RESH Discussion Papers

No. 8 / 2022

Lino Wehrheim/Bernhard Liebl/
Manuel Burghardt

Extracting Textual Data from Historical Newspaper Scans and its Challenges for "Guerilla-Projects"

Universität Regensburg
FAKULTÄT FÜR PHILOSOPHIE,
KUNST-, GESCHICHTS- UND
GESELLSCHAFTSWISSENSCHAFTEN

REGENSBURG ECONOMIC AND SOCIAL HISTORY (RESH)

Discussion Paper Series

Edited by

Prof. Dr. Mark Spoerer and PD Dr. Tobias A. Jopp

Processed by

PD Dr. Tobias A. Jopp

University of Regensburg

Faculty of Philosophy, Art History, History and Humanities

Department of History

Chair for Economic and Social History

The RESH Papers are intended to provide economic and social historians and other historians or economists at the University of Regensburg whose work is sufficiently intersecting with economic and social history with the possibility to circulate their work in an informal, easy-to-access way among the academic community. Economic and social historians from outside the University of Regensburg may also consider the RESH Papers as a means to publish informally as long as their work meets the academic standards of the discipline. Publishable in the RESH Papers is the following: *Work in progress* on which the author(s) wish to generate comments by their peers before formally publishing it in an academic journal or book; and English translations of already published, yet *non-English* works that may be of interest beyond the originally targeted (German, French, and so on) audience. Authors interested in publishing in the RESH Papers may contact one of the editors (Mark.Spoerer@ ur.de; Tobias.Jopp@ur.de).

Cover page photo: An announcement Berliner Börsenzeitung's evening-issue of 16 March 1872. Source: https://dfg-viewer.de/show?tx_dlf%5Bdouble%5D=0&tx_dlf%5Bid%5D=https%3A%2F%2Fcontent.staatsbibliothek-berlin.de%2Fzefys%2FSNP2436020X-18720316-1-0-0-0.xml&tx_dlf%5Bpage%5D=10&cHash=ee249263aed0998b7f6b132adefc1787
Final page photos: Own material.

Extracting Textual Data from Historical Newspaper Scans and its Challenges for "Guerilla-Projects"

Lino Wehrheim / Bernhard Liebl / Manuel Burghardt[+]

Abstract: In 2022, it is a common place that digital historical newspapers (DHN) have become increasingly available. Despite the undeniable progress in the supply of DHN and the methods to perform rigorous quantitative analysis, however, working with DHN still poses various pitfalls, especially when scholars use data provided by third parties, such as libraries or commercial providers. Reporting from a current project, we want to share our experiences and communicate the various problems we faced while working with DHN. After a short project summary, we present the main problems that we faced in our project and that we think might also be relevant for other scholars, particularly those who work in small research groups. We arrange these problems according to an archetype workflow, which is divided into the three steps of corpus acquisition, corpus evaluation, and corpus preparation. By raising some red flags, we want to call attention to what we think common DHN related problems, to raise awareness for potential pitfalls, and, this way, to provide some guidelines for scholars who consider using DHN for their research.

Keywords: Historical newspapers, OCR, layout detection, trouble shooting

JEL classification: C80, C02

[+] Contact 1 (corresponding author): Dr. Lino Wehrheim, University of Regensburg, Department of History, Economic and Social History, 93040 Regensburg; email: Lino.Wehrheim@ur.de. Contact 2: Bernhard Liebl, University of Leipzig, Institute of Computer Science, Computational Humanities, Augustusplatz 10, 04109 Leipzig; email: liebl@informatik.uni-leipzig.de. Contact 3: Prof. Dr. Manuel Burghardt, University of Leipzig, Institute of Computer Science, Computational Humanities, Augustusplatz 10, 04109 Leipzig; email: bernhard.burghardt@informatik.uni-leipzig.de.

Extracting Textual Data from Historical Newspaper Scans

and its Challenges for "Guerilla-Projects"

1. Introduction

In 2022, it is a common place that digital historical newspapers (referred to as DHN in the following) have become increasingly available (see, e.g., Beals and Bells (2020)). Accordingly, this type of historical source has received wide-spread interest by scholars, particularly from the humanities and the social sciences.[1] Large flagship projects such as Impresso, Newseye, or Oceanic Exchanges impressively illustrate the potential of this line of research. With libraries offering access to their collections, DHN also attract the attention of smaller research groups or individual researchers who want to use them for specific research questions. Despite the undeniable progress in the supply of DHN and the methods to perform rigorous quantitative analysis, however, scholars still have to deal with various DHN specific pitfalls. Poor OCR quality, missing pages, or missing article segmentation slow down or even prevent ambitious research projects from digital and economic history or cultural analytics. Many of these problems only become apparent after the project has begun, leading to oftentimes high additional costs or, in the worst case, a premature termination of the project. This can be particularly challenging for smaller research groups who are usually less well equipped in terms of staff and/or funding than the big consortia. In any case, when it comes to drafting the work programme and the time and personnel requirements for a novel DHN project, it is helpful to anticipate possible pitfalls in order to be able to plan alternative approaches at an early stage. By sharing the experiences that we made in a current DHN project, we want to provide exactly this kind of information, which we hope will benefit other scholars, especially those working in small "guerrilla-like" projects without substantial budgets. After a short project presentation, we present the main problems we faced in our project and which we think might also be relevant for other scholars. We arrange these problems according to an archetype workflow, which is divided into the three steps of corpus acquisition, corpus evaluation, and corpus preparation. We summarize this workflow in Table 1 at the end of the paper.[2]

---

[1] Among others, there were panels on DHN held at the last two DHd conferences, see Herzgsell et al. (2019) and Bunout et al. (2020).

[2] While DHN problems might be universal, solutions to solve them are highly project depended. Therefore, we outline our solutions only briefly in this paper.

2. More than a feeling: a project on historical financial sentiment

The DHN project from which we draw our experiences is located at the intersection of (media) history, behavioural finance, and computational humanities.[3] It is motivated by the observation that "soft" variables such as moods and stories influence the economy in general and financial markets in particular. In this respect, Nobel laureates George Akerlof and Robert Shiller promoted Keynes' concept of "animal spirits" (Akerlof and Shiller 2009, Keynes 1936) that is, non-economic, sometimes irrational motives, which influence people's behaviour and, thus, the economy. In this research area, newspapers are a source that is commonly used, particularly for capturing economic narratives (Shiller 2020; Wehrheim 2021) and measuring investor sentiment, that is, the mood of stock market participants (Raimondo 2019). This line of research started with the seminal work by Tetlock (2007) who studied the *Wallstreet Journal* to measure market sentiment. For a more recent example, see Frydman et al. (2021) and, in a more historical context, Hanna et al. (2020) and Kabiri et al. (2022).

Following this line of research, we launched an interdisciplinary project to study the role of news stories and financial sentiment at the Berlin stock exchange, which was the most important stock exchange in the German Empire, between 1872 and 1930. The aim of this project is to study how moods and stories, as expressed in a DHN, have influenced people's economic expectations. Specifically, we use the *Berliner Börsen-Zeitung* (BBZ), one of Berlin's leading newspapers, to extract data on the sentiment and the narratives expressed by financial journalists. Therefore, we are particularly interested in the BBZ's daily stock market report which give a verbal description of how the market has performed (see Figure 1 for an example report), particularly describing the market sentiment *(Marktstimmung)*, but also in other parts that provide information on further, non-economic topics. Using approaches such as named entity recognition, topic modelling, and sentiment analysis, we want to gain insights in the way the BBZ has reported on the socio-economic development of the German Empire, particularly on its stock market.

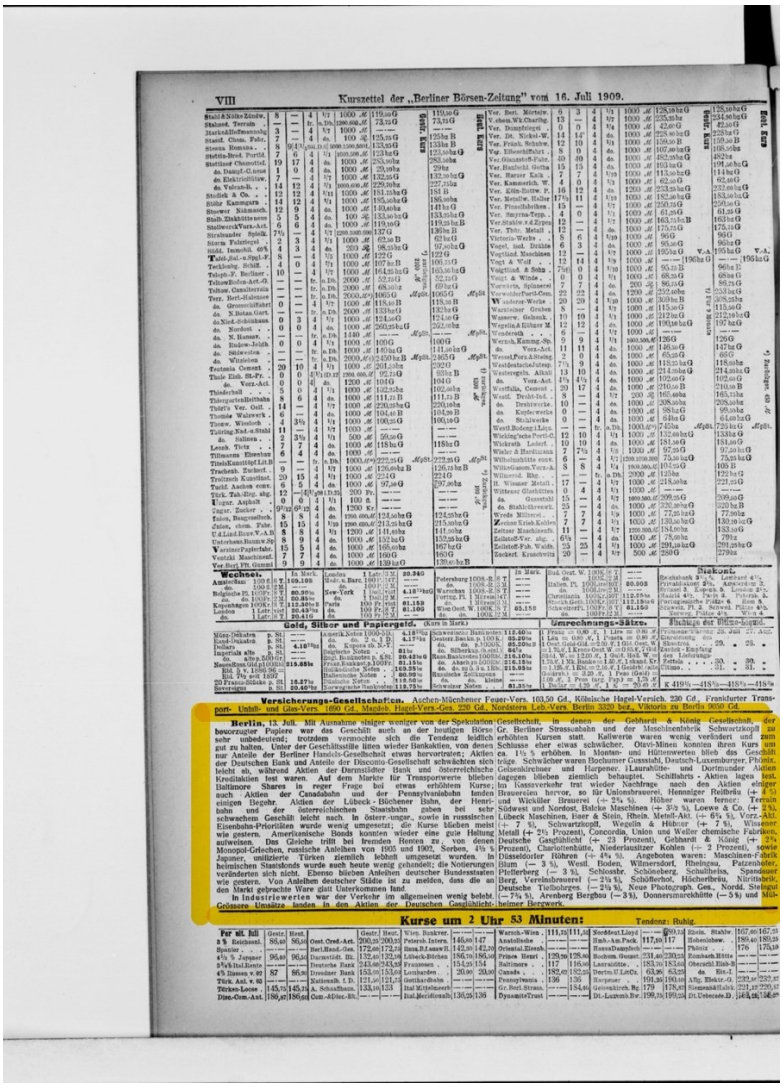
[3] https://media-sentiment.uni-leipzig.de. This project is part of the DFG priority programme 1859 "Experience and Expectation". See https://www.experience-expectation.de.

3. Obstacles and Workflow

The BBZ, which was published twice a day, is provided by the *Staatsbibliothek zu Berlin* (Stabi) for the period between 1872 and 1930. During the preparation phase of the project, we agreed that we would gain access to the whole BBZ collection. But even this first step – the *corpus acquisition* – turned out to be much less straightforward than we had expected. At first, we realized that, at that moment, the quality of the OCRed data provided by the Stabi was too low for our purposes, which is why we decided to perform our own OCR procedure to ensure sufficient OCR quality.[4] For this, we needed the original page images, what posed the problem of data transfer as the high amount of data volume did not allow for a network transfer. The solution consisted of a combination of cluster storage and physical transfer of hard drives, which was delayed by the Corona pandemic. The high data volume resulted from the fact that the Stabi provides the BBZ in a high resolution (20 MP) JPG format, which we needed to perform certain tasks.

[4] In the meantime, the Stabi has improved the OCR quality of much of its DHN stock.

Figure 1: Example Stock Market Report



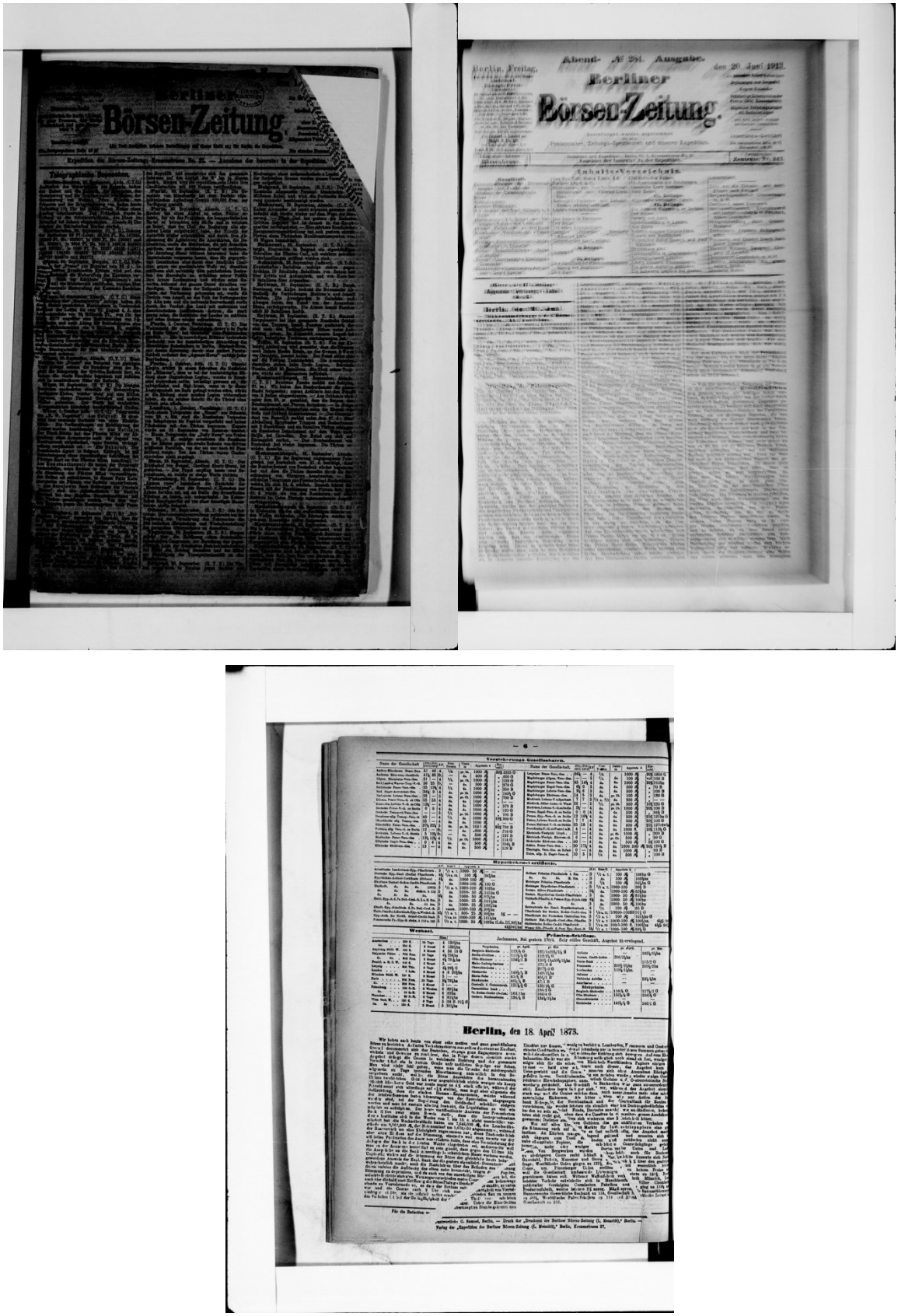
VIII

Kurszettel der „Berliner Börsen-Zeitung" vom 16. Juli 1909.

Source: Staatsbibliothek zu Berlin.

Although this point might seem trivial, it is important to consider it as a potentially time-consuming step, especially if, as in our case, one is required to transfer more data than initially planned. The lessons learned from this step were to double check the OCR quality, as standard OCR procedures do not automatically suffice the requirements of research projects that have text mining in mind. Furthermore, one should also check the availability of high-resolution primary sources and plan for data transfer of high amounts of data.

Figure 2: Examples of damaged scans



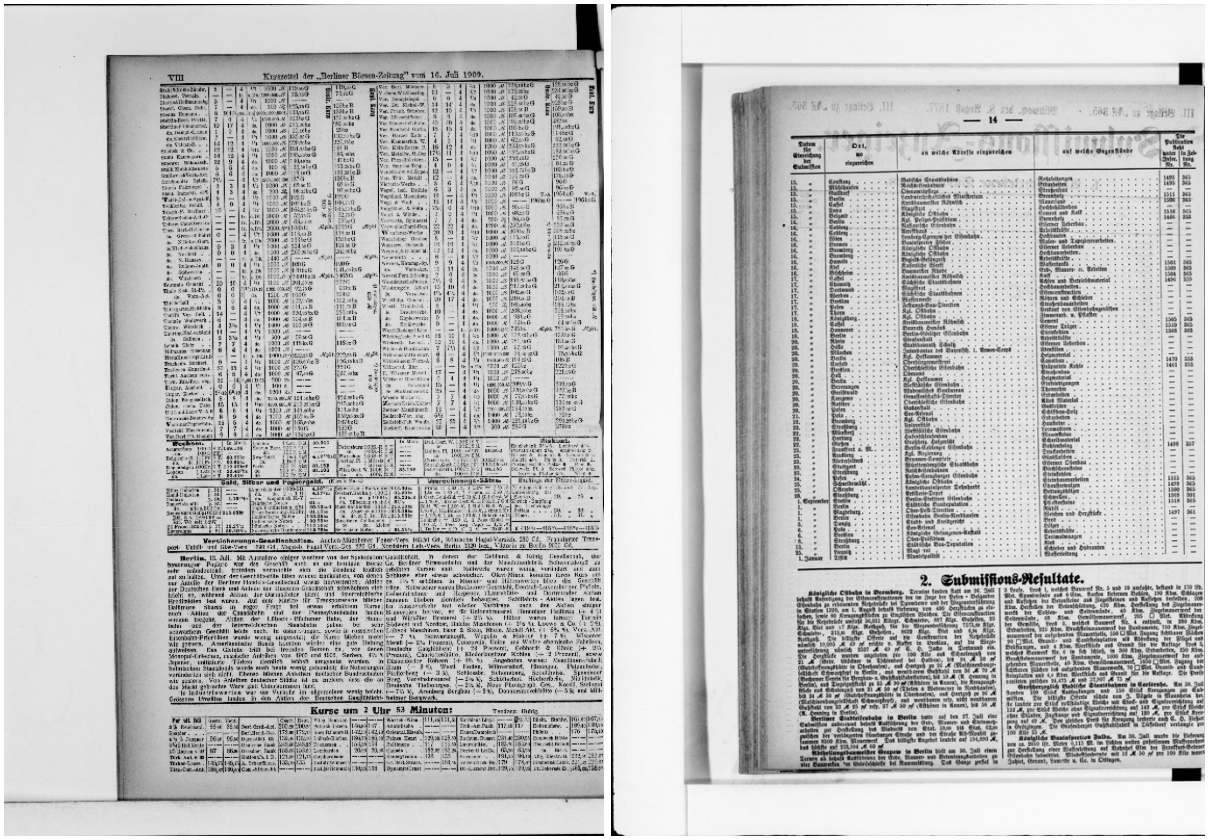
Source: Staatsbibliothek zu Berlin.

After having gained access to the available BBZ collection, it became necessary to evaluate its quality in a second step – *corpus evaluation* – as it turned out that the collection was far from being complete. Here, we faced problems such as missing and duplicate issues/pages, incomplete pages resulting from scan errors (see examples in Figure 2), and pages from different issues or even newspapers. Furthermore, we were confronted with messy meta data, for example erroneous date specifications and missing information on whether a certain issue was published as morning or evening edition – information that is relevant for many subsequent steps. All these issues required time consuming data cleaning, which partly involved non-trivial solutions such as layout detection of certain pages. Furthermore, it became evident that certain vital parts of the BBZ were missing, leading to considerable gaps in our time series. This concerns the *Kurszettel*, that is, the part of the BBZ containing stock market prices and, most importantly for our project, the daily stock market report. This problem was aggravated as the gaps mainly date in the 1870s and 80s, a particularly important period in German (financial) history. This is why we filled the gaps by using columns from a different newspaper, the *Vossische Zeitung*, which not only poses new troubles of its own but also doubles the problems of the third step, which we will address in a moment. The key lesson from this step is to rigorously evaluate one's corpus by drawing as many samples as possible, both regarding data and metadata, in order to identify potential gaps in the data right from the beginning.

The problems arising in the third phase – the *corpus preparation* – can be summarized under the term of "pre-preprocessing", that is, steps that had to be carried out before the preprocessing procedure commonly performed in text mining – lemmatizing, stop word removal, POS tagging, NER, etc. –, could be applied. The steps taken in this phase were the most expensive, both in terms of time and technical efforts. First and foremost, we had to run a new OCR procedure due to an insufficient OCR quality. Besides standard OCR problems such as poor scan quality, we were facing further difficulties, some being typical for 19th century newspapers, others being more BBZ specific. These difficulties, including, for example, a mixture of Gothic and Antiqua types, numerous non-German words, composita and technical terms, incoherent spelling, varying typesets, and idiosyncratic symbols, implied that existing standard OCR tools were insufficient. Additionally, we needed to separate page scans into individual articles, which again posed several issues, as historical newspapers normally have only very few layout markers, exhibit changing, complex, and inconsistent layouts, and contain also a lot of "noise", for example tables, small ads, and advertisement. Figure 3 illustrates the

challenge of identifying the correct article. Even though it is easy to define certain defining layout characteristics, such as “two columns” or “table on the same page”, the picture on the right-hand-side shows that relying only on layout-based identification heuristics poses the risk of producing many false positives. This why we used a combination of layout- and textual markers, such as date references.

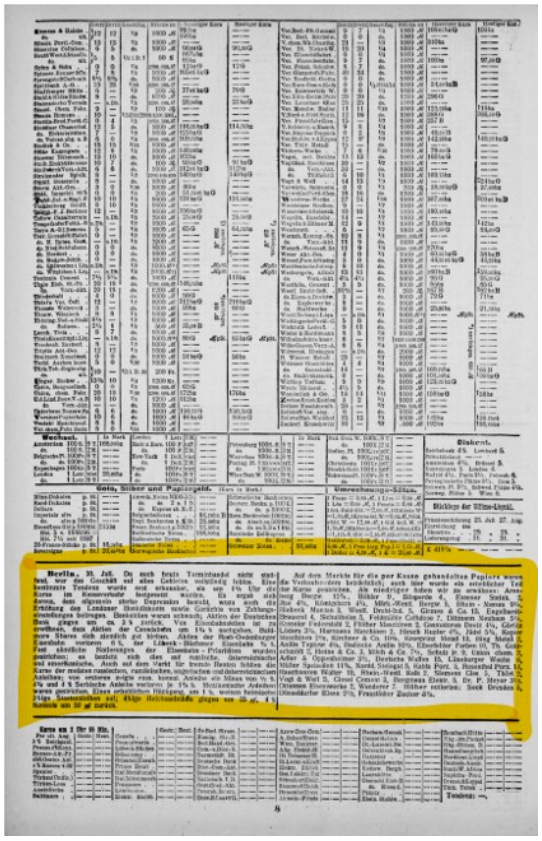
Figure 3: Confounding segments

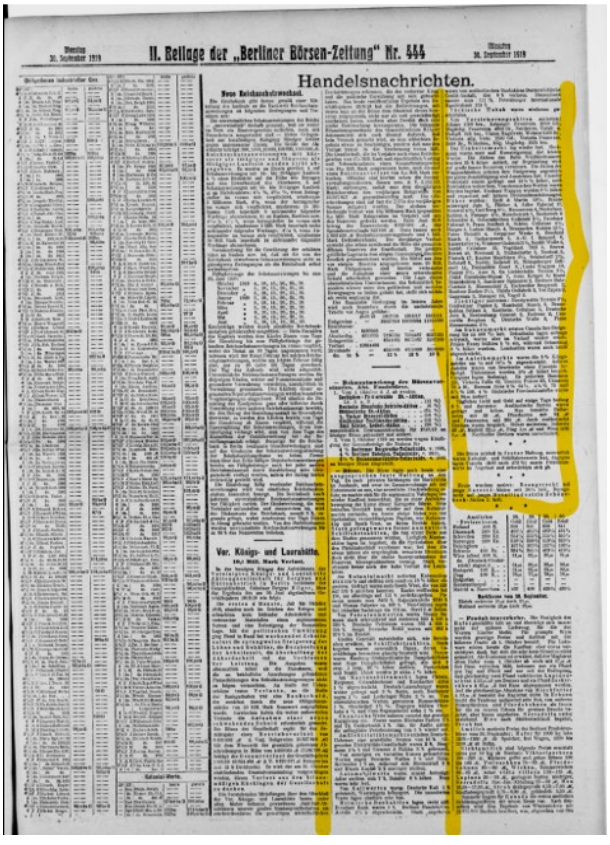


Correct segment False but visually similar segment

Source: Staatsbibliothek zu Berlin.

Figure 4: Types of layouts used for market reports





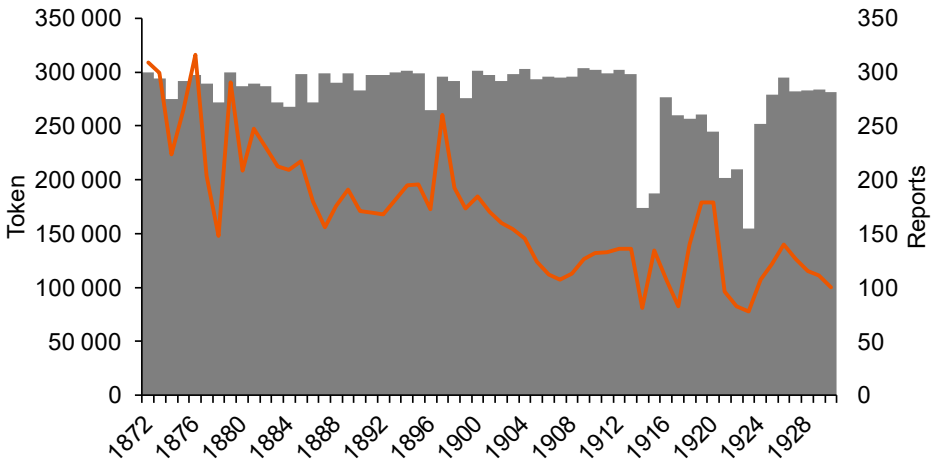
Source: Staatsbibliothek zu Berlin.

Having developed our own OCR pipeline *Origami* (Liebl and Burghardt 2020), we realized that we were facing a fundamental trade-off between OCR quality and processing time, the latter, despite of several optimizations, being rather high: even on modern CPU- and GPU clusters, the 9-step OCR computation took more than 10 minutes per page, with line-based OCR alone accounting for over 60% of that time. Therefore, we decided to concentrate on just two parts of the BBZ in the remainder of the project: title pages and the daily stock market reports. While the former was easy to identify, extracting the latter posed several challenges. First of all, we had to identify the right issue, that is, the evening issue. Second, we had to find the right page on which the column was located. Third, the layout of the reports changed several times, form a standardized two column text up until World War I to an inconsistent report, sometimes spanning one, sometimes spanning several columns. Figure 4 provides examples of the different types of layouts that were used for the stock market reports.

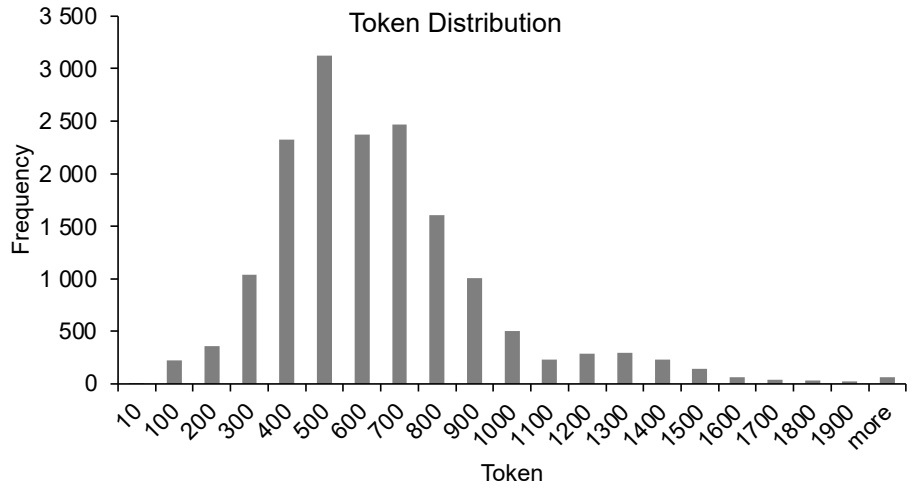
We identified three general types of reports, which we used to annotate about 10.000 pages by hand in order to train a classifier and building several rule-based extraction scripts. This way, we were able to extract about 16.400 market reports. Panel a of Figure 5 shows the number of extracted reports (grey bars, right-hand scale) and the number of tokens (orange line, left-hand scale) per year. The low number of reports during WWI and the early 1920s results from the fact that trading at the Berlin Stock Exchange was official prohibited during these periods. Panel b shows the distribution of report length in token. The median report has a length of 540 token.

Figure 5: Number of reports and token per year

a



b



Source: Own calculations.

After separating the title pages and markets reports (and hand-labelling about 2500 more pages to train a classifier to differentiate between two kinds of editions in the title page only to detect that the source data had more gaps) and performing OCR on these segments, we were finally confronted with several problems regarding the postprocessing of these texts. These included varying spellings, a plethora of now outdated names (e.g. *Planiawerke, Berzelius* or *Buschtiehrader Eisenbahn*), outdated terms (e.g. *Transkaukasier*), complex German compound words that are not included in standard dictionaries, such as *Beunruhigungsbacillus*, *Gasglühlichtgesellschaft*, or *Sprengbombenanarchismus*, inconsistent abbreviations, and false word divisions. We approached automated postprocessing by mapping terms to their closest (Levenshtein) pendant in an infinitely sized German dictionary (generated dynamically from a custom-built grammar and 1.3 million dictionary terms), however abandoned our efforts due to a high false-positive rate in terms that it got incorrectly corrected as OCR errors. Additional experiments with the German GottBERT also yielded no usable results. As this third step is the most complex, it is difficult to define a single key takeaway. In general, one should keep in mind that data provided by a third party is rarely in a text-mining ready state and that it will be necessary to spend a considerable amount of time on processing the corpus.

4. Conclusion

This list of problems and challenges might, at first glance, daunt scholars of starting a novel DHN based project, which is of course the contrary of what we intended with this paper. Instead, we wanted to call attention to what we think common DHN related problems, to raise awareness for potential pitfalls and, this way, to provide some guidelines for scholars who consider using DHN for their research. To send also a positive note: Although it was sometimes difficult and time-consuming, so far, we were able to find a solution to all problems mentioned above. The quality of our corpus is now sufficiently high, allowing us to use for example transformer-based language models to extract aspect-level sentiment data. We wish, however, that we had known about these issues in advance. Our conclusion is that, even in the year 2022, working with digital historical corpora is not all that simple and seems to require building custom data processing pipelines in order to tackle challenges that are often highly specific to the domain, data, and project.

Bibliography

Akerlof, George A., and Robert J. Shiller. 2009. *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton.

Beals, M H, and Emily Bell (eds.). 2020. *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough. https://www.digitisednewspapers.net/.

Bunout, Estelle, Marten Düring, Jana Keck, Bernhard Liebl, Sarah Oberbichler, Thorsten Roeder, and Lino Wehrheim. 2020. "What's in the News? (Erfolgs-)Rezepte für das wissenschaftliche Arbeiten mit digitalisierten Zeitungen." In: Christof Schöch (ed.). 2020. *DHd 2020: Spielräume. Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts, 7. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V*.

Frydman, Roman, Nicholas Mangee, and Josh Stillwagon. 2021. „How Market Sentiment Drives Forecasts of Stock Returns". *Journal of Behavioral Finance* 22 (4): 351–67.

Hanna, Alan J., John D. Turner, and Clive B. Walker. 2020. "News Media and Investor Sentiment during Bull and Bear Markets." *The European Journal of Finance* 26 (14): 1377–95.

Herzgsell, Teresa, Dario Kampkaspar, Jörg Lehmann, Clemens Neudecker, Claudia Resch, and Nanette Rißler-Pipka. 2019. "Zeitungen und Zeitschriften als multimodale, digitale Forschungsobjekte: Theorien und Methoden." In: Patrick Sahle (ed.). 2019. *DHd 2019 Digital Humanities: Multimedial & Multimodal. Konferenzabstracts,* Frankfurt am Main.

Kabiri, Ali, Harold James, John Landon-Lane, and Rickard Nyman. 2022. „The Role of Sentiment in the US Economy of the 1920s". *The Economic History Review* (Online First).

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London.

Liebl, Bernhard, and Manuel Burghardt. 2020. "From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline." *Proceedings of the 1st Workshop on Computational Humanities Research (CHR)*.

Raimondo, Carlo. 2019. "The Media and the Financial Markets: A Review." *Asia-Pacific Journal of Financial Studies* 48 (2): 155–84.

Shiller, Robert J. 2020. *Narrative Wirtschaft: Wie Geschichten Die Wirtschaft Beeinflussen - Ein Revolutionärer Erklärungsansatz*. Kulmbach.

Tetlock, Paul C. 2007. „Giving Content to Investor Sentiment: The Role of Media in the Stock Market". *The Journal of Finance* 62, (3): 1139–68.

Wehrheim, Lino. 2021. *Im Olymp Der Ökonomen. Zur Öffentlichen Resonanz Wirtschaftspolitischer Experten von 1965 Bis 2015*. Tübingen.

Table 1					
1. Corpus acquisition/storage		**2. Corpus evaluation**		**3. Corpus preparation**	
Problems	Solutions	Problems	Solutions	Problems	Solutions
High data volume: - no network transfer	Hard drives, cluster storage	Messy collection: - missing & duplicate issues/pages - scan errors - erroneous page allocations	- *ignore* - *manual search* - *ML-classifiers* - binary comparisons - perceptual hashing	OCR/layout detection: - failure of standard tools - mixture of Fraktur/Antiqua scripts - special typography (eg I vs. J) - multilingual vocabulary, distinctive vocabulary (financial terms) - bold print, letter spacing - multiple typesets - special symbols (eg Reichsmark) - low paper and print quality - several layout revisions - unstructured articles, few article markers (eg headlines) - tables, advertisements - combined model or separate models? - number of training lines - model evaluation	
Low OCR quality, no page segmentation	- *existing tools, eg OCR-D, Transcribus* - developing new tools	Dirty meta data: - missing dates - missing information on morning vs. evening edition		Limited resources: - applying OCR/layout detection on whole BBZ collection too time-consuming	Limitation on market reports and front pages
				Extraction of market reports: - varying layouts, inconsistent positioning	- *manual search* - manual sampling and annotation - ML-classifiers
				Post processing: - single- and multi-character errors - false word divisions and assembling - technical terms, proper nouns, compound words, inconsistent abbreviations and spelling, spelling/printing errors - low matching with standard dictionaries, no automatic correction	- manual and semi-automatic correction of high-frequent terms and systematic errors

RESH Discussion Paper Series

No. 1 / 2020 *Mark Spoerer*

Did Firms Profit from Concentration Camp Labour? A Critical Assessment of the Literature

No. 2 / 2020 *Tobias A. Jopp*

A Happiness Economics-Based Human Development Index for Germany (1920-1960)

No. 3 / 2020 *Tobias A. Jopp/Mark Spoerer*

Teaching Historical Statistics: Source-Critical Mediation of Aims and Methods of Statistical Approaches in Historiography

No. 4 / 2020 *Mark Spoerer*

The Short Third Reich: On Continuities in Socio-Economic Structures between the Weimar Republic, the Third Reich and the Federal Republic

No. 5 / 2020 *Tobias A. Jopp/Mark Spoerer*

How Political Were Airbus and Boeing Sales in the 1970s and 1980s?

No. 6 / 2020 *Jonas Scherner/Mark Spoerer*

Infant Company Protection in the German Semi-Synthetic Fibre Industry: Market Power, Technology, the Nazi Government and the Post-1945 World Market

No. 7 / 2020 *Michael Buchner/Tobias A. Jopp/Mark Spoerer/Lino Wehrheim*

On the Business Cycle of Counting – or How to Quantify Quantification. An Empirical Analysis of the Application of Quantitative Methods in German Historiography

No. 8 / 2022 *Lino Wehrheim/Bernhard Liebl/Manuel Burghardt*

Extracting Textual Data from Historical Newspaper Scans and its Challenges for “Guerilla-Projects”

DOI 10.5283/epub.53259 RESH Papers – ISSN 2701-2050