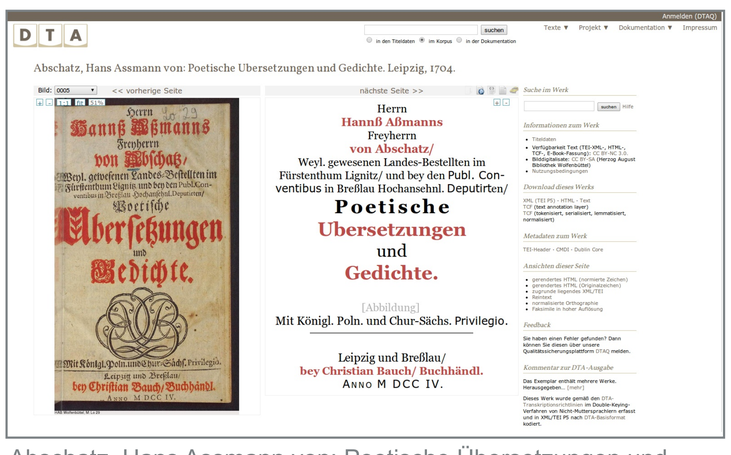
BERLIN-BRANDENBURG ACADEMY OF SCIENCES AND HUMANITIES

THE GERMAN TEXT ARCHIVE
DEUTSCHES TEXTARCHIV (DTA)

OPEN ACCESS TO MORE THAN 6400 HISTORICAL TEXTS

The project Deutsches Textarchiv (German Text Archive; DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) was funded by the German Research Foundation DFG from 2007 to 2016 to build 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 linguistics and various other disciplines of the humanities and sciences. 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 available under Creative Commons (CC) Licenses.

As of June 2020 and for the majority of the DTA texts, including the complete core corpus, the "Non Commercial"-restriction, hindering economic and various other reuses of the data, including Wikimedia's different projects, was dismissed.

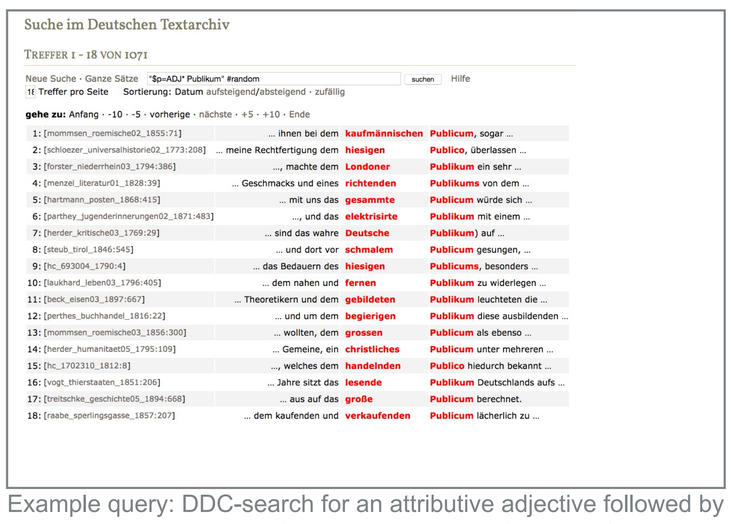


Abschatz, Hans Assmann von: Poetische Übersetzungen und Gedichte. Leipzig, 1704. [Title page] In: Deutsches Textarchiv <http://www.deutschestextarchiv.de/abschatz_gedichte_1704/5>.

The 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 historical spelling variants.

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 provides 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 truly interoperable. The DTABf is recommended as a best practice format for (historical) written corpora in the context of CLARIN-D, and is also recommended as a general baseline encoding by the German Research Foundation (DFG).

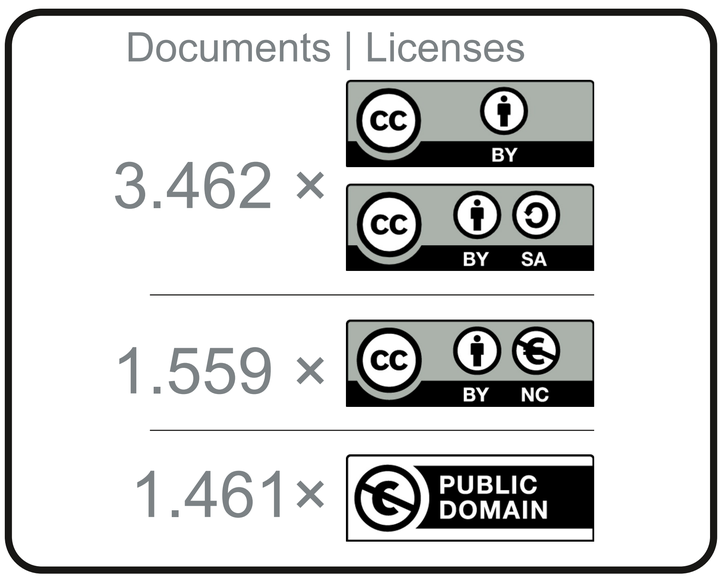


Example query: DDC-search for an attributive adjective followed by a morphological variant of the term 'Publikum' (audience). For more information on the DTA query engine cf. <http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe>.

DTA EXTENSIONS (DTAE)

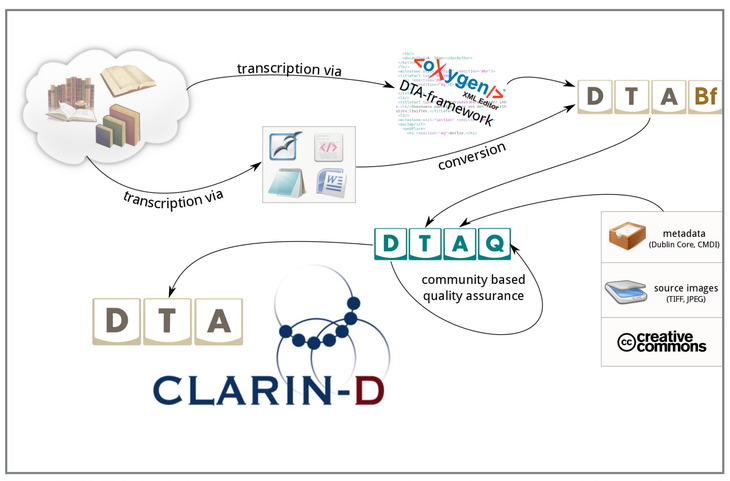
To broaden the text base, the DTA core corpus has been enriched by high-quality textual resources provided by other projects, which are curated in the context of the module DTA Extensions (DTAE). These supplementary corpus include large text collections such as 'Dingler's Polytechnisches Journal' (1820–1931; 370 volumes, ~78M tokens; supplied by Humboldt-Universität zu Berlin), the journal 'Die Grenzboten' (1841–1922; 270 volumes, ca. 180.000 pages and ca. 453M characters; supplied by the SuUB Bremen and the BBAW) and finally the Digitale Sammlung Deutscher Kolonialismus (DSDK) (1884–1920; 948 documents, ~170.000 pages).

THE DTA IN NUMBERS (JUNE 2020)



CLARIN-D

The European research infrastructure project CLARIN-ERIC is a web- and centers-based research infrastructure for the sciences and the humanities. Its German section, CLARIN-D, is funded by the German Federal Ministry for Education and Research (BMBF), and builds on the expertise of currently nine service centers in major research institutions. All texts from the DTA corpora are stored in the certified CLARIN-D repository at the BBAW, ensuring their further dissemination, long-term accessibility and preservation, as well as reliable addressability via Persistent Identifiers (PID).



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

Contact
dta@bbaw.de
www.deutschestextarchiv.de
@textarchiv

D T A

CLARIN-D

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN