About the Project

The DFG funded project Deutsches Textarchiv (2007–2013/14) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) aims at providing a text corpus of the historical New High German Language (1600–1900), which is balanced with respect to time and text genres.

- DTA ›base format‹ (DTABf): a strict subset of XML/TEI P5, 100 elements + corresponding attributes & values,
- corpus: 1,576 volumes, 443,185 pages, ~719M characters, +1 vol. every day,
- generally first editions,
- double keying, OCR,
- plan: 2,000 volumes until 2013/14; additional texts from external submitters (DTAE).

Problem Statement

Our quality assurance platform DTAQ has been running since March 2011. DTAQ users should be enabled not only to review texts but also to correct erroneous transcriptions and annotations, or add new annotations. Within DTAQ, each text is presented page by page alongside with its source images in various formats: XML, HTML, plain text etc. To produce this kind of view, the original XML/TEI P5 documents are splitted into several single page documents. This process is reversible, so modified single page documents can be reinserted losslessly into the original TEI document. Based on this page-oriented view, DTAQ provides several ways to change documents on the transcription or annotation level.

We differentiate between several kinds of changes and user experience levels of users:

- Changes to the text base, i.e. the plain transcribed text without any markup,
- Annotation of single tokens or groups of tokens, e.g. named entities, printing errors etc.,
- Modification of attribute values in existing XML elements, e.g. the values of @ref in <persName> elements to provide links to authority files,
- Editing of basic XML structures, e.g. adding quotation markup in citations (<cit>/<quote>/<bibl>),
- Editing of more complex XML structures, e.g. restructuring of paragraphs or even chapters.

Backend

The backend of DTAQ is built upon many open source packages.

Links and Contact

DTA team:
Susanne Haaf, Alexander Geyken, Bryan Jurish, Christoph Leijser, Christian Thomas, Matthias Schulz, Frank Wiegand

dta@bbaw.de
www.deutschestextarchiv.de

Integrated Online XML Editor

For mid-size changes on single pages, we provide an online XML editor. This tool is based on the Ajax.org Cloud9 Editor (ace). The editor window displays the syntax-highlighted XML for the corresponding text page. In addition, we provide several tools to support quick and efficient tagging (e.g. select a wrongly printed word like “errorneous”, press the “printing error” button, and an XML template like: `<corr>correct</corr>` is inserted into the editor). The editor also provides validation against the DTA ›base format‹ schema (via AJAX requests).

Integrated Formula Editor

As of October 2013, there are 27,000 formulas, yet untranscribed, which are referred to in the XML/TEI texts using the <formula/> element. DTAQ provides an integrated formula editor, which helps users to create &lt;formula\> transcriptions.

Instant Text Editor

Simple changes, like fixing transcription errors, may be carried out directly within the rendered HTML version of a document page, using the @contenteditable=“true” attribute. Users cannot produce invalid markup and don’t even have to bother with angle brackets.