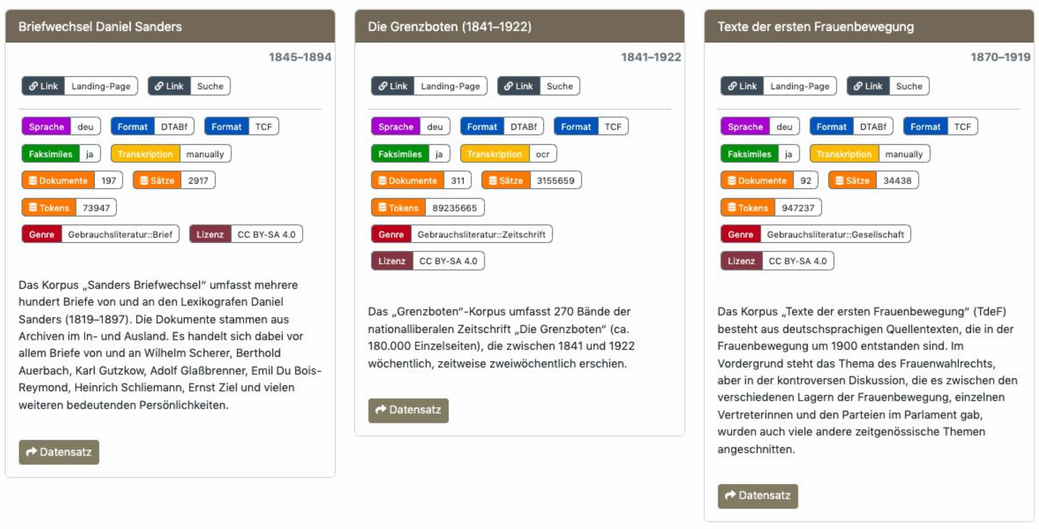
berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Historische Textsammlungen im Deutschen Textarchiv (DTA)

Das Deutsche Textarchiv (DTA) ist ein Archiv für deutschsprachige, historische Korpora und Sammlungen am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW). Es umfasst annotierte Volltexttranskriptionen von Drucken, Zeitungen und Zeitschriften sowie handgeschriebene Dokumente verschiedener Gattungen und Textarten. Für die Transkriptionen bietet das DTA mit dem DTABf – ein Subset der TEI-P5-Guidelines – ein etabliertes Basisformat an.

Spezialisierung	- deutschsprachige Texte - Zeitraum ca. 1600–1920 - hochwertige Transkriptionen - strukturiertes Textformat - kuratierte Metadaten - offene Lizenz - Dokumentation
Modalität	- geschrieben
Akzeptierte Datenformate	- DTABf - TEI - XML - gg.falls DOCX, EPUB
Ansprechperson(en)	- Marius Hug (marius.hug@bbaw.de) - redaktion@deutschestextarchiv.de

Forschungsdaten im DTA



Das DTA stellt aktuell etwa 40 Textsammlungen als Forschungsdaten zur Nachnutzung bereit. Im Zentrum steht das DTA-Kernkorpus, das mit rund 1.500 Werken die Grundlage für ein Referenzkorpus des Neuhochdeutschen vom 16. bis zum frühen 20. Jahrhundert darstellt.

Integration von Forschungsdaten

Projekte, die hochwertige Transkriptionen anfertigen und ein nachnutzbares Textformat verwenden, Metadaten bereitstellen, Lizenzfragen geklärt haben und eine detaillierte Dokumentation liefern können, finden im DTA eine etablierte Infrastruktur zur Bereitstellung ihrer Forschungsdaten. Das DTA berät zu allen Belangen, angefangen von Verfahren der Transkription, über die Annotation bis hin zur Dissemination der Forschungsdaten.

Das DTA im Kontext der NFDI

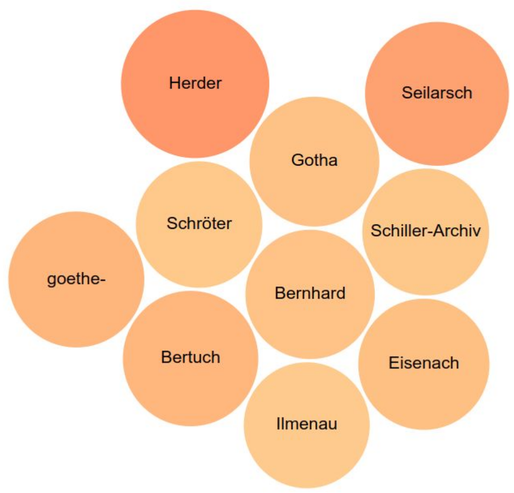
Durch die Bereitstellung der DTA-Forschungsdaten über die Schnittstellen des DWDS ist die Voraussetzung zur Integration der Sammlungen in das zentrale Nachweissystem von Text+ gegeben. Alle Ressourcen des DTA werden damit Teil der NFDI-Infrastruktur, was in Bezug auf Zugänglichkeit und Sichtbarkeit einen großen Mehrwert bedeutet.

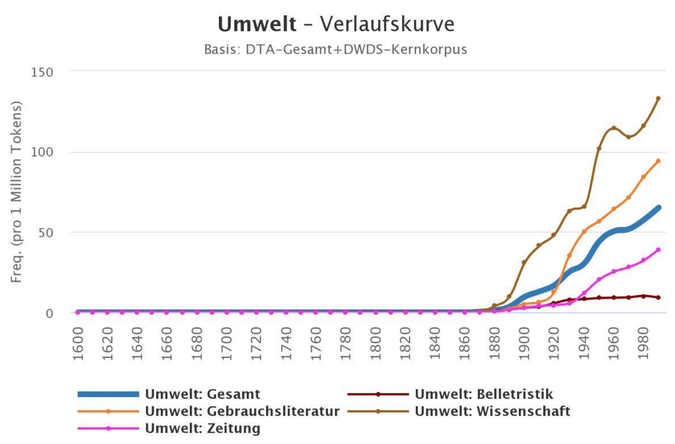
DTA www.deutschestextarchiv.de

DWDS www.dwds.de

Das DTA am Zentrum Sprache

Das DTA ist eng mit dem Digitalen Wörterbuch der deutschen Sprache (DWDS) verbunden und innerhalb einer gemeinsamen Infrastruktur zugänglich. Daraus ergibt sich ein Korpusbestand, der mehr als 500 Jahre umfasst, sowie die Möglichkeit, die vom DTA bereitgestellten Forschungsdaten über die Korpustools des DWDS zu explorieren und zu analysieren. Das Zentrum Sprache der BBAW dient als Kompetenzzentrum für historische Texte und Daten sowie für Formatspezifikationen und Standardisierungsaktivitäten in den internationalen Fachcommunitys.





DiBiPhil: Priorisierung der Kuration in Kooperation mit FID

In Absprache mit dem FID-Philosophie wurde in Q2/2023 ein neues Korpus von rund 270 philosophischen Werken kuratiert und in die Infrastruktur integriert. Alle diese Werke sind DTABf-kodiert und wurden umfangreich mit Metadaten angereichert. Sie sind mittels der linguistischen Suchmaschine des DWDS analysierbar und stehen unter einer CC BY-SA-Lizenz zur Nachnutzung bereit.

Ausblick auf zukünftige Datenaufnahme

Die Trennung von Daten und Metadaten soll bei der Aufnahme von Forschungsdaten auch zukünftig eine wichtige Rolle spielen. Aktuell wird an einer vereinfachten Möglichkeit gearbeitet, die den jeweiligen Texten zugehörigen bibliographischen Metadaten (unabhängig vom Dateiformat) so zu übergeben, dass anschließend ein möglichst automatisierbarer Kurationsworkflow angewendet werden kann.