OCR workshop with IMPACT:

- 300 ppi is good resolution for OCR
- Only small characters might need 600 ppi
- Greyscale is sufficient for OCR
- Abbyy binarisation works best
- Octopus binarisation, another binarisation from Athens (NCSR)
- Run all 3 binarisation methods and compare OCR results to find out what is the best one
- Special binarisation methods for reading, run on the fly and create them on demand for printing
- Sometimes scanning the borders is necessary
- We may need to do a survey what are the differences in scanning operations to see what tools may be appropriate
- Judith scans with borders, Tine does it not on purpose, but it happens when scientists scan books
- Border removal might be useful to work with (Fine Reader tools available)
- Skew, page curl – is a common problem also for us (Fine Reader tools available)
- Prototypes for letter deformation due to page curl
- Tag in the database the images and the defects occuring in the images – volunteers needed
- Warping of paper due to moisture, not easy to remove – prototype stage – maybe 20-30s for one page to process
- Gothic typeface: Abbyy and Tesseract supports these font types, 3000 different of these font types, Fine Reader trained to 200 currently
- “Long S” problem, needs a lot of training
- Complex layout
- Segmentation
- Functional extension parser: recognition structure – maybe used to extract page numbers from the text
- Extracting ToC or even creating ToC – but it needs training for the layout
- Collaborative correction tools: runs OCR first, return suspicious characters where the OCR engine is not sure, also a tool for suspiciuous words in a next step and then go to the full text overall page tool
- Android game to do text correction by playing
- Word spotting: mark key words and index them as graphics
- No Latin dictionary, no funding for it, tools to do them available
- Seach term expansion for old words for terms (world)
- Named entity recognition (author names, localities)
- Different processing pipelines to compare outcomes and evaluate them against each other, build sequential tool chain by drag and drop
- Abbyy FR 9, 10 & Tesseract, gives the best engine for the material