TEI, METS and ALTO, why we need all of them

Günter Mühlberger
University of Innsbruck
Digitisation and Digital Preservation

Agenda

- Introduction
- Problem statement
- Proposed solution

Starting point

- Mass digitisation takes place!
- Case 1
  - University of Innsbruck
  - Digitisation of 216.000 German dissertations from 1925 to 1988
  - 24 mill. pages
- Case 2
  - EU Newspaper Project
  - 8 mill. pages for Optical Character Recognition (OCR)
  - 2 mill. pages for structural enhancement (article tracking)
  - = factor 1:5-10 compared to books
- Case 3
  - Austrian National Library and Bavarian State Library: Probably 50-70% of all German language books from 1500 to 1870 digitised by Google
  - 1,3 Mill. books, 500 mill. pages
- Many more projects and programmes are going on
- Handwritten Text Recognition (HTR) will complete the picture
  - Takes off in 5-10 years
  - Mass sources will become available by automated processes

Objectives

- Make these highly important sources available to the CLARIN community
- Transform these sources into TEI in a meaningful and highly standardized way in order to support reusability and sustainability

Problem statement

- How shall this transformation be done?
  - TEI community is usually manually preparing TEI files
  - E.g. Deutsches Textarchiv (DTA): Instructions to generate (simplified) TEI based corpus material
  - Automation is used to assist manual preparation of the files
  - BUT: Simple calculation: to correct 1000 characters usually costs about 0,5 to 1 EUR (off-shore prices)
  - A book page has about 2500 characters, a journal page about 5000-8000 characters and a newspaper page from 5000 – 30.000 characters or even more.
  - Manual correction would cost at least 100 times more than the whole digitisation no one ever will pay for that
- Conclusion: Manual preparation is not usable for digitized mass sources – we need to find another way!

Availability of text and structure

- Mass digitisation of text results from Optical Character Recognition (OCR)
  - Text
  - Segmentation information: location of blocks, lines, words, characters on a page image
  - Discrimination of text vs. graphical blocks vs. tables vs. noise
- Enhanced processes
  - E.g. rule based approaches or machine learning approaches
  - Structural data such as paragraphs as logical unit, chapters, running titles, footnotes – depending on type of document
- Example of EU Newspaper Project
  - 8 Mill. pages OCR, 2 Mill. pages basic structural segmentation
  - Technical metadata (file type, file size, MD5 checksum, etc.) are generated automatically with File Analyzer Tool (FAT)
  - OCR and OLR transformation
  - METS container according to ENMAP (European Newspaper Mets Alto Profile)

ALTO format

- OCR Data
- Recorded in ALTO files
- ALTO introduced by University of Innsbruck within the METADATA ENGINE project (FP5 2000-2003)
  - Promoted by CCS Gmbh Hamburg/Germany
- Now hosted by Library of Congress as an addition to Metadata Encoding and Transition Standard (METS)
- National libraries and university libraries from all over the world (USA, Australia, Singapore, Norway, Finland, British Library, Biblioteque National de France, Austrian National Library, etc.) are using this format
- Industry support: ABBYY SDK provides native ALTO output since 2011
- Since 2012 also officially supported/requested from the German Science Funds (Digitisation guidelines)
- Format has some issues (discussion with board is going on) but is a de-facto standard and hardly to replace (there is only the Google format with more files but also much more concerns)

Structural data

- OCR is just the beginning!
  - IMPACT project
  - Development of a general purpose rule set to extract typical features from books
  - E.g. headings, running titles, page numbers, footnotes, table of contents pages, TOC entries, etc.
  - All this is done on top of the OCR results (stored as ALTO files)
  - Results are encouraging
  - Some features can be detected with rather good accuracy (e.g. running titles, page numbers) some are weaker but still rather helpful (e.g. headings, footnotes)
  - For specific document types (e.g. journals) higher accuracy rates can be reached due to their homogeniety

Crowd sourced correction

- Australian National Library
  - Newspaper digitisation programme
  - Simple editor to correct text on line level
- Reaction
  - Extremely positive
  - Ten-thousands of articles (partly) corrected
- Result
  - Not a solution to correct all historical Australian newspapers
  - But by sure an improvement and something which will be offered by many more libraries in the future
- Consequence
  - Situation becomes even more complex!

Conclusion

- Complex situation
  - A given document does not have “one” level of accuracy, but it may come with very different results on feature level (e.g. text, layout, structure,...) and some parts may even be corrected by human beings
- How to cope with this situation?

Proposed solution

- Situation of automatically produced “noisy” text data is relatively new - BUT for researchers (especially historians) this is not a “new” situation!
- Sources ALWAYS had to be assessed/verified in a critical way: This is exactly what we need to do for mass sources!

Prerequisites for automated assessment

- Ground truth
  - Evaluation of results in technical projects is often done via so-called “Ground truth” or a “Golden standard”
  - These data represent the “expected results” of a given process
  - The difference between “target” and “actual” rsults allows to figure out improvements, or to compare several results from different research groups
- ICDAR conference
  - Several competitions on e.g. “Historical newspaper structure recognition” or “Handwritten text recognition”, or eg. “Linking of table of contents entries with headings and page numbers”, etc.

Automated assessment

- ALETHEIA tool and PAGE format
  - Important work from PRIMA group (Apostolos Antonacopoulos, Stefan Pletschacher) at University of Salford
  - PAGE format = rich XML format to describe layout features of printed text
  - ALETHEIA = tool with GUI for manual production of “correct” text with layout features
- Evaluation scenarios
  - Evaluation is a complex task
  - Many dimensions and aspects have to be taken into account
  - Later usage strongly predicts “what needs to be evaluated to get meaningful results
  - E.g. evaluation of the quality of a full-text depends on what shall be done with the full-text: Does the reading order play a role, do structural elements play a role, etc.

Assessed feature list (example)

Feature	Assessed	Measure	Results (F-Measure)
Running text	1000	Randomly selected pages	92 (87-95)
Paragraphs	5000	Weighted random selection	67 (50-80)
Column titles	5 per book	Random selection	85 (81-89)
Footnotes	...		

Encoding

- METS
  - Provides exactly the structure to encode the feature list + assessment results
  - Container format with several predefined sections (descriptive, administrative, structural data)
  - Administrative metadata section
  - Comparable to technical metadata for image files
  - Information is linked to the appropriate sections in the files via ALTO coordinates (block, line, string)
- Assessment schema for textual and structural features of automatically processed documents
  - Description of features
  - Conventions how many items of a feature need to be assessed to get meaningful results
  - Assessment methods (e.g. algorithms for calculation)
- CLARIN?
  - TEI files can be produced from METS/ALTO – their usability for “CLARIN-based processes” will depend on the assessment results

Pilot project

- EU Newspaper Project
  - PRIMA group is partner
  - Evaluation of texts is foreseen in Document of Work (DoW)
  - ENMAP schema is currently developed by University of Innsbruck
- Pilot
  - In 2014 we should be able to come up with a first attempt

Thank you for your attention!

Günter Mühlberger
<guenter.muehlberger@uibk.ac.at>