Content providers, Researchers, Technology and the Crowd: Discovering the Best Possible Collaborative Strategies for Datafication and Publication of a Dutch Historical Newspaper Corpus

Depuydt, Katrien
katrien.depuydt@ivdnt.org
Instituut voor de Nederlandse Taal, Netherlands, The

van der Sijs, Nicoline
post@nicolinevdsijs.nl
Instituut voor de Nederlandse Taal, Netherlands, The

de Does, Jesse
jesse.dedoes@ivdnt.org
Instituut voor de Nederlandse Taal, Netherlands, The

de Jong, Ruud
ruud.dejong@ivdnt.org
Instituut voor de Nederlandse Taal, Netherlands, The

de Bonth, Roland
roland.debonth@ivdnt.org
Instituut voor de Nederlandse Taal, Netherlands, The

Fannee, Mathieu
mathieu.fannee@ivdnt.org
Instituut voor de Nederlandse Taal, Netherlands, The

Romein, Annemieke
annemieke.romein@huygens.knaw.nl
Huygens Instituut, Netherlands, The

van Zundert, Joris
joris.van.zundert@huygens.knaw.nl
Huygens Instituut, Netherlands, The

The National Library of the Netherlands (KB) makes over 130 million pages of newspapers, journals and books accessible via Delpher, and the collection is still growing. Delpher offers users the option to search the texts, to create selections based on metadata, and highlights results in the images (which is helpful because newspaper pages, for instance, can be a complex mix of articles). The digitisation and datafication is outsourced and omnifont OCR [1] is used.The advantage of this strategy is that large amounts of material can be made available quickly at low cost. The quality of the OCR for challenging material however, is very poor and unsuitable for most types of research. Moreover, the data is not structured in such a way that the material is suitable for the type of retrieval and research done in digital humanities and corpus linguistics.

A good example of challenging material is the collection of 17th century newspapers. They are an important source for historical (linguistic) research. However, the poor quality of the OCRed text, the inadequate structuring, and the available metadata in the KB version makes such research virtually impossible. Van der Sijs therefore proposed to work with volunteers [2] on a better datafication of the text and improved structuring and metadata. The idea was not only to deliver a better version to the KB for Delpher, but also to make the newspaper accessible in an alternative environment, suitable for both digital humanities researchers and corpus linguists.

The work was executed in two stages. A thorough evaluation of the workflow led to an entirely new approach for the second stage. In our contribution, we will present both approaches, discuss the problems we encountered, and the solutions we devised.

For both stages the starting point was a collection of images, the available OCR in ALTO format and associated metadata. In the ALTO files, region segmentation was already corrected and article segmentation was applied. The quality of the OCR was too low to serve as a basis for correction. Furthermore, the article segmentation was insufficiently detailed: newspaper sections were identified rather than newspaper articles.

In stage one, which ran from 2014 to mid 2022, a large group of volunteers rekeyed the text using a purpose built editor. A smaller group of more experienced volunteers corrected the 19 million words, after which interns added article segmentation and additional metadata in a database environment. The post processing of that work was done at the Instituut voor de Nederlandse Taal. The text was automatically enriched with part of speech and lemma and the dataset was put online in 2022.

After careful evaluation, it was decided to take a completely different approach for the second stage. One major unexpected issue was the fact that there were considerable problems with the metadata and the grouping of images in newspaper issues in the KB dataset. This led to unnecessary double keying (10% duplicates), incomplete issues [3], and a huge amount of suggestions for correction from both volunteers and interns, which had to be resolved in the postprocessing. We were also not happy with the way the article segmentation had been performed nor with the strategy for correction and extension of the metadata. We also wanted to benefit from the technological progress in the field of digitisation that has been made since 2014.

For more efficient processing, we now start with a check of the issues and the metadata of the KB set, to obtain a clean dataset with correct metadata on issue level. For that a special tool was developed. For the transcription of the text, we have used Transkribus' HTR+ and since 2023 Pylaia. The KB ALTO is converted to Page XML; images and XML are loaded into Transkribus, where new line segmentation and text recognition take place. Volunteers use Transkribus Lite to correct the layout analysis and the text. Training of the engine is done per newspaper and repeated at least once to get the best possible result. The error rate as well as an analysis of the type of errors in the automatic transcription help us to evaluate whether manual post-correction of the text is always necessary. Headers and subheaders in the text are tagged

so that article segmentation and article metadata are the result of a further conversion.

With the support of CLARIAH, we were able to realise a successful collaboration between content holder, researchers, software engineers, experts in digitisation and in corpus building. This has drastically improved both the digitisation workflow and the output.

Notes

1. Abbyy Finereader.
2. Volunteers of the *Stichting Vrijwilligersnetwerk Nederlandse Taal.*
3. The pages belonged to different newspaper issues.

Bibliography

P. Kahle, S. Colutto, G. Hackl and G. Mühlberger, "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 19-24, doi: 10.1109/ICDAR.2017.307.

Nockels J, Gooding P, Ames S, Terras M. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. Arch Sci (Dordr). 2022;22(3):367-392. doi: 10.1007/s10502-022-09397-0. Epub 2022 Jun 17. PMID: 35730063; PMCID: PMC9205146.

van der Sijs, N. (2019). De krant hoorde erbij. Geschiedenis Magazine, 3, 37-40.

Ströbel P, Clematide S, Improving OCR of black letter in historical newspapers: the unreasonable effectiveness of HTR models on low-resolution images. Paper presented at Digital Humanities 2019, Zurich. https://www.zora.uzh.ch/id/eprint/177164/1/Improving_OCR_of_Black_ Letter_in_Historical_Newspapers_The_Unreasonable_Effecti.pdf