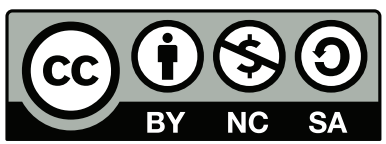
Module 3
Preprocessing: Practice

Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München (LMU)



2015-09-14

Practice session: Overview

The practice session consists of the following steps:

1. install software
2. download scanned document
3. split document into pages
4. preprocess with ScanTailor
5. save as tif images

Software installation

- software installation:
  - Linux: `sudo apt-get install ...` (Debian/Ubuntu) or use your distribution's package system
  - Mac: use your package system
  - Windows: download and install binary (.exe) file
- install PDFtk
  - Windows: download PDFtk free
  - Mac: download PDFtk Server or use your package system
- install ImageMagick
  - Windows, Mac: scroll to your OS version, download & install
- install ScanTailor
  - Windows: download .exe-File of latest version
  - on Mac, you may need to build from source

Download document

- Example document:
  Johann Wonnecke von Kaub (Johannes von Cuba), Gart der Gesundheit (1487)
- navigate to the data section of the workshop website and download the file `gdg.pdf` to your laptop
- this file contains 12 book pages + the calibration page (scanned ruler)

Preprocessing

Do the following:

- split the pdf into single page images
- convert the images into a format usable with ScanTailor (.tif, .tiff, .jpeg, .jpg, .png)
- determine the resolution in ppi
- use ScanTailor with the page images
- your end result should be “1-column text-only binarized tif images”

Some hints

- find out what the pdf contains:

```
pdfimages -list gdg.pdf
```

- write jpeg-images as jpeg-files:

```
pdfimages -j gdg.pdf gdg
```

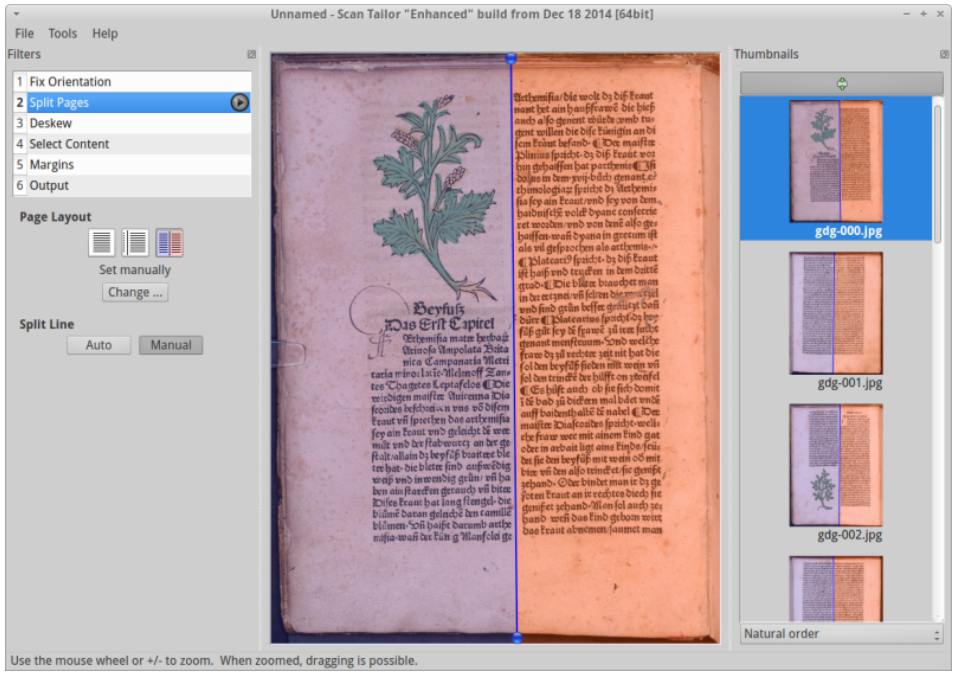
- otherwise split & convert using either of these methods:

```
convert gdg.pdf gdg.png
```

```
pdftoppm -png *.pdf gdg
```

- before using ScanTailor, measure the resolution using the ruler at the last page
- possible pitfall: the original jpeg has 1425x1872 pixels, so if your png turns out to have more pixels, you have inadvertently oversampled (pdftoppm) and need to adjust

Using ScanTailor



- ScanTailor:
  - run each command in its toolbar in sequence (1 to 6)
  - split pages into single columns
  - cut out the illustrations (best done after step 4 by adjusting the content boxes)
  - set output resolution to twice input resolution
- if you need help, ask your neighbors or the instructor