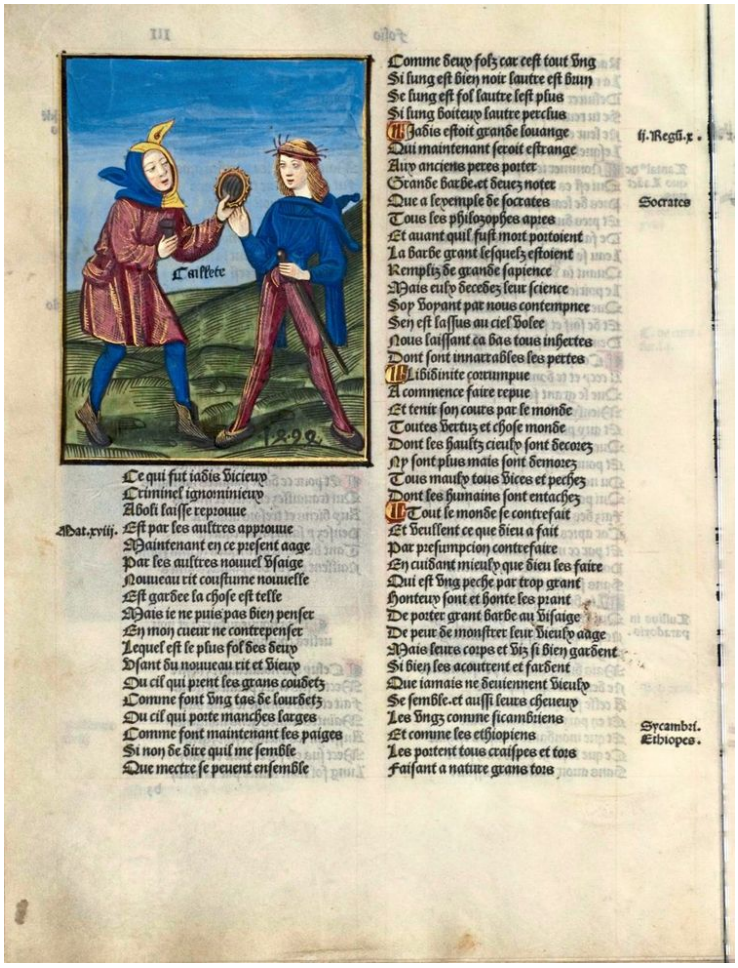
OPEN-SOURCE

Reliable text recognition tool for historical publications

German scientists presented free program for converting digitized documents into editable text

April 24, 2019, 08:00, 11 posts



Old printed works like this (from a French version of the "Ship of Fools") can be reliably converted into computer-readable text with OCR4all.

Photo: Dresden State and University Library [cc;4.0;by-sa]

Historians and linguists do not always have it easy. When working with centuries-old printed works, the effort to decipher writing is often great. If the pages are in bad condition, it becomes even more difficult. Many historically relevant documents are now available in digitized form. In order to be able to work with them, however, they have to be brought into a modern text form.

Researchers at the University of Würzburg have now presented an important further development in text recognition software: With the help of the free application OCR4all, digitized prints can be converted into computer-readable text with an error rate of less than one percent. The program offers a graphical user interface that requires no specialist knowledge to operate - a decisive improvement over many previous applications.

Well trained software

The development of OCR4all took place in close cooperation between computer scientists and humanities scholars, including German and Romance philology specialists in the "Narragonia digital" project. The aim there was to digitally process the "Ship of Fools" - a moral satire by Sebastian Brant from the 15th century, which was translated into many languages. The program is freely available **on the GitHub platform** [https://github.com/OCR4all] with instructions and examples. [https://github.com/OCR4all]

"One of the biggest problems was the typography," said Christian Reul, leader of the project. One of the reasons for this is that the first printers of the 15th century did not use uniform fonts. "Their printing stamps were all self-carved, each printing house practically had its own letters and symbols."

In order to automate text recognition, the software first had to learn to recognize subtle differences using sample material. In a case study with six historical prints from the years 1476 to 1572, the average error rate in automatic text recognition was reduced from 3.9 to 1.7 percent. A remarkable result for Reul: "The computer science behind it is extremely exciting." (red, 24.4.2019)

Link
OCR4all – An Open Source Tool Providing a Comprehensive But Easy to Use (Semi-)Automatic OCR Workflow for Historical Printings [https://github.com/OCR4all]