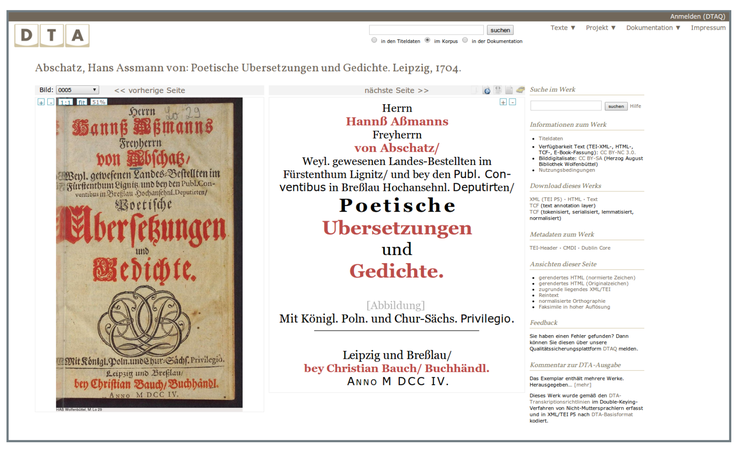
DEUTSCHES TEXTARCHIV (DTA)
THE GERMAN TEXT ARCHIVE

BERLIN-BRANDENBURG ACADEMY OF SCIENCES AND HUMANITIES

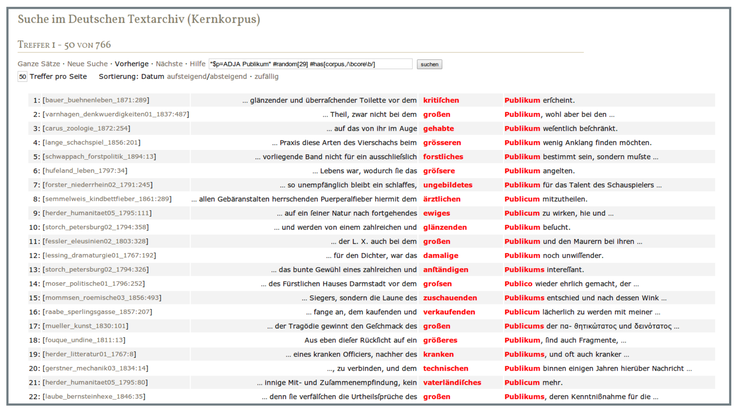
DEUTSCHES TEXTARCHIV (DTA)

The DFG-funded project Deutsches Textarchiv (German Text Archive; DTA) at the Research Centre Language of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) is building up a core corpus of ~1500 historical German texts (17th–19th century). This core corpus is balanced with regard to time of creation, text type, and thematic scope, thus serving as a basis for a reference corpus of the historical New High German language. This way, the DTA offers highly relevant primary sources for academic research in the linguistics and various other disciplines of the humanities and sciences as well as for legal scholars and economists.



Abschatz, Hans Assmann von: Poetische Übersetzungen und Gedichte. Leipzig, 1704. [Title page] In: Deutsches Textarchiv <http://www.deutschestextarchiv.de/abschatz_gedichte_1704/5>.

Text digitization within the DTA is based on the earliest edition accessible for each work, and is conducted closely to the underlying original text without any editorial interventions.
All texts are structured according to the TEI/P5 guidelines and are made freely available via the Internet in various formats (XML/TEI, HTML, plain text, etc.) along with their corresponding digital facsimiles as well as with comprehensive bibliographic metadata. The electronic full-texts are enriched with linguistic information gained through automatic tokenization, lemmatization, part-of-speech tagging, and modernization of the historical spelling. Thus, the DTA corpus can be explored with regard to structural as well as linguistic information.

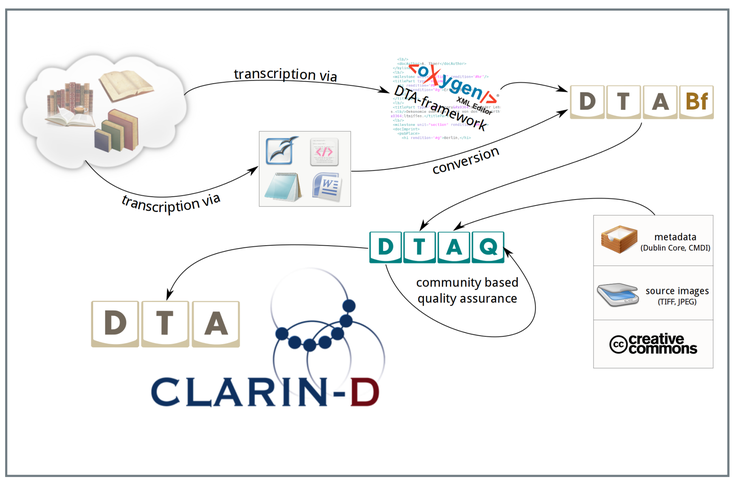


Example query: DDC-search for an attributive adjective followed by a morphological variant of the term 'Publikum' (audience). For more information on the DTA query engine cf. <http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe>.

THE DTA ›BASE FORMAT‹ (DTABf)

All DTA corpus texts are annotated according to the well-documented DTA ›base format‹ (DTABf), a strict TEI/P5 subset for the structuring of (historical) written corpora. The DTABf is designed to provide tagging solutions for a wide range of structural phenomena while avoiding ambiguities of the tagset in order to assure consistent tagging over the entire corpus. This way, all DTABf texts become interchangeable and truly interoperable. The DTABf is recommended as best practice format for (historical) written corpora in the context of CLARIN-D.

DTA EXTENSIONS (DTAE)



The DTAE Workflow, cf. <http://www.deutschestextarchiv.de/dtae>.

To broaden the text base, the DTA core corpus is gradually extended by high-quality textual resources provided by other projects, which are curated in the course of the module DTA Extensions (DTAE) and of a BMBF-funded 'curation project' in CLARIN-D, respectively. Additions include e.g. large text collections such as 'Dingler's Polytechnisches Journal' (1820–1931; 370 volumes, ~78M tokens; project at Berlin's Humboldt University) and the journal 'Die Grenzboten' (1841–1922; 270 volumes, ca. 180.000 pages and ca. 453M characters; project of the SuUB Bremen and the BBAW).

All external texts were converted from various formats into the DTABf and were enriched with linguistic information. This way, they could be made available on the DTA platform, where they can now be be exploited in the context of the whole DTA corpus.

DTA QUALITY ASSURANCE (DTAQ)

All DTA corpus texts are available in DTAQ, a web based platform for collaborative quality assurance. Within DTAQ, the transcription can be proofread, and misprints, transcription or annotation errors as well as erroneous metadata can be recommended for correction.

Storm, Theodor: Waldwinkel, Pole Poppenspäler. Novellen. Braunschweig, 1875, p. 173. In: Deutsches Textarchiv Quality Assurance <http://www.deutschestextarchiv.de/dtaq/book/view/30122?p=177>.

DTA IN NUMBERS (JAN. 2014)

- DTA core corpus: 1302 volumes (~97M tokens, ~684M characters)
- DTAE corpus: 1060 texts from external sources (~98M tokens, ~516M characters)
- DTA corpora in total: 2362 volumes (~195M tokens, ~1.2B characters)

D T A

Further reading:
www.deutschestextarchiv.de/doku/publikationen
@textarchiv

Contact
dta@bbaw.de
www.deutschestextarchiv.de

DFG Deutsche Forschungsgemeinschaft
CLARIN-D
berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN