political' became less consistent from the late 1950s (it takes more bigrams to achieve 50 percent of the total bigram frequency count), more consistent in the 1970s, and stabilized in the 1980s.

Fig. 4: The graph displays how many unique top bigrams (rank) it takes to reach 50 % of the total 'political' bigram occurrences for each year. A higher rank indicates an increased variety and vice versa. Some of the peaks can maybe be explained by election years (e.g. 1956, 1966, 1988), but others not (e.g. 1971, 1983).

Further exploration of the lists of bigrams can provide glimpses into cultural tendencies of how 'political' was used in the two newspapers. One way is to trace individual bigrams, which Fig. 5 illustrates as it displays how 'theatre', 'awareness' and 'language' became modified by 'political' over time.

Overall, the result from studying bigrams of 'political' confirms general observations made in previous research: the 1950s shows lower frequency and less diversity, while the usage increases in the 1960s and becomes more diverse. However, because of its limited word window approach, bigram extraction alone is not sufficient to explore the broader semantic contexts of the term 'political'. Hence, in

Fig. 5: Three political bigrams: 'political theatre', 'political awareness' and 'political language' (normalized frequency).

order to investigate the contexts of the 'political', we have turned to topic modeling, a method that takes every word in the extracted text blocks into account.

3 Analyzing Networks of Co-Occurring Topics

Based on our corpora of text blocks containing the term 'political', four topic models were initially produced with 50, 100, 200 and 400 topics (we used ParallelTopicModel, workers = 1, seed = 42). After examining the topics from each output, the model with 200 topics was chosen for the analysis. The models of 50 and 100 topics did not capture a satisfying representation of distinct themes. The 400 topic model, in turn, produced too specific topics, or topics that were too alike.

Topics were labeled manually in order to increase readability. This was done in a three-step iterative process, constantly shifting from topic model to newspaper texts. First, the authors worked together to manually interpret and label the topics from the 200 model, based on each topic's top list of most likely words. Often, topics could be labeled based on the first 20 words. If not, the top list was expanded to the top 100 words. Topics were deliberately assigned broad labels such as 'chile', 'parties' and 'middle east'. This was because a topic like 'greece', for instance, encapsulates different aspects related to Greece such as the 1946–1949 civil war or the 1967–1974 military junta. However, since all topics were built up from newspaper texts containing the term 'political', all labels imbue a political

dimension. Secondly, and as a way to deal with uncertainty and disagreement of which label a topic should be assigned, the authors traced such topics back to the actual text blocks in which ambiguous topics were most dominant. This was for example the case with topics such as 'human condition' (related to texts about how to be a human in today's world), 'analysis' (various texts with analytical and reasoning characteristics) and 'official statements' (mostly statements given by political figures). Third, to assert that our interpretations were robust, three multiple posteriors were computed with the same and slightly different number of topics as the original 200 model (with 190, 200, and 210 topics). The topics of these additional models were labeled according to the three-step hermeneutic process described above, which generated similar results.[3]

Our LDA topic model provides a static representation of all 'political' text blocks from 1945 to 1989. Topics' word lists do not change over time, but topics' presence in text blocks do, as well as their co-occurrences with other topics. Hence, an efficient way to explore changes over time is to examine co-occurring topics, corresponding to topics appearing in the same text block. We studied nine five-year periods of co-occurring topics between 1945 and 1989 (1945–1949, 1950–1954, et cetera). This was done in order to examine discursive changes in higher resolution. Here, we treated topics as nodes, and topic co-occurrences in text blocks as edges, with weaker edges filtered out. The graphs, modeled in Gephi, were based on topic pairs co-occurring in at least ten different text blocks, and where both topics made up at least 20 percent each of the text block. In the graphs, node weight corresponds to the sum of a node's edge weight, i.e. the number of documents where the topic is present.

The number of topics differ between the five-year periods, with networks being smaller in the early periods, and larger from the second half of the 1960s. Only nine topics are present in all the five-year periods: 'asylum', 'elections', 'names' (common Swedish names), 'negotiations/agreements' (with top words such as 'demand', 'proposition', 'sides'), 'official/succession' ('post', 'retirement', 'successor'), 'physical descriptions 2' ('standing', 'sitting', 'room', 'house'), 'uk 1' ('British', 'Labor', 'conservatives'), 'us presidents', 'visits/negotiations' ('meeting', 'representative', 'conference'). These topics could be seen as representing the core of the 'political' discourse. In order to study changes over time we manually examined co-occurring topics in the nine different graphs. In our interpretative process, we hermeneutically identified three thematic clusters evolving over time,

3 The topic model used in the analysis is accessible on github.com/welfare-state-analytics, see Topic model in references.

centered around international topics, domestic politics, and culture (see Appendix 1 for Fig. 10–15 the six graphs not presented in the main text).

3.1 The International Cluster

The co-occurring topics from the late 1940s and 1950s indicate that the 'political' in newspapers was essentially a foreign affair. The early post-war period was dominated by political issues related to Germany and the war (connected to topics labeled 'murder', 'prisoners' and 'trials'), and the division of Europe into two blocks (the central topic node in Fig. 6 is labeled 'soviet/cold war'). Related topics capture a succession of politicians and governments, and topics labeled 'negotiations and agreements', and 'recent developments' – the last topic having words such as 'situation', 'current', 'future' and 'development' among its top-ranked words. The cold war is still present in the network from the early 1960s, but focus shifts in the later decade to 'us/vietnam' and 'eec' (European Economic Community).

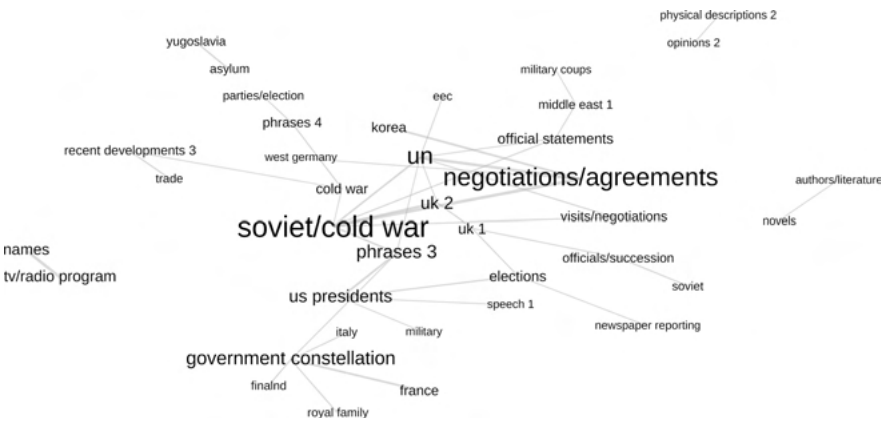


Fig. 6: Co-occurring topics 1950–1954. The graph is mostly dominated by the international cluster. The size of the nodes is proportional to the weighted degree.

Leading politicians, governments and organizations dominated most of the topics from the 1940s to the 1960s. Then, a gradual change emerged in the late 1960s and early 1970s, with topics such as 'political prisoners' and 'asylum' among those with the highest weighted degree but also topics labeled 'violence/terror' and 'human suffering'. These topics represent new perspectives in the newspaper text blocks. Apart from 'prisoners' and 'refugees', we find 'children' and 'families'

(in the topic 'asylum'); 'hostages' and 'innocents' (in the topic 'violence/terror'). Most of the groups made visible in these topics were victims rather than actors. Nevertheless, they were part of a widening of the reports on political issues. This shift can be understood as a signal for the professionalized and critical journalism in the 1960s and 1970s.[4] (Petersson & Carlberg 1990; Djerf-Pierre & Weibull 2001).

3.2 The Domestic Cluster

From 1945 to 1964, the cluster related to domestic affairs contains few nodes, but from 1965 to 1989 it expands and includes an increasing number of topics (e.g. see Figs. 7 and 8). During the first two deccades, the domestic cluster centers on politics from a top-down perspective, with topics labeled 'proposals/inquiry', 'parties' and 'taxes'. From the mid-1960s, the cluster is complemented by a bottom-up oriented perspective as the topic 'power/democracy' is introduced in the network, mostly related to participation and the condition of democracy in Sweden. The cluster in this later period also connects to 'work' and 'women', topics often representing an individual or personal perspective in the newspaper reporting, situated in or between the domestic cluster and culture. Another change is an increased local and regional focus on Stockholm, especially during the 1980s (both *AB* and *DN* are essentially Stockholm papers).

From 1945 to 1989, the domestic topic nodes with the highest weighted degree are often related to political parties, the public sector, and economic policy. The topics related to political parties mostly contain a generic vocabulary with words such as party names, 'election', 'opposition', 'majority' et cetera. From the late 1970s, however, we find topics that relate to individual politicians ('party leaders' and in the late 1980s 'social democrats') rather than to a generic vocabulary. Somewhat similar topics were also found in the earlier period, then connected to topics related to biographical-oriented content from obituaries and memorial days. In the 1980s, however, co-occurring topics are most often related to politicians' work in parties and parliament. This could again be understood as a signal of journalistic change, an increased individualization in political reporting (Petersson & Carlberg 1990; Ekecrantz & Olsson 1994).

4 Petersson & Carlberg 1990; Djerf-Pierre & Weibull 2001.

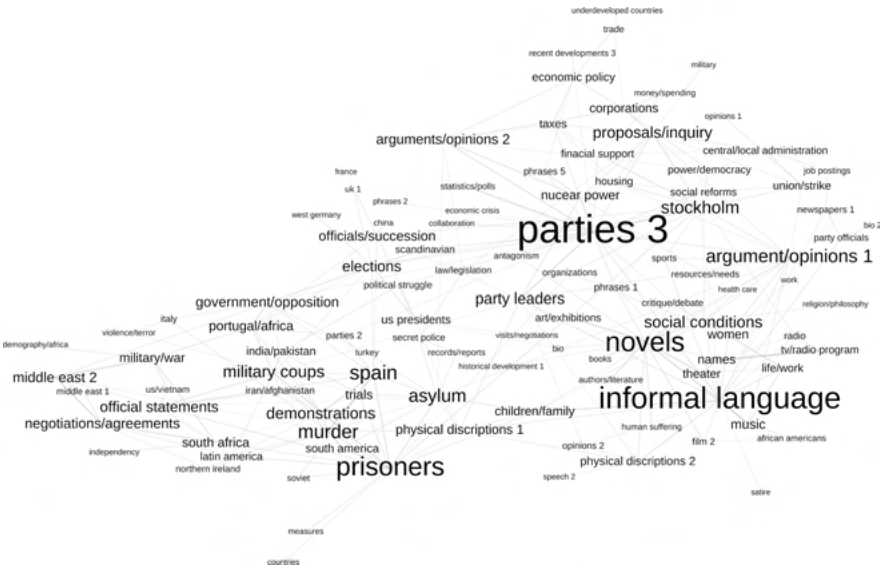


Fig. 7: Co-occurring topics 1975–1979. The international cluster is situated in the lower left part of the graph, the domestic cluster in the upper corner, and the cultural cluster in the lower right.

3.3 The Cultural Cluster

A third cluster relates to culture. From 1945 to 1964, this cluster consists of a few scattered topics related to literature ('authors/literature', 'novels', 'books') and plays ('theater'), as well as broadcasting ('tv/radio program'). In the late 1960s and onwards, a stable cluster emerges that also expands to include topics labeled 'art/exhibitions', 'music' and 'film', at the same time as 'novels' and 'theatre' increase in centrality in the cultural cluster. Furthermore, and somewhat surprising, 'novels' is one of the topics with the highest weighted degree after 1965 (e.g. see Figs. 7 and 8). In fact, *the* highest in the network of 1965–1969, and in the other later networks it is among the top five, equal to topics labeled 'us/vietnam' (period 1970–1974) and 'elections' (period 1985–1989). This indicates a politicization of literary culture as well as culture playing an important part in political discourse (see Appendix 1).

Furthermore, from the late 1960s, culture-oriented topics tend to co-occur with non-culture topics such as 'social condition' (with top words such as 'society', 'human', 'economic', 'development'), 'human suffering' ('world', 'life', 'death', 'people') and 'political prisoners' ('chile', 'amnesty', 'torture', 'international'). This

development can be understood as a signal of how culture – 'novels', 'theatre', 'art/exhibitions' et cetera – was mobilized in a political discourse beyond the cultural sphere. Authors, directors, and musicians took political positions, and political and social issues were frequently discussed in the arts, literature and culture sections of newspapers. These patterns confirm what researchers have described as a turn in the cultural sphere during the 1960s and 1970s, from aesthetics to culture as a political tool for social critique and societal change (Frenander 1999: 138–144; Östberg 2008: 341).

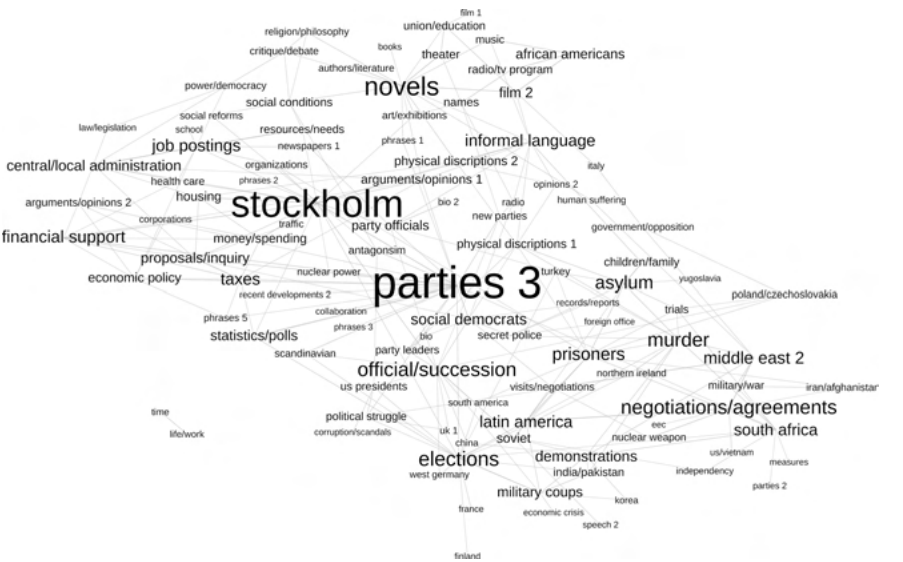


Fig. 8: Co-occurring topics 1985–1989. The international cluster is situated in the right corner of the graph, the domestic cluster in the left part, and the cultural cluster in the upper corner.

3.4 The Rise of 'Women' within Political Reporting

Some topics are located in between clusters. The one we have labeled 'women' is one example, co-occurring with 'novels' and 'art/exhibitions' as well as domestic issues such as 'social reforms' and 'central/local administration' (see Appendix 2). If we follow the co-occurrences over time, and manually examine the section and genre for those texts where the probability for 'women' is at least 0.3, we can trace how the framing of the topic shifts. LDA generates topics with top words that are static over time. Hence, 'women' stays the same, but

since it occurs on different pages and sections, the contexts and meaning of the topic change over time.

When the topic becomes more frequent in the 1960s (see Fig. 9), it is in the contexts of arguments and opinions (see Appendix 2). Our manual examination of text blocks shows that the topic is found in letters to the editors, editorials and reviews in the arts, literature and culture section. Articles called for women to participate in parties and unions, and for men to accept women as political actors. An editorial in *DN*, for instance, explained that 'we [will] achieve full gender equality in the political process only when women [will] really dare to have a finger in the political pie' (1963-07-14). Furthermore, hardly any articles written by regular news journalists are found among the texts in which the topic of women is significant in the 1960s.

Fig. 9: The mean weight of the topic 'women' in text blocks over time.

Editorials and letters to the editors continued to be published in the 1970s, but the topic is also present in other contexts, most notably in new sections titled 'The everyday' (*DN*) and 'Women' (*AB*). The presence of everyday life is visible among co-occurring topics such as 'social conditions', 'children/family' and two topics related to work (see Appendix 2). The personal ads quoted in the introduction indicate something similar: political views had become important on a personal level, while various aspects of everyday life were contextualized as 'political'. Separate newspaper sections for women and household matters were established already in the nineteenth century, but the sections in the 1970s were different. 'The everyday' was where new research on gender equality and gender roles were reported and where journals published by women's movements were referred. Ordinary women and activists were interviewed about their everyday

life and explained the difficulties of combining regular jobs, household tasks and political activities. This was also where a new movement that organized men was reported: 'We don't want to be the oppressors of women anymore' (*DN* 1975-02-18). Similar interests are visible in *AB*. As a tabloid, it had a tradition of covering the everyday life of various celebrities, but in the 1970s it also started to interview ordinary women about their daily lives. 'How do you do in your workplace, Bettie Liljeqvist?', a headline asked in 1978. Liljeqvist was then the only female steelworker at the company Asea. In the interview, she told the reporter about her fight for gender equality at work (*AB* 1978-06-28). The co-occurring topic 'informal language', capturing words commonly used in spoken Swedish, frames the issues of women as less formal than regular party politics.

Everyday life as a context and framing is less prominent in the 1980. Instead, the topic 'women' co-occurs with topics such as 'organizations', 'stockholm' and 'central/local administration' (see Appendix 2). Editorials and letters to the editors were still published, but the topic was also present in regular news articles in the 'Politics' section, covering party politics on national and local arenas. Women's issues were part of the established political discourse and those being interviewed were professional politicians rather than ordinary women. When *DN* printed the statement 'Sex is political too' in 1984, for instance, it was as a headline for an article referring to the new 'sexual-political program' of the social democrats (*DN* 1984-05-23).

The significance of gender issues in the 1960s and 1970s is well documented in previous research (Östberg 2008; Bjereld and Demker 2018). What our analysis indicates, however, is that the issues moved within the newspapers and that different genres and newspaper sections framed the issues in different ways. In the 1960s 'women' was a topic discussed in editorials, letters to the editor and in articles in the art, literature and culture section. The texts represented individual voices outside of regular news journalism. In the 1970s it was news journalists covering the issues, but in the 'everyday' and 'Women' sections of the newspaper. The sources represented actors outside of the political establishment: ordinary women, activists, researchers. In the 1980s, the topic was instead part of established party politics, covered in articles published in the 'Politics' section. What was once in the margin had become an established part of the political sphere, broadening what 'political' could mean.

4 Concluding Remarks

An attempt to capture the transformation of what is considered political, faces the apparent risk of being either too general or far too narrow. In this paper, we have defined the political as what is explicitly described as 'political'. To some extent, this limitation is compensated by the analytical scale of using all text blocks that contain the keyword 'political', in two major Swedish newspapers from 1945 to 1989.

The extraction of bigrams confirms some of the claims made in previous research: the total frequency hits a low point in the 1950s, the decade when 'political' bigrams were less diverse. Figures from the 1960s and 1970s indicate that 'political' was used in a broader sense. Examining co-occurring topics reveals similar patterns: new issues became political in the 1960s and 1970s. Still, by combining different approaches, our paper also paves the way for a deeper understanding of changing political discourses in post-war Swedish newspapers. For example, most of the discourses dominating in the 1950s continued to dominate in the 1960s and 1970s: international conflicts, party politics, elections, economic policy. However, and more interestingly, these discourses of domestic and foreign politics were sometimes challenged in dominance by topics such as 'novels', indicating a perhaps more vital, or at least different role of culture in the newspaper material than previous research has indicated. Furthermore, following a topic such as 'women' through the newspaper pages makes it evident that genre and section are as important as content for the establishment of new political issues. A female reader complaining about the lack of representation in a letter to the editor is one thing, the same issue reported in the regular section for 'Politics' is quite another. Hence, to conclude, new issues entered the political sphere without the old ones leaving, suggesting that the 'political' was broadened, but not entirely redefined.

In this paper, we manually traced the movement of one specific topic between different sections of the newspapers. Our analysis indicates that the context of publication is vital to the historical significance of a topic. What was a marginal topic when it was part of letters to the editors became accepted as part of established journalism when it was published in the news sections. Thus, topic distributions in the newspaper corpus would be more valuable if we had the ability to calculate its distribution in different sections of the newspapers. An important research task in the future is to train models for the automatic detection of sections in order to enrich the metadata and scale up the analysis.

Appendix 1: Networks of co-occurring topics 1945–1989

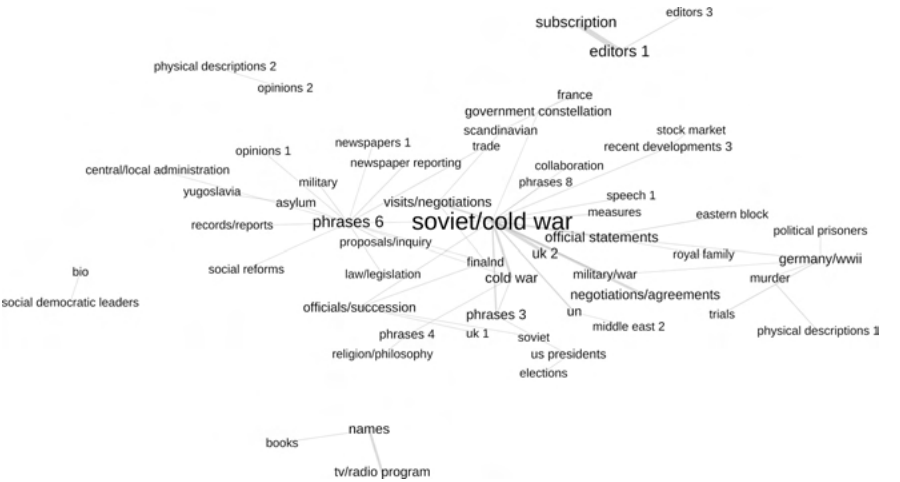


Fig. 10: Network graph of co-occurring topics 1945–1949.

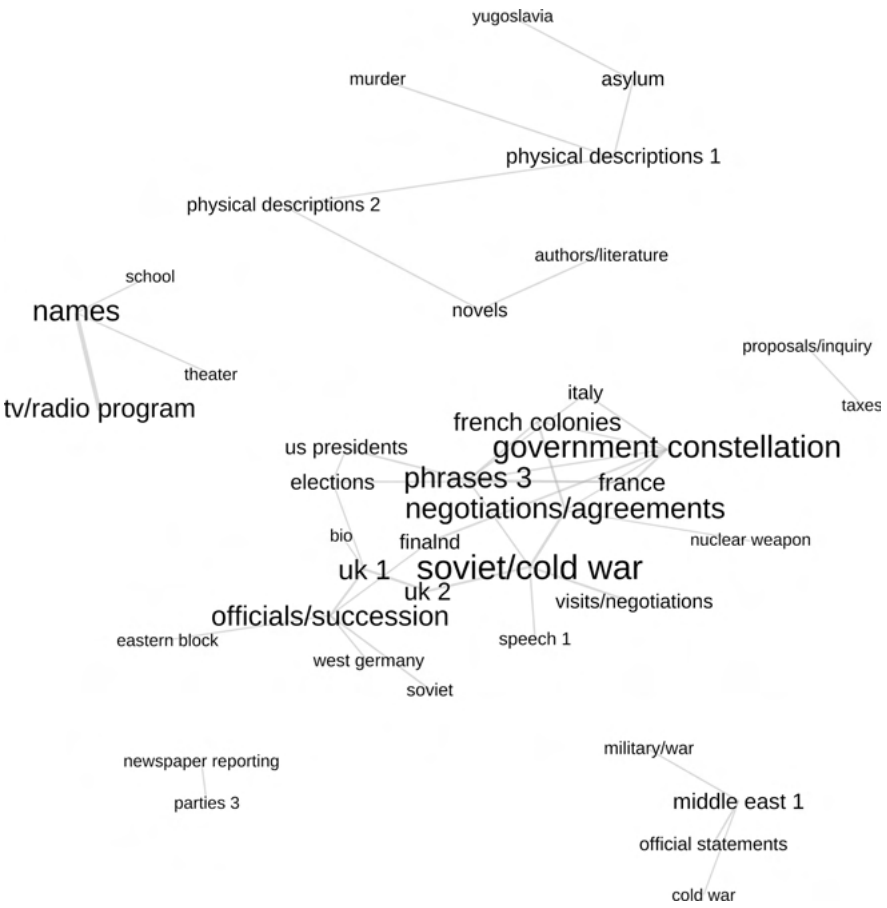


Fig. 11: Network graph of co-occurring topics 1955–1959.

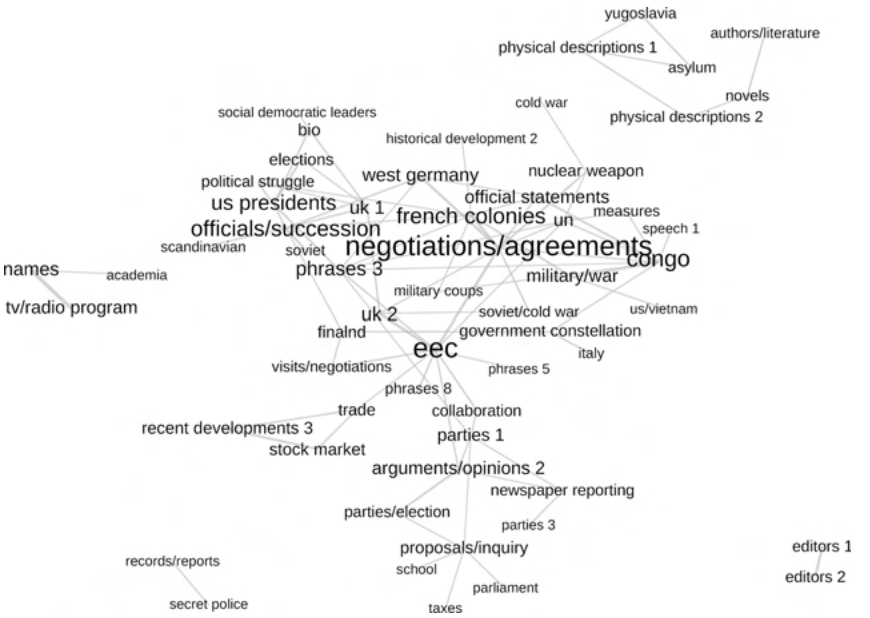


Fig. 12: Network graph of co-occurring topics 1960–1964.

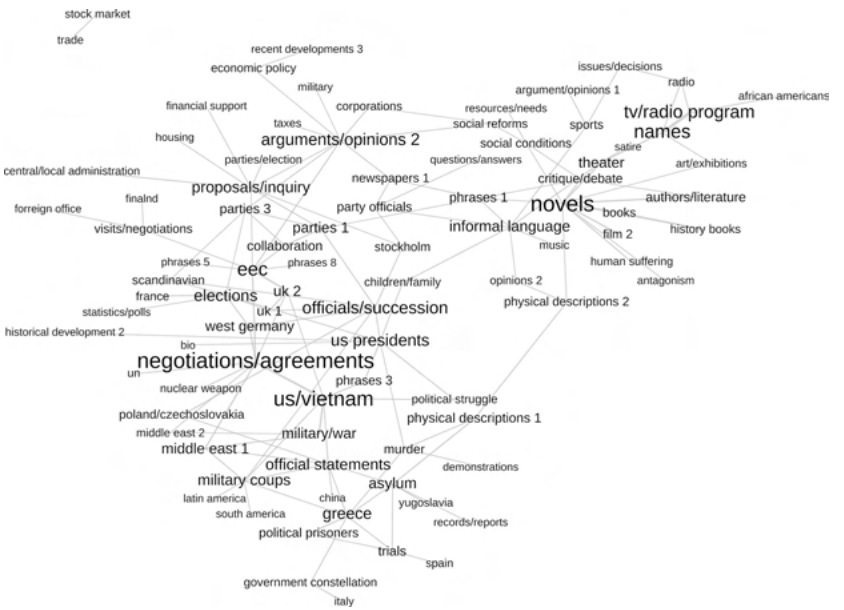


Fig. 13: Network graph of co-occurring topics 1965–1969.

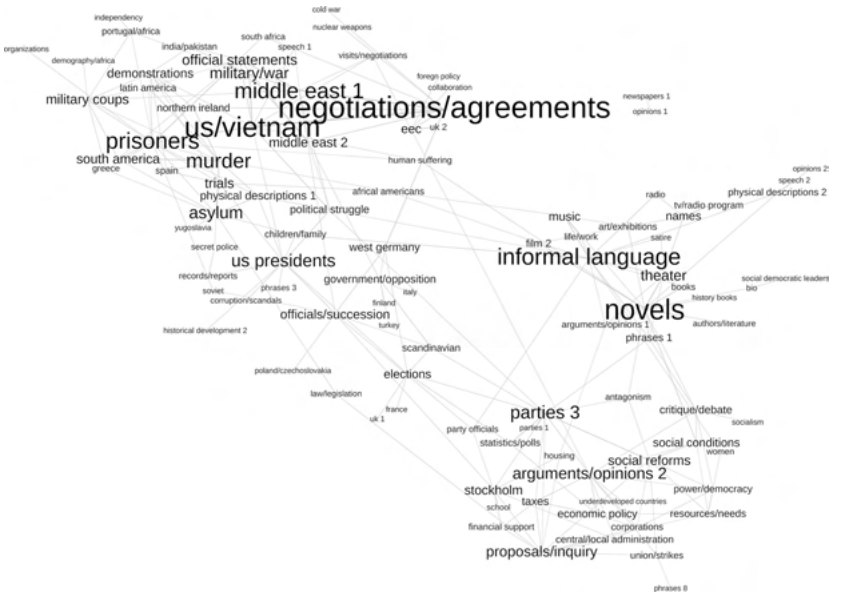


Fig. 14: Network graph of co-occurring topics 1970–1974.

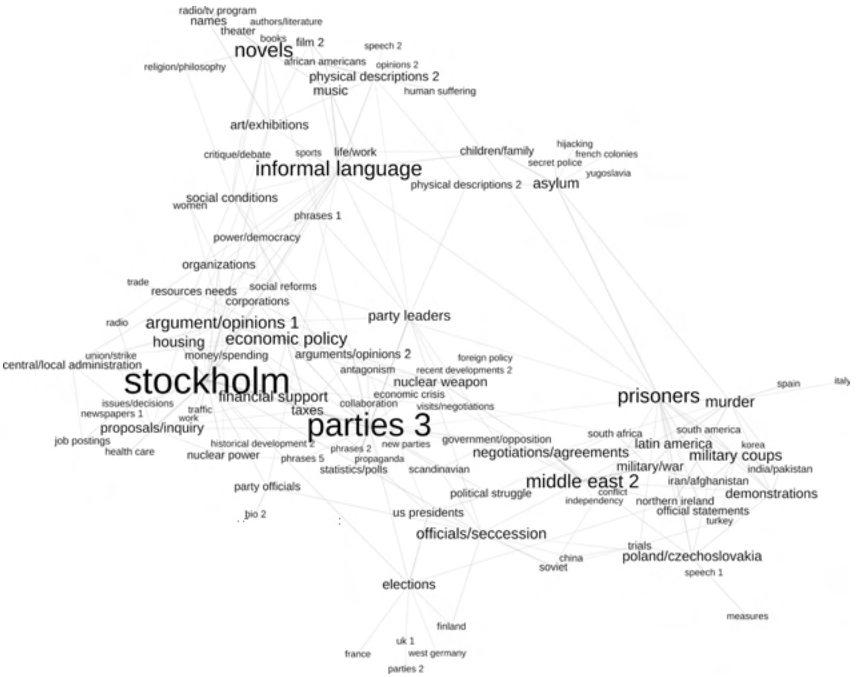


Fig. 15: Network graph of co-occurring topics 1980–1984.

Appendix 2: Co-occurring topics with 'women' over time

Topics co-occurring with 'women' (5 times or more; threshold 0.2).

1965–69
arguments/opinions 2, *should, issue, reason*, 7
phrases 1, *surely*, *seem*, *of course*, 6

1970–74
social conditions, *society*, *people*, *social*, 12
resources/needs, *resources*, *possibilities*, *needs*, 11
informal language, *they* (dom), *oneself* (sej), *me* (mej), 10
social reforms, *economic*, *social*, *reforms*, 7
argument/opinions 1, *think, know, consider*, 6
children/family, *children*, *family, parents*, 5
life/work, *people*, *work*, *life*, 5

1975–79
informal language, 28
arguments/opinions 1, 24
life/work, 20
social conditions, 15
organizations, *organizations*, *members*, *activity*, 13
resources/needs, 12
work, *work*, *time*, *working hours*, 10
critique/debate, *article*, *analysis*, *debate*, 9
phrases 1, 8
arguments/opinions 2, 7
novels, *reality*, *life*, *novel*, 7
art/exhibitions, *exhibition*, *art*, *museum*, 6
party officials, *chairman, vice, parliamentarian*, 6
children/family, 5
opinions 1, *opinion*, *view*, *position*, 5
power/democracy, *power*, *democracy*, *politician*, 5

1980–84
informal language, 14
social conditions, 13
arguments/opinions 1, 12
organizations, 12
stockholm, *stockholm*, *city hall*, *municipality*, 6
novels, 6

1985–89
social conditions, 8
statistics/polls, *percent*, *shows*, *number*, 8
arguments/opinions 1, 7
central/local administration, *government, state, municipality*, 6
informal language, 6

Bibliography

Referenced newspaper articles from *Aftonbladet* and *Dagens Nyheter* are findable through the Swedish National Library's search interface for Swedish digitized newspapers (https://tidningar.kb.se/). To be able to read the articles you have to have access to a computer with license access provided by the National Library.

Bjereld, U. & Demker, M., 1968: När allting började (Stockholm: Hjalmarson & Högberg, 2018).

Blei, D., Carin, L. & Dunson, D, 'Probabilistic topic models', IEEE Signal Processing Magazine 27:6 2012.

Blei, D., Ng, A. Y. & Jordan, M. I., 'Latent Dirichlet Allocation', The Journal of Machine Learning Research 3:1 2003.

Boréus, K., Högervåg: Nyliberalismen och kampen om språket i svensk debatt 1969–1989 (Stockholm: Tidens förlag, 1994).
Djerf-Pierre, M. & Weibull, L., Spegla, granska, tolka: Aktualitetsjournalistik i svensk radio och TV under 1900-talet (Värnamo: Fälth & Hässler, 2001).
Ekecrantz, J. & Olsson, T., Det redigerade samhället: Om journalistikens, beskrivningsmaktens och det informerade förnuftets historia (Stockholm: Carlsson, 1994).
Frenander, A., Debattens vågor: Om politisk-ideologiska frågor i efterkrigstidens svenska kulturdebatt (Göteborg: Institutionen för idé- och lärdomshistoria, 1999).
Gustafsson, K. E. & Rydén, P., A History of the Press in Sweden (Göteborg: Nordicom, 2010).
Hurtado Bodell, M., Norén, F., Edoff, E., Jarlbrink, J. & Magnusson, M., 'Curating the Swedish National Newspaper Corpus', draft, 2020.
Ifversen, J., 'About Key Concepts and How to Study Them', Contributions to the History of Concepts, 6:1, Summer 2011: 65–88.
McCallum, A. K., 'MALLET: A Machine Learning for Language Toolkit', http://mallet.cs.umass.edu, 2002.
Möller, T., Svensk politisk historia: Strid och samverkan under tvåhundra år (Lund: Studentlitteratur, 2019).
Palonen, K., 'Four Times of Politics: Policy, Polity, Politicking, and Politicization', Alternatives, 28 2003: 171–186.
Palonen, K., 'Two Concepts of Politics', Distinktion: Scandinavian Journal of Social Theory, 7:1 2006: 11–25.
Petersson, O. & Carlberg, I., Makten över tanken: En bok om det svenska massmediesamhället (Stockholm: Carlsson, 1990).
Riegert, K. & Roosvall, A., 'Cultural Journalism as a Contribution to Democratic Discourse in Sweden', In N. Nørgaard Kristensen & K. Riegert (eds.), Cultural Journalism in the Nordic Countries (Göteborg: Nordicom, 2017).
Stråth, B., 'Ideology and Conceptual History', In M. Freeden, L. T. Sargent & M. Stears (eds.), The Oxford handbook of political ideologies (Oxford: Oxford University Press, 2013).
Topic model: https://github.com/welfare-state-analytics/papers/tree/main/the-transformation-of-the-political-in-post-war-sweden
Östberg, K., 'Sweden and the Long '1968': Break or Continuity?', Scandinavian Journal of History 33:4 2008.

List of Contributors

Amstutz Irene
Schweizerisches Wirtschaftsarchiv
University of Basel
Basel, Switzerland
irene.amstutz@unibas.ch

Beavan David
The Alan Turing Institute
London, UK
DBeavan@turing.ac.uk
ORCID: 000-0002-0347-6659

Beelen Kaspar
The Alan Turing Institute
London, UK
kbeelen@turing.ac.uk
ORCID: 0000-0001-7331-1174

Bekesi Janos
University of Wien
Vienna, Austria
janos.bekesi@univie.ac.at

Borg Alexandra
Department of Scandinavian Languages
Uppsala University
Uppsala, Sweden
alexandra.borg@nordiska.uu.se

Bunout Estelle (editor)
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
Esch sur Alzette, Luxembourg
estelle.bunout@uni.lu
ORCID: 0000-0003-1009-3426

Clavert Frédéric (editor)
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
Esch sur Alzette, Luxembourg
frederic.clavert@uni.lu
ORCID: 0000-0002-0237-2532

Edoff Erik
Department of Culture and Media Studies
Umeå University
Umeå, Sweden
erik.edoff@umu.se

Ehrmann Maud (editor)
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
maud.ehrmann@epfl.ch
ORCID: 0000-0001-9900-2193

Gaillard Claire-Lise
Centre d'Histoire du XIXe siècle
Université Paris 1 - Panthéon-Sorbonne
Paris, France
Claire-Lise.Gaillard@u-paris.fr

Guilbaud Perez Malorie
Centre de Recherche Interdisciplinaire en Histoire, Histoire de l'Art et Musicologie
University of Poitier
Poitier, France
malorie.perez@ac-poitiers.fr

Hanzig Christoph
Hannah Arendt Institute for Totalitarianism Studies
Technical University of Dresden
Dresden, Germany
christoph.hanzig1@mailbox.tu-dresden.de

https://doi.org/10.1515/9783110729214-020

Hosseini Kasra
Departement of Earth Sciences
University of Oxford
Oxfortd, UK
kasra.hosseinizad@earth.ox.ac.uk

Jarlbrink Johan
Department of Culture and Media Studies
Umeå University
Umeå, Sweden
johan.jarlbrink@umu.se

Kergomard Zoé
Department of Modern History
University of Zurich
Zurich, Switzerland
zoe.kergomard@hist.uzh.ch
ORCID: 0000-0002-9184-7738

Korman Rémi
Centre d'études sociologiques et politiques Raymond Aron
Ecole des Hautes Etudes en Sciences Sociales
Paris, France
remi.korman@ehess.fr

Kovarova-Simecek Monika
Media and Digital Technologies
St. Pölten University of Applied Sciences
St. Pölten, Austria
monika.kovarova-simecek@fhstp.ac.at

Kreyenbühl Elias
Zentralbibliothek Zürich
Zurich, Switzerland
Elias.Kreyenbuehl@bs.ch
ORCID: 0000-0002-7893-9873

Krivulskaya Suzanna
History Department
California State University San Marcos
San Marcos, USA
skrivulskaya@csusm.edu

Langlais Pierre-Carl
Numapresse project
Université Paul Valéry
Montpellier, France
(now at OpinionScience, France)
pierrecarl.langlais@gmail.com
ORCID: 0000-0001-9035-1127

Lawrence Jon
History Department
University of Exeter
Exeter, UK
J.Lawrence3@exeter.ac.uk

Måns Magnusson
Department of Statistics
Uppsala University
Uppsala, Sweden
mans.magnusson@statistik.uu.se

McDonough Katherine
The Alan Turing Institute
London, UK
kmcdonough@turing.ac.uk

Munke Martin
Saxon State and University Library Saxon State and University Library
Dresden, Germany
Martin.Munke@slub-dresden.de

Norén Fredrik
Digital Humanities Center (Humlab)
Umeå University
Umeå, Sweden
fredrik.noren@umu.se

Oberbichler Sarah
Department of Contemporary History
University of Innsbruck
Innsbruck, Austria
sarah.oberbichler@uibk.ac.at
ORCID: 0000-0002-1031-2759

Paju Petri
School of History, Culture and Arts Studies
University of Turku
Turku, Finland
petpaju@utu.fi

Pfanzelter Eva
Department of Contemporary History
University of Innsbruck
Innsbruck, Austria
Eva.Pfanzelter@uibk.ac.at

Rantala Heli
Department of European and World History
University of Turku
Turku, Finland
hemara@utu.fi

Reisacher Martin
Basel University Library
Basel, Switzerland
martin.reisacher@unibas.ch

Resch Claudia
Austrian Centre for Digital Humanities and Cultural Heritage
Austrian Academy of Sciences
Vienna, Austria
claudia.resch@oeaw.ac.at

Robinet Francois
Institut d'Etudes Culturelles et Internationales
Université de Versailles Saint-Quentin-en-Yvelines
Saint Quentin-en-Yvelines, France
francois.robinet2@uvsq.fr

Salmi Hannu
Department of European and World History
University of Turku
Turku, Finland
hansalmi@utu.fi

Thoß Michael
Hannah Arendt Institute for Totalitarianism Studies
Technical University of Dresden
Dresden, Germany
michael.thoss1@mailbox.tu-dresden.de

Tolfo Giorgia
British Library
London, UK
giorgiatolfo@bl.uk
ORCID: 0000-0002-2821-4049

Torget Andrew
Department of History
University of North Texas
Denton, USA
torget@unt.edu

Vane Olivia
British Library
London, UK
olivia.vane@bl.uk
ORCID: 0000-0002-3777-4910

von Waldkirch Tobias
Department of Languages and Literatures
University of Basel
Basel, Switzerland
tobias.vonwaldkirch@unibas.ch

Wevers Melvin
Faculty of Humanities
University of Amsterdam
Amsterdam, Netherlands
m.j.h.f.wevers@uva.nl
ORCID: 0000-0001-8177-4582