23.04.2014: OCR für historische Drucke (Tutorial, Uwe Springmann, LMU München)

Preparations for the Tutorial

For a hands-on experience, you will at minimum need to install Tesseract + gImageReader and to download the example file.

1. Installation of Tesseract

Go to the website of Tesseract and download/install the binary for your OS or compile it from source. Don't forget to download some training files for the languages you want to recognize as well! The instructions are here:

https://code.google.com/p/tesseract-ocr/wiki/ReadMe

Training Files:

http://code.google.com/p/tesseract-ocr/downloads/list

For this tutorial, we specifically need the German-Fraktur training set (deu-frak.traineddata.gz).

There is a 3rd party GUI for Tesseract which I recommend:

http://sourceforge.net/projects/gimagereader/

Try it with some example image (tif, png, jpg, pdf).

2. Installation of Ocropus (recommended)

The installation of Ocropus requires a Unix-like OS (Linux or Mac are ok). If you are running Windows, you could install a virtual Linux OS using virtualbox (www.virtualbox.org) and run Ocropus from there.

Installation instructions for Ocropus are here:

https://code.google.com/p/ocropus/

3. Installation of ScanTailor (recommended)

ScanTailor is a very useful tool for preprocessing scans that we are going to OCR afterwards and can be found here:

http://http://scantailor.org/

4. For Windows users: Install Cygwin (optional)

Many of the more advanced features of this tutorial make use of the Unix command line and its many excellent freely available tools. You can enjoy the same privilege on Windows by installing Cygwin (http://cygwin.com).

5. Example file we are going to use

The example file is an extract dealing with herbs of a 1557 book by Adam von Bodenstein on podagra. You can download the scan here (we will treat pp. 65-104 of the pdf).

https://download.digitale-sammlungen.de/BOOKS/pdf_download.pl?id=00015627&nr=1