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

- 1,259 volumes, 351,526 pages, ∼565M characters (one more vol. every day)
- generally first editions
- Double Keying, OCR
- XML/TEI P5 (DTA “base format”)
- Plan: 1,800 volumes until 2013/14; additional texts from external submitters (DTAE)

### Problem Statement

Though the transcription accuracy of the double keying method is generally very high (at least 99.95%), considerable transcription errors may still occur alongside with other error types. Quality assurance (QA) therefore has to take into account different error sources, namely:

- transcription errors
- printing errors in the text source
- errors concerning metadata
- XML annotation errors
- (HTML) presentation errors
- problems within the workflow

### Backend

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

### Ticketing System

In DTAQ, errors can be reported in tickets and by that be classified, commented, and assigned to a certain user, like in a software bugtracking system. To keep track of the reports, administrators can create importance levels, blockers, and milestone lists. Work with DTAQ started in June 2011. Since then ∼48,000 tickets were created (33,700 solved), and ∼27,700 pages were proofread.

### Statistics and Analyses

All tickets and proofread pages are stored within a database, thus DTAQ provides in-depth analysis and visualisation about the accuracy of the DTA corpus (cf. Haaf et al., jTEI 4, 2013).

### Integrated Annotation Editor

An annotation editor is currently being implemented, which will allow for the stand-off annotation of inline phenomena, such as named entities. Moreover, comments may be added on page level.

### Integrated Formula Editor

As of April 2013, there are ∼25,400 formulæ marked as such in the XML/TEI texts using the `<formula/>` element. DTAQ provides an integrated formula editor, which helps users to create \( \LaTeX \) transcriptions.

\[
\lambda_T = \frac{1}{\tau} + 4\pi^2 - e^{-x} dx
\]

Boltzmann: Vorlesungen, Bd. 1, Leipzig, 1886, p. 73.

The linguistic tools provided by the DTA (CAB, Part-of-Speech Tagger) are integrated into the DTAQ environment, where they may support the retrieval of errors.