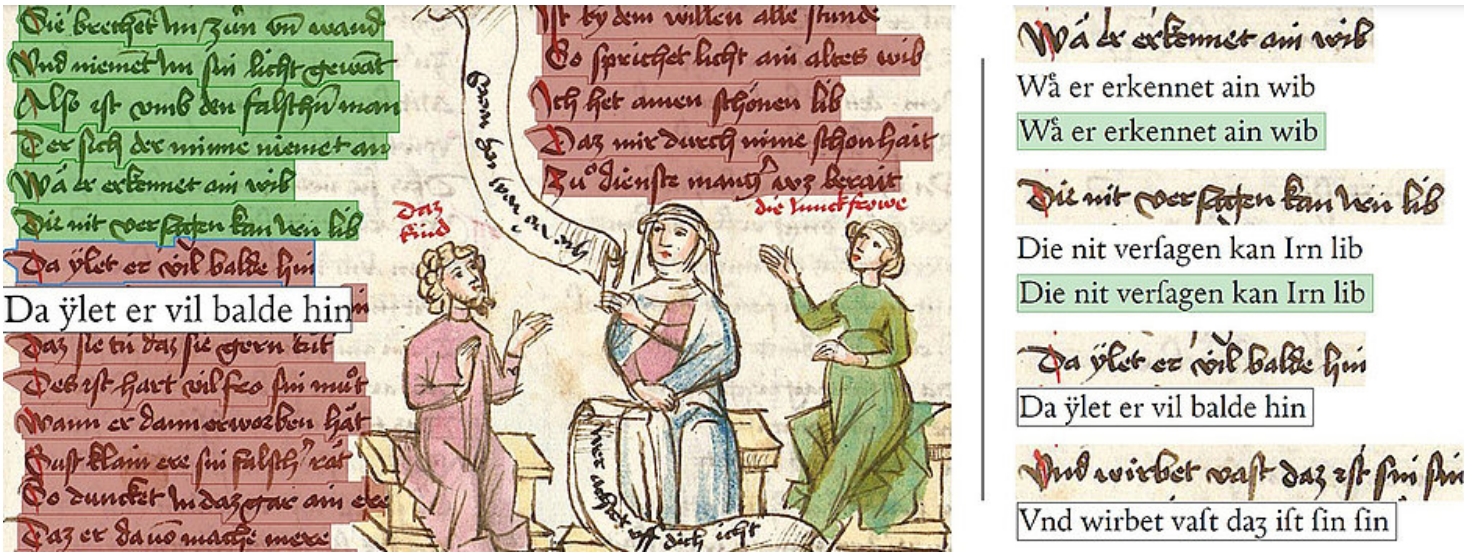
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Recognizing historical writings digitally

01/18/2022

The text recognition software OCR4all is used with great success for historical prints. Now she is being trained on old manuscripts.



The source text of a historical manuscript can be compared line-by-line in various views with the transcription into computer-readable text and, if necessary, corrected. This is just one of the numerous OCR4all functions. (Image: Christian Reul / University of Würzburg)

Reading today's standard fonts such as Calibri or Times New Roman is no problem for modern text recognition software, or OCR for short. It becomes more difficult with historical prints. Because the further you look back in history, the more varied the typefaces become - right up to a time when every printer carved his own typeface sets.

So there is good news for everyone who works with such historical material: The OCR4all program is text recognition software that recognizes historical printed matter and converts it into computer-readable text. No programming knowledge is required to use it.

OCR4all has been available worldwide for free on the web since 2019. It has now been downloaded around 5,000 times; to date there has not been a comparable offer in the open source area. The tool was developed by an interdisciplinary team led by Dr. Christian Reul, head of the digitization unit at the Center for Philology and Digitality "Kallimachos" (ZPD) at the Julius Maximilian University (JMU).

OCR4all emerged from the JMU's Kallimachos joint project funded by the Federal Ministry of Research. This project built bridges between the humanities, computer science and the digital humanities. In the beginning, OCR4all was about digitally processing Sebastian Brant's Ship of Fools, a moral satire from the 15th century, in the Narragonia sub-project.

Factory specific models are very accurate

Since then, the project has grown significantly and is also known in specialist circles abroad. "The nice thing about open source projects is that there's always give and take," says Reul. Models are trained so that the software later recognizes certain fonts as precisely as possible. This requires as much training material as possible, consisting of line images and the correct transcription of the text to be seen on them, and this is often provided by the software users themselves.

This form of cooperation is bearing fruit, as Reul explains: With so-called work-specific models, very precise recognition results can now be achieved, even on the oldest existing prints from the incunable period (before 1500). These are models that, as in the case of the Ship of Fools, are specially trained to recognize a print type.

Sponsored by the Vogel Foundation

The ZPD is now working harder to further develop mixed models, which ideally can be applied to as many print types as possible. For example, while there were already very good models for German-language Fraktur typefaces from the 19th century, there has so far been no broader model that can be applied to prints from several centuries with a clear conscience. According to Reul, this required more training data.

He was therefore happy to receive funding from the Vogel Foundation Dr. Eckernkamp (Würzburg): "There were gaps in the training data, especially in the case of historical Fraktur typefaces, which we were able to close with the funding," says the computer scientist.

Best Paper Award

At the HIP'21 (6th International Workshop on Historical Document Imaging and Processing) conference in September 2021 in Lausanne (Switzerland), Reul presented for the first time a publication on a mixed model covering Latin script from the period 1450 to 1900.

"At the time, we ended up with a drawing accuracy of more than 98 percent, which clearly surpassed the previous state-of-the-art," says the JMU computer scientist. It is hardly surprising, then, that the publication received the Best Paper Award from the HIP conference.

350,000 euros from the DFG

Reul also describes the two-year OCR4all-libraries project, which was approved by the German Research Foundation (DFG) in July 2021 and funded with 350,000 euros, as a milestone. "We are now marrying OCR4all with OCR-D," he says happily.

The main goal of the DFG-funded OCR-D project is the conceptual and technical preparation of the full text transformation of the prints published in the German-speaking area from the 16th to 18th centuries. For this purpose, the automatic full text recognition is broken down into individual process steps, which can then be processed with different tools. This aims to create optimal workflows for the old prints to be processed and thus to generate scientifically usable full texts.

An additional benefit of the software from Würzburg in the course of the full text recognition of the historical collection: OCR4all enables the application by technically less experienced users and also serves as a tool for more experienced users to analyze and optimize the workflow.

In the course of the OCR4all-libraries project, Reul hopes for a comprehensive further development of the software, especially due to the rapidly growing number of available tools. The ZPD will work together with the Leibniz Institute for Educational Media | Georg Eckert Institute in Braunschweig and the JMU chair for human-computer systems.

Historical manuscripts: a challenge

Text recognition software for old prints is one thing. But what about historical manuscripts?

"In principle, the approach is similar, but due to the irregularity of the fonts, it is usually much more demanding," says Reul. In addition, manuscripts can be considerably older than prints, thus covering an even longer period of time and are often poorly preserved.

No reason for the ZPD not to face this challenge as well. "The need for manuscripts is huge - here you can find book typefaces that look like they were printed, right through to texts that are almost illegible," says Reul.

In the face of this challenge, he remains calm: "First of all, we need a lot of training to lay a solid foundation." Stefan Tomasek from the JMU Chair for German Philology, older department: In the course of his new edition of the childhood of Jesus Konrad von Fußesbrunnen, he made data available to the ZPD for model training. Since then, the stock of training data and thus the model has been continuously developed in cooperation between the ZPD and the chair. Excellent results have already been achieved on medieval manuscripts. The first models will be made freely available online in the coming weeks and the associated paper will be submitted in January 2022. A joint DFG application is also in preparation.

Other researchers are now increasingly using the know-how of the Würzburg ZPD and OCR4all for their third-party funded projects. In addition to the already ongoing DFG project Camerarius digital, numerous project applications are in preparation, both for the recognition of manuscripts and prints.

New building for the Center for Philology and Digitality

The Center for Philology and Digitality "Kallimachos" (ZPD) is a central scientific institution of the University of Würzburg. Since it was founded in 2019, it has been pursuing the purpose of providing the best possible support and further development of humanities research in the digital age.

The ZPD is getting a three-storey new research building with 2,500 square meters of floor space on the north campus. It is expected to be ready for occupancy by the end of 2022. Then as many research projects as possible should move in in order to spatially bundle knowledge and skills from the humanities, cultural studies and computer science.

Nevertheless, the ZPD is already fully operational and available for cooperation. Other focal points, in addition to machine text recognition on historical prints and manuscripts, are the collaborative creation of digital editions in a virtual research environment and the modeling and realization of semantic databases.

web links

- **Center for Philology and Digitality ZPD**
- **OCR4all**
- **manuscript project**

Contact

dr Christian Reul, Center for Philology and Digitality, University of Würzburg, T +49 931 31-80722, **christian.reul@uni-wuerzburg.de**

Back

Subscribe to einBLICK

a VIEW as PDF

archive