Contribution ID: **30** Type: **Talk (15min + 5min)**

OCR(-D)4all - An easy to use and highly adaptable open-source solution for automatic text recognition of historical printings and manuscripts

Wednesday, March 6, 2024 1:30 PM (20 minutes)

Despite the various challenges in automatic text recognition for printed (OCR) and handwritten (HTR) material, great progress has been made during the last decade. Several milestones have been reached regarding the actual text recognition step but also in layout analysis and pre- and postprocessing. Additionally, free open-source implementations of related tools and algorithms are released constantly. While these allow tackling highly heterogeneous use-cases ranging from mass full-text digitisation in libraries to the processing of individual documents, including specific production of training data (Ground Truth, GT) and subsequent training of deep learning OCR/HTR models, they rarely offer standardized interfaces and a low barrier of entry for non-technical users.

To solve the issue of easy-to-use, flexible, connectable, and sustainable combination and application of current and future individual technical OCR/HTR solutions we introduce the open-source tool OCR4all, which in turn leverages the open-source OCR/HTR framework OCR-D. In the following we discuss how users can benefit from the powerful combination of the two.

Whereas OCR-D concerns on the standardized implementation of single-step processors, OCR4all aims at enabling any user –even those without technical background –to perform OCR/HTR completely on their own and in great quality, while also offering tools to manually generate training data in order to train more performant work-specific models and consequently improve the output of the fully automated OCR/HTR processors.

To combine both approaches we engineered interfaces between OCR-D tools and OCR4all based on the OCR-D specifications. For a very flexible integration of different OCR and especially OCR-D processors, OCR4all relies on the Java Service Provider Interface. OCR-D processors, which are written in Python, are plugged into OCR4all as containerized service providers that implement the required interfaces, either through a simple manual configuration or fully automatically. The latter is achieved by leveraging the OCR-D-Tool-JSON which contains all necessary information about input/output relationships and available parameters and is mandatory for all OCR-D processors.

Due to the above described adaptations and the thereby extended flexibility, OCR4all can now be applied to an even more heterogeneous selection of materials and use cases. Its main focus still lies on the interactive high quality processing of challenging early printings and manuscripts by non-technical users, including correction and GT production and consequently material-specific training. However the applicability of OCR4all for mass digitization, e.g. in libraries and archives, has vastly improved.

Slot length

Primary authors: Dr REUL, Christian (Zentrum für Philologie und Digitalität); BAIER SAIP, Herbert (Zentrum für Philologie und Digitalität); NÖTH, Maximilian (Zentrum für Philologie und Digitalität)

Presenters: BAIER SAIP, Herbert (Zentrum für Philologie und Digitalität); NÖTH, Maximilian (Zentrum für Philologie und Digitalität)

Session Classification: Research Software for Computing and Visualising Text

Track Classification: Research Software: Research Software for Computing and Visualising Text