Attention

The site is currently being revised. If you encounter any problems , please **contact us.**

2. Preparation of scans and image files (ScanTailor)

Works for which an OCR is to be carried out are often only available as facsimiles. Although their individual images usually have a good to very good quality, their overall system is rather unsuitable for direct import into OCR4all. This is the case, for example, when image files show parts of the book cover or back as well as parts of a support surface in addition to the actual page content. If such images are binarized during the workflow, the different contrasts in the originals result in black dividing lines, which, in addition to the actual OCR, are particularly problematic for segmentation. The rotation of scans or the display of two pages per scan are also common problems. However, these can easily be avoided by appropriate pre-processing of image files: The goal must therefore be to use scans for working with OCR4all that only show the content of a single page that is intended for recognition. At the same time, in addition to the actually relevant so-called content, these pre-processed images should also have sufficient unprinted or blank side surfaces in order not to complicate certain segmentation processes, for example. So it makes sense to remove exactly those parts of the image that do not belong to the actual printed page and therefore do not need to be recorded, but also to retain as much of the original printed page as possible (ie not completely removing page margins, for example). Theoretically, all image editing programs can be used for this (GIMP, Photoshop, etc.). If the work to be edited is available as a PDF, it is also possible Batch crop or rotate all pages within Adobe Acrobat DC. However, we recommend working with ScanTailor at this point, since large amounts of images can be processed easily and in a standardized manner in a relatively short time. Detailed instructions and video material can be found here. This step is optional and not part of the OCR4all workflow, which is why no support can be provided here. Each user must decide for himself whether additional pre-processing of this kind would be profitable for his work or is even necessary. Detailed instructions and video material can be found here. This step

profitable for his work or is even necessary. Detailed instructions and video material can be found here. This step is optional and not part of the OCR4all workflow, which is why no support can be provided here. Each user must decide for himself whether additional pre-processing of this kind would be profitable for his work or is even necessary.

←1.2 Setup and Folder Structure

3.1 Start OCR4all→