Towards a *NAvigator* Tool for Dutch *Verbaal*-Archives: Leveraging Nineteenth-Century Archival Logic for Keyword Search

Sebastiaan Peeters[1], C. Annemieke Romein[1,2,3], and Andreas Weber[1]

[1] University of Twente, Enschede, The Netherlands
[2] Walter-Benjamin Kolleg, University of Bern, Bern, Switzerland
[3] READ-COOP SCE, Innsbruck, Austria

Abstract

Recent advances in machine learning and Automated Text Recognition (ATR) have encouraged efforts to digitize historical archives. However, processing text with an ATR-engine is not (necessarily) making archives accessible to researchers and the broader public; additional datafication steps are needed. This paper adds to this conversation by introducing and evaluating a simple navigation tool, the Dutch National Archives navigator, or in short NAvigator, that capitalizes on the structure of a specific historical archival management method, the verbaalstelsel-1823 ('verbal system'), which was common in Dutch government archives between the early nineteenth century and 1950. The NAvigator provides structured access to the information within such archives by recreating the historical workflow of search using the archives of the Dutch Ministry of Colonies (1850-1900) as a case study. The paper outlines the steps towards the development of the NAvigator and the challenges posed along the way, reports the results in terms of information retention throughout the search process, and discusses the potential for future research based on the intermediate output of the tool. As an early result, it finds that 84.7 percent of glossary entries are successfully connected to their corresponding documents.

Keywords: historical archives; digital history; document structure; archival structure; information retrieval; Dutch Ministry of Colonies.

1 Introduction

Over the last decades, the Dutch National Archives (Nationaal Archief, NA) in The Hague, as well as other archives around the world, have begun digitizing their collections to democratize access to the information they contain [17]. In the case of handwritten archives, this mainly involves scanning and possibly transcribing the documents, with minimal post-processing or enrichment, which limits the accessibility of these digitized collections [16]. To improve their

accessibility and make them available to researchers and broader audiences, further datafication and enrichment steps [21] are necessary. For instance, recent research has employed named entity recognition as a foundation for creating linked datasets [5; 12; 15] and developing domain-specific ontologies [1] that capture the complex relationships between historical actors, institutions, and places. While these approaches have advantages, they require a substantial commitment of time and resources, and often necessitate the involvement of citizen volunteers for labor-intensive tagging tasks [4; 18]. Using the digitized archives of the Dutch Ministry of Colonies (Ministerie van Koloniën) as a case study, this paper describes and evaluates a lightweight and inexpensive

Sebastiaan Peeters, C. Annemieke Romein, and Andreas Weber. "Towards a *NAvigator* Tool for Dutch *Verbaal*-Archives: Leveraging Nineteenth-Century Archival Logic for Keyword Search ." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1153–1163. https://doi.org/10.63744/ByKY4LADHCBb.

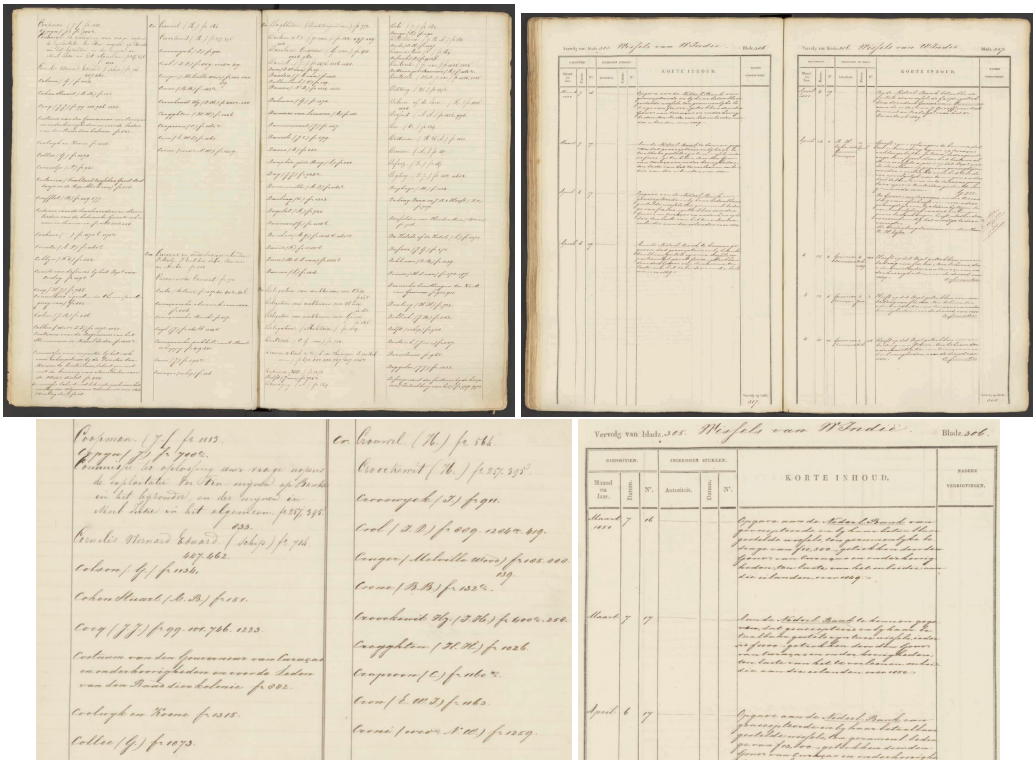

Figure 1: Examples of a klapper (left) and index (right) page and a close up of each. Source: NL-HaNA_2.10.02_5654_0016.jpg and NL-HaNA_2.10.02_5654_0220.jpg

method of unlocking an archive by leveraging its textual structure, in an approach that takes inspiration from earlier work by Koolen et al. [10; 11] and Colavizza et al. [3]. The goal here is to improve accessibility with a tool that opens new avenues for further research, building on this navigation.

Many Dutch government archives, including those of the Ministry of Colonies, utilize the *verbaalstelsel-1823* [7], which, by government decree, became the prescribed system for organizing archives in 1823 and remained in use until the middle of the twentieth century [19]. Following the rules of the verbaalstelsel-1823, archivists and administrators bundled completed dossiers and stored them by date. This means that if the Minister of Colonies or any other leading civil servant within the ministry had decided on a specific matter, all relevant documents (e.g., letters, reports, and earlier decisions) were compiled into a dossier and archived in a folder labeled with the date of the final decision. These dossiers are known as *verbalen*. The *verbalen* were later grouped by half-year and summarized in indices, which, besides a summary of the *verbaal*, also list dates, dossier numbers, and related persons or institutions. As these indices themselves are extensive, archivists also created *klappers* (glossaries), which provide access to the indices (see fig. 1). At the national level alone, there exists an estimated 9.25 km of this kind of archive out of a total of 12.4 km of archives. Although there are no comparable estimates for lower levels of government, available research mentions that a similar proportion of 19th and 20th centuries' Dutch archival materials is organized in the same way [6].

The approach to searching verbaalstelsel-1823 archives, as intended by their original creators, is to locate the relevant keyword, most commonly a name or location, in the *klappers* and then navigate to the corresponding index entry, which provides information on where to find the right *verbaal*. Right now, this is the only feasible search strategy for the physical archives. Mimicking this analog approach is also the primary way to navigate digitized images of the same archive. While

functional, this method requires users to have a good understanding of the underlying archival structure and historical ordering principles. Additionally, such navigation is rather time-consuming due to the disjointed nature of the image galleries, the absence of hyperlinks between individual pages, and the lack of keyword search for users accessing the archive through a web interface.

The NAvigator (an acronym for Nationaal Archief Navigator) addresses these issues by making the *klappers* searchable and interlinking them to associated indices, and linking the index to its *verbaal*. This improves accessibility by providing paths of navigation through the archive while respecting the original principles of archival arrangement and management. The latter makes it easier to contextualize the search results, as opposed to a simple keyword search across the entire archive. This additional context derived from the archival structure also helps in disambiguating common names, as the linked *klappers*, indices, and *verbalen* will all refer to the same person.

The following sections first introduce the dataset, then provide more details on the necessary steps to build the navigator, highlighting the challenges and reporting the loss of information due to errors along the way. Finally, the paper concludes by indicating avenues for future research and opportunities opened by the NAvigator.

2 Dataset

The verbaalstelsel-1823 archive used in this study covers the affairs of the Ministry of Colonies from 1850 to 1900 and is stored at the Nationaal Archief in The Hague. The full scans of the archive, comprising over five million images, are accessible as a collection of image galleries on the Nationaal Archief's website.[1] They were transcribed using Loghi [9] in 2024, with layout detection using Laypa [8], and both scans and transcriptions, in PAGE XML format, were made available to researchers of work package 2 [22] of the HAICu project.[2] The character error rate, without applying any text pre-processing, for a random sample of 63 files was 17.8%, which is well above the state-of-the-art [20]. This combination of high CER and historical language complicates the implementation of many NLP techniques [14], which is why the method in this paper relies mainly on language-agnostic techniques.

This paper uses only the main series of *verbalen* and associated indices, as well as *klappers*. It excludes the so-called agendas,[3] as they are not part of the primary search process, and the secret *verbalen*, whose layout and structure deviate significantly from the rest of the archives. Together, these excluded pages make up around 820,000 pages or 15.2% of the total Ministry of Colonies archive. While future work may attempt to include these pages in the NAvigator, they are beyond the scope of this prototype.

As this research concerns a colonial archive, it is important to acknowledge the issue of colonial biases. Studying such biases and how to mitigate them has become the subject of much research in recent years [2; 13; 23] and remains an open question in the field. However, at this stage in the research, we will hold off on engaging with these issues and limit ourselves to acknowledging them.

3 Methodology

The NAvigator uses PAGE XML transcriptions and line pixel coordinates as points of departure, and, except for cropping image snippets for visual support, it does not utilize the original images. This means that, besides the initial ATR by Loghi, the NAvigator requires no computer vision

[1] https://www.nationaalarchief.nl/onderzoeken/archief/2.10.02

[2] www.haicu.science

[3] Agendas are registers of received documents. In *verbaal* archives, these documents are not sorted by entry date into the agenda, but by date of the decision concerning the associated dossier. This makes the agendas less relevant in the search process.

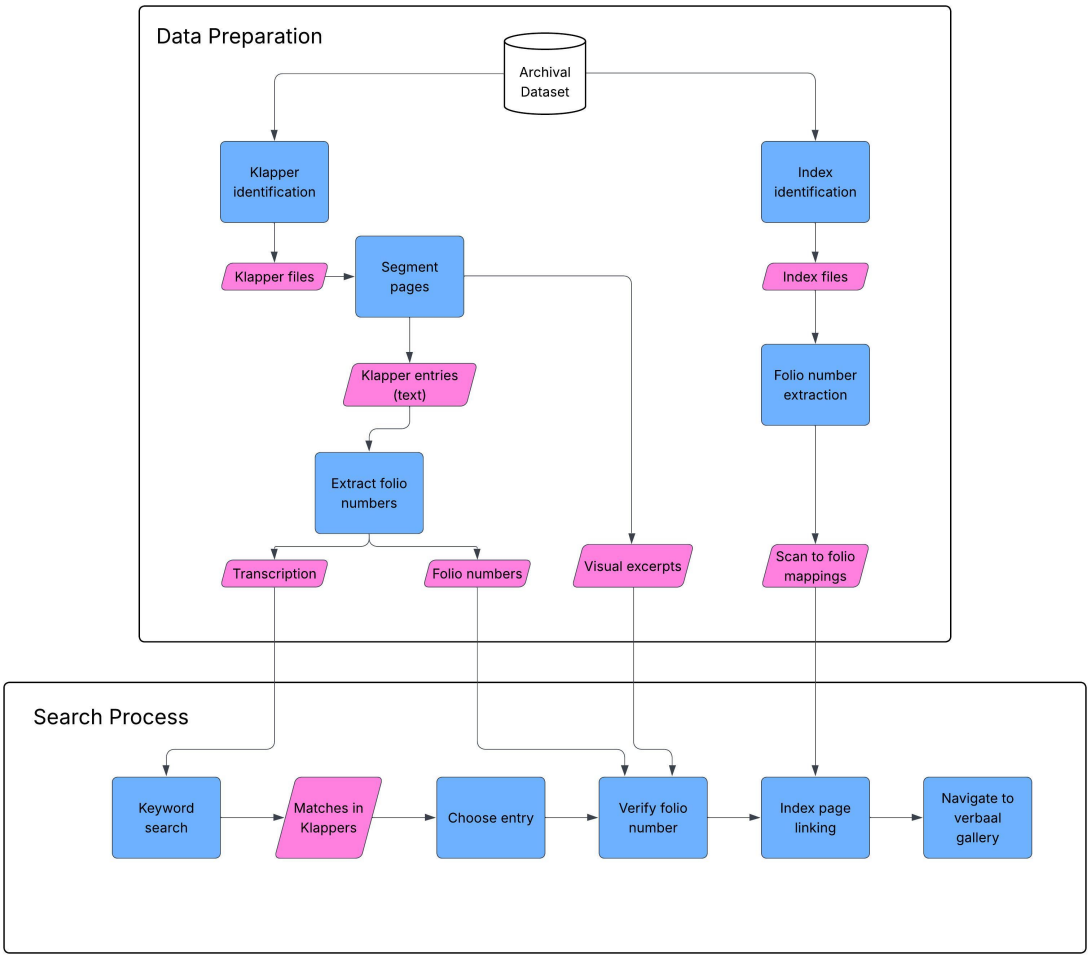

Figure 2: Schematic representation of the NAvigator workflow, showing the preparatory data processing (top) and the search process (bottom).

techniques; instead, it relies on computationally lightweight, rules-based algorithms and classical machine learning techniques, which were executed in Jupyter Notebooks on a consumer-grade laptop.

Figure 2 offers an overview of the entire NAvigator construction process and the search process. First, the *klapper* and index pages are extracted using a classifier. The *klapper* pages are segmented into individual entries, which in turn are split into transcription text and folio numbers using a regular expression. In the process, a visual excerpt (see figure 3) is extracted from the source image to provide visual support during the search. From the index pages, folio numbers are extracted to create a mapping of scan numbers to folio numbers. The search process starts with a fuzzy keyword search through the extracted *klapper* entries. After selecting an entry from the results, the user will be prompted to correct the folio number based on the visual excerpt, allowing navigation to the associated index page. Finally, the user enters the relevant date on the index page into the NAvigator, which opens the corresponding image gallery.

Two separate random forest classifiers were built for the extraction of the *klappers* and indices. Indices were easily detectable as they were written on pages with pre-printed table structures, featuring fully capitalized headers. Simply detecting these headers in the transcription proved sufficient to classify the indices with 100% accuracy. The classifier for *klappers* leveraged the

unique properties of these pages. Each scan (i.e., two pages) contains up to four text columns, each with margins starting at predictable positions. Text outside of these columns is very sparse. The available textual content within the column structure also provides additional information: as it is structured in the form of an alphabetic list, most text lines start with the same letter, which is highly unusual in running text, and most text lines end with a numeral (the folio number). Although using these features did not yield a perfect result, primarily due to near-empty *klapper* pages that were more difficult to detect, it classified the pages with an accuracy of 99.1 percent.

The segmentation of the *klapper* pages into a searchable list of individual entries using a textual approach posed several challenges, including ATR-noise, extraneous textual elements, incorrect text line detection, and multi-line entries, which complicate boundary detection. It was accomplished by first grouping the text lines by column, sorting them based on their y-coordinates, and finally segmenting them based on the line endings, where lines ending on a number were assumed to be the end of an entry.

The further separation of textual content and the folio number of individual entries necessary for linking to the index pages was performed using a regular expression. While the presence of a folio character (f) theoretically simplifies this, the ATR model struggled to identify the character correctly. Instead, the transcriptions showed a great deal of variation, complicating the regular expression. The text and number were stored as key-value pairs alongside the positional information of the entry. For each entry, a visual excerpt from the original image was made based on its coordinates, functioning as a visual verification mechanism.

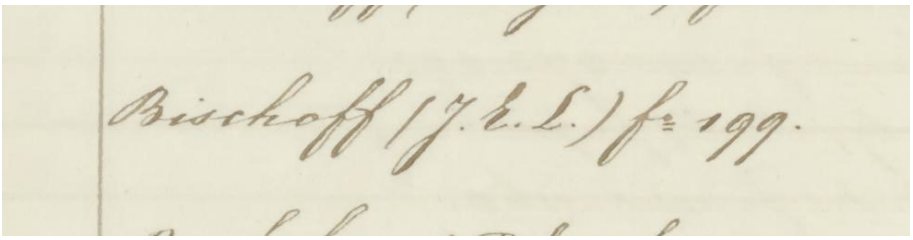

Figure 3: Example of an excerpt. Source: NL-HaNA_2.10.02_5654_0013.jpg

To navigate to the index, it was first necessary to map the folio number used in the physical materials to the scan numbers assigned during the digitization process. Although the folio numbers are sequential, this is challenging, as numerous anomalies complicate the numbering system. Firstly, the ATR failed to detect some numbers correctly. Other numbers have lettered suffixes, which means that several consecutive pages have the same folio number with a different suffix. Some pages are empty and not numbered at all, while others were accidentally duplicated during the scanning process. Finally, sometimes clerks made a mistake and used the same page number twice.

The algorithm for mapping scan numbers to folio numbers starts with the ATR as a basis and attempts to infer the missing values. It uses a rolling average to detect suspiciously high or low values, erasing them before searching for sequences of consecutive or identical numbers, and then attempts to extend these patterns with the missing values. For instance, if a missing value separates the numbers three and five, the system will assume the value is four.

The algorithm deliberately excludes morphological suffixes from its predictive processes, recognizing that these display structural inconsistencies. To compensate for this limitation, the search protocol incorporates a disambiguation procedure employing fuzzy matching to locate the original search term within the selected pages. Thus, it identifies the entry with the highest degree of lexical correspondence. Furthermore, the system enhances user navigation by automatically highlighting the textual passage within the original manuscript scan, thereby facilitating an immediate visual reference to the source material.

The current iteration of the navigation system does not yet facilitate automated linking between the indices and the corresponding *verbalen*. Instead, the interface requires manual user intervention, in which the researcher inputs the date retrieved from the index entry into the navigation module. Subsequently, this process generates a hyperlink to the relevant image gallery, which hosts the associated *verbaal*.

4 Results and Evaluation

The efficacy of the linking operation depends upon the successful completion of each of the following constituent stages: 1) the accurate detection of the *klapper* and index page, 2) the precise extraction of the *klapper* entry, and 3) the subsequent mapping of the identified folio number to the corresponding index scan. By calculating how errors in these steps compound, it is possible to estimate the proportion of *klapper* entries that successfully link to their index and *verbaal*.

The first key metric is the recall of detecting *klapper* and index pages. While precision is also important, for search, the ability to find as many pages as possible is essential. For *klappers*, the classifier showed a recall of .98 and a precision of 1.0. However, the recall of information in terms of individual entries is likely higher, as a paucity of textual content appears to be the primary contributor to misclassifications. For indices, both precision and recall are perfect, due to the pre-printed column headers.

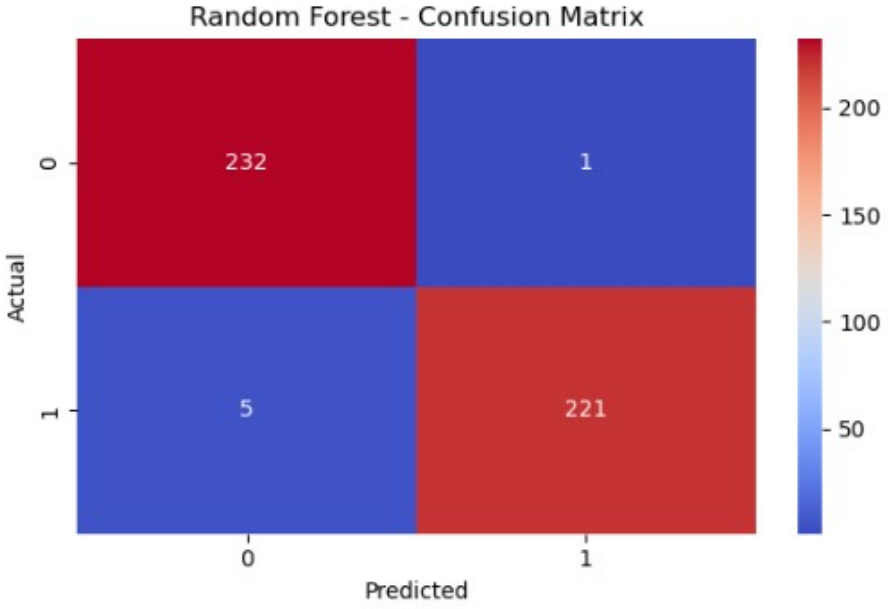

Figure 4: Confusion matrix for the *klapper* classification.

The second metric concerns the extraction of *klapper* entries. This is more challenging to measure, as one of five things can happen: the entry can be a) correctly extracted, b) extracted with the wrong folio number, c) merged with another entry during segmentation, d) composed of noise, or e) missing altogether. A sample of a dozen randomly selected pages containing 564 entries was manually corrected as a test. 384 entries (68%) were correctly transcribed, and a further 159 (28%) were extracted with an incorrect folio number. As the latter can be manually corrected during search thanks to the visual support of the clippings, this means that 96% of all entries are usable in the search. Of the remaining entries, eleven were merged into another, eight contained noise, and two entries were not picked up during extraction.

Extracting page numbers from the index pages presents a technical challenge due to the com-

plex and variable characteristics of the numbering system. The Ministry of Colonies archive contains almost 200,000 index pages, of which over 20,000 proved to be blank. In addition to those, 0.36% of the pages were found to be duplicates accidentally introduced during the digitization process. Excluding those pages, the algorithm failed to identify 4.0% of the remaining pages. A manual verification of the identification accuracy on a sample of 300 pages showed that 53% were correct, another 45% needed further disambiguation due to lettered suffixes, which the algorithm excluded, and 2% were incorrect. Considering the unidentified and incorrectly identified pages, we can expect 90.0% of the searches to arrive at the correct page.

	Page Recognition	**Klapper Entry Extraction**	**Index Linking**	**Total**
Retention (%)	98.0	96.3	90.0	84.7

Table 1: Percentage of data points retained at each step and throughout the full process.

To calculate the total percentage of linkage operations to be successful, calculate the propagation of errors considering the success rates of these three consecutive steps. By multiplying the ratios of correct entries at each step, we can expect around 84.7% of the *klapper* entries to be fully linked to their *index* (see table 1). This number likely represents a conservative estimate, as the misidentified *klapper* pages likely contain substantially fewer entries than the average *klapper* page.

5 Discussion and Future Research

Although the NAvigator is in a prototype stage, preliminary findings for the Ministry of Colonies archive already demonstrate considerable potential in establishing paths of navigation through the *verbaalstelsel-1823* structure. This establishes an important precondition for successful keyword search, which takes the digitized *klappers* as a starting point. It achieves this without relying on computationally expensive methods or labor-intensive human annotator tasks. While the NAvigator is still far removed from being fully automated, it nonetheless achieves substantial acceleration of the search process compared to manual methods as described in section 1. Our paper, therefore, demonstrates the potential of utilizing the archive's structure for unlocking its contents.

One avenue of future research involves further automating the NAvigator, as well as improving information retention throughout the process. The scan-to-page number mapping is the biggest bottleneck in the current process, with only 90% of searches arriving at the correct page, even when ignoring lettered suffixes. Given the complex nature of the page numbering system and the numerous irregularities, further tweaks to the algorithm are unlikely to yield substantial improvements. There is a clear correlation between the number of folio numbers successfully extracted from the XML and the final number of predicted pages (see fig. 5). Hence, the solution is likely to be found in the extraction part of the process. Future research will need to determine the most effective approach to address this problem.

Both the page detection and segmentation steps leave little room for improvement, and any additional gains will likely require improvements in the ATR process, as most errors result from problems in the layout detection and ATR noise. Likewise, the extraction of folio numbers is hampered by the ATR quality. Although 95% of the entries had recognizable folio numbers when combined with the visual verification through the image clippings, the identification rate of the folio numbers does not allow full automation. Resolving the above issues is likely impossible without either repeating the ATR with an improved model or manually correcting the transcriptions, both of which would negate the main advantage of the methodology described in this paper.

Although initial attempts have been made to segment individual indices beyond the page level and link them with their respective *verbaal*, as of the time of writing, the NAvigator is not yet capa-

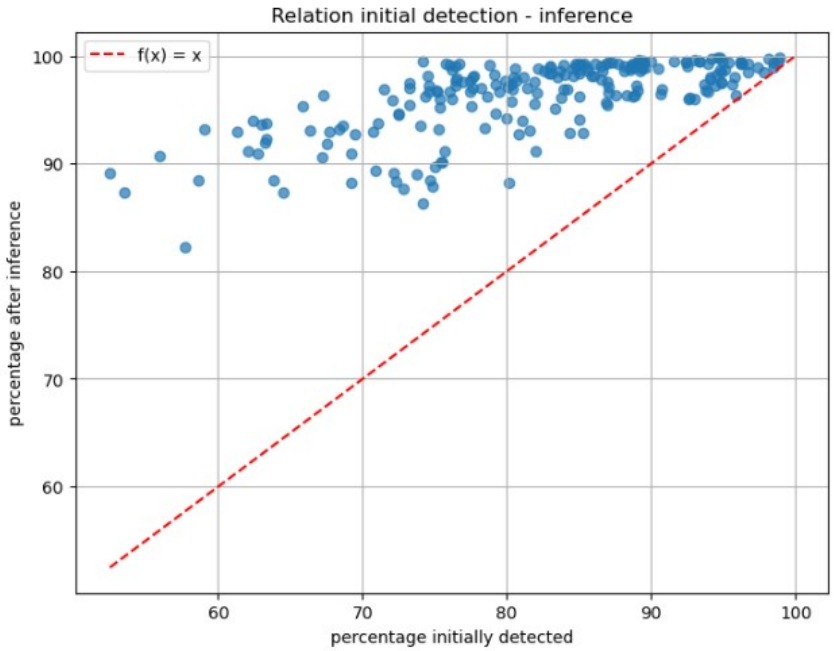

Figure 5: Relationship between the initial percentage of page numbers extracted and the percentage of numbers after inference. Points (volumes) above the dotted line benefited from inference.

ble of this. It can only link to the correct image gallery based on the date the user manually enters on the index page. The segmentation task for indices is much more complicated than for *klappers* due to the complex and inconsistent tabular structure of the indices, and automatic date extraction is hindered by the frequent use of quotation marks or 'idem' to denote repeated dates. Furthermore, additional text-based matching between indices and *verbalen* requires access to transcriptions of the full archive. While these are available for the archive used in this study, this might drastically increase the costs for any project trying to replicate this method, as the number of *verbaal* pages far exceeds the number of *klappers* and indices [24]. In the future, particular attention will be directed towards addressing this issue, as it will form the basis for further research into how multiple instances of the same named entity can be identified and linked together using linked open data.

As of the time of writing, it is unclear how well the NAvigator generalizes to other *verbaalstelsel-1823* archives. While the overall workflow can likely be repeated, the implementation may need to be adjusted to accommodate differences in the specific layout of various archives. The particularities of specific archives might also cause unexpected problems. However, the NAvigator is modular by design, allowing flexibility in implementation, which broadens its applicability. With some adjustments to the computationally light-weight workflow, the tool could be repurposed for unlocking verbal archives other than those using the 1823 system, as they have a similar navigational flow. This would significantly increase the temporal range of archives where the NAvigator is applicable. In general, the tool shows the potential to make the verbaalstelsel-1823 archives accessible to researchers and broader audiences by capitalizing on their historical ordering and management principles.

Acknowledgements

We want to express our gratitude to Rutger van Koert for his work on digitizing the archives. We would also like to thank Liesbeth Keijser of the Dutch National Archives for providing additional materials and information about the archives.

Funding

This research is part of the digital Humanities - Artificial Intelligence - Cultural heritage (HAICu) research project. The HAICu project is funded by the Dutch Research Council/Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)/Nationale Wetenschapsagenda (NWA), project number: NWA.1518.22.105.

CRediT

- S. Peeters: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Software; Visualization; Writing - original draft; Writing: review and editing.
- C.A. Romein: Supervision; Writing - review and editing.
- A. Weber: Funding acquisition; Supervision; Writing - review and editing.

Code and Data

- The NAvigator code is available at https://github.com/sebastiaanpeeters/NAvigator
- The intermediate data, needed to run the NAvigator, is available on Zenodo: https://doi.org/10.5281/zenodo.17453406
- For access to the transcription files, please get in touch with the Nationaal Archief.

References

[1] Candela, Gustavo, Pereda, Javier, Sáez, Dolores, Escobar, Pilar, Sánchez, Alexander, Torres, Andrés Villa, Palacios, Albert A., McDonough, Kelly, and Murrieta-Flores, Patricia. “An Ontological Approach for Unlocking the Colonial Archive”. In: *J. Comput. Cult. Herit. 16*, no. 4 (Nov. 2023), 74:1–74:18. ISSN: 1556-4673. DOI: 10.1145/3594727. (Visited on 05/22/2025).

[2] Colavizza, Giovanni, Blanke, Tobias, Jeurgens, Charles, and Noordegraaf, Julia. “Archives and AI: An Overview of Current Debates and Future Perspectives”. In: *J. Comput. Cult. Herit. 15*, no. 1 (Dec. 2021), 4:1–4:15. ISSN: 1556-4673. DOI: 10.1145/3479010. (Visited on 07/15/2025).

[3] Colavizza, Giovanni, Ehrmann, Maud, and Bortoluzzi, Fabio. “Index-Driven Digitization and Indexation of Historical Archives”. English. In: *Frontiers in Digital Humanities 6* (Mar. 2019). Publisher: Frontiers. ISSN: 2297-2668. DOI: 10.3389/fdigh.2019.00004. (Visited on 08/26/2025).

[4] Guralnick, Robert et al. “Humans in the loop: Community science and machine learning synergies for overcoming herbarium digitization bottlenecks”. en. In: *Applications in Plant Sciences 12*, no. 1 (2024). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aps3.11560, e11560. ISSN: 2168-0450. DOI: 10.1002/aps3.11560. (Visited on 04/15/2025).

[5] Hawkins, Ashleigh. “Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web”. en. In: *Archival Science 22*, no. 3 (Sept. 2022), pp. 319–344. ISSN: 1573-7500. DOI: 10.1007/s10502-021-09381-0. (Visited on 05/21/2025).

[6] Jeurgens, Charles. “Schetsboek | 1 januari 2016 | pagina 55”. nl. Jan. 2016. URL: https://kvan.courant.nu/issue/SB/2016-01-01/edition/null/page/55 (visited on 07/01/2025).

[7] Ketelaar, Eric. “Veranderingen in archiefvorming en archiefgebruik in een veranderende samenleving 1747-18471”. en. In: *Tijdschrift voor Geschiedenis 133*, no. 3 (Nov. 2020). Publisher: Amsterdam University Press, pp. 435–456. ISSN: 0040-7518, 2352-1163. DOI: 10.5117/TVGESCH2020.3.002.KETE. (Visited on 05/28/2025).

[8] Klut, Stefan, Koert, Rutger van, and Sluijter, Ronald. “Laypa: A Novel Framework for Applying Segmentation Networks to Historical Documents”. In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. HIP ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 67–72. ISBN: 979-8-4007-0841-1. DOI: 10.1145/3604951.3605520. URL: https://dl.acm.org/doi/10.1145/3604951.3605520 (visited on 07/15/2025).

[9] Koert, Rutger van, Klut, Stefan, Koornstra, Tim, Maas, Martijn, and Peters, Luke. “Loghi: An End-to-End Framework for Making Historical Documents Machine-Readable”. en. In: *Document Analysis and Recognition – ICDAR 2024 Workshops*, ed. by Harold Mouchère and Anna Zhu. Cham: Springer Nature Switzerland, 2024, pp. 73–88. ISBN: 978-3-031-70645-5. DOI: 10.1007/978-3-031-70645-5_6.

[10] Koolen, Marijn and Hoekstra, F.G. “Detecting Formulaic Language Use in Historical Administrative Corpora”. English. In: *Proceedings of the Computational Humanities Research Conference 2022*, ed. by Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo. CEUR Workshop Proceedings 3290. 2022, pp. 127–151.

[11] Koolen, Marijn, Hoekstra, Rik, Oddens, Joris, Sluijter, Ronald, Van Koert, Rutger, Brouwer, Gijsjan, and Brugman, Hennie. “The Value of Preexisting Structures for Digital Access: Modelling the Resolutions of the Dutch States General”. In: *J. Comput. Cult. Herit. 16*, no. 1 (June 2023), 1:1–1:24. ISSN: 1556-4673. DOI: 10.1145/3575864. (Visited on 05/01/2025).

[12] Linhares Pontes, Elvys, Cabrera-Diego, Luis Adrián, Moreno, Jose G., Boros, Emanuela, Hamdi, Ahmed, Doucet, Antoine, Sidere, Nicolas, and Coustaty, Mickaël. “MELHISSA: a multilingual entity linking architecture for historical press articles”. en. In: *International Journal on Digital Libraries 23*, no. 2 (June 2022), pp. 133–160. ISSN: 1432-1300. DOI: 10.1007/s00799-021-00319-6. (Visited on 06/10/2025).

[13] Luthra, Mrinalini, Todorov, Konstantin, Jeurgens, Charles, and Colavizza, Giovanni. “Unsilencing Colonial Archives via Automated Entity Recognition”. arXiv:2210.02194 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.02194. (Visited on 07/15/2025).

[14] Manjavacas Arevalo, Enrique and Fonteyn, Lauren. “Non-Parametric Word Sense Disambiguation for Historical Languages”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, ed. by Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 123–134. DOI: 10.18653/v1/2022.nlp4dh-1.16. (Visited on 07/09/2025).

[15] Ngo, Vuong M., Munnelly, Gary, Orlandi, Fabrizio, Crooks, Peter, O’Sullivan, Declan, and Conlan, Owen. “A Semantic Search Engine for Historical Handwritten Document Images”. en. In: *Linking Theory and Practice of Digital Libraries*, ed. by Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen. Cham: Springer International Publishing, 2021, pp. 60–65. ISBN: 978-3-030-86324-1. DOI: 10.1007/978-3-030-86324-1_7.

[16] Nockels, Joseph. *Making the past readable: a study of the impact of handwritten text recognition (HTR) on libraries and their users*. en. PhD Doctor of Philosophy. Edinburgh, UK: The University of Edinburgh, 2025. DOI: 10.7488/era/5988.

[17] Nockels, Joseph, Gooding, Paul, and Terras, Melissa. “The implications of handwritten text recognition for accessing the past at scale”. en. In: *Journal of Documentation 80*, no. 7 (Apr. 2024). Publisher: Emerald Publishing Limited, pp. 148–167. ISSN: 0022-0418. DOI: 10.1108/JD-09-2023-0183. (Visited on 05/01/2025).

[18] Oort, Thunnis van, Prats López, Montserrat, Ganzevoort, Wessel, and Galen, Coen van. “Citizen Science and Participatory Engagement with a Contentious Past”. en. In: *The Palgrave Encyclopedia of Cultural Heritage and Conflict*. Palgrave Macmillan, Cham, 2025, pp. 1–7. ISBN: 978-3-030-61493-5. DOI: 10.1007/978-3-030-61493-5_319-1. (Visited on 07/15/2025).

[19] Otten, F.J.M. “Gids voor de archieven van de ministeries en de Hoge Colleges van Staat, 1813-1940”. 2004. URL: https://resources.huygens.knaw.nl/retroboeken/archiefgids_overheid/#page=0&accessor=toc_1&view=homePane (visited on 05/02/2025).

[20] Romein, C. A., Rabus, A., Leifert, G., and Ströbel, P. B. “Assessing advanced handwritten text recognition engines for digitizing historical documents”. en. In: *International Journal of Digital Humanities* (May 2025). ISSN: 2524-7840. DOI: 10.1007/s42803-025-00100-0. (Visited on 05/13/2025).

[21] Romein, Christel Annemieke, Veldhoen, Sara, and Gruijter, Michel de. “Entangled Histories Ordinances Low Countries”. en. Jan. 2020. DOI: 10.5281/ZENODO.3567844. (Visited on 07/01/2025).

[22] Schuijlenburg, Koen van, Romein, C. A. (Annemieke), Wolf, Ben, Weggeman, Sjors, Peeters, Sebastiaan, Dhali, Maruf A., Dijkstra, Klaas, Weber, Andreas, and Schomaker, Lambert. “The HAICu Project (WP2)”. July 2025. DOI: 10.5281/zenodo.15829129. URL: https://zenodo.org/records/15829129 (visited on 10/28/2025).

[23] Stoler, Ann Laura. *Along the archival grain: epistemic anxieties and colonial common sense*. eng. Princeton, NJ: Princeton Univ. Press, 2009. ISBN: 978-0-691-14636-2.

[24] Vriend, Nico. “An archive in numbers: the pulse of the Dutch Ministry of Colonies, 1813–1900”. en. In: *Archival Science 25*, no. 2 (June 2025), p. 17. ISSN: 1573-7500. DOI: 10.1007/s10502-025-09483-z. (Visited on 07/01/2025).