Marius Hug • Susanne Haaf

DAS DEUTSCHE TEXTARCHIV IM KONTEXT DER NFDI

1. Einleitung

Im Kontext der Nationalen Forschungsdateninfrastruktur (NFDI) bringt die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) mit dem Zentrum Sprache eine etablierte Infrastruktur in das Konsortium Text+ ein. Diese Infrastruktur umfasst einerseits Entwicklungen aus dem Akademienvorhaben »Digitales Wörterbuch der deutschen Sprache« (DWDS) und dem Projekt zum »Aufbau eines Zentrums für digitale Lexikographie der deutschen Sprache« (ZDL). Andererseits fungiert das »Deutsche Textarchiv« (DTA) als Archiv für Sammlungen und strukturierte, vor allem deutschsprachige Texte aus dem Zeitraum von etwa 1650 bis 1900. Dieser Beitrag möchte die Rolle des DTA sowie das Zusammenspiel mit dem DWDS und dem ZDL im NFDI-Konsortium Text+ beleuchten, und zwar bezugnehmend auf Fragestellungen der Interoperabilität und der Nachnutzung von Forschungsdaten.

2. Das Deutsche Textarchiv an der BBAW

Das DTA startete im Jahr 2007 als DFG-gefördertes Langfristvorhaben.[1] Ziel war es, ein möglichst nach Zeitschnitten, Fachbereichen und gedruckten Textsorten ausbalanciertes Korpus von digitalisierten deutschsprachigen Texten aus einem Zeitraum von etwa 1600 bis 1900 zu erstellen. Im Ergebnis entstand das sogenannte DTA-Kernkorpus als Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Es zeichnete sich durch eben diese ausbalancierte Textauswahl, die Verwendung von Erstausgaben bei der Digitalisierung, die Wahrung des Sprachstandes sowie durch eine einheitliche Abbildung von Textstrukturen aus.

1 Vgl. Wolfgang Klein: Deutsches Textarchiv (DTA) – Aufbau eines Aktiven Archivs deutscher Texte und Entwicklung entsprechender Werkzeuge, 2018 (https://gepris.dfg.de/gepris/projekt/37149321/, Zugriff: 31. Mai 2023).

DOI https://doi.org/10.15499/kds-005-019 | CC BY-SA 4.0

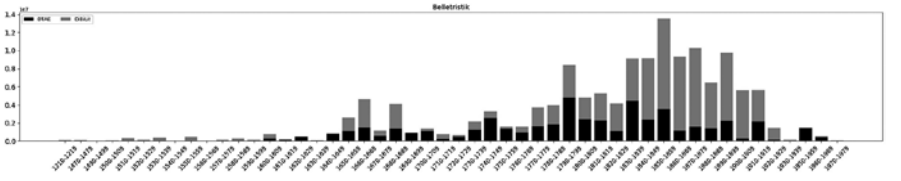

Abb. 1: Visualisierung des Korpusaufbaus durch CLARIAH-DE/DiBiLit, hier für die Textsorte Belletristik.

In zwei weiteren Förderphasen (2010-2016) wurde das Kernkorpus vor allem um Texte aus der Zeit von circa 1650 bis 1780 erweitert. Außerdem wurde das DTA als »Aktives Archiv«[2] etabliert und damit zunehmend zur Community hin geöffnet. Diese Entwicklung beinhaltete, dass im Rahmen eines eigenen Moduls für Erweiterungen zum Kernkorpus (DTA-Erweiterungen: DTAE) bereits digitalisierte, historische Quellen aus externen Projekten entsprechend den Richtlinien des DTA aufbereitet und in die Infrastruktur integriert wurden. Im Laufe der Jahre konnten so ungefähr 4.000 weitere Texte für die Text-Bild-Ansicht des DTA und die DTA-Korpora akquiriert werden. Auch diese Texte stammen schwerpunktmäßig aus Quellen des 17. bis 19. Jahrhunderts. Gerade in jüngerer Zeit wurden außerdem größere Korpora in den Bestand aufgenommen, die nicht mehr in der Text-Bild-Ansicht präsentiert werden, aber im Zusammenhang der historischen Korpora recherchierbar und in der Regel als Volltexte zugänglich sind.

Ein wichtiger Teil der im Rahmen von DTAE aufgenommenen Texte stammt aus umfangreichen Kurationsarbeiten aus dem CLARIN-Kontext.[3] Im Kontext des Projekts CLARIAH-DE – dem Zusammenschluss der beiden Forschungsdateninfrastrukturen CLARIN-D und DARIAH-DE – erfolgte zwischen 2019 und 2021 mit der Einbindung von Texten aus der Digitalen Bibliothek erneut eine beträchtliche Erweiterung des Datenbestandes (siehe Abb. 1).[4]

2 Vgl. Alexander Geyken u.a.: Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv, in: Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010. Beiträge der Tagung, hg. von Silke Schomburg u.a., 2., erg. Fassung, Hbz, 2011, S. 157-161.

3 Vgl. Alexander Geyken u.a.: Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN, in: Digitale Infrastrukturen für die germanistische Forschung (Germanistische Sprachwissenschaft um 2020 6), hg. von Henning Lobin, Roman Schneider und Andreas Witt, Berlin und Boston 2018, S. 219-248 (DOI: 10.1515/9783110538663-011, Zugriff: 6. Juli 2023).

4 Vgl. beispielsweise Marius Hug u.a.: Über bereichernde Anreicherung. Wechselsei-

Für den Aufbau der Sammlung wurde konsequent auf die XML-Kodierungsrichtlinien der Text-Encoding-Initiative (TEI P5) und damit auf den de-facto-Standard zur Kodierung geisteswissenschaftlicher Texte zurückgegriffen.[5] Da dieser seinem Anspruch nach für sämtliche Bedürfnisse bei der Textaufbereitung eine Lösung bieten möchte, sind die zugehörigen Guidelines sehr umfangreich und vielfältig. Das kann im konkreten Fall dazu führen, dass ein bestimmtes Phänomen verschiedenartig beschrieben werden kann. Ein einfaches Beispiel ist die Auszeichnung von Orten, wofür die TEI-Richtlinien verschiedene Möglichkeiten bereitstellen:

```
‹rs type="propNounPlaceName"›Marbach‹/rs›,
‹name type="place"›Weimar‹/name› und
‹placeName role="venue"›Wolfenbüttel‹/placeName›
sind Orte in Deutschland.
```

Derartige Variationsspielräume bei der Textauszeichnung schränkt das vom DTA entwickelte DTA-Basisformat (DTABf) ein.[6] Dies sorgt schließlich für eine größere Interoperabilität der entsprechend annotierten Texte:

```
‹placeName›Marbach‹/placeName›,
‹placeName›Weimar‹/placeName› und
‹placeName›Wolfenbüttel‹/placeName›
sind Orte in Deutschland.
```

Mit den DTA-Annotationsrichtlinien ist das DTABf ausführlich dokumentiert. Die DTABf-Steuerungsgruppe – zusammengesetzt aus Expert:innen für die TEI-Auszeichnung und -Anpassung mit Verankerung in verschiedenen Communities[7] – kümmert sich um Pflege und Weiterentwicklung des von der DFG und CLARIN-D zur Nachnutzung empfohlenen Formats.[8]

tige Annotation von Dramen als Subkorpus der Digitalen Bibliothek zwischen Zeno.org, TGRep, GerDraCor und DTA, in: Im Zentrum Sprache (blog), 12. Juni 2020 (https://sprache.hypotheses.org/2234, Zugriff: 31. Mai 2023).

5 TEI P5 Guidelines (https://tei-c.org/guidelines/p5/, Zugriff: 31. Mai 2023).

6 Susanne Haaf, Alexander Geyken und Frank Wiegand: The DTA ›Base Format‹. A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources, in: jTEI 8, 2014/15 (https://doi.org/10.4000/jtei.1114, Zugriff: 6. Juli 2023).

7 DTABf-Steuerungsgruppe (https://www.deutschestextarchiv.de/doku/basisformat/steuerungsgruppe.html, Zugriff: 31. Mai 2023).

8 Vgl. den CLARIN-D User Guide. Part II (Linguistic resources and tools), ch. 6

Neben der Interoperabilität verfolgte das DTA von Beginn an das Ziel, die Daten über persistente Identifikatoren leicht und zuverlässig auffindbar zu machen. Die Zugänglichkeit zu den Sammlungen ist gesichert, und die qualitativ herausragenden Standards im Bereich der Metadaten garantieren die Nachnutzbarkeit und maximale Wiederverwendbarkeit.[9]

Aktuell stellt das DTA rund 6.500 Volltexte und 370 Millionen Token bereit. Dieser Bestand an historischen Korpora wird derzeit für den Aufbau des Zentrums für digitale Lexikographie der deutschen Sprache (ZDL) fortlaufend erweitert. Außerdem dient das DTA im Cluster »Historische Korpora« des NFDI-Konsortiums Text+ als Archiv für historische, deutschsprachige Korpora, wie untenstehend noch erläutert wird. So kann auch weiterhin eine möglichst große Zugänglichkeit und Reichweite der Forschungsdaten gewährleistet werden.

3. Korpusanalyse historischer Texte am Zentrum Sprache

Das DTA arbeitet am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) seit jeher sehr eng mit dem Akademienvorhaben »Digitales Wörterbuch der deutschen Sprache« (DWDS) und seit 2019 mit dem ZDL zusammen. Das DWDS ist ein Wortinformationssystem, das für die Belegarbeit und die Beschreibung einzelner Wörter auf große Textkorpora zurückgreift. In diesem Kontext sind natürlich auch historische Korpora relevant, dann beispielsweise, wenn es darum geht, die historisch frühesten Verwendungen von Begriffen nachzuweisen. Das Teilprojekt »Wortgeschichte digital« des ZDL untersucht die Entwicklungen in der Wortschatzgeschichte von 1600 bis heute und greift dabei ebenfalls auf die im DTA bereitgestellten historischen Textsammlungen zu, die hierfür gezielt erweitert werden.

Jedes Korpus und jeder Bestand bringt auch seine Schwierigkeiten mit sich. Während zum Beispiel Urheberrechte für historische Quellen nicht in gleichem Maße eine Herausforderung darstellen, wie das für die immer rechtebewehrten, gegenwartssprachlichen Texte der Fall ist, können historische

(Types of resources), section »Text Corpora«, hg. von CLARIN-D AP 5, Berlin 2012 oder die Handreichung »Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora«, hg. vom Fachkollegium Sprachwissenschaften der Deutschen Forschungsgemeinschaft (DFG). Bonn 2015.

9 Das DTA stellte damit seit Beginn des Projekts seine Forschungsdaten gemäß der FAIR-Guidelines bereit, auch wenn es diese damals (in der heutigen Form) noch gar nicht gab.

Abb. 2: Korpusübergreifende Suche im Metakorpus Historische Texte; hier die Suche nach »Wetter« unter Einbindung des Thesaurus »Germanet«. Dabei ist die Suche beschränkt auf fünf Sammlungen sowie die Textgattung »Brief«.

und gegebenenfalls uneinheitliche Schreibweisen dagegen ein echtes Problem darstellen. Für die Bereitstellung der historischen Texte als möglichst breit nachnutzbare Forschungsdaten wurde daher im Kontext des DTA mit CAB (Cascaded Analysis Broker) eine eigene Software zur orthografischen Normierung und linguistischen Analyse entwickelt.[10] Diese Software wird mittlerweile über das DWDS bereitgestellt und im Rahmen von Text+ weiterentwickelt.

Die automatisierte Normierung der historischen deutschen Sprachdaten und die darauf aufbauende (computer-)linguistische Analyse auf Token-Ebene stellt ein bislang einzigartiges Angebot für das historische Deutsche dar, das auch bereits in anderen Kontexten außerhalb des DTA nachgenutzt wird[11] und dessen Ergebnisse zum Beispiel in die Entwicklung von Sprach-

10 Bryan Jurish: Finite-state Canonicalization Techniques for Historical German. PhD thesis, Universität Potsdam, 2012. URN: urn:nbn:de:kobv:517-opus-55789; CAB (https://deutschestextarchiv.de/public/cab/, Zugriff: 31. Mai 2023).

11 Vgl. Tobias Kraft und Stefan Dumont: »The Humboldt Code. On creating a hybrid

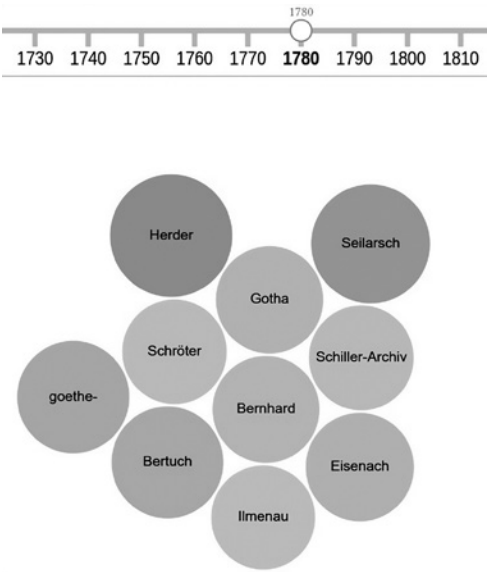

Abb. 3: Diachrone Untersuchung typischer Wortverbindungen mittels DiaCollo, hier die Suche nach Wortverbindungen zu »Weimar« um 1780.

analysesoftware für das Deutsche in der Python-Bibliothek spaCy eingeflossen sind.[12] Damit kann für die bereitgestellten Textressourcen beispielsweise eine schreibweisentolerante Korpussuche[13] inklusive Anbindung an Thesauri ermöglicht werden (siehe Abb. 2). Mit DiaCollo[14] und den Wortverlaufskurven stellt das Zentrum Sprache der BBAW zudem zwei wichtige Korpusana-

digital scholarly edition of a 19th century globetrotter«, in: Wiener Digitale Revue 1, 2020, S. 40 (https://doi.org/10.25365/wdr-01-03-02, Zugriff: 6. Juli 2023).

12 Tillmann Dönicke u.a.: MONAPipe. Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy, in: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022) (https://aclanthology.org/2022.konvens-1.2.pdf, Zugriff: 6. Juli 2023).

13 Eine sehr ausführliche Dokumentation der Korpussuche stellt das DWDS bereit (https://www.dwds.de/d/korpussuche, Zugriff: 31. Mai 2023).

14 Bryan Jurish, Alexander Geyken und Thomas Werneke: DiaCollo. diachronen Kollokationen auf der Spur, in: DHd 2016. Modellierung – Vernetzung – Visualisierung (Leipzig, 7.-12. März, 2016), hg. von Digital Humanities im deutschsprachigen Raum e.V., Leipzig 2016, S. 172-175.

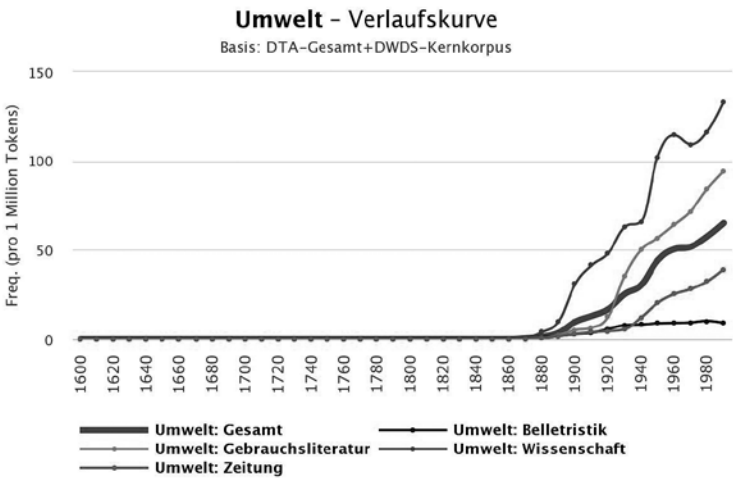

Abb. 4: DWDS-Wortverlaufskurven für die Korpusanalyse, hier die Suche nach »Umwelt« in den DTA- und DWDS-Korpora.

lysewerkzeuge sowohl für die diachronen Daten des DTA als auch für die synchronen Korpora des DWDS bereit (siehe Abb. 3 und 4). Dabei zeichnet DiaCollo Entwicklungen von Kollokationen (typischen Wortverbindungen) im zeitlichen Verlauf nach und unterstützt so Betrachtungen zu Bedeutungswandel und Begriffsgeschichte. Auch eine direkte Anbindung an die Voyant Tools, denen die einzelnen Texte in normierten und lemmatisierten Fassungen übergeben werden können, stellt eine wichtige Analyseoption für die DTA-Daten dar.[15]

4. Das DTA im Kontext des NFDI-Konsortiums Text+

Im Rahmen der Nationalen Forschungsdateninfrastruktur (NFDI) werden aktuell vier geisteswissenschaftliche Konsortien gefördert.[16] Im Verbund Text+[17] wird die langfristige Nachnutzbarkeit von text- und sprachbasierten

15 Für weiterführende Informationen zu den genannten Analysewerkzeugen und Szenarien ihrer Nutzung vgl. Geyken u.a., Das Deutsche Textarchiv (Anm. 3), S. 231-237.

16 Vgl. dazu: Die Konsortien der NFDI (https://www.nfdi.de/konsortien/, Zugriff: 31. Mai 2023).

17 Text+ (https://www.text-plus.org/, Zugriff: 31. Mai 2023).

Forschungsdaten ermöglicht. Text+ agiert als verteilte Infrastruktur mit den verschiedenen daran beteiligten Datenzentren und organisiert sich entlang der drei Datendomänen Editionen, lexikalische Ressourcen sowie text- und sprachbasierte Sammlungen.

Aufgrund der Herausforderung, die mit der Verfügbarmachung von sehr heterogenen Beständen einhergeht, ist die Datendomäne »Sammlungen« noch einmal in verschiedene Texttypen geclustert. So gibt es neben den beiden Clustern »Unstrukturierte Texte« und »Gegenwartssprachliche Texte« mit den »Historischen Texten« ein drittes Cluster, dem wiederum das DTA als Archiv für historische, deutschsprachige Texte dient. Dabei werden die etablierten Workflows des DTA konsequent weiterentwickelt, wobei neben der Transformation der Ausgangsdaten in das Zielformat TEI P5/DTABf der Bereich Metadaten eine zentrale Rolle spielt. Im Bereich der digitalen Volltexte stellt sich das DTA zunehmend auf den Umgang mit und die Kuration von mittels Optical Character Recognition (OCR) automatisiert erfassten Daten ein. Naturgemäß sind damit ganz neue Herausforderungen im Bereich der Qualitätssicherung verbunden, die aber zum Beispiel das Projekt OCR-D, das ebenfalls einen Standort an der BBAW hat, adressiert.[18]

Neu im Kontext der NFDI ist der Fokus des DTA auf die *Bereitstellung* von Sammlungen. Über eine menschen- und maschinenlesbare (schemaorientierte) Beschreibung können neue Sammlungen – wenn die darin enthaltenen Objekte den DTA-Standards entsprechen, mit Metadaten versehen sind und eine Dokumentation vorhanden ist – in die Infrastruktur integriert werden. Dadurch wird eine neue Sammlung zunächst einmal Teil der DTA-Korpusübersicht und enthält eine eigene Adresse. Alle Texte innerhalb der Sammlung sind dann so aufbereitet, dass sie als Forschungsdaten zur Nachnutzung bereitgestellt werden können. Ein erstes (*inhouse*) Nachnutzungsszenario besteht im Zentrum Sprache der BBAW immer in der linguistischen Aufbereitung der Texte und der Möglichkeit der Korpusanalyse im DWDS.

Da die Sammlungsbeschreibung andererseits als Grundlage für eine Integration in die Text+-Registry – das sich im Aufbau befindende zentrale Text+-System zum Nachweis der dezentral bereitgestellten Ressourcen – dienen soll, ist mit einer Aufnahme in die DTA-Infrastruktur auch die Grundlage

18 Clemens Neudecker u. a.: Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten, in: Qualität in der Inhaltserschließung, hg. von Michael Franke-Maier u. a., Berlin und Boston 2021, S. 137-166 (https://doi.org/10.1515/9783110691597-009, Zugriff: 6. Juli 2023). Zum Projekt OCR-D vgl. Elisabeth Engl u. a.: Volltexte für die Frühe Neuzeit, in: Zeitschrift für Historische Forschung 47.2, 2020, S. 223-250 (https://doi.org/10.3790/zhf.47.2.223, Zugriff: 6. Juli 2023).

für eine konsortienweite Sichtbarkeit der Sammlung im Kontext der NFDI geschaffen, wobei dann neben der großen Sichtbarkeit der Ressource natürlich wieder weitere Analysemöglichkeiten, beispielsweise eine übergreifende Suche, bereitgestellt werden.

Ein weiteres Ziel im Kontext der NDFI besteht darin, die Community für das Thema Forschungsdaten zu sensibilisieren und dafür – vor allem im Hinblick auf die aktuellen Entwicklungen im Bereich der künstlichen Intelligenz – eine gegebenenfalls erweiterte Lesart vorzuschlagen. Verschiedene Disziplinen haben teilweise ein sehr unterschiedliches Verständnis von Forschungsdaten: Während es in der naturwissenschaftlichen oder medizinischen Forschung eine lange Tradition von der Erstellung, Auswertung und Veröffentlichung von Forschungsdaten gibt, stellt das Wissen darum in den Geisteswissenschaften mitunter noch immer ein Desiderat dar.[19] So ist es auch 2023 noch kein Einzelfall, wenn im Rahmen einer sprachwissenschaftlichen Dissertation historische Quellen zwar transkribiert und annotiert, diese dann aber nicht für weitere Forschungszwecke bereitgestellt werden. Hier fungiert Text+ als geeigneter Ort, um diese Daten nach bestimmten Kriterien langfristig nachnutzbar zu machen. Die Idee ist, kleine wie große Bestände an historischen Daten des Deutschen als Forschungsdaten (das heißt in strukturierter Form mit sehr guter Erfassungsqualität) mit dem DTA an einem Ort zusammenzuführen, sodass daraus neue und individuelle Korpora von Wissenschaftler:innen ganz verschiedener Disziplinen gebildet werden können. Beispielsweise kann die Korpuslingustik von dieser Zusammenführung und Bereitstellung aggregierter Ressourcen stark profitieren. Denn die Aufbereitung historischer Daten des Deutschen ist noch immer nicht trivial und ein arbeitsintensiver Prozess. Umso mehr kann die wissenschaftliche Community hier von einer Praxis des Teilens profitieren.

Schließlich stellen die im DTA bereitgestellten, hochwertigen Transkriptionen historischer Texte ein wachsendes Korpus an potenziellen Trainingsdaten für Machine-Learning-Algorithmen dar. Letztere sind in neuester Zeit so gut geworden (wie uns Anwendungen wie ChatGPT und DeepL zeigen), dass hier auch für die Entwicklung der automatischen Texterfassung für historische Texte des Deutschen ein echtes Potenzial gesehen werden kann – das DTA kann somit auch im Bereich der KI eine wichtige Grundlage liefern.

19 Zu den disziplinären Unterschieden in Bezug auf Forschungsdaten siehe beispielsweise Susanne Blumesberger: Forschungsdaten in den Geisteswissenschaften. Bereits selbstverständlich oder doch noch etwas exotisch?, in: o-bib. Das offene Bibliotheksjournal 8.4, 2021, S. 1-8; hier S. 2 (https://doi.org/10.5282/o-bib/5739, Zugriff: 6. Juli 2023).

Nicht zuletzt helfen für die hier konturierten Nachnutzungsszenarien die freien Lizenzen, unter denen die Daten des DTA konsequent bereitgestellt werden und die deren Weiterverwendung in den unterschiedlichsten Kontexten erst ermöglichen.

5. Zusammenfassung

Das DTA verwendet Standards, die sowohl menschen- als auch maschinenlesbar sind. Neben der gezielten Kuration der eigentlichen Textdaten wird ein besonderes Augenmerk auf die Objekt- und Sammlungs-Metadaten gelegt, die die Voraussetzung für eine möglichst breite Nachnutzung der Ressource darstellen. Nachnutzbar sind aber nicht nur die im DTA veröffentlichten Daten, sondern natürlich das für die Annotation verwendete und bereitgestellte und für bessere Interoperabilität sorgende Format DTABf sowie die im Zentrum Sprache bereitgestellten Tools.[20]

Im Kontext von Text+ fungiert das DTA als Archiv für historische, deutschsprachige Texte. Projekte, die hochwertige Transkriptionen anfertigen und dabei ein weiterverarbeitbares Textformat verwenden, Metadaten – sowohl auf Objekt- wie auch auf Sammlungsebene – bereitstellen, eventuelle Lizenzfragen geklärt haben und im besten Fall eine detaillierte Dokumentation liefern können, finden im DTA eine etablierte Infrastruktur zur nachhaltigen Bereitstellung ihrer Textdaten. Das DTA berät zu allen Belangen, angefangen von Verfahren zur Transkription über die Annotation bis hin zur Dissemination der Forschungsdaten.

Durch das Zusammenspiel von Deutschem Textarchiv (DTA) und Digitalem Wörterbuch der deutschen Sprache (DWDS) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) kann für integrierte Ressourcen eine umfängliche Korpusanalyse-Infrastruktur bereitgestellt werden. Neben der linguistischen Aufbereitung als Voraussetzung für die Korpussuche und die Verwendung weiterer Analysewerkzeuge stellt besonders auch die Integration in die erweiterte Infrastruktur von Metakorpora einen wichtigen Mehrwert dar. Dadurch wird eine korpusübergreifende Suche ermöglicht. Nicht zuletzt können die Forschungsdaten über die Schnittstellen des DWDS zur Nachnutzung bereitgestellt werden. Das stellt einerseits eine Voraussetzung zur Integration der Sammlung in das zentrale Nachweissystem von Text+ und die NFDI-Infrastruktur dar, wodurch die

20 An dieser Stelle sei exemplarisch verwiesen auf DiaCollo für GEI-Digital (https://diacollo.gei.de/, Zugriff: 31. Mai 2023).

Sichtbarkeit der Ressource noch einmal deutlich vergrößert wird. Andererseits ist damit die Voraussetzung geschaffen, dass aus Forschungsdaten Trainingsdaten werden können.