OPEN ACCESS TO MORE THAN 3000 HISTORICAL TEXTS

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, 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 as well as for legal scholars and economists. 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 as OA-Documents under Creative Commons Licenses.

The texts are structured according to the TEI/PS 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 corpus can thus be explored with regard to structural as well as linguistic information.

THE DTA ‘BASE FORMAT’ (DTABf)

All DTA corpus texts are annotated according to the well-documented DTA ‘base format’ (DTABf), a strict TEI/PS 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 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).

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) 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; supplied by Humboldt-Universität zu Berlin) and the journal ‘Die Grenzboten’ (1841–1922; 270 volumes, ca. 180.000 pages and ca. 453M characters; supplied by the SuUB Bremen and the BBAW).

CLARIN-D

The European research infrastructure project CLARIN-ERIC is developing a web- and centers-based research infrastructure for the sciences and the humanities. Its German section, CLARIN-D (www.clarin-d.de), 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 corpus 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 DTA IN NUMBERS (SEPT. 2016)


Selection of partners contributing to the DTAE programme.

2.144 documents
614 documents
271 documents

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


Poster available under the Creative Commons Attribution 4.0 Licence.