An Optical Character Recognition Software Benchmark for Old Dutch Texts on the EYRA Platform

Mirjam Cuper[1], dr. Adriënne Mendrik[2], Maarten van Meersbergen[2], Tom Klaver[2], Pushpanjali Pawar[2], Dr. Annette Langedijk[3], Lotte Wilms[1]

[1] National Library of the Netherlands (KB), [2] The Netherlands eScience Center, [3] SURF

Digitized collections of printed historical texts are important for research in Digital Humanities. However, acquiring high-quality machine readable texts using currently available Optical Character Recognition (OCR) methods is a challenge. OCR Quality is affected by old fonts, old printing techniques, bleedthrough of the ink, paper quality, old spelling, multiple columns and so on. It is unclear which OCR methods perform best. Therefore, we are currently in the process of setting up a benchmark to enable the evaluation of the performance of OCR software on old Dutch texts. The benchmark is being set-up on the EYRA benchmark platform (eyrabenchmark.net) developed by The Netherlands eScience Center and SURF.

For the pilot version of the benchmark a data set containing 2055 Dutch book pages (1630- 1796) and 1024 Dutch newspaper pages (1618-1945) is made available by the National Library of the Netherlands (KB). This data set contains both scanned pages (OCR method input data) and machine readable text (ground truth that can be used to assess the quality of the OCR method output). This dataset is split in training and validation data. The training data can be downloaded and used by algorithm developers to train their OCR algorithms or tune their workflows (pre-processing, layout segmentation, character recognition, post-processing). The EYRA platform offers algorithm developers the opportunity to submit their OCR algorithm or workflow to the EYRA platform in a docker container. The docker container will, in turn, be run on the validation data in the cloud on the Dutch national infrastructure of SURF. The advantage of this set-up, is that it prevents over-tuning on the validation data and therefore provides a fair comparison of the performance of the OCR methods. Also, if new validation data is available and added to the benchmark later on, the OCR methods can easily be re-run on the new data.

Various metrics could be used to assess the performance of the OCR methods in comparison to the ground truth. In the pilot we will use the most commonly used metrics (Character Error Rate and Word Error Rate). However, we are planning to add more metrics later on, that address different aspects of the OCR method performance. The EYRA platform uses Observable (observablehq.com) to visualize algorithm results on the platform, to gain more insight into algorithm performance. These visualizations can easily be integrated in a journal paper, which promotes replication of result visualizations. Furthermore the OCR benchmark provides an easy way for OCR method developers to compare their method to other existing methods, by providing the data, metrics, ground truth and algorithms for comparison, replicating algorithm validation in the experiment and results section of a journal paper. For the National Library of the Netherlands, this benchmark provides a way to gain insight into the performance of OCR methods and to select the best available OCR method for their problem of digitizing old Dutch texts. This in turn will provide higher quality digitized texts for Digital Humanity research.