Quality Assurance in a Large Corpus of Historical Texts

German Text Archive — Deutsches Textarchiv (DTA)

DTAQ

Deutsche Forschungsgemeinschaft DFG

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

About the Project

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.

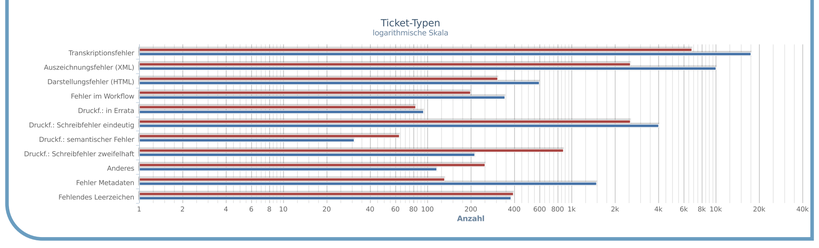
Perl 5 · DBIx::Class · git · T2T template-toolkit.org · jQuery write less, do more. · catalyst Web Framework

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).

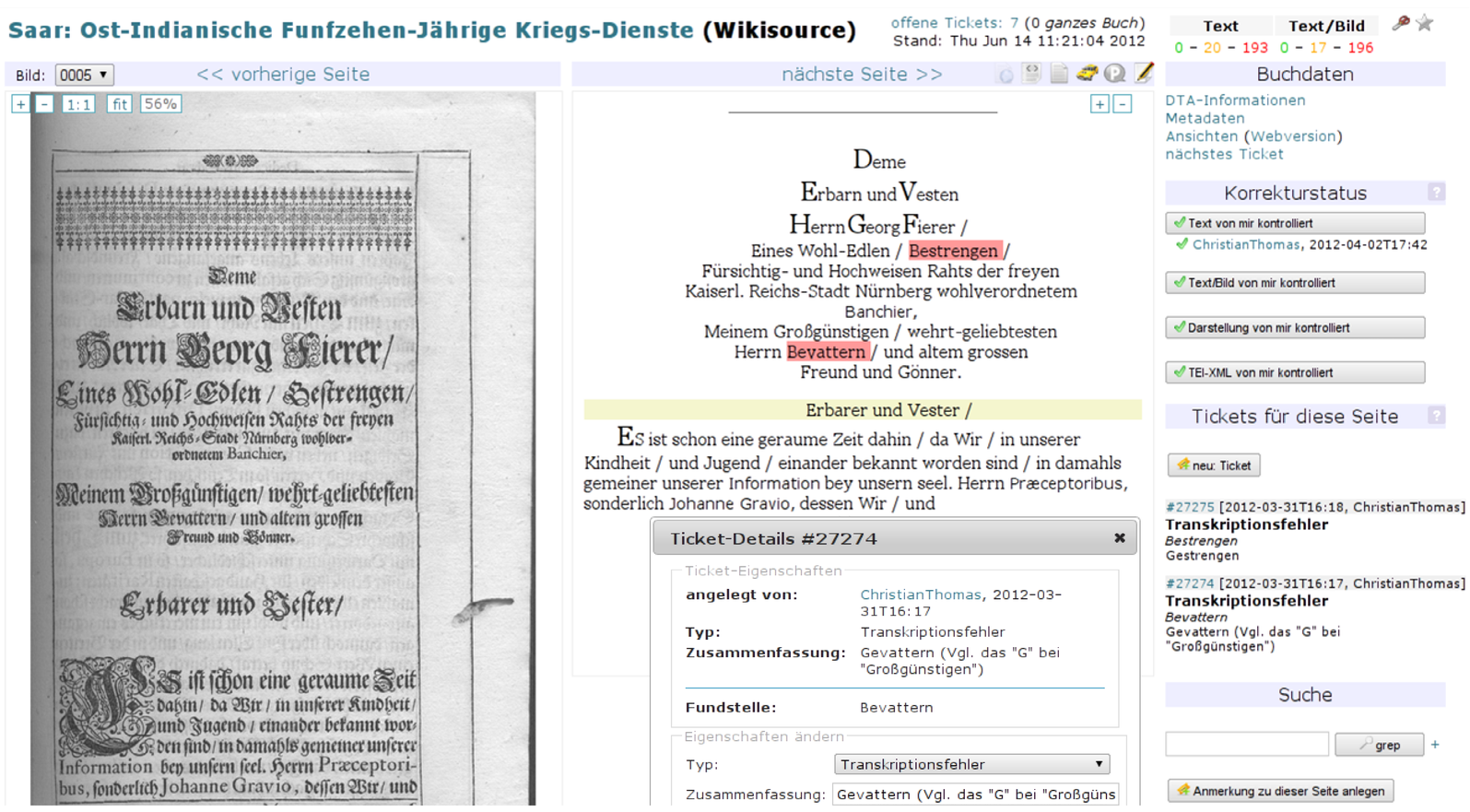


Links and Contact

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

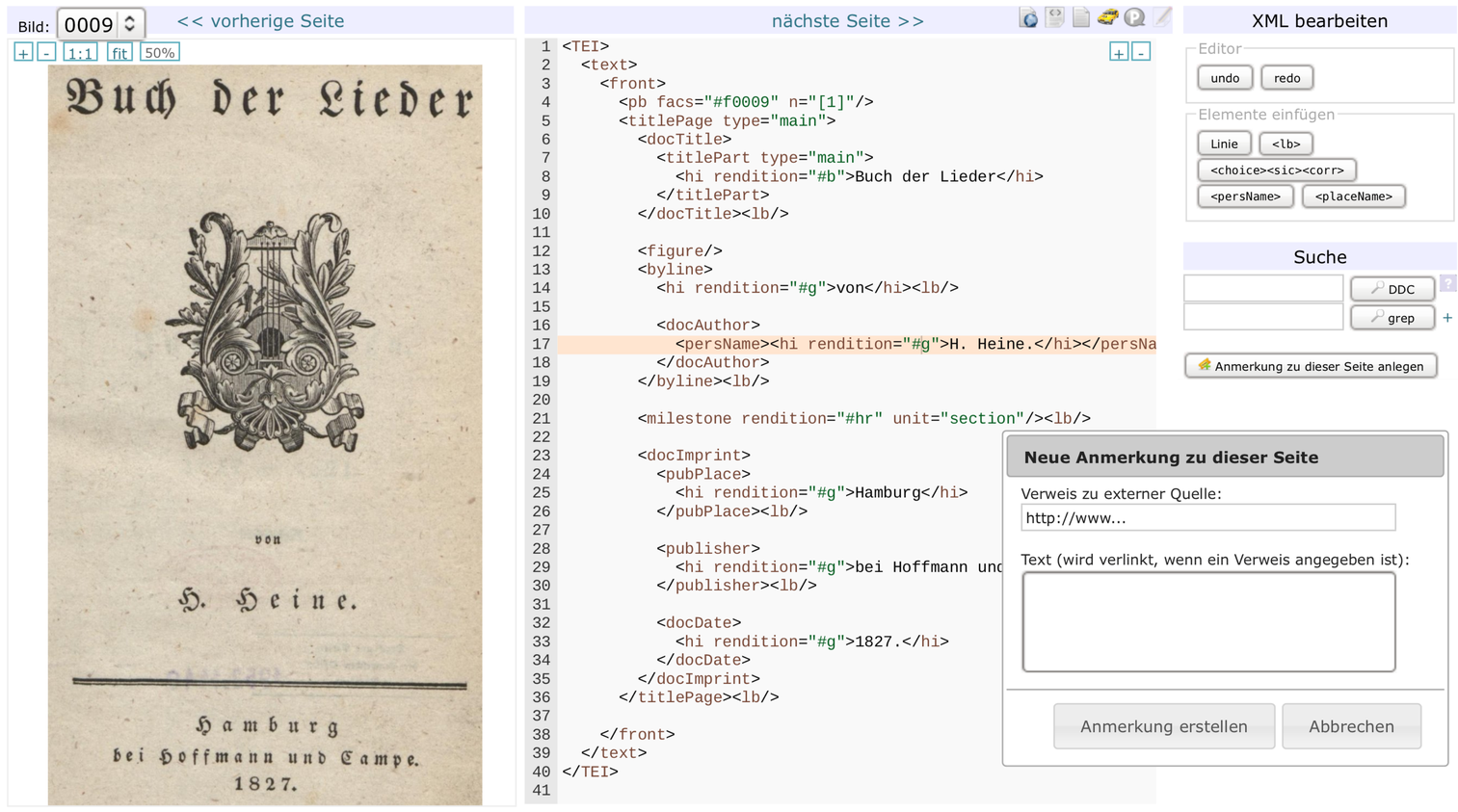
dta@bbaw.de
www.deutschestextarchiv.de

Quality Assurance in a Collaborative System — DTAQ



Quality assurance for all DTA corpus texts takes place within the quality assurance platform DTAQ. In DTAQ texts may be proofread pagewise in comparison with their source images. The GUI is highly customizable, so that different views of the transcriptions (XML/TEI, HTML, raw text, linguistic analyses) can be offered. An authentification and an access control system help managing different users within DTAQ. Users may categorize errors, report them for correction, and, in the future, correct them within DTAQ. To avoid repetitions in proofreading, pages can be marked as proofread. Using this technique, different quality levels of the DTA texts can be specified.

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 formulae 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.

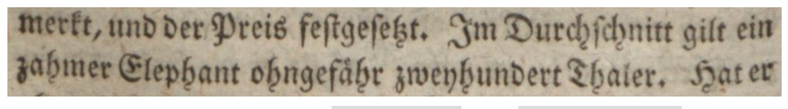
$$\lambda_T = \frac{1}{\pi n s^2} \int_0^\infty \frac{4\,x^2\,e^{-x^2}\,dx}{\psi(x) + \frac{n_1\sigma^2}{n s^2}\psi\left(x\sqrt{\frac{m_1}{m}}\right)},$$

$$\lambda_T = \frac{1}{\pi n s^2} \int_0^\infty \frac{4x^2 e^{-x^2}\,dx}{\psi(x) + \frac{n_1\sigma^2}{n s^2}\psi\left(x\sqrt{\frac{m_1}{m}}\right)}$$

Boltzmann: Vorlesungen, Bd. 1, Leipzig, 1896, p. 73.

Integrated CAB View

The Cascaded Analysis Broker (CAB; cf. Jurish, JLCL 2010, 25/1) provides a normalization of historical forms in order to allow for orthography-independent and lemma based corpus searches.


merkt, und der Preis festgesetzt. Im Durchschnitt gilt ein zahmer Elephant ohngefähr zweyhundert Thaler. Hat er

angemerkt, und der Preis festgesetzt. Im Durchschnitt gilt ein zahmer Elefant ungefähr zweihundert Taler. Hat er

Thunberg: Reisen, Bd. 2, Berlin, 1794, p. 472.

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.