OCR4all - Open-Source OCR and HTR Across the Centuries

Langhanki, Florian
florian.langhanki@uni-wuerzburg.de
University of Wuerzburg, Germany

Wehner, Maximilian
maximilian.wehner@uni-wuerzburg.de
University of Wuerzburg, Germany

Roeder, Torsten
torsten.roeder@uni-wuerzburg.de
University of Wuerzburg, Germany

Reul, Christian
christian.reul@uni-wuerzburg.de
University of Wuerzburg, Germany

Topic and Aims of the Workshop

Automated text recognition is an ever-present task in the humanities. A multitude of manuscripts and prints from all epochs and cultures have neither been edited nor fully indexed. Even though a considerable amount of them have been digitized, they are usually only accessible as images or PDF files. Machine-actionable transcriptions are often not available, despite being imperative for full-text search, annotation, scholarly editions and text analyses. Consequently, public institutions – including (digital) humanities scholars – require easy to use software solutions which enable them to perform high-quality OCR (Optical Character Recognition) [1] .

This proposed full-day workshop introduces the participants to the completely open-source and free of charge software OCR4all (https://www.ocr4all.org) which, unlike other available platforms like Transkribus [2] , eScriptorium [3] , or PERO-OCR [4] project [5] workflows exactly to the specific needs of the material at hand.

OCR4all

OCR4all provides a fully automated OCR workflow by combining different OCR solutions. At pretty much any stage of the workflow the user can interact with the results in order to minimize consequential errors and optimize the end result. Initially adapted to the difficulties of early modern prints from the 15th and 16th century (project Narragonien digital [6] projects. The software has undergone continuous development, resulting in significant advances in layout analysis and highest recognition accuracy through robust OCR models. [7]

Working with OCR4all

OCR4all integrates and uses existing software solutions, which are combined according to their strengths to form a coherent OCR workflow. The typical steps of such a workflow are **Preprocessing**, **Layout Analysis**, **Recognition,** and **Post-Correction** (Fig. 1).

Fig. 1: Main components of a typical OCR workflow. From left to right: original image, Preprocessing, Layout Analysis (regions & lines), Text Recognition, Post-Correction

During **Preprocessing**, the images are deskewed and binarised (OCRopus [8]). This is followed by the **Layout Analysis** step (Kraken [9] the user can intervene to correct markup using LAREX [10] recognition.

After the preliminary steps, machine-actionable text can be generated from the available line images (by the integrated OCR engine Calamari [11]). OCR4all already provides excellent *mixed models* for this, which have been trained on a very large number of text lines of different fonts in order to be able to recognize a wide range of different typefaces.

The accuracy of the **Recognition** depends heavily on the material. To calculate the error rate of the text recognition, the evaluation module compares the originally recognized text with the correct transcription produced by the user.

If the mixed model does not meet the requirements, it is possible to comfortably train a so-called *book-specific model*. For this, it is necessary to provide the machine with a sufficient amount of training material. While the general approach never changes the amount of required training data and resulting recognition accuracy depends on the underlying material (printing or manuscript, typeface, script, quality of the scan, ...). [12]

The overall manual effort required strongly depends on various factors (material, use case, user requirements, ...) and varies between almost no manual effort at all (fully automated run) and very extensive manual interventions (fine-grained semantic markup and, above all, extensive GT creation or even complete correction of the text) [13] (**Post-Correction**).

Particularly Worth Mentioning

2.2.1 Full OCR-D Compatibility

OCR-D is a large DFG (German Research Foundation) funding initiative whose stated goal is to “prepare the mass processing of prints published in German-speaking areas from 1500 to 1800 to make them accessible as high-quality research data”. [14]

To achieve this, OCR-D places great emphasis on interoperability. Thus, a framework has been created in which currently over fifty open-source OCR solutions, so-called *processors*, are compatible with each other and can be freely combined into workflows. It is important to note that a processor is not Kraken, Calamari or Tesseract as a whole, but for example tesseract-binarize kraken-segment, or calamari-recognize. [15]

Since the idea of bringing different OCR solutions together in one workflow is very similar to the approach of OCR4all it stood

to reason to bring both projects, OCR4all and OCR-D, closer together. This is currently happening in the DFG-funded project OCR4all-libraries [16] solutions via OCR4all, even for less-technical or non-technical users. A second goal is to further optimize the OCR result for example by comparing and improving workflow configurations. At the time the workshop takes place the support of OCR-D solutions in OCR4all will be completed and fully operational allowing the creation of even more versatile and powerful workflows.

2.2.2 LAREX

Originally developed for interactive segmentation and markup of particularly demanding early prints, LAREX meanwhile represents a comprehensive correction tool. Almost all (partial) results of a typical OCR workflow (region coordinates and types, line coordinates, text) can be corrected comfortably and precisely if required, thus generating high-quality training material for a wide variety of applications.

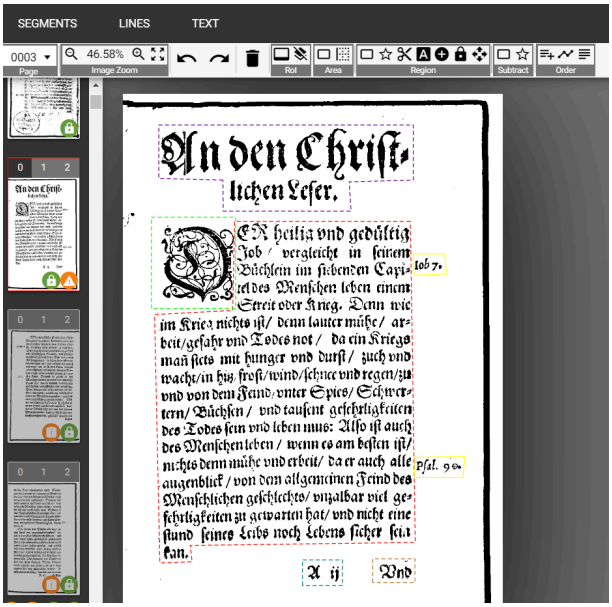

Fig. 2: The Region Segmentation in LAREX.

Within LAREX, different region types can be defined and labeled, either automatically or manually (see Fig. 2). The regions drawn in this way can be subsequently adapted and modified (split, joined, reassigned) using a wide range of tools. Additionally, a reading order can be defined, which is especially relevant for complex layouts.

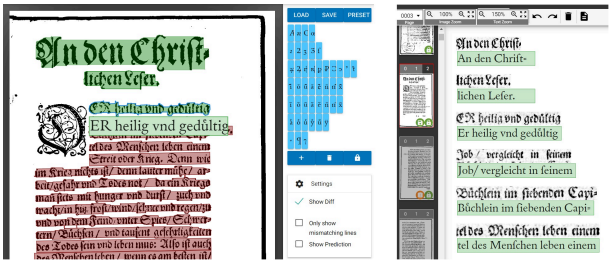

Fig. 3: Ground Truth Production in LAREX: page-based view (left) with Virtual Keyboard (middle); line-based view (right).

Regarding textual correction, LAREX offers two separate corrections views, one page- and one line-based (Fig. 3). A customizable Virtual Keyboard enables the insertion of special characters.

2.2.3 Mixed Models

The progress in development and research is reflected by the mixed models that have been created and continuously developed to this point. Thanks to extensive preliminary work, OCR4all offers a large number of highly performant and free-to-use OCR models. [17] amounts of heterogeneous training material. Currently, OCR4all already contains various mixed models ranging from medieval bastard scripts to Fraktur types of the 19th century. [18] unnecessary for a wide variety of materials.

Workshop Concept

In this workshop, participants will learn to independently perform the entire OCR workflow from the image file to the finished text. Suitable material will be provided but participants are also invited to bring their own. An OCR4all instance dedicated to the workshop will be hosted on servers of the University of Würzburg and can be accessed and worked on by the participants using their own laptops. This format has been performed successfully at various instances.

For this course we estimate a full day to thoroughly introduce the subject of OCR, its problems, its possibilities and most importantly our software. For a face-to-face event, we consider a number of 25 participants to be appropriate.

The workshop will start with a short introduction to the topic and the software (0.5h). Afterwards, the participants will perform the first steps independently and familiarize themselves with the software (0.5h). Afterwards, the layout analysis is carried out in a longer session with the support of the tutors (2h). The recognition and the subsequent correction will also be done in individual or group work (2h). Between the sessions there will be short input phases. If necessary, help and short instructions will be provided by the tutors at any time. Finally, there will be enough room for further questions (1h) and breaks taken at regular intervals.

Questions to be addressed during the workshop:

- What kind of data is OCR4all applicable to?
- How does OCR4all's workflow change depending on the underlying material and how does working on prints differ from working on manuscripts?
- What manual effort can be expected at different processing phases of the material?
- To what extent can the workflow be automated depending on the underlying material, intended use and the expectation for the result?
- How and according to which specifications can work-specific text recognition models be trained? What recognition accuracies can be expected?
- What is OCR-D and what are the additional benefits of integrating OCR-D solutions?

After the workshop, each participant will be able to independently perform OCR tasks with OCR4all. No prior knowledge is required. It will be suitable for any audience - “4all”!

Workshop Tutors

Florian Langhanki, M.A. works as a research associate at the ‘Center for Philology and Digitality’ (Würzburg, Germany) and

as lecturer at the Department for German Philology at the University of Würzburg. His research interests are early modern literature and the OCR/HTR of medieval and early modern prints and manuscripts.

Maximilian Wehner is a research associate at the Department of German Philology at the University of Würzburg. His research interests include early modern literature, the OCR or HTR of medieval and early modern prints and manuscripts.

Dr. Torsten Roeder works as a research associate at the 'Center for Philology and Digitality' (Würzburg, Germany) with a focus on digital edition and the curation of database resources.

Dr. Christian Reul heads the digitization unit of the 'Center for Philology and Digitality' at the University of Würzburg (Germany). His research focuses on OCR/HTR on historical material and the further development of the corresponding software.

Notes

1. Since processing printed and handwritten (HTR) material is very similar on a technical level we use the term "OCR" as a general term which relies to both application scenarios.
2. Kahle et al. 2017.
3. Kiessling et al. 2019.
4. Kodym / Hradiš 2021.
5. DFG-funded initiative for Optical CHaracter Recognition development, https://ocr-d.de/en
6. https://www.narragonien-digital.de
7. https://www.ocr4all.org
8. https://github.com/ocropus
9. https://kraken.re/master/index.html
10. https://www.uni-wuerzburg.de/zpd/larex/
11. https://github.com/Calamari-OCR/calamari
12. Reul et al. 2022.
13. Reul et al. 2019.
14. https://ocr-d.de/en/about
15. Neudecker et al. 2019.
16. https://gepris.dfg.de/gepris/projekt/460665940?language=en
17. https://github.com/OCR4all/ocr4all_models
18. Reul et al. 2022.

Bibliography

Kahle, Philip / **Colutto, Sebastian** / **Hackl, Günter** / **Mühlberger, Günter** : Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 19-24. URL: https://doi.org/10.1109/ICDAR.2017.307

Kiessling, Benjamin / **Tissot, Robin** / **Stokes, Peter** / **Stökl Ben Ezra, Daniel** : eScriptorium: An Open Source Platform for Historical Document Analysis. In: International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, NSW, Australia, 2019, pp. 19-19. URL: https://doi.org/10.1109/ICDARW.2019.10032

Neudecker, Clemens / **Baierer, Konstantin** / **Federbusch, Maria** / **Boenig, Matthias** / **Würzner, Kay-Michael** / **Hartmann, Volker** / **Herrmann, Elisa** : OCR-D: An end-to-end open source OCR framework for historical printed documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019). Association for Computing Machinery, New York, NY, USA, 2019, pp. 53–58. URL: https://doi.org/10.1145/3322905.3322917

Reul, Christian / **Christ, Dennis** / **Hartelt, Alexander** / **Balbach, Nico** / **Wehner, Maximilian** / **Springmann, Uwe** / **Wick, Christoph** / **Grundig, Christine** / **Büttner, Andreas** / **Puppe, Frank** : OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. In: Applied Sciences 2019. (9) 22. URL: https://www.mdpi.com/2076-3417/9/22/4853

Reul, Christian / **Tomasek, Stefan** / **Langhanki, Florian** / **Springmann, Uwe** : Open Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning. In: Uchida, S., Barney, E., Eglin, V. (eds) Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science, vol. 13237. Springer, Cham. URL: https://doi.org/10.1007/978-3-031-06555-2_28