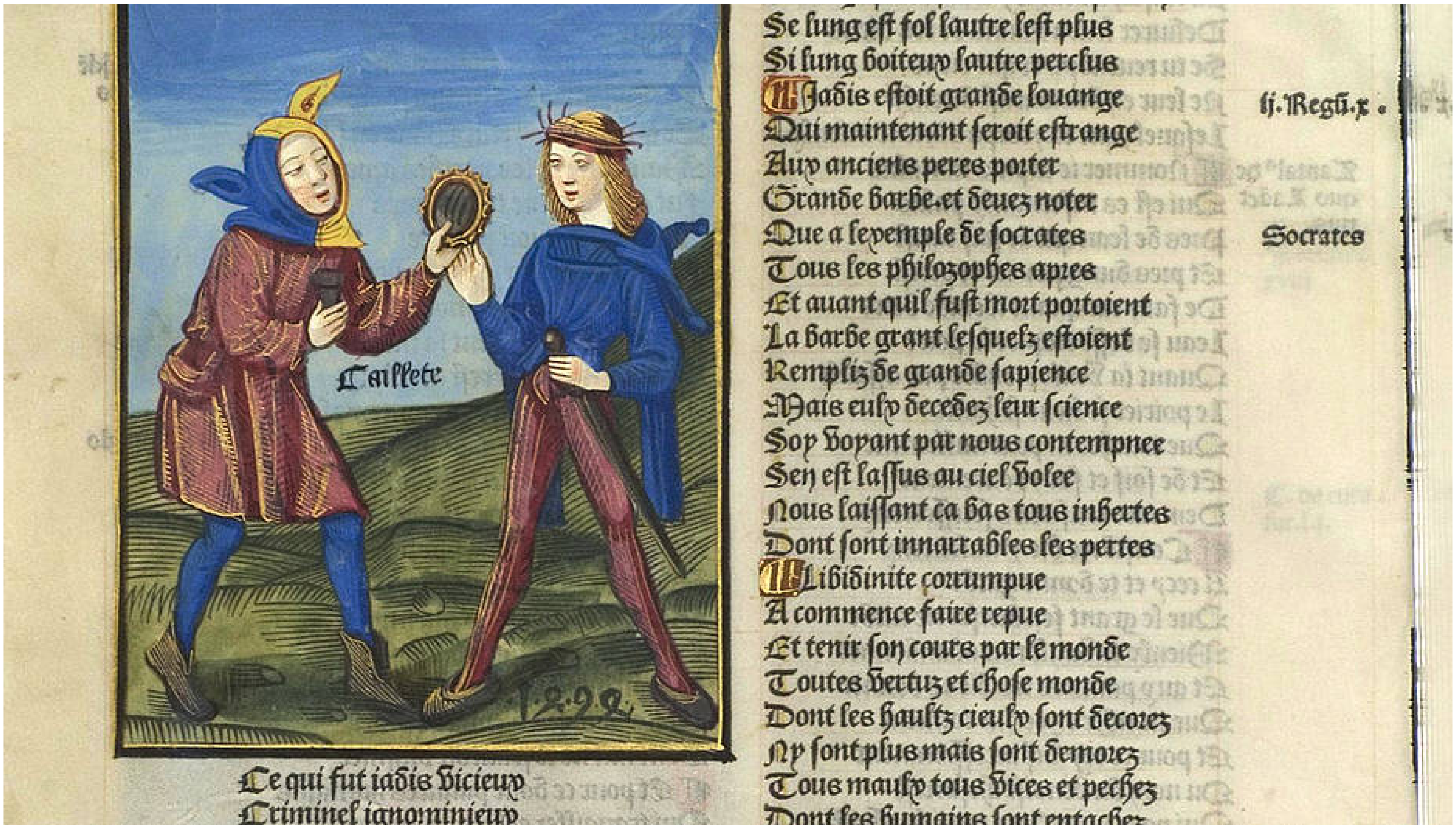
Modern tool for old texts

04/23/2019

Converting historical printed matter into computer-readable text: The OCR4all tool, which works very reliably, is easy to use and freely available, takes care of this. Scientists from the University of Würzburg developed it.



Ce qui fut iadis vicieux
Criminel ignominieux

Se lung eſt fol lautre leſt plus
Si lung boiteux lautre perclus
Jadis eſtoit grande louange ij. Regũ.x.
Qui maintenant ſeroit eſtrange
Aux anciens peres porter
Grande barbe.et deuez noter
Que a lexemple de ſocrates Socrates
Tous les philozophes apres
Et auant quil fuſt mort portoient
La barbe grant leſquelz eſtoient
Rempliz de grande ſapience
Mais eulx decedez leur ſcience
Soy voyant par nous contempnee
Sen eſt laſſus au ciel volee
Nous laiſſant ca bas tous inhertes
Dont ſont innarrables les pertes
Libidinite corrumpue
A commence faire repue
Et tenir ſon cours par le monde
Toutes vertuz et choſe monde
Dont les haultz cieulx ſont decorez
Ny ſont plus mais ſont demorez
Tous maulx tous vices et pechez
Dont les humains ſont entachez

Page from a French version of The Ship of Fools. Such old fonts can be reliably converted into computer-readable text with OCR4all. (Image: Dresden State and University Library, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0/deed.de)

Historians, Germanists and other humanities scholars often have to deal with difficult research objects: with centuries-old printed works that are not easy to decipher and are often poorly preserved. Many of these documents have now been digitized - usually photographed or scanned - and are available online worldwide. This is a step forward for research.

However, there is still a challenge to overcome: using text recognition software to bring the digitized old writings into a modern form that is also readable by non-specialists and computers. In this field, scientists from the Center for Philology and Digitality at the Julius Maximilian University of Würzburg (JMU) have ensured significant further development.

With **OCR4all** , the JMU research team is making a new tool available to experts. It converts digitized historical prints into computer-readable text with an error rate of less than one percent. And it offers a graphical user interface that does not require any IT expertise to operate. With previous tools of this type, user-friendliness was not particularly pronounced, and programming commands usually had to be used.

Developed in cooperation with Humanities

The new tool OCR4all was developed under the direction of Christian Reul with his computer science colleagues Professor Frank Puppe (Chair of Artificial Intelligence and Applied Computer Science) and Christoph Wick as well as with Uwe Springmann, specialist in digital humanities, and numerous students and assistants.

OCR4all has its roots in the JMU's Kallimachos joint project, which is funded by the Federal Ministry of Education and Research. This cooperation between the humanities and computer science will be continued and institutionalized in the newly founded Center for Philology and Digitality (ZPD).

In the development of OCR4all, the computer scientists worked closely with the humanities disciplines at JMU - including with German and Romance studies in the "Narragonia digital" project. The aim there was to digitally process the “Ship of Fools” – a moral satire by Sebastian Brant, a 15th-century bestseller that was translated into many languages. OCR4all was and is also used in the "Middle Ages and Early Modern Age" college at the JMU.

OCR4all is open to the public **Platform GitHub**(with instructions and illustrative examples) freely available.

Every printer had its own typeface

Christian Reul explains what was a challenge in the development of OCR4all: The automatic text recognition (OCR = Optical Character Recognition = optical character recognition) has been working very well for modern fonts for a long time. However, this has not yet applied to historical writings.

"One of the biggest problems was the typography," says Reul. 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 printer practically having their own letters and symbols."

Error rate reduced to less than one percent

Whether e or c, whether v or r - this is often not easy to distinguish in old prints. However, software can learn to recognize such subtleties. But to do this, it must first be trained on example material. In his work, Reul has developed methods to make this training more efficient. 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.

But not only the methodology has been improved. JMU computer scientist Christoph Wick has also decisively advanced the technical component by developing the OCR tool Calamari, which is also freely available and has now been fully integrated into OCR4all. All in all, the results were even better: meanwhile, even for the oldest printed works, error rates of less than one percent can usually be achieved.

Lexical Projects

Reul has also convinced external partners of the quality of Würzburg's OCR research. Together with the "Center for Digital Lexicography of the German Language" (Berlin), Daniel Sanders' "Dictionary of the German Language" was digitally indexed; a publication on this is on the way. This work often contains different fonts per line of text, each of which stands for different semantic information. Here, the existing approach to character recognition was expanded in such a way that, in addition to the text, the typography and thus the complex content structure of the lexicon can be mapped very precisely.

The Würzburg computer scientist will soon be completing his doctoral thesis, but he also wants to work with OCR in the future: "The computer science behind it is extremely exciting," he says. A possible project for the near future: the creators of the "Idiotikon", a dictionary of the Swiss-German language, signaled to him that they could really use the Würzburg expertise.

Contact

Christian Reul, Acting Head of the Digitization Unit, Center for Philology and Digitality at the University of Würzburg, **christian.reul@uni-wuerzburg.de**

- **Center for Philology and Digitality**

Christian Reul website

web links

- **OCR4all on GitHub**
- **Calamari on GitHub**
- **The project "Narragon digital"**
- **Link to publication**(case study with six historical books)
- **Publication combining methodological and technical improvements**

Center for Philology and Digitality

The Center for Philology and Digitality at the University of Würzburg is the result of an initiative that came from Professors Dag Nikolaus Hasse, Fotis Jannidis and Ulrich Konrad. It bridges the gap between humanities, computer science and digital humanities. It represents the first building block for a new humanities center on Campus North.

A new building for the ZPD is to be built there, close to the canteen and the Graduate School building. Around 100 people are expected to work in the new ZPD building on a total of 2,700 square meters from 2022. In the planning, total costs of 15 million euros are set for the building. A digital lab, research rooms and lecture halls are planned on the ground floor of the ZPD. Offices and communication rooms are to be built on the upper floors.

Back

Subscribe to einBLICK

a VIEW as PDF

archive