



Deutsches Textarchiv

Deutsches Textarchiv, Berlin-Brandenburgische Akademie der Wissenschaften (ed.), 2017. <http://www.deutschestextarchiv.de/> (Last Accessed: 16.08.2017). Reviewed by Dario Kampkaspar (Herzog August Bibliothek / Austrian Centre for Digital Humanities an der Österreichischen Akademie der Wissenschaften), [dario.kampkaspar \(at\) oeaw.ac.at](mailto:dario.kampkaspar@oeaw.ac.at).



Abstract

Owing to its well documented TEI subset and highly accurate transcriptions usually based on the first edition of a text, the German Textarchive (= Deutsches Textarchiv, DTA) is currently one of the best corpora for historical German texts (1600-1900), albeit not necessarily the most extensive. To date (August 2017) it contains around 3260 works. The full texts of the corpus are enhanced by digital facsimiles and encoded with regard to their visual features (e.g. layout, fonts) as well as annotated linguistically (e.g. PoS-tagging). The search engine focuses on linguistic features and allows for searching both exact spellings as well as spelling variants of a word. Due to its goal of providing a corpus for historical linguistics and the underlying selection criteria of the collection, no further comments or variants in other editions of one work than the first print are given. Even though there are minor points of criticism, due to the quality of its textual sources and the accurate documentation the DTA can be seen as a point of reference for other corpora.

Einleitung

1 Das Deutsche Textarchiv (DTA) entstand als Projekt an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und wurde von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft (DFG) unter der Leitung von

Wolfgang Klein und Alexander Geyken gefördert. Mit derzeit (August 2017) über 620.000 Seiten aus mehr als 3.260 verschiedenen Werken aus der Zeit zwischen ca. 1600 und 1900¹ gehört das DTA zwar nicht zu den größten Textkorpora, doch dies ist gar nicht sein Anspruch. Die übergreifende Leitlinie wird dem Nutzer gleich auf der Startseite mitgeteilt:² „Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache“ oder, detaillierter, in den Leitlinien:³ „Ziel des Deutschen Textarchivs (DTA) ist die Erstellung eines disziplinenübergreifenden Volltextkorpus deutschsprachiger Texte“. In diesem Sinne geht es mehr um die Erreichung und Einhaltung einer verlässlichen Qualität der Texterfassung, insbesondere was die sprachlichen Eigenarten eines Textes angeht. Neben dieses ‚Kernkorpus‘ treten noch die sogenannten DTA-Erweiterungen.⁴ Hierbei handelt es sich um Texte, die teils unter anderen Gesichtspunkten ausgewählt und digitalisiert wurden (siehe Abschnitt zum DTA Erweiterungskorpus DTAE).

Das Kernkorpus

2 Aus der Zielsetzung des Kernkorpus ergeben sich viele der Eigenschaften des Deutschen Textarchivs, allen voran der Umstand, dass bei der Textauswahl und -aufbereitung hauptsächlich linguistische Gesichtspunkte im Vordergrund standen. Die Werke im DTA wurden nicht nach thematischen Gesichtspunkten ausgewählt, sondern, dem Ziel des Korpus entsprechend, wurde auf eine ausgewogene Zusammenstellung nach Textgenres geachtet und auf eine historisch-kritische Kommentierung der Texte verzichtet. Zur bestmöglichen Dokumentation des Sprachstandes wird überwiegend die erste verfügbare Ausgabe wiedergegeben.

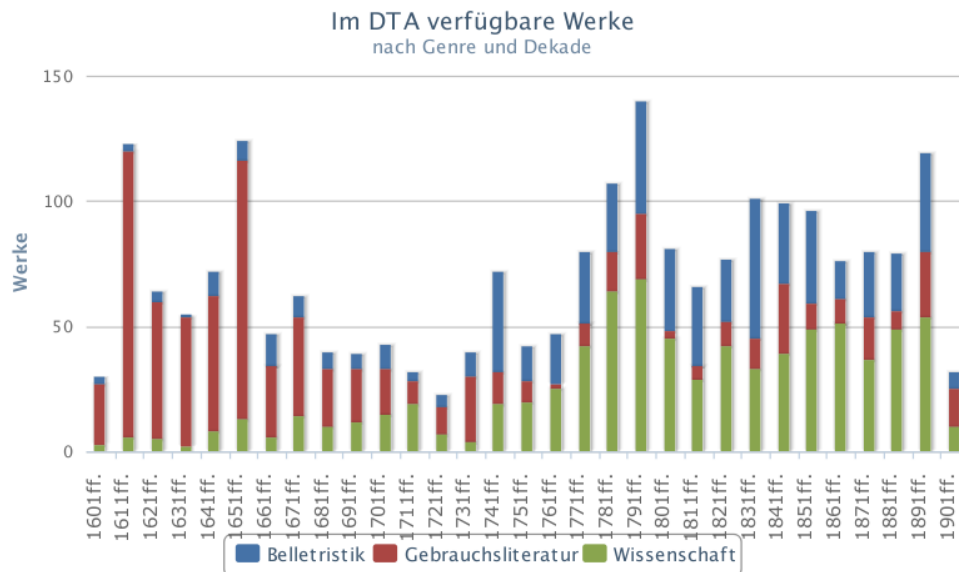


Abb. 1: Zeitliche Verteilung der Texte.

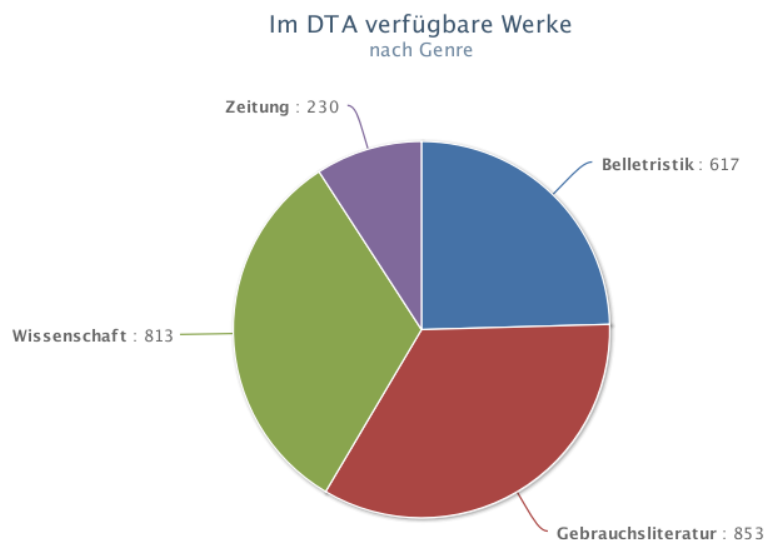


Abb. 2: Texte nach Genre.

3 Auch wenn der Anspruch einer ausgewogenen Textauswahl zur Abbildung des Neuhochdeutschen Sprachstandes und seiner Entwicklung nachvollziehbar ist, ist die Dokumentation, in der die Kriterien dieser Auswahl beschrieben werden,⁵ etwas dünn. So heißt es: Es wurden „mehrere Quellen herangezogen“, wobei „ausgewählte Literaturgeschichten“ und „Empfehlungen der Mitglieder der BBAW“ explizit genannt werden. Die Dokumentation der Textauswahl enthält zwei Graphiken: eine zur zeitlichen Verteilung der Werke ([Abb. 1](#)), die zur Zeit nur bedingt als gleichmäßig angesprochen

werden kann, und eine zur Verteilung der Texte nach Genre ([Abb. 2](#)), wobei hier ‚Belletristik‘, ‚Gebrauchsliteratur‘, ‚Wissenschaft‘ und ‚Zeitung‘ genannt werden.⁶

4 Scheint die Verteilung unter den ersten drei Gruppen noch recht gleichmäßig zu sein, stellt eine detailliertere Betrachtung⁷ ein etwas anderes Bild dar: So enthält die Gruppe ‚Belletristik‘⁸ beispielsweise eine Untergruppe ‚Briefe‘ mit mehreren Bänden Briefsammlungen⁹ (sowie einem Tagebuch!), während die Untergruppe ‚Novelle‘ deutlich mehr Werke verschiedener Personen umfasst (mit einem auffälligen Übergewicht auf den Werken Theodor Storms). Auch andere Untergruppen wie etwa ‚Lyrik; Prosa‘ zeigen, dass hier gegebenenfalls noch etwas Klassifizierungs- wie Aufteilungsarbeit geleistet werden könnte.

Texterfassung und Formate

5 Das DTA versucht, bei den Texten des Kernkorpus eine möglichst hohe Erfassungsgenauigkeit zu erreichen. Neben der initialen Erfassung der Texte, die laut den Leitlinien in der Regel per *Double Keying* erfolgt, steht eine eigene browserbasierte Umgebung zur Qualitätskontrolle (DTA-Qualitätssicherung, DTAQ¹⁰) zur Verfügung. Über DTAQ können registrierte Benutzer Korrekturen und Annotationen am Text anbringen, die dann von Mitarbeitern des DTA geprüft und gegebenenfalls in die Texte eingearbeitet werden. Detailliert ausgearbeitete und sehr gut dokumentierte Regeln zur Transkription erlauben eine sehr verlässliche Auswertung der Texte, was auch durch das zugrunde gelegte Datenmodell unterstützt wird. Hierbei wurde basierend auf TEI P5 das DTA-Basisformat (DTA-Bf)¹¹ erstellt, das ebenfalls vorbildlich dokumentiert ist. Dieses Format wurde dementsprechend auch von der DFG als Format für historische Texte empfohlen.¹²

6 Der Fokus des Formates und der Erfassungsarbeit liegt auf der exakten Erfassung der Textstruktur sowie der originalen Schreibweisen und damit dem Aufbau eines Korpus historischer Wortformen. Für die weitergehende linguistische Bearbeitung wird auf das im Kontext von CLARIN-D entwickelte Text Corpus Format (TCF)¹³ zurückgegriffen.

7 Wie bereits angesprochen kann die Dokumentation der Texterfassung und Modellierung des DTA in weiten Teilen als vorbildlich angesehen werden. Neben verschiedenen thematischen Zugängen (so zum Beispiel Metadaten und formale wie inhaltliche Erschließung) gibt es Übersichten zu den verwendeten Elementen und eine Suche innerhalb der Dokumentation. Da die Einhaltung dieser Regeln im Rahmen der

Qualitätskontrolle DTAQ gesichert wird, sind die Texte des Kernkorpus in vielen Belangen der Goldstandard im Hinblick auf die weitere Verwendbarkeit. Ebenfalls positiv anzumerken ist, dass auch die Bestandteile der Dokumentation durch die Lizenz CC BY-SA 3.0 frei verfügbar sind.

Generelle Benutzerführung

8 Der Zugriff auf die Texte ist auf mehreren Wegen möglich. Am auffälligsten ist dabei direkt auf der Startseite die Suchmöglichkeit, die bei einer Volltextsuche innerhalb des Korpus auch linguistische Verfeinerungen erlaubt.

9 Als linguistische Suchengine wird DDC („Dialing/DWDS-Concordancer“) verwendet, das umfangreiche linguistische Abfragen bietet. Unterstützt wird diese durch CAB („Cascaded Analysis Broker“), ein Tool zur Findung moderner Wortformen zu historischen Schreibungen. Auch die (im Vergleich zu einfachen Volltextsuchen natürlich umfangreichere) Syntax der Engine ist mit Beispielen dokumentiert.¹⁴ Einige Features sind dabei die lexembasierte Suche (durch CAB gefundene Formen zu einem Grundlexem), Suche nach einer genauen Form mit oder ohne Trunkierung und die Suche von Phrasen (genaue Phrasen oder auch Wörter mit bestimmten Abständen zueinander) oder eine Suche aufgrund der Wortart. Diese Suchmöglichkeiten können für komplexe Anfragen miteinander kombiniert werden.

The screenshot displays the DTAQ (Deutsches Textarchiv) search interface. At the top, there's a navigation bar with 'DTAQ' logo and links for 'Anmelden (DTAQ)', 'Suchen', 'Hilfe', 'Texte', 'Projekt', 'Dokumentation', and 'Impressum'. Below this, a search bar contains the query 'wegen @dem'. The main section, titled 'Suche im Deutschen Textarchiv', shows 'Treffer 1 - 10 von 195'. It includes a search bar with 'Neue Suche - KWIC' and 'Suchen' button. Below the search bar, there's a list of search results. Each result includes a thumbnail of a document page, a title, and a snippet of text. The snippets show the search term 'wegen dem' highlighted in red. To the right of the search results, there's a 'Verlaufsdiagramm DTA-DWDS' (Line graph DTA-DWDS) showing the frequency of the search term over time. Below the graph, there's a legend explaining the data series: 'w: Originaltext, UTF-8-kodiert', 'w: approximierter Latex-1-Text', 'w: CAB-normalisierte Wortform', 'l: Lemma (unfalte Form)', and 'p: Part-of-Speech-Analyse'.

Abb. 3: Ergebnisansicht der Suche.

10 In der Ergebnisansicht ([Abb. 3](#)) werden die Treffer im Kontext ihres Satzes angezeigt, wobei jeder Eintrag in der Liste einem Treffer entspricht (mehrere Treffer auf der gleichen Seite erzeugen standardmäßig jeweils eigene Einträge).¹⁵ Ebenfalls angeboten wird eine Verlaufskurve der gesuchten Form innerhalb des (gewählten Teil-)Korpus, wobei die Zahl der Treffer je einer Million Tokens über der Zeit angegeben wird. Ein Klick auf diese Graphik führt zu einer detaillierteren Ansicht.

11 Neben der Volltextsuche gibt es in der Seitenspalte der Startseite die Möglichkeit zum ‚Stöbern im DTA‘, unter der alle Texte nach dem Jahrhundert der Veröffentlichung sowie nach der Textgattung (hier sehr weit gefasst in die drei Gruppen ‚Belletristik‘, ‚Gebrauchsliteratur‘ und ‚Wissenschaft‘; eine feinere Aufteilung erfolgt auf den jeweiligen Unterseiten) gruppiert werden. Dazu kommt im Menü ‚Texte‘ eine Zeitleiste, die die Titel nach ihrem Erscheinungsjahr zusammengefasst aufführt. Wer Werke eines bestimmten Verfassers sucht, muss entweder die Suche in den Titeldaten bemühen oder – weniger offensichtlich – in der Seitenspalte unter ‚Neue Werke im DTA‘ dem Link ‚alle Titel ...‘ folgen.

Textansichten

The screenshot shows the DTA (Deutsches Textarchiv) interface. At the top, there's a navigation bar with 'DTA' and search options. The main content area is divided into several sections:

- Thumbnail:** A small image of the title page of the book 'Leib-Medicus der Studenten' by Heinrich Kaspar Abel.
- Bibliographische Angaben:** A section containing bibliographic data such as URN, title, author, and publication details.
- Informationen zum Werk:** A section providing more details about the work, including its genre and availability.
- Inhaltsverzeichnis:** A table of contents listing the chapters and sections of the book.
- Suche im Werk:** A search bar for finding specific content within the text.
- Ansichten für dieses Werk:** A section showing different views of the text, including 'Text-Bild-Ansicht', 'alle Faksimiles', and 'DTAQ (Qualitätssicherung)'.
- Download:** A section offering download options for the text in various formats like XML, HTML, and TCF.
- Metadaten:** A section displaying metadata such as TEI-Header, Dublin Core, and statistics.
- Wortwolken:** A section showing word clouds generated from the text.

Abb. 4: Einstiegsseite eines Textes.

12 Die Einstiegsseite eines Textes ([Abb. 4](#)) bietet bibliographische Angaben, weitere Metadaten, den Zugang zu den Ansichten des Textes, verschiedene Download-Optionen sowie Statistiken und Links zu einigen Tools (zur Zeit Wortwolken sowie die Übergabe des transliterierten, normalisierten oder lemmatisierten Textes an *Voyant Tools*). Die bibliographischen Angaben enthalten die üblichen Titelangaben, die URN des Textes zur permanenten Adressierung sowie Informationen zu dem für die Transkription verwendeten Exemplar. Die Metadaten sind neben anderen Formaten als TEI-Header und Dublin Core verfügbar, während für den Volltext verschiedene Versionen angeboten werden: erfasster Volltext als XML (im DTA-Bf), eine HTML-Ansicht, Plain Text, sowie linguistische Annotationen verschiedenen Umfanges als TCF. Die Download-Option ‚HTML‘ gibt zwar den gesamten Text als HTML wieder, bietet diesen aber tatsächlich zum Speichern an; den Volltext in einer einzigen HTML-Datei online einzusehen, scheint nicht vorgesehen zu sein.

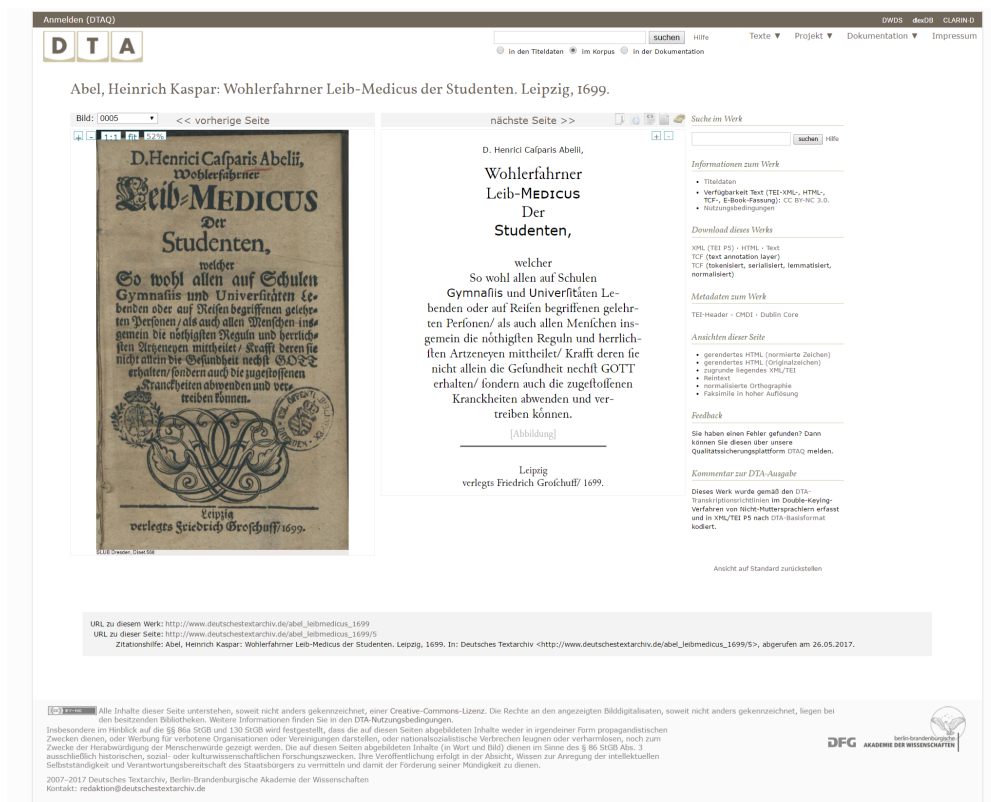


Abb. 5: Parallelansicht einer Seite.

13 Der Standard-Zugang zum Text ist die Text-Bild-Ansicht, die eine seitenweise Parallelansicht bietet (Abb. 5). Die auf der Einstiegsseite gebotenen Optionen sind hier (bei ausreichender Fensterbreite) am rechten Seitenrand zu finden. Die parallele Ansicht stellt Digitalisat und Transkription nebeneinander, wobei für die Transkription verschiedene Anzeigeformate gewählt werden können: das originale XML, eine HTML-Ansicht der getreuen Abschrift oder verschiedene Normalisierungen. Leider sind die Größen der Seitenbestandteile fix vorgegeben, sodass teils sogar bei kurzen Seiten innerhalb der Transkription gescrollt oder die Schriftgröße angepasst werden muss, um den gesamten Text einer einzelnen Textseite zu lesen. Die Informationen zum gesamten Werk sind auch jederzeit abrufbar und es werden außerdem der Link zum gesamten Werk, zur einzelnen Seite (nicht aber jeweiligen Ansicht) und eine Zitationshilfe mitgegeben. Verschiedene Versionen des Textes, der sich durch die Qualitätssicherung DTAQ ändern könnte, sind nicht abrufbar. Zwar ist durch die bereits vor der Veröffentlichung greifenden Kontrollmechanismen nicht von großen Änderungen auszugehen, doch ist gerade bei den oft feinfühligsten linguistischen Suchen hier das Problem der fehlenden Nachvollziehbarkeit einer Aussage, die auf einem älteren Datenstand basiert, nicht von der Hand zu weisen. Bislang noch nicht umgesetzt, aber eine denkbare Erweiterung wären, sind Möglichkeiten zur Anbringung privater

Annotationen oder Lesezeichen sowie Möglichkeiten zur Adressierung von Textauszügen, zum Beispiel unter Nutzung der Tokenisierung.

DTA Erweiterungskorpus (DTAE)

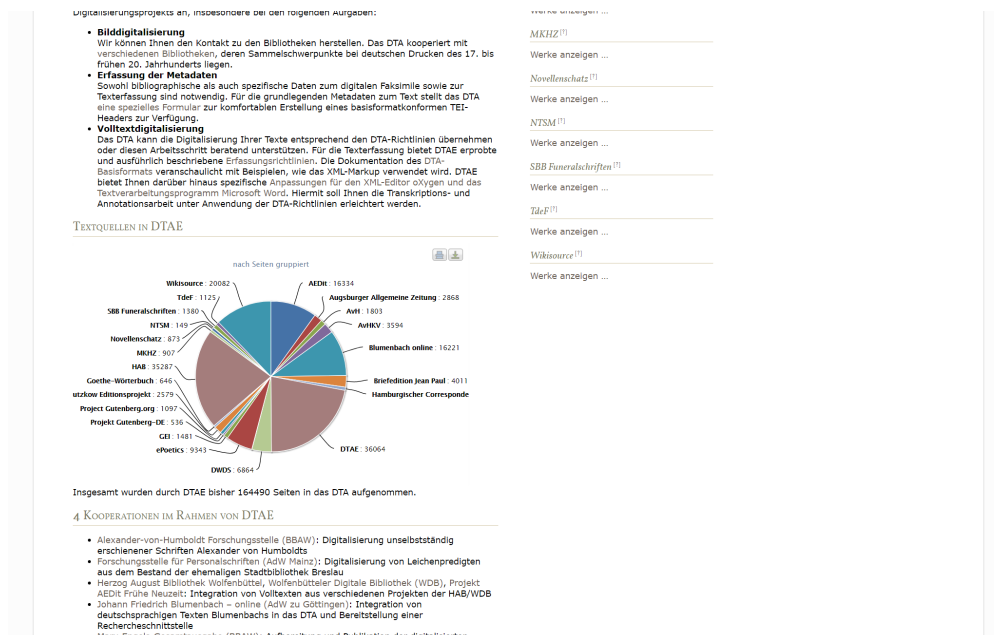


Abb. 6: Ausschnitt der Seite zum DTAE.

14 Die derzeit (August 2017) knapp 165.000 Seiten im Erweiterungskorpus¹⁶ stammen aus externen Quellen, wobei es sich um Korpora einzelner Institutionen wie der Wolfenbütteler Digitalen Bibliothek (WDB), der Herzog August Bibliothek Wolfenbüttel (HAB), Kooperationsprojekte wie ‚AEDit‘ (HAB, BBAW und Akademie der Wissenschaften und Literatur in Mainz) oder Einzelprojekte wie ‚Semantic Blumenbach‘ an der Akademie der Wissenschaften in Göttingen handeln kann (Abb. 6). Ergänzt wird das DTAE durch weitere Volltexte, die aus digitalen Archiven (überwiegend Wikisource) übernommen und in das DTA-Bf überführt wurden.

15 Ausführlichere Informationen zu den einzelnen Quellen der übernommenen Texte sind nicht auf derselben Seite zu finden – unter der Graphik sind nur kurze Informationen zu den Kooperationen aufgeführt –, sondern unter ‚Textquellen‘ in der Dokumentation.¹⁷ Hier erst findet man die Erklärung, dass sich unter dem Beiträger ‚DTAE‘ (wie er in der Grafik in Abb. 6 auftritt) kein Zirkelschluss verbirgt, sondern eine Zusammenfassung kleinerer Einzelquellen ohne genaue Zuordnung zu einer Institution.

16 In der rechten Spalte lässt sich auch eine Übersicht der Texte aufrufen, die aus den einzelnen Projekten übernommen sind. Eine Angabe, zu welchem Zeitpunkt und in

welcher Form (Bilddigitalisat, Volltexte) Daten von einer Quelle übernommen wurde, sucht man hier wie auch in der Kurzbeschreibung vergebens. Die Liste enthält alle übernommenen Texte, auch wenn sie noch nicht kontrolliert wurden. Es ist nicht erkennbar, welche Links zu einem Volltext führen und welche zum DTAQ. Eine deutlichere Kennzeichnung wäre für den Benutzer hilfreich.

Zugänglichkeit und Nachnutzung

17 Die Texte werden in der Regel unter verschiedenen freien Lizenzen zur Verfügung gestellt. Genauere Angaben lassen sich auf den jeweiligen Seiten einzelner Werke entnehmen. Die verschiedenen zur Verfügung gestellten Downloadformate und Lizenzen erlauben eine Weiterverwendung in verschiedenen Forschungskontexten.

18 Als einfachste Variante der Übernahme ist ein Download des gesamten Korpus (nur Kernkorpus oder Kern- und Ergänzungskorpus) wie auch der einzelnen Gruppen, wie sie unter ‚Stöbern‘ zu finden sind, möglich. Eine Gruppierung der Texte durch den Nutzer ist zurzeit nicht umgesetzt. Es werden außerdem einzelne ‚Versionen‘ dieser Download-Optionen angeboten, wenngleich diese eher den Charakter von Snapshots mit schwankendem zeitlichen Abstand von teils fast zwei Jahren haben und so nur als grobe Versionierung der Texte dienen können.

19 Die bei der Erstellung des Korpus verwendete Software wird im Rahmen der Dokumentation ebenfalls knapp beschrieben,¹⁸ wobei aber nicht immer klar ist, ob und inwiefern es sich um Eigenentwicklungen handelt. Teile der Software sind für die eigene Nutzung verfügbar, doch ist bei den externen Links oft nicht ersichtlich, in welchem Umfang und unter welcher Lizenz diese Software-Stücke genutzt werden können. Für die Verwendung im XML-Editor oXygen wird eine auf das DTA-Bf zugeschnittene Anpassung zur Verfügung gestellt, die mit einer Veröffentlichung im November 2013 aber auf einer nicht mehr aktuellen Version von oXygen basiert.

20 Erfreulich ist das Angebot an APIs. Neben einer OAI-PMH Schnittstelle wird auch eine API zu CAB angeboten.¹⁹ Etwas dürftig ist allerdings die ‚Dokumentation‘ der APIs ausgefallen, beschränkt sie sich doch auf eine einfache Aufzählung von Links. Ohne weitere Unterscheidung stehen hier Atom-Feeds neben einem Link zur OAI-PMH-Schnittstelle oder einem OpenSearch-Wrapper für die oben beschriebene Suche. Ergänzende Informationen könnten an dieser Stelle hilfreich sein.

Abschließende Bemerkungen

21 Das Deutsche Textarchiv bietet seinem Anspruch entsprechend qualitativ hochwertig aufbereitete Texte, die insbesondere, aber nicht nur für die linguistische Nachnutzung einen ausgezeichneten Standard bilden. Die Dokumentation vor allem des verwendeten Datenformates kann mit Fug und Recht als vorbildlich bezeichnet werden. Kleinere Eigenarten der Benutzerführung oder der zur Verfügung gestellten Ansichts-, Browsing- oder Downloadoptionen stellen keine gravierenden Probleme dar und werden vom Projektteam im Rahmen ihrer Möglichkeiten sicherlich auch angegangen werden.

22 Es steht zu hoffen, dass auch in Zukunft das hohe Niveau der Textaufbereitung gehalten und das Textkorpus sukzessive weiter ausgebaut werden kann. Leider ist der Webseite nicht zu entnehmen, wie sich die weitere Zukunft des DTA gestalten wird, nachdem die Förderung durch die DFG 2016 ausgelaufen ist. Genauso fehlen Angaben, ob die Vorhaltung der Texte dauerhaft gesichert ist. Zwar ist davon auszugehen, dass ein solch prominentes Projekt einer Akademie nicht sang- und klanglos verschwinden wird, doch sollten nach Ablauf eines Projektes zumindest kurze Hinweise zu Fragen der Archivierung und langfristigen Verfügbarkeit gegeben und möglichst sichtbar platziert werden.

23 Der weitere Ausbau des Korpus auch durch externe Kooperationspartner lässt hoffen, dass sich die Datengrundlage noch deutlich erweitern wird. Kleinere hier angeführte Wermutstropfen schmälern die Leistung nicht und werden durch das Projektteam sicher geprüft werden. Durch die Qualität der Texte, aber insbesondere auch durch die Dokumentation des DTA-Basisformates hat das DTA Vorbildcharakter für viele andere Projekte.

Anmerkungen

1. Die Angaben sind der auf der Startseite recht prominent platzierten Statistik entnommen, die einen guten Überblick über die laufende Arbeit am Korpus gibt.

2. DTA, Startseite: <https://web.archive.org/web/20170830180801/http://deutschestextarchiv.de/>.

3. DTA, Leitlinien: <https://web.archive.org/web/20170526105849/http://www.deutschestextarchiv.de/doku/leitlinien>.

4. DTA, DTAE: <http://www.deutschestextarchiv.de/dae>.

5. DTA, Dokumentation: <https://web.archive.org/web/20170526085322/http://www.deutschestextarchiv.de/doku/textauswahl>.

6. ‚Zeitung‘ als ‚Genre‘ taucht allerdings in der gattungsorientierten Auswahl auf der Startseite nicht auf.

7. Eine Übersicht über die Texte der Gruppen ist allerdings von hier aus nicht möglich; sie kann stattdessen von der Startseite aus erreicht werden.

8. DTA, Genre Belletristik <https://web.archive.org/web/20170526103123/http://www.deutschestextarchiv.de/list/browse?genre=Belletristik>.

9. Unter anderem Bettina von Arnims ‚Goethes Briefwechsel‘, eher ein Briefroman als eine unverfälschte Briefsammlung, und Ludwig Börnes ‚Briefe aus Paris‘, die auf Basis der früheren Edition der BBAW wiedergegeben wurden.

10. DTA, DTAQ: <http://www.deutschestextarchiv.de/daq>.

11. DTA, Basisformat: <https://web.archive.org/web/20170526111458/http://www.deutschestextarchiv.de/doku/basisformat/>.

12. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf und http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf.

13. Weblicht, TCF-Format: http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

14. DTA, DDC-Suche: https://web.archive.org/web/20170526123024/http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe.

15. Nach der Dokumentation der Suche sollten die Ergebnisse „alphabetisch nach dem Titel der Werke, in dem die Ergebnisse gefunden wurden, sortiert“ werden (vgl. https://web.archive.org/web/20170821082531/http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe). Die Kopfzeile einer einfachen Suche nennt eine Sortierung nach Datum. Ohne weitere Eingriffe scheint jedoch, wie im abgebildeten Beispiel zu sehen, keine

dieser Aussagen korrekt zu sein. Über die auf der Seite gegebenen Links oder über Parameter in der Suche kann zumindest eine Sortierung nach Datum erzwungen werden, eine alphabetische Anordnung ließ sich im Test nicht erreichen.

16. DTA, DTAE: <http://www.deutschestextarchiv.de/dtae>.

17. DTA, Textquellen: <https://web.archive.org/web/20170526092945/http://www.deutschestextarchiv.de/doku/textquellen>.

18. DTA, Software: <https://web.archive.org/web/20170526122208/http://www.deutschestextarchiv.de/doku/software>.

19. DTA, Dokumentation der APIs: <https://web.archive.org/web/20170529084103/http://www.deutschestextarchiv.de/api>, API zu CAB im Detail: <http://odo.dwds.de/~moocow/software/DTA-CAB/>.

Bibliographie

DTA: Deutsches Textarchiv. 2007-2017. Berlin-Brandenburgische Akademie der Wissenschaften.

<http://deutschestextarchiv.de/>.

Factsheet

Resource reviewed	
Title	Deutsches Textarchiv
Editors	Berlin-Brandenburgische Akademie der Wissenschaften
URI	http://www.deutschestextarchiv.de/
Publication Date	2017
Date of last access	16.08.2017

Reviewer	
Surname	Kampkaspar
First Name	Dario
Organization	Herzog August Bibliothek / Austrian Centre for Digital Humanities an der Österreichischen Akademie der Wissenschaften
Place	Wolfenbüttel, Germany / Wien, Austria
Email	dario.kampkaspar (at) oeaw.ac.at

General Information		
Bibliographic description	Can the text collection be identified in terms similar to traditional bibliographic descriptions (title, responsible editors, institution, date(s) of publication, identifier/address)? (cf. Catalogue 1.1)	no
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	yes
Contacts	Is contact information given? (cf. Catalogue 1.4)	yes
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	yes
Purpose	What is the purpose of the text collection? (cf. Catalogue 2.2)	General purpose
Kind of research	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Quantitative research

Self-classification	How does the text collection classify itself (e.g. in its title or documentation)? (cf. Catalogue 2.3)	Corpus, Digital Archive
Field of research	To which field(s) of research does the text collection contribute? (cf. Catalogue 2.2)	Linguistics
Content		
Era	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Early Modern, Modern
Language	What languages are the texts in? (cf. Catalogue 2.5)	German
Types of text	What kind of texts are in the collection? (cf. Catalogue 2.5)	Literary works, Essays, Newspaper/journal articles, Scientific papers, Speech transcripts
Additional information	What kind of information is published in addition to the texts? (cf. Catalogue 2.5)	Facsimile
Composition		
Documentation	Are the principles and decisions regarding the design of the text collection, its composition and the selection of texts documented? (cf. Catalogue 3.1.1-3.1.3)	yes
Selection	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Linguistic characteristics
Size		
Texts/records	How large is the text collection in number of texts/records? (cf. Catalogue 3.1.4)	> 1000
Tokens	How large is the text collection in number of tokens? (cf. Catalogue 3.1.4)	> 10 Mio.
Structure	Does the text collection have identifiable sub-collections or components? (cf. Catalogue 3.1.5)	yes
Data acquisition and integration		
Text recording	Does the text collection record or transcribe the textual data for the first time? (cf. Catalogue 3.1.6)	yes

Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	Full texts, Metadata
Quality assurance	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	yes
Typology	Considering aims and methods of the text collection, how would you classify it further? For definitions please consider the help-texts. (cf. Catalogue 3.1.8)	Diachronic corpus
Data Modelling		
Text treatment	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	Diplomatic transcription
Basic format	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	XML
Annotations		
Annotation type	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Structural information
Annotation integration	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Embedded
Metadata		
Metadata type	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	Descriptive, Administrative
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	Individual texts
Data schemas and standards		
Schemas	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	Customized standard schema
Standards	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	TEI, Dublin Core, CMDI
Provision		
Accessibility of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	yes

Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	yes
Technical interfaces	Are there technical interfaces which allow the reuse of the data of the text collection in other contexts? (cf. Catalogue 4.2)	OAI-PMH, General API
Analytical data	Besides the textual data, does the project provide analytical data (e.g. statistics) to download or harvest? (cf. Catalogue 4.3)	yes
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	yes
User Interface		
Interface provision	Does the text collection have a dedicated user interface designed for the collection at hand in which the texts of the collection are represented and/or in which the data is analyzable? (cf. Catalogue 5.1)	yes
User Interface questions		
Usability	From your point of view, is the interface of the text collection clearly arranged and easy to navigate so that the user can quickly identify the purpose, the content and the main access methods of the resource? (cf. Catalogue 5.3)	yes
Access modes		
Browsing	Does the project offer the possibility to browse the contents by simple browsing options or advanced structured access via indices (e.g. by author, year, genre)? (cf. Catalogue 5.4)	yes
Fulltext search	Does the project offer a fulltext search? (cf. Catalogue 5.4)	yes
Advanced search	Does the project offer an advanced search? (cf. Catalogue 5.4)	yes
Analysis		
Tools	Does the text collection integrate tools for analyses of the data? (cf. Catalogue 5.5)	yes

Customization	Can the user alter the interface in order to affect the outcomes of representation and analysis of the text collection (besides basic search functionalities), e.g. by applying his or her own queries or by choosing analysis parameters? (cf. Catalogue 5.5)	no
Visualization	Does the text collection provide particular visualizations of the data? (cf. Catalogue 5.6)	Charts
Personalization	Is there a personalisation mode that enables the users e.g. to create their own sub-collections of the existing text collection? (cf. Catalogue 5.7)	no
Preservation		
Documentation	Does the text collection provide sufficient documentation about the project in general as well as about the aims, contents and methods of the text collection? (cf. Catalogue 6.1)	yes
Open Access	Is the text collection Open Access? (cf. Catalogue 6.2)	yes
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	yes
License	Under what license are the contents released? (cf. Catalogue 6.2)	CC-BY-NC
Persistent identification and addressing	Are there persistent identifiers and an addressing system for the text collection and/or parts/objects of it and which mechanism is used to that end? (cf. Catalogue 6.3)	Persistent URLs
Citation	Does the text collection supply citation guidelines? (cf. Catalogue 6.3)	yes
Archiving of the data	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	no
Institutional curation	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	no

Completion	Is the text collection completed? (cf. Catalogue 6.4)	no
Personnel		
Designers	BBAW	