HERZOG AUGUST BIBLIOTHEK
WOLFENBÜTTEL

OCR for manuscripts and early prints

Torsten Schaßan (HAB Wolfenbüttel)

ESF Exploratory Workshop *Digital Palaeography*

Würzburg – July 21, 2011

Institut für Dokumentologie und Editorik

Experiences

- Brief report on OCR in libraries
- Results of a meeting
  of the „OCR workgroup“
  of the UAG (=sub working group) *Altes Buch*
  of the DBV (=German libraries association)
  - Exchange of experiences
  - Evaluate whether this topic is „permanent“
  - Decide whether „best practice guidelines“ could be published
- Experiences from HAB

The scope

- Participants
  - StaatsB Berlin
  - SLUB Dresden
  - German Certral library for the blind, Leipzig
  - BSB Munich
  - HAB Wolfenbüttel
- Except Berlin and Wolfenbüttel most of these were dealing with modern printed materials

Starting point

- Libraries didn't develop new software but have applied existing
  - Abbyy FineReader
  - BIT Alpha
  - [Omnipage]
  - I.R.I.S → no experiences yet
  - OCRopus/tesseract → engine might change

Abbyy FineReader

- Do not train! (Results tend to get worse!)
- Has major problems with mixed font types (Gothic / Roman)
- The version that is prepared to read Gothic script used to be expensive!
  (licensing according to masses of material to OCR'ed)

BIT Alpha

- Originally shipped without dictionary
  → Needs to be trained (heavily!)
  → Can be trained usefully
- Extensive communication needed (almost weekly updates, wishlists for features possible)
- Parametrisation is complex

HAB experiences

- Step 1: Research-cooperation with BIT Alpha
  - Basic training and parametrisation explored
- Step 2: Project „Helmstedt imprints“
  - Digitisation of 5.000 prints (c17)
  - 120.000 pages OCR'ed
- Will be used mainly under service conditions

[We wonder whether our training efforts could/should be re-used commercially]

Helmstedt imprints

- Basic ideas
  - Printers in Helmstedt might have used similar typefonts
  - Paper quality homogenuous
  - Recognition of the mix of Gothic / Roman typefaces, and different languages (Latin, Greek) successfully tested
- Pricing
  - 25c per page (double-keying = 1,50 Euro per page)

OCR results

- Export of Searchable PDF and ALTO-XML
- One XML file (001.xml, …) per page
- TEI fragments
  - div/p
  - Each „word“ wrapped by <w>
  - <w facs="#drucke_131-helmst-dr-52s
    _00001
    _ulx691uly359lrx1261lry484
    mw2433mh3516">Programma</w>

Processing / Use

- Automated upload of an compiled XML file to an eXist-server for searching
- Highlighting of search results based on @facs
  - On-the-fly generated images (ImageMagick)
  - So far only one hit per page highlighted

A word on typefaces

- Typefaces used in Helmstedt seem to be homogenuous
- Lack of research on typesetters or their trade of matrices
- So far no attempt to make assumptions about the grade of similarity

A word on writing hands

- During processing each character is assigned with a singular value, describing the characteristics like shape etc
- Turning the workflow around, it would be possible to extrapolate from the similarity of these values the distinctiveness of hands

OCR quality

- Image quality is crucial for OCR quality
- Factors to reduce image quality are
  - Intrinsic: print quality, paper quality, staining, annotations, etc.
  - Extrinsic: scan quality, bending of the page, low resolution, compression artefacts, scan from film instead of scan from original, etc.)
- Resolution and/or image size might be to high
  - BIT Alpha expects 300dpi
  - JPEG2000 so far not supported

OCR quality

- So far no „objective“ criteria developped to measure OCR quality
  - character-level? word-level? whitespaces relevant?
- How do we measure error frequency?
  - In the project „Helmstedt imprints“ for certain pages of each print the lines 4/5 are examined, errors counted
  - Result will be extrapolated

Error frequency

- Needed are 99,95% error-free texts to be used for scholarly purposes
  - below that, results are useful „only“ for searching
- Approx. error ratio
  - Abbyy: 90% for modern prints
  - BIT Alpha: 95-99% also for early prints, depending on the training

Representation of OCR results

- PDF is no choice
- Preferred is an XML format
  - TEI
  - hOCR
- Especially important are text coordinates
  - ALTO
  - hOCR

What to find

- OCRed texts are important as full-texts, but
  - Entities are of special interest
    → how to find them automatically?
- Post-processing needed

How to use

- Citeability and granularity of OCR results are an issue
  - What will be cited/citeable → What is a word? Abbreviations?
  - How to represent what is cited → again, coordinates?!
- Will re-processed documents generate the same OCR results?
- Under what legal conditions can OCRed texts be made available?

Resumé

- Problems and issues for OCR for manuscripts and early prints do not differ to much from those for other prints
- Font types are recognised with high probability
  → but image quality derogates easier achieved / better results
  → bent pages disturb the OCR processing most
- Lots of training not possible for mass digitisation

Finally …

… the most important questions seem to be:

- What is an error?
- How do we recognise errors?