LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Presentation by Benjamin Leak

From facsimile to TEI-5 compliant XML document

Bachelorthesis 2014

Synopsis

1. Optical character recognition (OCR)
2. XML file format
3. Text Encoding Initiative (TEI)
4. Choice of tags
5. Preprocessing through EOS tagging
6. PERL programs
7. Test run
8. Problems
9. Conclusion

Optical character recognition

- Conversion of digital image (facsimile) to machine-readable text
- Roughly 5 steps:
  - Image acquisition
  - Segmentation
  - Extraction
  - Classification
  - Corrections

Optical character recognition

- OCR works well for standardized documents
- OCR does not work well for:
  - Documents with physical damage (discoloration, rips, dirt)
  - Documents with unstructured or uncommon layout
  - Anything handwritten

XML

- XML is a widely used file format
  - Many tools available (creation, extraction, transformation)
  - Though using the tools can be complicated
- Based on the Ordered Hierarchy of Content Objects (OHCO) Model to structure data
- Easy to verify and validate (well-formed, DTD, XSD)

Text Encoding Initiative

- Defines standard for encoding digital humanities data
  - Started 1990 with guidelines for SGML
  - Since 2002 XML is used
- Over 500 XML tags are defined by the TEI
- Those tags were highly influenced by the Humanities and their requirements
- For a single project a small subset of tags is usually sufficient

Tags used in the thesis

- Rather minimal TEI Header with information about:
  - Title
  - Author
  - Publication details
  - Source description
  - Responsibilities of the involved parties
  - Optional note
  - Creation date of the source document

Tags used in the thesis

- Breakdown of the TEI text body:
  - Division ( ) for each facsimile (with attribute for linking)
  - Abstract Block (<ab>) for paragraphs and similar structures
  - Sentence (<s>) for each sentence tagged

End of sentence tagging

- Sentence boundaries are recognized & tagged by CIS tool 'Satzenderkennung (eos) Version 3'
  - Does not recognize every end of sentence reliably:
    - Numbers at the end of a sentence
    - Words accidentally interpreted as abbreviations
    - Contracted points of an abbreviation
  - Works well for identifying any other boundaries

PERL programs

- Two programs created
  - One for the TEI Header & one for the TEI text body
- The information in the header is inputted via command line
  - Query if optional parts should be added or if parts are unknown
- After the header is created the body can be added
  - At least one EOS tagged plain text must be specified
  - Multiple texts can be specified and added (one at a time)
  - Text is parsed, split, and inserted accordingly into the appropriate tags

Test run

- Tested with dummy texts
  - No problems, since everything is ‘perfectly’ formatted and tagged
  - Problems only arise due to certain characteristics of documents
- Well-formed
- Validated with (full) TEI DTD and modified CISWAB DTD for TS-213

Problems

- Overlapping Hierarchies
  - Sentence does not necessarily end with the page (facsimile)
  - TEI tries to solve this with:
    - Multiple documents
    - Milestones
    - Fragmented markup
- Interoperability
  - Most of the time unable to reuse data and markup

Problems

- Multiple documents
  - How to compare?
  - Complicated to remove all tags
- Milestones
  - Must choose one of the concurrent hierarchies as main hierarchy
  - Works well if only two layers are considered (e.g. physical & logical)
- Fragmented Markup
  - Virtual elements (through linking) destroy the tree structure of XML
  - Can be hard to validate (worst case: cyclic linking)

Problems

- A possible solution through stand-off markup
  - Complete separation of data and structure
  - Plain text data and certain markup easily reusable (interoperability)
  - Multiple choices (insertions, deletions, etc.) managed independently from markup

Conslusion

- Easy to use TEI-5 XML for small or ‘uncomplicated’ projects
- Becomes increasingly difficult for big projects (one document for all needs)
- Same information can be encoded in different ways
- Hard to automate all aspects
- Sometimes not completely verifiable
- Other structures promise more interoperability

Links

- Satzenderkennung (eos Version 3):

  http://demomax.cis.uni-muenchen.de/home_demos/eosv3/index.html

- Information about TEI:

  http://www.tei-c.org/Guidelines/

- Interoperability and problems with TEI-5 markup:

  Desmond Schmidt, The Role of Markup in the Digital Humanities
  http://www.cceh.uni-koeln.de/files/Schmidt_final.pdf