

The Wartime Propaganda Puzzle

Mari Wigham¹, Rana Klein¹, Vincent Kuitenbrouwer², Marjet Brolsma², and Roeland Ordelman^{1,3}

¹ Netherlands Institute for Sound and Vision

² University of Amsterdam

³ University of Twente

1 Abstract

This paper presents the efforts of the Media War Matching project to combine the puzzle pieces of Second World War propaganda collections, namely audio recordings of radio broadcasts, transcripts of radio broadcasts and newspapers. To give researchers a richer and broader picture of propaganda dynamics in the Nazi-occupied Netherlands, the project published digitised radio transcripts in the CLARIAH Media Suite, an online platform for humanities research, so that they could be analyzed both in isolation and in combination with other wartime media collections. In this paper, we first introduce the wartime propaganda landscape, as well as the Media War project, which combined quantitative and qualitative approaches to analyzing propaganda in pro- and anti-Nazi Dutch language media. Subsequently, we discuss the Media War Matching project, which emanated from the challenges and opportunities that the Media War research team encountered during their work. Then, we relate how we published a specific set of radio transcripts in the Media Suite so that researchers could optimally search and browse them, including the development of an algorithm to automatically date the individual pages. Next, we explain how we attempted to link the audio of radio broadcasts to the relevant transcripts, using Automatic Speech Recognition (ASR). Finally, we demonstrate how the publication of propaganda collections in the CLARIAH Media Suite, despite the limitations of the source material and the lack of success of linking, still enables the identification and comparison of quantitative patterns that can act as a starting point for further qualitative analysis of narratives. In conclusion, this paper reflects on the lessons learned and aims to inspire others to digitise, publish and link media collections.

2 Introduction

The study of historical wartime propaganda is increasingly relevant in the light of current events (Connelly et al., 2019, pp. 1-12). This is underlined by the fact that national leaders embroiled in conflicts refer to incidents from previous wars in their propaganda (Wijdeven (2022)). Understanding propaganda in the past can therefore contribute to understanding the present day.

A large number of sources of World War II propaganda in the Netherlands has survived. Yet these sources are scattered over isolated silos, residing in fragmentary collections of different media types preserved by different institutions and published on different platforms. This scattering makes it hard for propaganda researchers to find sources, let alone to see the relationships between them or discover patterns. The MediaOorlog (Media War) project, financed by the Mondriaan Fonds within the programme '75 jaar Vrijheid' (75 years of freedom), focused on making various propaganda sources available to researchers to enable them to study transnational dynamics over time. To achieve these objectives researchers within this project worked with relevant collections in the digital humanities platform CLARIAH Media Suite.

During the project, it became clear that there are significant imbalances in the availability of different types of material. In Kuitenbrouwer and Brolsma (2023) the authors argue that 'these imbalances can be explained by the historical context in which these sources were created as well as by archival policies after 1945'. For example, there are far more Nazified newspaper articles available than anti-Nazi (Kuitenbrouwer and Wijfjes (2022)), and newspaper sources are available in far larger numbers than radio recordings (Kuitenbrouwer and Brolsma (2023)). These imbalances present researchers with additional challenges, as it is hard to determine if an apparent trend is meaningful, or merely an artefact of the imbalance. The limited availability of radio recordings led the authors of Kuitenbrouwer and Brolsma (2023) to argue for this imbalance to be partially addressed by the digitisation of archives of related documents such as radio transcripts and monitoring reports, and for their publication in the CLARIAH Media Suite. Additionally this could also bring together the bulky 'radio on paper' collections and the fragmented audio which has remained from the war to reconstruct the production archives of the broadcasting organisations that have been separated over the past decades (Kuitenbrouwer (2022)).

The Media War Matching project took up this call to arms by addressing the availability, findability and usability of wartime media. More specifically, the project worked to publish the BNO (Berichtendienst Nederlandsche Omroep)¹ collection in the CLARIAH Media Suite so that it could be easily searched and browsed. This collection contains transcripts of broadcasts by the BNO radio broadcaster, a radio news service which spread Nazi propaganda. These transcripts are particularly relevant to researchers as they can provide insight into the overall broadcasting strategies and also serve as a proxy to better search and analyze surviving audio recording of radio broadcasts. In a broader goal, the project sought to combine the puzzle pieces of audio recordings of radio broadcasts, their transcripts and wartime newspapers from the same date to give researchers a richer and broader picture of wartime propaganda in the Netherlands. Finally, the project looked at how analysis of patterns over these sources could offer added value for propaganda researchers.

In this paper, we first introduce the wartime propaganda landscape. Then we delve into the Media War project, the propaganda collections involved, the approaches developed for working with them, and the challenges that gave birth to the Media War Matching project. Then we describe the publication of the BNO collection, including our efforts to date the pages and to support data and tool criticism by users of this collection. Next, we explain how we attempted to link the audio of radio broadcasts to the relevant transcripts, using Automatic Speech Recognition (ASR). After that, we demonstrate how the publication of propaganda collections in the CLARIAH

¹ Nederlands Instituut voor Oorlogs-, Holocaust- en Genocidestudies, Amsterdam, archief 103 Nederlandse Omroep, inv.no, 302-460

Media Suite enables quantitative analysis. Finally, we conclude by summarizing our experiences in publishing the BNO propaganda collection, and detail both lessons learned and plans for future work.

3 Background

3.1 Multimedia wartime propaganda

During World War II, all official media in the Netherlands were under the control of the Nazis. This included the two Nazified radio stations in Hilversum, operated from March 1941 onwards by the Nederlandsche Omroep, and the existing national and regional newspapers which were brought under Nazi control after the invasion in May 1940 (Verkijk (1974), Vos (1988), Wolf and van Vree (2019)). However, there were also media available from the opposing side which challenged the Nazi propaganda narratives. Illegal newspapers were produced within the Netherlands, and the radio programmes De Brandaris and Radio Oranje were broadcast from London on the BBC transmitters to reach Dutch audiences (Sinke (2005), Sinke (2009), Winkel (1954), van den Heuvel and Mulder (1990)).

There are many imbalances in the primary sources left from the years 1940-1945 which have to be taken into consideration when analysing this propaganda battle between pro-Nazi and pro-Allied Dutch language media. The digitised war newspapers collection of the Royal Library is by far the largest propaganda archive. In the context of the Heritage of War [Erfgoed van de oorlog] project all available wartime newspapers have been digitised. The collection is available via the Delpher² portal and also in the Newspaper collection of the Media Suite.³ It consists of no fewer than 133.000 digitised newspapers segmented into 6.942.271 articles. The sheer size of this collection allows media scholars and historians to employ quantitative research methods to identify semantic patterns.

Compared to the Dutch newspaper media landscape, radio broadcasting was in various respects more dynamic due to the faster interaction that the medium offered. On both sides of the English Channel people listened intensively to the broadcasts of the adversary and produced monitoring reports that were actively used as 'ammunition' for their own propaganda. These practices continued even after the Nazi authorities confiscated wireless radio receiving sets in Occupied Netherlands in May 1943. In the last years of the war, broadcasts from London remained an important source of news about the war for the underground press. People who clandestinely listened to the BBC and Radio Oranje transcribed and reproduced the texts of these broadcasts and disseminated them in periodicals with telling titles such as *Here is the B.B.C.* (*Hier is de B.B.C.*) and *The Aether Courier* (*De Aetherkoerier*) (Kuitenbrouwer and Wijfjes, 2022, p.192).

Despite the historical importance of radio as a propaganda tool, the digitised audio archive held by The Netherlands Institute for Sound and Vision (NISV) is fragmented and relatively small. It contains 2271 audio items, of which 2040 have been digitised. Of Radio Oranje broadcasts, for instance, only 223 audio fragments remain, out of an estimated total of 3500 full broadcasts. Due to the specific circumstances of the war, including censorship, a relatively large amount of paper radio sources has been

² <https://www.delpher.nl/>

³ <https://mediasuite.clariah.nl/tool/single-search?queryId=13e0c19e-0d92-4d36-9b36-db899f118ec2>

preserved compared to the period before and after. These include a large body of radio texts from the *Nederlandsche Omroep*: the complete transcripts of *Radio Oranje* and *De Brandaris*, and more than 20,000 monitoring reports (summaries and verbatim transcripts) of the broadcasts of the Hilversum stations created by the listening service of the Dutch government-in-exile. All of these analogue collections, consisting of tens of thousands of pages, are kept at the NIOD Institute for War Genocide and Holocaust Studies.

Digitizing this large 'radio on paper' archive and making it accessible to researchers together with the audio collection would offer an important extension of the available material. Reuniting these collections that have been separated and scattered over various archival institutions would also undo the artificial divisions between them. Finally, it would enable historians to conduct academic research, using both qualitative and quantitative methods, into the content of the pro-Nazi and pro-Allied propaganda and the war of words between these adversaries. This type of research, that was pioneered by the Media War project, is an important intervention in the existing scholarly literature on Dutch language media during the Second World War, which mainly examines the institutional side of wartime radio (Verkijk (1974), Sinke (2009)).

To achieve its goals the Media War research fellows made use of digital humanities techniques to study the newspaper collection in conjunction with the radio audio collection, employing both distant reading and close reading methodologies. These digitised wartime collections of *Delpher* and *NISV* are available in the online, digital humanities platform *CLARIAH Media Suite*. In addition, the fellows also consulted the paper radio archive at NIOD to fill in gaps. In the following sections we will first discuss the Media Suite, the various relevant source collections, and the challenges they posed. Subsequently, we will focus on the pilot Media War Matching project, which was created to tackle these challenges, adding a sub-collection of digitised Nazified radio transcripts to the existing collections in the Media Suite and exploring the possibilities and challenges for analysing these sources in combination with the available audio and newspapers.

3.2 The Media Suite

The Media Suite is a suite of tools developed within the *CLARIAH*⁴ project. The Media Suite brings together diverse media heritage collections, and enables researchers to search, view, bookmark and annotate items from these collections. Search results can be visualised over time, potentially revealing patterns, and these visualisations can be compared between different collections.

The Media Suite supports basic quantitative analysis with statistics and visualisations of the occurrence of search hits over time and the frequency of metadata field values. To allow more detailed analysis, the Media Suite APIs provide direct data access, so that the data can be processed by a researcher in their own software. A popular tool for this is a Jupyter Notebook⁵, an easy-to-use programming environment that combines text, code and visualisations. Direct data access is currently only available onsite at NISV for selected projects, due to copyright restrictions.

The Media Suite is already used by Dutch researchers from various humanities disciplines. At the same time, it is under active development. The Media Suite development team cooperates with users in projects to explore and develop new functionality that

⁴ <https://www.clariah.nl/>

⁵ <https://jupyter.org/>

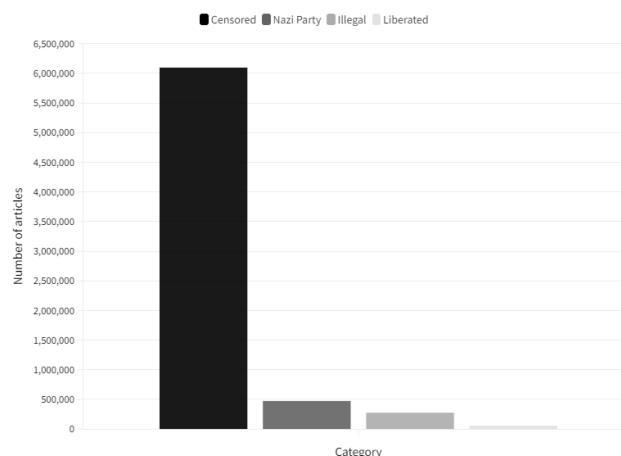


Figure 1: Imbalance in categories within wartime newspaper collection

may be added to the Media Suite, immediately or in the future.

These characteristics make the Media Suite uniquely placed to play a key role in breaking down the barriers between propaganda collections and enabling their investigation and analysis.

3.3 The Newspaper collection

From the onset of the Media War project it was clear that the bulky newspaper collections were a rich resource but also provided various challenges for historical research on Second World War propaganda. On the basis of the existing metadata no divisions could be made between papers supporting the Nazi occupation of the Netherlands and those resisting it. Therefore, the Media War team set out to enrich the collection by categorizing the newspapers into the following categories:

- Pro-Nazi, subdivided into censored newspapers, Nazi party newspapers;
- Anti-Nazi, subdivided into illegal newspapers and newspapers which appeared in the liberated south of the Netherlands from September 1944 onwards).

The first quantitative visualisations of these categories showed significant imbalances within the wartime newspaper collections, as Figure 1) shows. The biggest discrepancy is between 'censored', which includes more than 6 million articles, and 'illegal', which contains not even 300 thousand articles. This discrepancy, however, changed over time. As the war progressed, more and more illegal papers were published. In turn, both the number and size of the censored newspapers decreased. Around September 1944 the anti-Nazi press outnumbered the pro-Nazi press. This general trend can be explained by the historical context, including the dwindling supplies of the occupying regime which was partly caused by the successful raids of the resistance to steal paper to print their own publications.

To further analyse the war of words between pro- and anti-Nazi media, the research fellows developed keyword searches to generate semantic patterns over time. The Compare tool in the Media Suite allowed them to simultaneously visualise these patterns in the categories mentioned above. To tackle the quantitative imbalances

the Media War team added a new functionality to the Compare tool to show query results relative to their own category - or in Media Suite terms, 'relative to query facet'. Despite these efforts, however, researchers occasionally encountered problems generating semantic patterns, particularly in the illegal newspaper category, which consists of a lot of newspaper titles which had relatively few articles. Considering these limitations the researchers used the visualisations not as an end product of their examination, but rather as a starting point for further qualitative analysis, as they discussed in a 'data story' about the research.⁶

3.4 The Radio audio collection

The wartime radio collection maintained by the Netherlands Institute for Sound and Vision is published in the Media Suite as part of the larger Sound and Vision TV and Radio collection.⁷ The category 'broadcaster' in the metadata of this collection allowed the Media War team to easily make a distinction between pro-Nazi and anti-Nazi radio items. Due to the fragmentary nature of this collection, visualisations of this material did not yield significant results. Another obstacle for employing quantitative methodologies is the poor audio quality of many of the preserved and later digitised recordings which made Audio Speech Recognition (ASR) transcriptions not sufficiently reliable. Nonetheless, the research fellows made use of selected audio fragments in their work, focusing on important 'media moments', which they identified using the visualisations of semantic patterns in the wartime newspapers.

3.5 The Radio transcript collection

Confronted with the fragmentary nature of the wartime audio sources, the Media War team became increasingly aware of the importance of the radio paper archive to gain insight into the content of both the pro-Nazi and anti-Nazi radio broadcast during the war. While these paper collections, which seem to be nearly complete, provided them with much relevant information, studying these archival sources was also very time consuming. Moreover, because of the enormous amount of radio transcripts and monitoring reports, it was only possible to study very specific, well-demarcated moments in time. Obtaining an overall picture of certain propaganda narratives throughout the war was therefore not possible. Digitizing these collections could help to provide this overall picture.

One such paper collection of interest consists of digitised radio transcripts produced by the radio news service Berichtendienst van de Nederlandsche Omroep (BNO), which made news broadcasts for the centralised and Nazi-controlled Nederlandsche Omroep. For researchers the BNO collection is particularly relevant as it allows them to study how the occupying regime and its Dutch supporters used radio news for propaganda purposes and employed certain frames to interpret events along the lines of the Nazi ideology.

The BNO was set up on 1st April 1941, less than a month after the occupying authorities ordered a concentration of Dutch broadcasting corporations into one national broadcaster. The BNO replaced the radio news service of the press agency Algemeen Nederlandsch Persbureau (ANP) that had been brought under German control immediately after the start of the occupation. It was located in The Hague and had 55

⁶ <https://mediasuitedatastories.clariah.nl/mediaoorlog/>

⁷ <https://mediasuite.clariah.nl/tool/single-search?queryId=70967a36-4b36-40a1-a51e-a056836d0d61>

employees, mostly from the Drahtlose Dienst (Wireless Service). The direction was in the hands of G. Noordhuis, a fanatic Nazi who had worked for Radio Bremen before the war. In October 1943, the BNO moved to Hilversum and the number of employees was reduced to 33 (Verkijk, 1974, pp. 637-638).

The goal of the BNO was to create propaganda in the form of news broadcasts. Aside of the approximately five daily news bulletins that were broadcast simultaneously on both Hilversum stations, the BNO also produced two daily radio talks about economy, agriculture or culture. In addition to this, the BNO was responsible for the radio items of the pundits 'De Jordaner' and 'Jan Hollander', for broadcasts to the Dutch Indies and an infamous programme aimed at Dutch seafarers that was vehemently opposed from London by Radio Oranje and De Brandaris. According to the Dutch journalist and historian Dick Verkijk, who wrote a monumental work about Radio Hilversum during the Second World War, approximately 35% of all texts spoken on the Nederlandsche Omroep were produced by BNO employees (Verkijk, 1974, p.629). The service made programmes that partially overlapped with the topics covered in the Nederlandsche Omroep broadcasts. According to Verkijk, the BNO functioned as a 'broadcaster within a broadcaster', that in fact competed with the Nederlandsche Omroep. That the BNO was successful in this competition is apparent from the fact that the service could carry on broadcasting until the end of the war, whereas the Nederlandsche Omroep was forced to cease most programmes in 1944 (Verkijk, 1974, pp. 632-634).

The BNO collection is held by the NIOD institute for War, Holocaust and Genocide studies. In the context of the Media War project a large number of BNO transcripts were digitised, according to guidelines established in *Metamorfoze*.⁸ Two parts of the BNO collection were made available for digitisation: the 'BNO reports, 1940-1945' (84 folders of mainly news transcripts) and the 'Texts of broadcasts about economy and politics' (75 folders).⁹ After digitisation, the digital scans (Figure 2) were processed by OCR (Optical Character Recognition) using Abbyy software¹⁰ to produce searchable text transcripts. The available metadata for these transcripts was rudimentary, consisting of a division into folders, with each folder assigned to a date interval, ranging from two weeks to one or more months.

The NIOD institute made the transcripts available on their website. There, researchers can browse the folders, search for words in the OCR text, and read the scanned reports. It is not, however, possible to visualise search results to discover patterns or carry out any quantitative analysis. Finding out which transcripts were recorded on which date requires browsing through the folder of transcripts to find the broadcast. This is a cumbersome task, particularly as the transcripts are not always in date order, not all pages are dated and occasionally transcripts have been put in the wrong folder.

3.6 The Media War Matching project

The main aim of the Media War Matching project was to make the digitised BNO transcripts available in the Media Suite so that they could be analyzed by researchers, both in isolation and in combination with other wartime media collections, using the digital tools that the Media Suite offers. In addition, the project aimed to investigate

⁸ <https://www.metamorfoze.nl/front>

⁹ https://www.archieven.nl/mi/298/?mivast=298&mizig=210&miadt=298&micode=103&milang=nl&mizk_alle=bno&miview=inv2

¹⁰ <https://www.abbyy.com/ocr-sdk/>

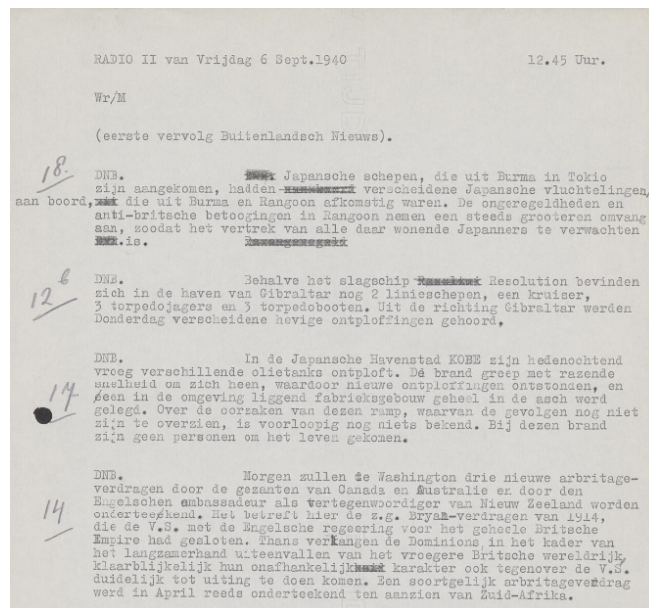


Figure 2: Example transcript from the BNO collection

and develop new methods for publishing similar 'radio on paper' collections that can be used for future projects.

4 Method and results

4.1 Publication of radio transcripts

Publishing the BNO radio transcript collection in the Media Suite required special care and new techniques, due to the lack of metadata.

4.1.1 Granularity and date/title metadata

To allow the Media Suite tools to be optimally used, a certain degree of metadata is necessary, so that the data can be split into units of a useful granularity. Of particular importance is the date field, as this metadata allows researchers to examine how the occupying authorities framed specific events and appropriated and incorporated news in their broader propaganda narratives over time. Moreover, accurate date fields are also essential for comparison. For example, the source material in the newspapers collection is separated into individual articles, which are annotated with the newspaper title and the date of publication, enabling researchers to find articles from a given newspaper on a given date, and compare trends between newspapers and over time.

For the BNO collection, as already discussed, the metadata is rudimentary. One additional piece of metadata was derived from background knowledge of the BNO. It was known that the radio programmes were broadcast under the name 'ANP' from the start of the occupation up until the BNO was officially set up in April 1941. By checking the transcripts around this date, a researcher could pinpoint the moment at which the name BNO was first used in the radio announcement, in the 8:30am broadcast on 9th April 1941. This enabled us to automatically annotate all previous broadcasts with the broadcaster 'Nazified ANP', while later broadcasts were annotated



Figure 3: Examples of different date formats in the BNO collection

with the broadcaster 'BNO'.

Ideally the transcripts would be divided up into the texts of the individual radio broadcasts, and each broadcast annotated with the broadcast title and the broadcast date. The broadcast title and broadcast date, while absent from the metadata, are printed at the start of the relevant section of the transcript in the majority of cases.

For this reason, it was opted to extract date and title information from the OCR text. The assumption was that the presence of a date and broadcast title at the top of a page would indicate the start of a new broadcast.

4.1.2 Date and title recognition by Named Entity Recognition

The OCR text from the top of each page was passed through the SPACY¹¹ named entity recogniser, with the aim of detecting the dates and titles. During this process, we discovered an error in the underlying metadata: the page text had been produced by concatenating the individual blocks of sorted text, but this sorting was incorrect, meaning that the words at the start of the page text were not necessarily the words from the top of the page. When attempting to correct this by generating a new page text from correctly sorted blocks, we discovered that short blocks of text had apparently been discarded from the metadata, meaning that the resorted text missed some words. Crucially, the date/broadcast title was often recognised as a short, isolated block, so precisely this essential information had a higher risk of being left out.

Even allowing for these losses, SPACY missed a large number of dates and titles in the text. The reason for this was not entirely clear, but looking at the missed dates, we noted the large variation in date formats used in the text (Figure 3). The date/title combinations were also usually printed in isolation, without the context that would help NER. The BNO data may also differ from the type of data on which this SPACY model is trained. Our hypothesis is that these factors are the cause of the poor NER performance in this case.

In the course of the NER tests, we discovered that our assumption that a date would signal the start of a new broadcast was incorrect. A date and/or date/title combination was also sometimes repeated partway through a broadcast.

In our tests, we used the SPACY model out of the box, so we didn't train it on the BNO data or fine-tune its parameters. This would require a large amount of work to annotate and train the model, which was out of scope for this project. Retraining could

¹¹ <https://spacy.io/>, model `nl_core_news_lg` (3.4.0)

be an option for future work.

4.1.3 Alternatives to Named Entity Recognition

Attempts to find algorithms online that were capable of detecting dates with a wide range of formats met with no success. As a result, we chose to develop our own algorithm for date detection, which we could tailor to the formats used in the BNO collection.

For the titles of the news bulletins, we faced the problem that the broadcast titles varied only minimally - Radio 1, Radio 2, Radio 3, etc. Difficulties were compounded by Roman numerals also being used: Radio I, Radio II etc. To distinguish broadcasts required accuracy of a single character, while the variable quality of the recognised text would require us to allow more character errors in order to successfully detect the word 'radio'. This paradox led us to abandon the attempt to detect program titles within this project.

4.1.4 Custom date detection

We treated short numerical dates - e.g. 10-01-41, 10.01.1941 - separately to textual dates - e.g. '10 januari 1941'. We parsed short numerical dates using existing Python functionality.¹² For textual dates, the task was more complicated, so we created various aids to help us. We constructed lists of possible date parts: days (numbers 1-31 with and without leading zeros), months (full names and abbreviations) and years (1940-1945, 40-45). As the transcripts turned out to use English and German dates in addition to Dutch, we also listed the month names and abbreviations in these languages. We developed functions tailored to matching text to days, months or years. For example, for days we conducted a strict matching (as '20' must not be allowed to match to '21'), whereas for months we used fuzzy matching¹³ to compensate for small typing or OCR errors (e.g. september). Finally, we listed date formats that occurred in the transcripts, e.g. 'day month year', 'day month'. The initial version of the algorithm used these aids in the following steps:

- Extract the start of the page text
- Split the text on spaces
- Clean up each part of the text to correct frequently occurring OCR errors. E.g. '194I' to '1941', '10' to '10'
- Try to match each part of the text to a short numerical date. If a match is found, stop
- Otherwise, match each part of the text to the lists of months and years
- For good matches: For the list of possible date formats, extract the relevant surrounding parts of the text. E.g., given the text 'Vrijdag 13 eptember 94I', we have a good match for the month 'september'. A possible date format involving the month is 'day month year'. So we extract the word before our month match and the word after, giving us the candidate text '13 eptember 94I'.

¹² <https://www.programiz.com/python-programming/datetime/strptime>

¹³ <https://pypi.org/project/fuzzywuzzy/>

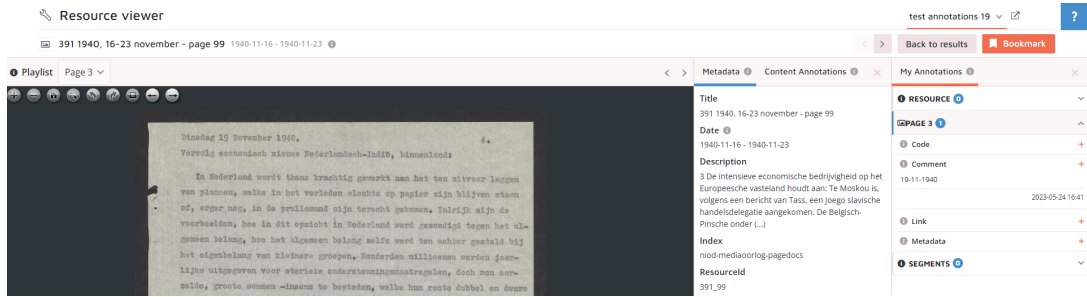


Figure 4: Example of a date annotation in the Media Suite (Under 'Comment' in 'My Annotations')

- For each candidate text, we calculated the scores for matching each date part with possible days, months and years
- For each possible format, we combined the relevant scores of each date part. For example, when testing the format 'day month year' on '13 eptember 94I', we combined the scores for matching '13' with days, 'eptember' with months, and '94I' with years.
- The highest scoring candidate text was selected. If this scored higher than a given threshold, then it was selected as a match for the transcript date

Note that not all date formats used in the transcripts were complete. So correctly detecting 'jan 1941' still did not give us a full date for the transcript.

We also experimented with generating a list of all possible dates, in all possible formats, and matching this with the candidate texts via fuzzy matching. This worked less well than comparing the individual parts of the date. For example, it regarded the match of '13 september' with '15 september' as being equally good as the match '13 september' with '13 s3ptember', whereas the latter is clearly better. The large number of possible dates also led to this method being very slow.

4.1.5 Performance of custom date algorithm

To evaluate the performance of our algorithm, we displayed the results as annotations in the Media Suite (Figure 4). A researcher viewing a page of a transcript was shown an annotation with the extracted date, if present. If this date was wrong, the researcher could add their own annotation with the correct date. This was done for three complete folders of transcripts, and for the first and last 50 pages of a random selection of 2 folders per year, one from the news transcripts, and one from the economy and politics transcripts. Based on the corrected results, we could measure the algorithm performance, analyse the source of the errors, and try to improve it.

The performance of the custom date algorithm, as evaluated manually by the researcher, varied strongly per folder. The worst folder scored only 3%, whereas the best was 100%.

Manually analysing the source of the errors gave the following two main causes:

- OCR text not readable by a human 62%
- Punctuation or spaces in the date, e.g. '1/1 2/1941' 14%

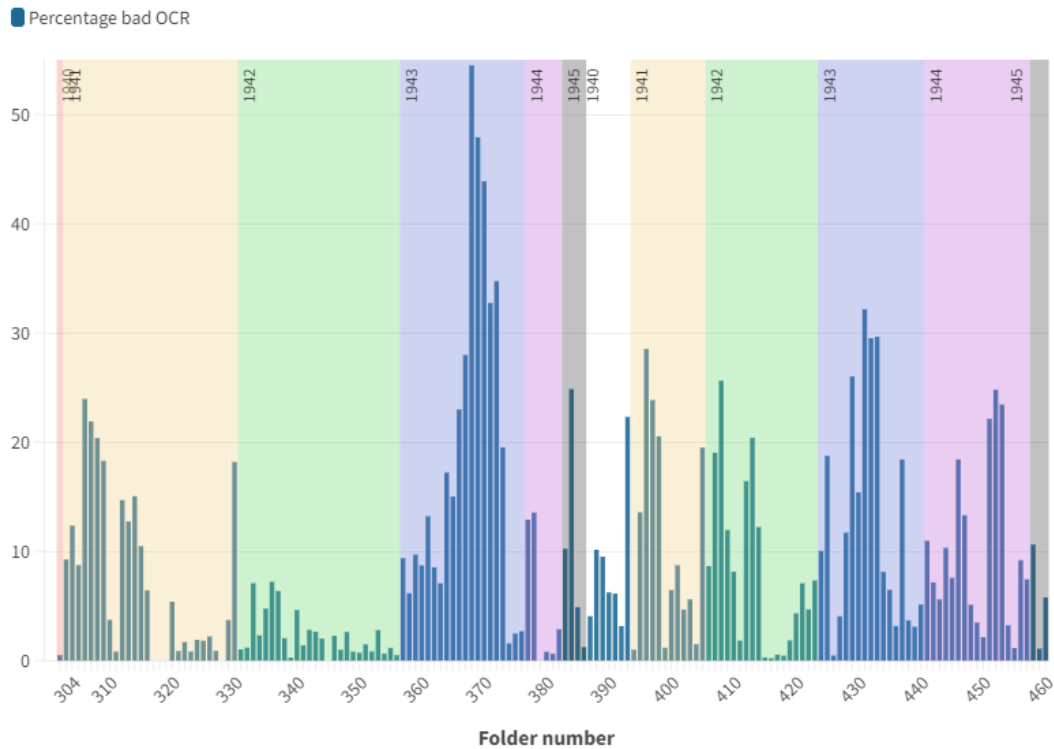


Figure 5: Percentage 'bad' OCR per folder

For the next 20% of errors, there were again two main different causes. For example, bad quality OCR sometimes led to letters being split up into multiple punctuation marks. Also, the previously noted problem of short blocks of text having been discarded from the metadata led to some dates being missed. Finally, the last 4% of the errors were due to a variety of small errors or unknown reasons.

Initial impressions of the OCR quality when starting the project were positive. The availability was high: there was OCR present for 99.97% of the transcripts. When browsing through the transcripts by hand, the quality generally appeared good. The analysis of the date algorithm errors, however, showed up a subset of the transcripts that had very poor OCR. A proper analysis of the quality would require manual evaluation, or perhaps analysis using language models. As a rough approximation, we plotted the number of transcripts per folder that had OCR transcripts of 50 characters or fewer (Figure 5). We regarded such short texts as an indication of 'bad' OCR.

The results show a strong dependency on the folder, explaining why the date extraction results were also strongly dependent on the folder. For the economy and politics broadcasts, there does not seem to be a particular pattern. For the news broadcast transcripts, quality clearly deteriorates over the course of 1943, possibly linked to poor paper quality in that phase of the war.

A brief test with Tesseract OCR¹⁴ on a few example transcripts showed significantly better results. For example, the date of one transcript was OCR'd as 'SQj3d@rda< IB 1944' in the current transcript, but as 'Rad4ô) 3 aa Dondeniag 18 1et 1944' by Tesseract, and when zoomed in to just the first line: 'Radio

¹⁴ <https://github.com/tesseract-ocr/tesseract>

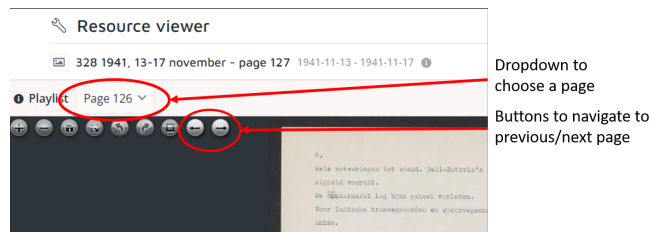


Figure 6: Navigation within the folder of a transcript page

3vaù Donderdag 18 Mei 1944'. This would suggest a possible gain by repeating OCR processing of the transcripts with a different algorithm. This, however, lay outside the scope and remit of the Media War Matching project.

Experiments were conducted to correct the other sources of errors, for example by extending the list of common OCR errors to be corrected during cleanup, removing punctuation, or examining longer sections of text. Each method produced improvements for some transcripts while other transcripts performed worse. The only method that produced an overall improvement on average was correcting the OCR errors. This resulted in a marginal performance increase from 54.6% to 55% on average.

This performance is far too low for reliable use of these dates in the Media Suite. However, the errors occurred mostly in dates that were only partially recognised (e.g. day and month with no year, month and year with no day). When we used the date information of the folder to determine the year for dates with only a day and a month, and otherwise discarded partial dates, then we achieved a precision of 90-100%. This is far more usable. However, this meant we had dates for 36% of the transcript pages in total. This was a trade-off for higher accuracy at the expense of coverage.

4.1.6 Final version in the Media Suite

The ideal situation was to publish the transcripts to the Media Suite as separate texts for individual broadcasts. As identifying individual broadcasts did not work, we had the choice between publishing the transcripts per folder or per page. We opted to publish the transcripts per page, as this meant researchers could quickly pinpoint relevant material instead of having to browse an entire folder, often consisting of around 600 pages. Moreover, this choice meant that plots of search results over time gave more insight into how often a word occurred in different broadcasts. However, it is also valuable for researchers to be able to easily read the context of a found page, seeing the pages before and after it in the folder. For this reason, we added a button to the viewer, to allow researchers to easily browse through the folder around their found page, and a dropdown to allow them to select other pages in the folder (Figure 6).

We had three types of date information available to annotate the pages. The first was the folder date information, which was available for every page, but was imprecise, indicating as it did a range of weeks. The second was the automatically extracted date information, which was precise to the day and also had a high accuracy, but was only available for about a third of the pages. Finally, we had the ground truth date information entered by the researcher. This was extended for the folders 310 and 397 which contain materials from the first month of the BNO's existence, and for the folders 323-330, which contain transcripts of news bulletins from September-December 1941 a period which was of particular interest to the Media War research fellows. This

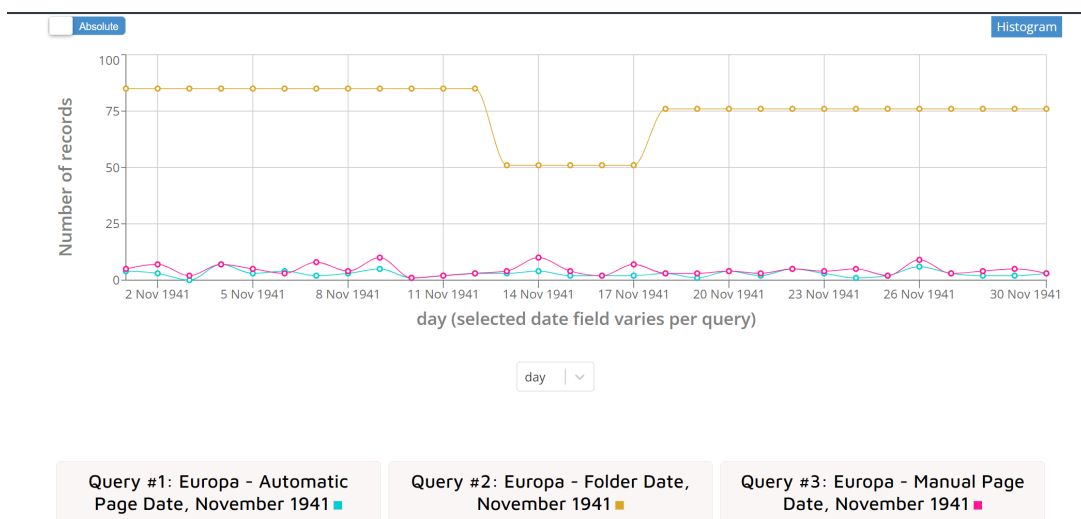


Figure 7: Combining the strengths of the different dates in the Compare Tool. Yellow: Folder dates, pink: manual dates, blue: automatic dates

date information was precise to the day, highly accurate, but only available for 6% of the collection.

To get the most out of this date information, we decided to publish it all. Researchers can then switch between the dates depending on their needs. To find as much material as possible, they can use the folder date, whereas to pinpoint found pages more accurately, they can use the manual or automatic dates. The user can even utilise the Compare Tool to combine the strengths of both.


With the graph shown in Figure 7, the researcher can use the folder date line to see that there was a dip in the middle of November, with far fewer hits for 'Europa' (Europe) than in the latter half of the month. Perhaps more interesting for researchers, though incomplete, are the automatic and manual date lines, which suggest dates for further investigation, such as 9th and 26th November, perhaps caused by the publicity surrounding Hitler's speech on the 18th anniversary of the Beer Hall Putsch on the 8th of November, or the renewal of the Anti-Comintern pact on the 25th, which was presented as a milestone in the history of European unity. Obviously, given the incomplete coverage of the automatic and manual dates, this functionality must be used with care. Such hypotheses require further research into the propaganda content, for example with close reading.

In general, the characteristics of the BNO collection have a large impact on how a researcher can work with it. The poor OCR quality influences search results, as search terms are less likely to be successfully found in transcripts with poor OCR quality. Furthermore, the incomplete availability of dates influences both what can be found when searching by date, and what is visualised in the Compare tool, as only items with dates can be plotted on the Compare tool graph. To allow the researcher to work with the collection and the Media Suite tools appropriately, it is essential that they are properly informed about the limitations of both. For this reason, we documented the BNO collection, with particular emphasis on the processing that was applied to generate dates (Figure 8). The documentation was produced by a team of data engineers and propaganda researchers to ensure that it included relevant information presented in such a way that researchers could understand the implications. This


**Wartime radio news bulletins:
Berichtendienst
Nederlandsche Omroep (BNO)**


Followers
0

Organization

 **NIOD**
Kwesties die met oorlogsgeweld te maken hebben, wekken veel maatschappelijke belangstelling en vragen om onafhankelijk en wetenschappelijk onderzoek. Het NIOD verricht en... [read more](#)




Social

 Twitter

 Facebook

License

Other (Not Open)

 Dataset
 Groups
 Activity Stream

Wartime radio news bulletins: Berichtendienst Nederlandsche Omroep (BNO)

This collection consists of transcripts made of wartime radio bulletins from the Netherlands, broadcast by the BNO (Berichtendienst Nederlandsche Omroep). The BNO operated under the aegis of the occupation regime in the Netherlands during the Second World War and actively made propaganda to promote national-socialism in its news broadcasts. The collection is provided by the NIOD (the Dutch Institute for War, Holocaust and Genocide Studies), collection 103, Nederlandsche Omroep.

The collection consists of images of the scanned transcripts. The transcripts have been processed by the NIOD using OCR (Optical Character Recognition) to make them searchable. The quality of the OCR is variable. The transcripts are stored per page, and have only very basic metadata. Each page belongs to a folder for which a year and a range of dates is known. Page dates have been automatically extracted from the transcript, and dates have been manually added for pages in certain folders. See section below for more information about dates.

This collection was imported into the Media Suite as part of the Mediaoorlog project.

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

The entire BNO collection has been imported

What years does the collection cover?

The collection covers 1940-1945

Date fields

The original metadata contains no page dates, but every page belongs to a folder, for which a year and a range of dates within that year are specified.

Individual page dates have been added both manually and automatically for subsets of the collection.

Five date fields are available:

Figure 8: Documentation page for the BNO collection

documentation is included in the Media Suite and can be reached via a link in the collection (Figure 9)¹⁵.

The tools available in the Media Suite are also documented, and this information is available via icons in the tool itself and the Help section of the Media Suite.

In this way, the researcher is supported in both data and tool criticism, equipping them to work with the BNO collection in an informed manner.

¹⁵ <https://mediasuitedata.clariah.nl/dataset/wartime-radio-bulletin-transcripts>



Figure 9: Collection Selector showing the BNO collection and its documentation link

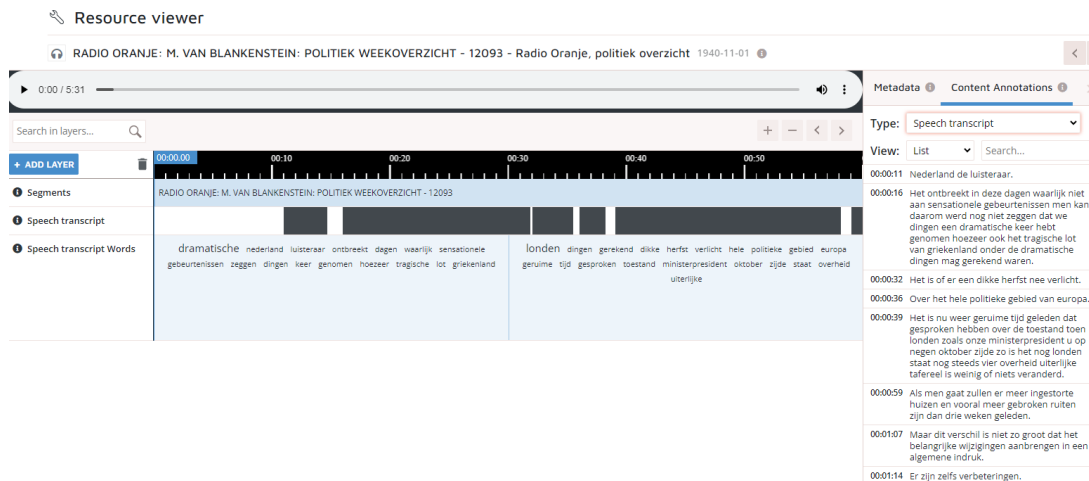


Figure 10: An example of an audio recording with an ASR transcript in the Media Suite

4.2 Matching of radio audio with transcripts and newspapers

Publishing the BNO transcript collection in the Media Suite gives unprecedented possibilities to explore the collection. But to truly start putting the propaganda puzzle together, we want to link the collections. Then a researcher could find a radio transcript discussing a given topic, listen to the audio, and look at what the newspapers were saying about that topic on the same day.

The radio transcripts can also be used to fill some of the holes in the audio collection, and vice versa. There are far more radio transcripts (to give an indication, the digitised BNO collection alone comprises 68,986 pages) than there are audio recordings (2262). At the same time, there are cases where an audio recording is present while the transcript is missing, such as the news bulletin of 8am on 25th July 1943.

In order to link transcripts to the relevant audio recording, we proposed the following steps:

- Generate speech transcripts of the audio recordings
- For each audio recording, find the transcripts within a small date range of the broadcast date of the audio (to allow some margin for errors)
- Match the words in the speech transcript with the words in the candidate transcripts
- With a sufficient number of matching words, link the transcript to the audio recording

For the first step, we used Kaldi NL¹⁶ to generate speech recognition transcripts of the audio recordings. For 27% of the collection, no transcript could be produced. For another 16%, the produced transcripts contained fewer than 50 words, which could indicate poor quality. The remaining 57% of the collection has transcripts that are longer than 50 words. All transcripts were added to the radio audio collection in the Media Suite, and can be viewed there (Figure 10).

However, a review of the transcripts quickly showed that the quality is very poor in many cases, so much so that it is often impossible to make sense of it (Figure 10).

¹⁶ https://github.com/opensource-spraakherkenning-nl/Kaldi_NL

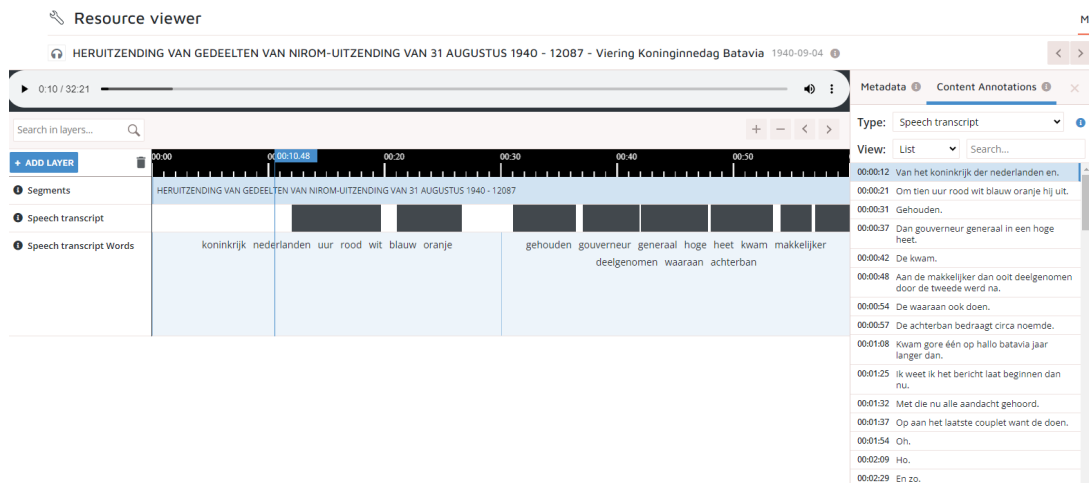


Figure 11: An example of an audio recording with a poor quality ASR transcript

The poor performance is most likely due to the poor audio quality of the recordings. Given this poor quality, combined with the poor coverage of the date information in the BNO, it was decided to abandon attempts to link the audio with the BNO transcripts within this project.

Despite the fact that the items within the collections could not be explicitly linked, their publication together in the Media Suite still enables quantitative analysis of combined collections, as we indicated in the section on 'the radio audio collection'. To demonstrate this, we proceeded to compare semantic patterns in the BNO transcripts with the wartime newspaper collection in the Media Suite.

4.3 Quantitative analysis of combined collections

Via the visualisations provided in the Compare tool, the Media Suite supports some basic quantitative analysis of the combined collections. In the example in Figure 12, the search results for the term 'Europa' (Europe) are shown per month for the Nazi party newspapers (pink line) and the BNO transcripts (blue for folder date, yellow for the automatic page date). It can be seen that the two lines for the BNO transcript follow the same pattern, showing that despite the incompleteness of the automatic date information it is still usable to show trends. All three graphs show noticeable peaks in July 1941 and February 1943. The first peak reflects the increasing propaganda value of the 'Europe' construct for the Nazis after the start of the attack on the Soviet Union on 22nd June. The Nazis framed the fight against the 'godless Bolshevism' and the 'Asian' Soviet Russia as a common European fight or crusade to preserve the 'Christian European civilisation'. The second peak shows the urgency of 'Europe' in the propaganda offensive of the Nazis after the loss of the Battle of Stalingrad, in which Goebbels emphasised that only total war could neutralise the acute Bolshevik danger and that the fight against the Soviet Union was a matter of life and death, a struggle for the continued existence of Germany and Europe (Brolsma (2022)).

As explained in Section 3.3, the Compare Tool offers an option to provide relative graphs, compensating for the imbalance between collection sizes and therefore making it easier to obtain a meaningful comparison. Figure 13 shows the query results of the keyword term 'Europa' (Europe) relative to their own category. This graph confirms,

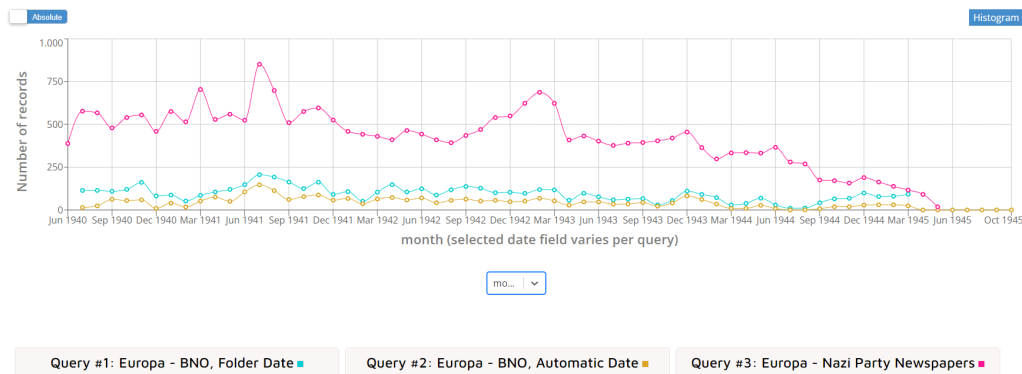


Figure 12: Comparing search results over time for the search term 'Europa'. Pink: Nazi party newspapers, yellow: BNO with automatic page dates, blue: BNO with folder dates

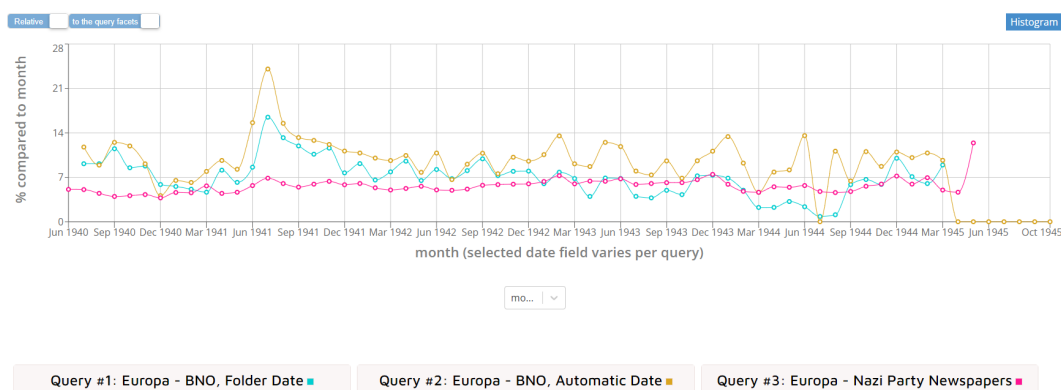


Figure 13: Comparing search results over time for the search term 'Europa', relative to the category/-collection size. Pink: Nazi party newspapers, yellow: BNO with automatic page dates, blue: BNO with folder dates

and indeed foregrounds, the urgency of Europe as a propaganda concept in the context of Nazi Germany's invasion of the Soviet Union. However, compared to Figure 12 it suggests that the peak of February 1943 was less significant, as Europe continued to be frequently used as a propaganda frame, also in the last two years of the war. These visualisations invite researchers to reflect on the continuities and discontinuities of propaganda narratives throughout the Second World War.

This sort of data visualisation is therefore useful both to identify important moments for a closer investigation of the media content, and to map out the similarities and differences between various categories of wartime media, such as between pro-/anti-Nazi media and/or between newspapers and radio broadcasts. In other words, semantic patterns in quantitative visualisations can guide researchers to interesting 'media moments' which merit further investigation.

The texts produced by OCR and ASR assist researchers in searching collections. However, they are also a rich resource for quantitative analysis. We wrote a Jupyter Notebook, using the Media Suite APIs, that extracted the words occurring just before or just after a given search term.¹⁷ This provides a tool for researchers to investigate in what context a word is used.

¹⁷ Stop words such as 'de', 'het', 'van', 'in', etc were excluded from this analysis

Table 1: The top ten most frequently occurring words before the word 'Europa', per (sub)collection

	1	2	3	4	5
Anti-nazi	west	oost	oorlog	rijk	geheel
Pro-nazi	nieuwe	geheel	strijdt	nieuw	west
BNO	nieuwe	geheel	oost	west	landen
Radio audio	heel	oost	west	landen	we
	6	7	8	9	10
Anti-nazi	landen	bevrijd	heel	vasteland	nieuwe
Pro-nazi	oost	landen	oorlog	midden	vasteland
BNO	oorlog	nieuw	vrijheid	vasteland	heel
Radio audio	zuid	binnen	Nederland	vanuit	rest

This notebook was used to conduct an analysis of the use of the word 'Europa' (Europe) in the anti- and pro-Nazi newspapers, the BNO collection (via OCR) and the radio audio collection (via ASR). The results are shown in Table 1

Interesting is that the most frequently occurring word in both the pro-Nazi and BNO collections is 'nieuwe' (new). The variant 'nieuw' also appears high in the ranking for these collections. These results show that the construct 'Europe' was very important for the occupying authorities to promote a national-socialist vision of the future in the nazified printed press and via Radio Hilversum. Of course, the poor - and variable - quality of the ASR and OCR means that this analysis is at best incomplete, and may also be inaccurate. For example, if a frequently occurring word is poorly recognised, then it will be counted much less often, causing it to appear to occur less frequently than it actually does. However, the fact that the results in the BNO and pro-Nazi categories are so similar seems to indicate that the quality is sufficient to discover important Nazi media frames via digital research methods, such as the frame of a 'new Europe'. The results in the ASR of the (both pro- and anti-Nazi) radio seem to consist mainly of geographical indications.

Due to the poor quality of the texts, such analyses cannot provide reliable conclusions, but can be used to discover possible interesting trends for further qualitative investigation.

5 Conclusion

During the Second World War in the Netherlands, the Berichtendienst Nederlandse Omroep (BNO) made an important contribution to the propaganda of the Nazi occupying regime through their news broadcasts. The publication of the digitised BNO transcripts in the CLARIAH Media Suite enables researchers to more effectively search through the transcripts. We did not achieve the goal of linking these transcripts to the audio recordings also stored in the Media Suite. Yet despite the absence of explicit links, the publication of collections of newspapers, radio transcripts and radio audio recordings in the same online environment means that researchers can analyse the BNO collection in conjunction with other categories of war media (categories per political-ideological signature and media type) for the first time. The availability of the OCR and ASR transcripts make it possible for them to combine qualitative analysis of the media content with various quantitative research techniques.

For both quantitative and qualitative research of the BNO transcripts, date information is essential. Making the transcripts searchable by date enables historians and media scientists not only to effectively analyse the reporting around various events - as demonstrated in the examples above - but also to make optimum use of the digital tools that the Media Suite offers. For example, they can compare semantic patterns in various media categories in the Compare Tool to identify important 'media moments' or study the use of certain frames over time. With the algorithm we developed, we succeeded in extracting dates for part of the collection.

The biggest obstacle to the automatic recognition of date information is the poor quality of the OCR. A priority in the future is therefore to improve the OCR, so that more pages can be reliably dated. This could be done by using new OCR tools, and potentially by re-scanning difficult pages. This should be discussed with the company that produced the original scans, as they may be able to suggest improvements. Alternatively, a more labour-intensive approach would be to manually annotate more pages with the date. As was demonstrated in this project, both approaches can be combined.

The large variation in date formats used in the collection also presented a challenge. An interesting option for the future could be to finetune or fully train a NER model on part of the BNO collection. This would ensure that the model is trained on dates in the formats and context such as they typically appear in the BNO collection (including the typical OCR errors) and hence can recognise them with more accuracy. This would require manual effort to identify all the variations and tag these to use as training material.

In addition to a better date recognition, improving radio ASR transcript quality could also help to make linking the radio broadcasts with the radio transcripts feasible. New speech recognition models are now available, which offer the potential for improvement. These should be tested on the radio broadcasts to see if it is possible to achieve better ASR quality.

To increase the value of the collection of wartime media in the CLARIAH Media Suite for future media-historic research into propaganda in Dutch newspapers and radio, a further expansion of this wartime media collection is desirable. This can be achieved by digitizing and publishing (parts of) the sizable radio archive of the years 1940-1945, such as the Radio Oranje transcripts and the reports of the listening service of the Dutch government in London exile. Such an addition would allow for research that will deepen our understanding of the dynamics in radio propaganda during the Second World War, and more specifically of the transnational war of words between Nazi-controlled transmissions from Hilversum and pro-Allied broadcasts from London.

In the course of the Media War Matching project, we learned a number of valuable lessons about the publication of such collections. The first was that quality starts at the source, and we would recommend that any future projects intending to publish a collection first conduct a thorough investigation of the data to identify quality issues up front. For large collections this is a challenge, as manual checking of everything is impossible, but checking samples alone can be misleading, as was the case in this project. The second lesson concerns the importance of data and tool criticism when dealing with such collections. The limitations of both data and tools must be clearly communicated with researchers to allow proper use of the collection. The third lesson is that cooperation between data engineers and historical experts is essential both to achieve this clear communication and to publish the collection (including development of tools and processing of data) in such a way that it is optimally usable by researchers.

This ongoing dialogue between experts in the various Digital Humanities disciplines brings us to our final and pivotal point. While it is important to keep in mind the ideal situation of digitised collections, each with complete metadata and linked to related collections, it is possible to achieve a great deal while falling far short of that ideal. In this particular project, we didn't put together the complete puzzle of the three propaganda collections, because of both computational challenges and the wartime circumstances that affected the production and preservation of these sources. However, what we did achieve has already proven to be of great value to researchers, showing how quantitative visualisations of media collections can serve as an entry point for further qualitative research. We would therefore encourage others to embark on the challenge of publishing such collections, to discover the limitations of the data and tools, and to use these to assist transparent and usable publication of the data, rather than letting them prevent it.

6 Acknowledgement

This work was enabled by the CLARIAHPLUS project funded by NWO (Grant 184.034.023), and the Mondriaan Fonds (75 jaar vrijheid).¹⁸

References

- Marjet Brolsma. Propagandaslag om europa: Wisselwerkingen tussen de nederlandse genazificeerde en antinazistische pers na operatie barbarossa. *Tijdschrift voor Geschiedenis*, 135, 2022. doi: <https://doi.org/10.5117/TvG2022.2/3.003.BROL>.
- Mark Connelly, Jo Fox, Stefan Goebel, and Ulf Schmidt. Prologue. power and persuasion': Propaganda into the twenty-first century. *Propaganda and Conflict. War, Media and Shaping the Twentieth Century*, 2019.
- Vincent Kuitenbrouwer. The traces of a media war: Archives of dutch broadcasts from london during the second world war. *Journal of Media History*, 25, 2022. doi: <https://doi.org/10.18146/tmg.821>.
- Vincent Kuitenbrouwer and Marjet Brolsma. Audio on paper: The merits and pitfalls of the dutch digital media archive for studying transnational entanglements during the second world war. *Journal of European Television History & Culture*, 12, 2023. doi: <https://doi.org/10.18146/view.306>.
- Vincent Kuitenbrouwer and Huub Wijfjes. Media war. *Journal of History*, 135, 2022. doi: <https://doi.org/10.5117/TvG2022.2/3.001.KUIT>.
- Onne Sinke. Onderling strijdend voor de goede zaak: Radio oranje en de brandaris. *Journal of Media History*, 8:97–109, 2005. doi: <https://doi.org/10.18146/tmg.540>.
- Onno Sinke. *Verzet vanuit de verte. De behoedzame koers van Radio Oranje*. Augustus, 2009.
- Hans van den Heuvel and Gerard Mulder. *Het vrije woord. De illegale pers in Nederland*. SDU Uitgevers, 1990.

¹⁸ <https://www.mondriaanfonds.nl/subsidie-aanvragen/regelingen/open-oproep-75-jaar-vrijheid/>

- Dick Verkijk. *Radio Hilversum 1940-1945. De omroep in de oorlog*. Uitgeverij de Arbeiderspers, 1974.
- René Vos. *Niet voor publicatie. De legale Nederlandse pers tijdens de Tweede Wereldoorlog*. Uitgeverij Sijthoff, 1988.
- Ivo van de Wijdeven. *De macht van het verleden. Geschiedenis als politiek wapen*. Unieboek Het Spectrum, 2022.
- Lydia E. Winkel. *De ondergrondse pers 1940-1945*. Martinus Nijhoff, 1954.
- Mariëtte Wolf and Frank van Vree. *De krant. Een cultuurgeschiedenis*, chapter Oorlog herstel en vernieuwing 1940-1950, pages 205–218. Boom, 2019.