CLARIAH-DE-Arbeitsbericht 6

Integration von Textdaten aus der Community in bestehende Infrastrukturen

Matthias Boenig, Marius Hug und Simon Sendler

CLARIAH-DE 2021

DOI: 10.14618/ids-pub-10497

Dieser Bericht steht unter der Creative-Commons-Lizenz
Namensnennung (CC BY 4.0),
vgl. *https://creativecommons.org/licenses/by/4.0*.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

FÖRDERKENNZEICHEN
01UG1910 A bis I

1 Einleitung

Forschungsprojekte erschließen, erfassen und publizieren eine große Menge digitaler Daten. Bis zur Publikation entstehen häufig Vorarbeiten oder auch Nebenprodukte des beabsichtigten Ergebnisses (beispielsweise Transkriptionen einzelner Texte oder Textzeugen, die die Grundlage z.B. für eine Edition bilden).

CLARIAH-DE[1] bietet verschiedene Möglichkeiten zur Integration von Angeboten und Inhalten aus der Community, die deren längerfristige Sicht- und Nachnutzbarkeit sicherstellt.

1. Das Erweiterungsmodul des Deutschen Textarchivs (DTAE)[2] ermöglicht es Wissenschaftlerinnen und Wissenschaftlern, einschlägige Volltexte, die im Rahmen ihrer Forschungsarbeiten digital aufbereitet wurden, nachhaltig in die DTA[3]-Infrastruktur an der Berlin-Brandenburgischen Akademie der Wissenschaften zu integrieren.
2. Mit dem TextGrid Repository[4] steht ein Langzeitarchiv für geisteswissenschaftliche Forschungsdaten zur Verfügung, das den Forschenden eine nachhaltige Publikation von zitierfähigen Forschungsdaten und Metadaten sichert.
3. Das DARIAH-DE Repository[5] ermöglicht die selbständige dauerhafte und referenzierbare Speicherung von Forschungsdaten mithilfe standardisierter deskriptiver Metadaten. Es integriert zudem die Generische Suche zum Auffinden der Forschungsdaten.

Die vorliegende Handreichung befasst sich mit den Fragen, welche Textangebote wo und auf welche Weise archiviert werden können, sowie welche Kriterien verschiedene Arten von Daten erfüllen müssen, um grundsätzlich für eine Übernahme in den CLARIAH-DE-, Forschungsdatenmanagement- oder NFDI-Kontext geeignet zu sein.

Weitere Infrastrukturen zur Integration einer Vielzahl von Datenressourcen wie bspw. multimodale oder multilinguale Daten und Objekte werden von den jeweiligen Partnern aus den Ursprungsprojekten[6] weiter betrieben und sind nicht Bestandteil dieser Handreichung.

2 Grundsätzliches und Standards

Um für eine Integration in bestehende Infrastrukturen in Frage zu kommen, muss ein Datenangebot grundlegende Kriterien erfüllen:

- Es muss allgemeinen wissenschaftlichen Standards entsprechen,
- mit möglichst geringer rechtlicher Einschränkung nutzbar sein,[7]
- und grundlegend dokumentiert sein.

[1] *https://www.clariah.de/*
[2] *http://www.deutschestextarchiv.de/dtae*
[3] *https://www.deutschestextarchiv.de/*
[4] *https://textgridrep.org/*
[5] *https://de.dariah.eu/repository*
[6] Bspw. das „Bayerisches Archiv für Sprachsignale“ (2021) der LMU München (CoreTrustSeal[CTS]-zertifiziert), das „Hamburger Zentrum für Sprachkorpora“ (2021) der Universität Hamburg (CTS-zertifiziert) oder das „Forschungsdaten-Repositorium“ (2021) des Leibniz-Institut für Deutsche Sprache (IDS) Mannheim (CTS-zertifiziert). Zum CoreTrustSeal vgl. die „CoreTrustSeal Homepage“ (2021).
[7] Im Falle, dass Daten einer Nutzungsbeschränkung unterliegen, die aufgrund von Persönlichkeits- und Datenschutzrechten, Urheberrecht, Lizenzvereinbarungen, z.B. mit Verlagen bestehen, müssen zumindest Metadaten dauerhaft öffentlich zur Verfügung stehen. In diesem Falle soll über die Anbindung an eine Authentication and Authorization Infrastructure (AAI) der Zugang für eine möglichst große Gruppe innerhalb der wissenschaftlichen Community sichergestellt werden. Eine entsprechende Lizenz, z. B. auf Basis der Creative Commons („Creative Commons Homepage“ (2021)), muss vergeben werden, um die Nutzung des Daten-Angebotes zu regeln.

Die CLARIAH-DE-Partner betreiben verschiedene Repositorien[8], in denen Forschungsdaten entsprechend den FAIR-Prinzipien veröffentlicht und nachhaltig archiviert werden können. Diese zeichnen sich jeweils durch unterschiedliche fachliche Schwerpunkte bzw. die Spezialisierung auf bestimmte Datentypen aus. Entscheidend für die Auswahl des geeigneten Repositoriums für die eigenen Forschungsdaten kann weiterhin der jeweils gebotene Funktionsumfang sein. Die genauen Workflows zur Integration von Daten sowie die Qualitätsstandards, die diese erfüllen müssen, unterscheiden sich zwischen den einzelnen Repositorien. Aus diesem Grund wird hier zunächst ein allgemeiner Ablauf der Integration dargestellt.

Der im Folgenden beschriebene CLARIAH-DE *Data Integration Cycle* orientiert sich an gängigen FDM-Workflows[9]. Nach diesem durchläuft ein Datenpaket folgende Stadien:

1. Pre-Processing
2. Processing
3. Quality Assurance
4. Publishing
5. Storage

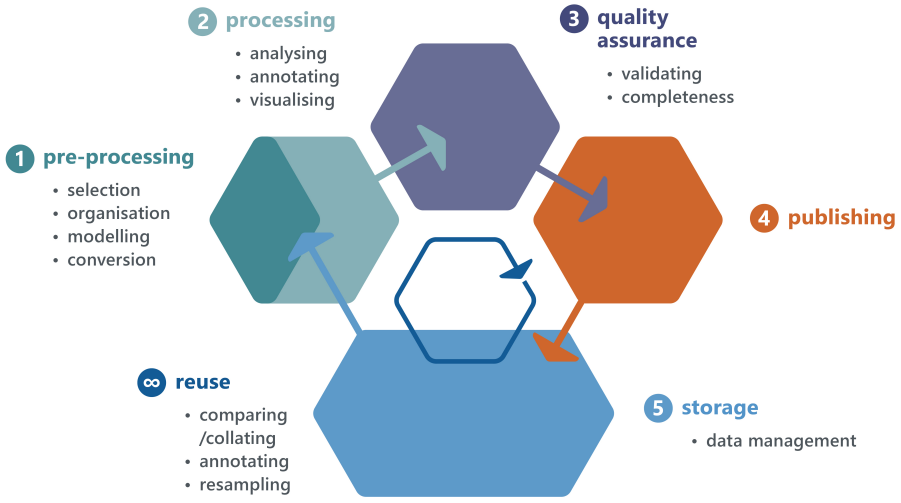


Abbildung 1: Der CLARIAH-DE *Data Integration Cycle*

Das **Pre-Processing** (1) steht am Anfang eines Projektes und beinhaltet die Entwicklung von spezifischen Datenmanagementplänen sowie die Modellierung der Daten von der Meta- bis zur Mikroebene. In der Phase der Integration in eine Forschungsdateninfrastruktur erfolgt ebenfalls eine Phase des preprocessing. Hierbei sind die am Anfang erarbeiteten Pläne und die Datenmodellierung zu überprüfen und zu ergänzen sowie u. a. eine Transformation der Daten in ein einheitliches Format vorzunehmen. Für die Integration sind folgende Informationen notwendig:

- **Kontaktinformationen** und institutionelle Zugehörigkeit (Titel, Partner, Koordinator und CLARIAH-DE-Kontaktperson).
- Die **Beschreibung des Datenpakets** enthält eine Darstellung des Umfanges, des Inhalts, möglichst detaillierte technische Informationen sowie eine Nutzungslizenz.
- Wünschenswert ist auch die Vorlage von Anleitungen zum Gebrauch der Daten. Bezüglich der weiteren Aufbereitung und Kuration sind potenzielle Abhängigkeiten zu anderen Daten zu bedenken, wie auch Überlegungen zu Nachnutzbarkeit und Nachhaltigkeit.

[8] Einen Überblick über verschiedene Repositorien zur Archivierung von Forschungsdaten bietet die vom Projekt CLARIAH-DE gelistete Übersicht: *https://clariah.de/publizieren-archivieren*.
[9] Zum Beispiel: Wissik & Ďurčo (2015) oder auch der „DARIAH-DE Research Data Life Cycle“ (2021).

Für das Stadium des **Processing** (2) erfolgt eine Aufbereitung der Daten, wobei diese bspw. eine Analyse und Annotation umfassen kann. Damit wird der Zugriff bzw. die Recherche nach dem gesamten Paket oder Teilen innerhalb des Forschungsdatenmanagementsystems möglich. Eine Visualisierung der Daten – entweder als spezifischer Stylesheet oder der auch in Form eines gängigen Präsentationsformates – kann ebenfalls Bestandteil des processing sein.

Bei der **Quality Assurance** (3) wird die Qualität der Daten anhand verschiedener Kriterien ermittelt bzw. nachgewiesen:

- Das Datenformat entspricht wissenschaftlichen Standards,
- die Annotation erfolgt kohärent nach entsprechendem Standard,
- eine evtl. Nähe der Auszeichnung zum gemeinsamen Standard für die Weiterverarbeitung wird abhängig vom „Zielrepositorium“ geprüft.

Im Stadium des **Publishing (4)** sollten den Daten persistente Identifier zugewiesen werden. Über eine entsprechende Lizenz werden sie den Nutzenden möglichst Open Access verfügbar gemacht. Das **Storage (5)** kann als ein Endpunkt angesehen werden. Nachdem die Daten erschlossen, aufbereitet und qualitätsgeprüft sind, werden sie im gewählten Zielrepositorium gespeichert.

Im Optimalfall endet der *Data Integration Cycle* nicht mit der Speicherung und Archivierung der Daten, stattdessen erlaubt die Einhaltung entsprechender Standards und die Verpflichtung zu den FAIR-Prinzipien die langfristige Nachnutzbarkeit der Ressourcen, d.h. einen **Reuse**.

Voraussetzung für den hier beschriebenen Integrationsprozess ist die Verwendung von Standards, die von den jeweiligen Anbietern von Diensten und Repositorien vorgegeben werden. Grundlegende Informationen zu den verschiedenen Datentypen finden sich hier:

Textdaten:

- TEI-XML: Der Quasi-Standard für digitale Texte, aktuell in ihrer fünften Ausgabe (kurz: TEI-P5)[10]
- Standards für spezifische Anwendungsfälle, z.B. das TEI-Subset DTA-Basisformat (DTABf)[11]
- ISO 24624:2016 für Transkripte gesprochener Sprache: Language resource management – Transcription of spoken language[12]
- Empfehlungen des DFG-Fachkollegiums 104 „Sprachwissenschaften“[13]

Audioformate, **Videoformate**; **Bildformate**:

- DFG-Praxisregeln „Digitalisierung“[14]
- IIIF für Bildressourcen im TGRep[15]

3 Repositorien zur Aufnahme von Daten

Für die Aufnahme v. a. deutschsprachiger Textdaten als Forschungsdaten stehen aus den beiden Infrastrukturprojekten CLARIN-D und DARIAH-DE verschiedene Repositorien zur Verfügung. Diese Handreichung stellt einen Workflow zur Übernahme von Daten exemplarisch anhand zweier für die Langzeitspeicherung für Texte im TEI-XML-Format etablierter Infrastrukturen vor. Neben dem TextGrid Repository ist das v.a. das Erweiterungsmodul des Deutschen Textarchivs, das DTAE.[16]

[10] Burnard & Bauman (2020)
[11] Deutsches Textarchiv (2020a)
[12] ISO/TC 37/SC 4 Language resource management (2016)
[13] DFG-Fachkollegium Sprachwissenschaften (2015)
[14] Deutsche Forschungsgemeinschaft (2016)
[15] DARIAH-DE (2019a)
[16] Deutsches Textarchiv (2021a)

Das Deutsche Textarchiv (DTA) als ein Schwerpunkt des CoreTrustSeal-zertifizierten CLARIN-Servicezentrums[17] des Zentrums Sprache an der BBAW bietet mit seinem Erweiterungsmodul DTAE die Möglichkeit, historische **deutschsprachige Texte und lexikalische Ressourcen** aus der Community dem Korpus des DTA beizustellen. Während das DTA mit seinem Kernkorpus – das in der Regel aus Transkriptionen historischer Erstausgaben besteht – einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte mit Schwerpunkt frühes 16. bis frühes 20. Jahrhundert zur Verfügung stellt, wird durch diese angelagerten Textsammlungen die Vielfalt und der Umfang der Ressource erhöht. Bei der Integration der Daten in die Infrastruktur wird eine **automatische linguistische Analyse** vorgenommen. Diese Analyse ist sowohl Voraussetzung für den dadurch möglich gewordenen Einsatz der DDC-Suchmaschine, also auch einen (korpusübergreifenden) Textvergleich. Die Texte im DTAE sind DTABf-valide[18] kodiert.

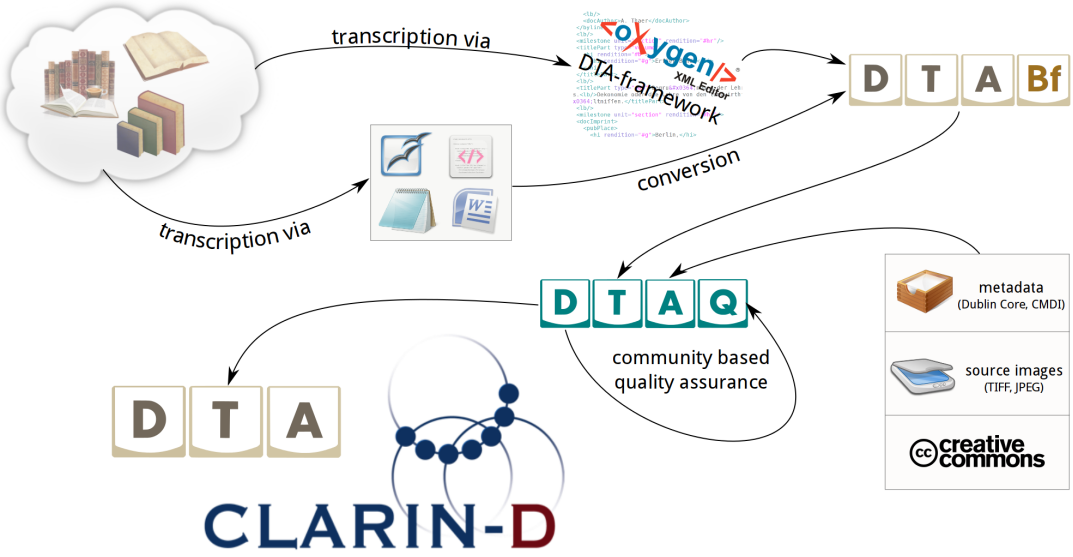


Abbildung 2: DTAE-Workflow.

Eine inhaltliche Voraussetzung für die Aufnahme von Texten in das DTAE ist ein Entstehungszeitraum zwischen dem frühen 16. und dem frühen 20. Jahrhundert. Es wird empfohlen, den Transkriptionen die digitalen Faksimiles beizulegen. Wenn es sich zudem um Schlüsseltexte innerhalb wichtiger Diskurse handelt, sind die Digitalisate eine interessante Ressource für eine Integration in das DTAE.

Eine technische Voraussetzung besteht in der Verfügbarkeit der den Textdigitalisaten zugrunde liegenden Bildquellen in hoher Bildqualität (unkomprimiertes TIFF) und der lizenzrechtlichen Möglichkeit, diese auch weiter verwenden zu können. Die Annotation der Volltexte sollte im DTA-Basisformat erfolgt sein oder alternativ vergleichsweise leicht in das DTABf konvertierbar sein. Das DTAE bietet für die Texttranskription ausführlich beschriebene Richtlinien.[19] Die mit vielen Beispielen veranschaulichte Dokumentation des DTABf[20] zeigt, wie das XML-Markup verwendet wird.

Neben der Erfassung bibliographischer Metadaten sind auch spezifische Daten zum digitalen Faksimile sowie zur Texterfassung notwendig. Für die Erstellung von basisformatkonformen Metadaten und die Generierung eines entsprechenden TEI-Headers stellt das DTA ein spezielles Web-Formular zur Verfügung.[21] Über die ebenfalls im DTA-Kontext etablierte Umgebung DTAQ[22] können die in die In-

[17] Berlin-Brandenburgische Akademie der Wissenschaften (2021)
[18] Zum Begriff „DTABf-valide“ vgl. Fisseni et al. (2021).
[19] Deutsches Textarchiv (2020b)
[20] Deutsches Textarchiv (2020a)
[21] Deutsches Textarchiv (2021b)
[22] Deutsches Textarchiv (2021c)

frastruktur integrierten Texte bei Bedarf für die wissenschaftliche Nutzergemeinschaft zur Korrektur freigegeben werden.

Die Integration von Textressourcen in das DTA ermöglicht:

- eine Text-Bild- sowie eine Leseansicht, die per Einbettung eines Inlineframes (`<iframe>`) auf einer eigenen Webseite erfolgen kann,
- eine ausführliche Präsentation aller Metadaten,
- die schreibweisentolerante Texterschließung über die DTA-Suchmaschine,
- einen eigenen Suchindex für die Textkollektion,
- die Textanalyse mittels integrierter Tools wie Wortverlaufskurven oder DiaCollo,
- die Anbindung an externe Analysetools wie bspw. die Voyant Tools[23] oder das Language Resource Switchboard (LRS)[24],
- den Zugriff auf standardisierte Schnittstellen zum Austausch der Metadaten,
- den Download verschiedener Datenformate[25],
- die Konvertierung in etablierte CLARIN-D-Formate (CMDI, TCF),
- die Integration in die CLARIN-D-Infrastruktur (Recherchierbarkeit über das Virtual Language Observatory und die Federated Content Search), wodurch die Texte zusätzlich mit einer ePIC-PID[26] versehen werden.

Wenn für die Datenbestände keine Bilddateien vorliegen, besteht die Möglichkeit der Integration eines reinen Textkorpus in die sogenannte dstar-Infrastruktur.[27] Auch diese Korpora werden automatisch indiziert und können anschließend mit der DTA-Suchmaschine (DDC) erschlossen werden. Der Zugriff auf die Metadaten im TEI-Header der Texte ist ebenfalls möglich. Es besteht außerdem die Möglichkeit einer korpusübergreifenden Suche.

4.1 TextGrid Repository

Das TextGrid Repository[28] bietet sich als Langzeitarchiv für geisteswissenschaftliche Forschungsdaten für eine Vielzahl von textuellen Forschungsdaten insbesondere im TEI-XML-Format an. An den Grundsätzen von Open Access und den FAIR-Prinzipien orientiert, ist das TextGrid Repository seit 2020 CoreTrustSeal-zertifiziert. Hierbei ist das Repository nicht auf Texte aus einer bestimmten historischen Epoche, einer bestimmten Sprache oder Textgattung festgelegt. Entsprechend liefert es einen umfangreichen, durchsuch- und nachnutzbaren Bestand an Texten und Bildern. Auch bietet das TextGrid Repository die Möglichkeit, auf Bildressourcen IIIF-konform zuzugreifen. Die zitierfähigen Forschungsdaten können von der Community nachhaltig, dauerhaft und selbstständig veröffentlicht werden.

Das TextGrid Repository ist Teil der virtuellen Forschungsumgebung TextGrid[29], die mit dem TextGrid Laboratory eine Open- Source-Softwarelösung für das kollaborative Bearbeiten, Analysieren und Publizieren von Text- und Bilddaten bietet. Während die Workflows von TextGrid für die Erstellung von digitalen Editionen im TEI-XML-Format optimiert sind, ist auch eine von diesem Format unabhängige Publikation sowohl über das Laboratory als auch über die TextGridRep-API[30] möglich. Die im TextGrid

[23] Sinclair, Stéfan and Rockwell, Geoffrey (2016)
[24] CLARIN (2021)
[25] TEI-XML: *https://www.deutschestextarchiv.de/book/download_xml/alexis_ruhe03_1852*,
HTML: *https://www.deutschestextarchiv.de/book/download_html/alexis_ruhe03_1852*,
Text: *https://www.deutschestextarchiv.de/book/download_txt/alexis_ruhe03_1852*
[26] ePIC Persistent Identifier Consortium for eResearch (2016)
[27] Berlin-Brandenburgische Akademie der Wissenschaften (2016)
[28] *https://textgridrep.org/*
[29] DARIAH-DE (2019b)
[30] DARIAH-DE (2019c)

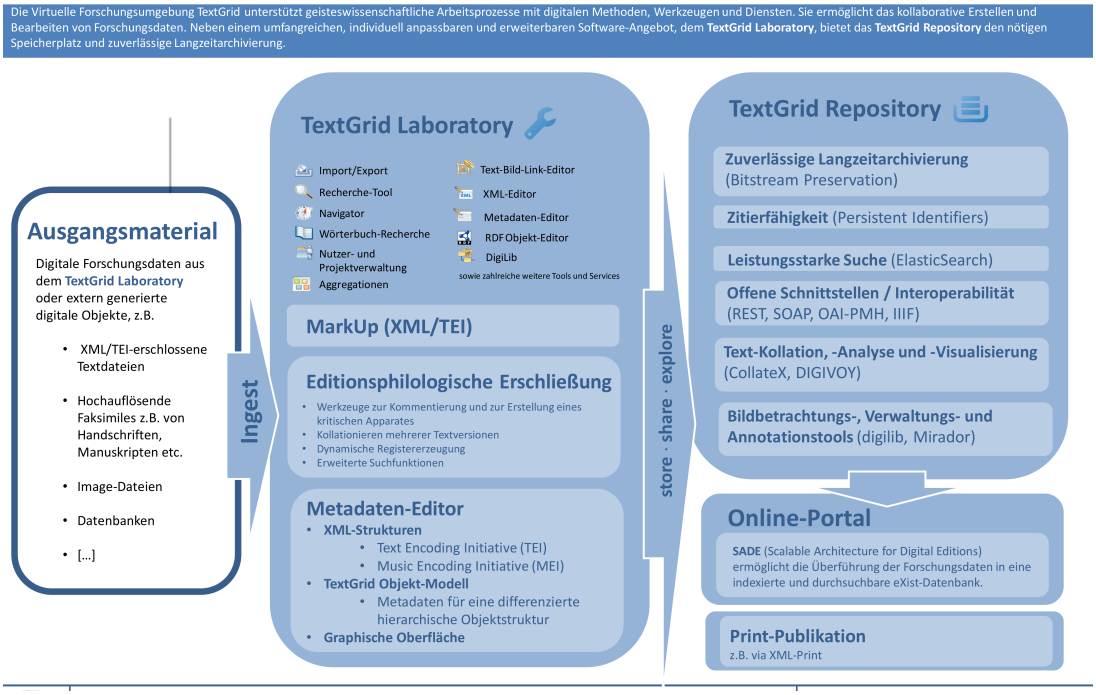


Abbildung 3: TextGrid, die virtuelle Forschungsumgebung

Repository veröffentlichten Bilddateien können mittels TG-digilib[31] angezeigt werden. Digilib unterstützt die digilib-Scaler-API[32] sowie die IIIF-Image-API[33]. Für Projekte, deren Bild- und Textdateien die entsprechenden Voraussetzungen erfüllen, ist eine Freischaltung weiterer IIIF-Funktionalitäten möglich.[34]

Im TextGrid Repository publizierte Objekte erhalten ebenfalls ePIC-PIDs, um eine nachhaltige Referenzierbarkeit sicherzustellen. Objekte im Repository können vielseitig nachgenutzt werden. Über die Bereitstellung der ursprünglichen Bild- und Textdateien hinaus können Textdateien in anderen Formaten, wie beispielsweise *plain text* heruntergeladen werden. Die Anbindung der Voyant Tools[35] und des Language Resource Switchboard[36] ermöglicht einfache explorative und analytische Zugänge zu Objekten im Repository. Nutzerinnen und Nutzer haben darüber hinaus die Möglichkeit, über das Regal[37] eigene Korpora zusammenzustellen und diese ähnlich wie einzelne Dateien weiter zu nutzen.

5 Verschiedenen Datentypen

Die An- und Herausforderungen für die Integration von Daten in bestehende Infrastrukturen können sehr unterschiedlich sein. Im Folgenden wird der Fokus auf drei konkrete Datentypen gelegt: Textsammlungen, lexikalische Ressourcen und Editionen. Da die Einspielung von Daten in das Deutsche Textarchiv im Vergleich zum TextGrid Repository voraussetzungsreicher ist, legt die folgende Darstellung einen entsprechenden Fokus. Die Ausführungen sind als Fallbeispiele zu verstehen.

[31] DARIAH-DE (2019d)
[32] digilib Community (2021)
[33] Appleby, Crane, Sanderson, Stroop & Warner (2021)
[34] DARIAH-DE (2019a)
[35] DARIAH-DE (2019e)
[36] CLARIN (2021), vgl. DARIAH-DE (2019f)
[37] DARIAH-DE (2019g)

5.1 Textsammlungen

Die in einem Projekt zur Digitalisierung von Texten entstandenen Sammlungen können als Ausgangsmaterial zur Integration in ein Repository von sehr unterschiedlicher Form und Qualität sein. Das soll an zwei Beispielen verdeutlicht werden.

5.1.1 Beispiel 1: XML/TEI

Mit der Digitalen Bibliothek wurde im Projekt CLARIAH-DE eine sehr umfangreiche Textsammlung im Hinblick auf die Konversion in das Basisformat des Deutschen Textarchivs (DTABf) evaluiert. Datengrundlage der Konversion[38] waren verschiedene XML-Derivate der ursprünglich durch die Editura GmbH & Co. KG digitalisierten Sammlung gemeinfreier Texte. Die Originaldaten, die als Basis der Veröffentlichung auf www.zeno.org dienten, entsprechen autorbasierten XML-Dateien und einem zugehörigen externen Metadatensatz. Der TextGrid-Verbund hat diese Daten 2009 erworben und große Teile davon (den Literatur-Ordner, 135 Mio Token) nach TEI-P5 konvertiert. Die ursprünglich extern gespeicherten bibliographischen Metadaten wurden dabei, wenn möglich, den einzelnen Texten zugeordnet und in den TEI-Header integriert.

Bei derart großen Datenmengen sind Inkonsistenzen in den Ausgangsdaten natürlich nicht auszuschließen. Standardmäßig werden also schematische und vergleichsweise unspezifische Heuristiken angewandt, die im Einzelfall zu Fehlinterpretationen führen können. Darauf hat TextGrid bereits hingewiesen. Wichtig in diesem Fall ist die Einbindung von Evaluierungsroutinen in den Workflow aus *Pre-Processing*, *Processing*, *Quality Assurance*, *Publishing* und *Storage*.

Grundsätzlich ist das TEI-P5-konforme TextGrid Baseline Encoding natürlich näher am Zielformat des DTABf, bei dem es sich ja bekanntlich um ein TEI-Subset handelt. Die Grundstruktur der beiden Formate ist identisch. Eine einfache und zielführende Herangehensweise zur Evaluierung der Ausgangsdaten besteht darin, die TEI-Daten des TextGrid Repositories an Hand des DTABf-Schemas zu validieren. Mit diesem Verfahren können Unterschiede und Abweichungen zum DTABf-Schema ermittelt werden. In der Interpretation dieses Prüfverfahrens werden Ambiguitäten oder Tag-Abuse sichtbar. Diese Probleme lassen sich jedoch kaum automatisiert beheben, sondern stellen eine typische Aufgabe für erfahrenes wissenschaftliches Personal dar. Was sich dagegen wieder vergleichsweise einfacher evaluieren lässt, sind die Metadaten im TEI-Header. Interessant sind dabei v.a. die werkspezifischen bibliographischen Metadaten in der sogenannten `<sourceDesc>`. Bei einer derart umfangreichen Sammlung wie der Digitalen Bibliothek wurden dafür kleine Skripte geschrieben, die die Daten der Einzeldateien überprüfen und das Ergebnis als kommaseparierte Liste (CSV) abspeichern. Die CSV-Datei lässt sich anschließend mit jeder üblichen Tabellenkalkulationssoftware öffnen und verschiedene Sortierroutinen lenken schnell den Blink auf eventuelle Lücken oder Ausreißer. In der Praxis hat sich im übrigen gezeigt, dass genau darin sowie in der Aufsplittung der Ausgangsdateien in die kleinstmögliche Texteinheit – das kann ein Gedicht oder gar eine Widmung sein – eine sehr große Herausforderung für die werksspezifische Transformation der im TextGrid Repository enthaltenen Texte der „Digitalen Bibliothek“ nach DTABf bestand.

Etwas anders stellt es sich mit den Originaldaten der Digitalen Bibliothek dar. Diese Daten entsprechen einer XML-Datei pro Autor sowie einem bibliographischen Metadatenkatalog, der die Quellenangaben enthält. Hier muss das einzelne Werk aus einem Werkverbund extrahiert werden. Die entsprechenden Metadaten sind zuzuordnen und das extrahierte Werk abschließend nach TEI bzw. DTABf zu transformieren. Aufgrund der großen Textmenge ist ein möglichst *automatisierter* Workflow zu entwickeln, der im Idealfall auch nachgenutzt werden kann. Ein *generisches* Vorgehen jedoch, ist bei derartigen Textsammlungen mit hohen Diversität von Textgattungen und der daraus resultierenden hohen Komplexität der Textkodierung in der Regel nicht möglich. Mit der Trennung von Daten und Metadaten

[38] DARIAH-DE (2019h)

kommt ein Workflow zum Einsatz, der in mehreren voneinander getrennten Schritten die werkspezifische Zusammenführung der Texte, die Generierung der TEI-Header sowie die Konvertierung der eigentlichen Texte umfasst.

Dabei kann es an mehreren Stellen notwendig sein, manuell in den Workflow einzugreifen. Zwar wurden die Ausgangsdaten kuratiert, d.h. bekannte Inkonsistenzen wurden dort direkt korrigiert, andere evtl. auftauchende Probleme werden jedoch erst im Laufe des Workflows sichtbar, weswegen manchmal ein Eingriff von außen unumgänglich ist.

Erst zu einem vergleichsweise späten Zeitpunkt werden die bis dahin getrennten Metadaten im TEI-Header und die eigentlichen Texte im `<text>`-Element in einer dann DTABf-konformen Datei zusammengeführt. Das Ergebnis der Kuration und Konversion der Daten ist ein eigenes dstar-Korpus.[39]

5.1.2 Beispiel 2: Microfilm/MS-Word

Im Vergleich zur Digitalen Bibliothek könnte das Ausgangsmaterial zur Integration in die DTA-Infrastruktur beim sogenannten „Lefèvre-Korpus“ kaum unterschiedlicher sein: Es handelt sich dabei um elektronische Transkripte im Word-Format (doc), die auf Basis von Mikrofilmen erstellt und dem DTA von Prof. Dr. Michel Lefèvre (Université Paul Valéry Montpellier 3) zur Verfügung gestellt wurden. Das Ziel war in diesem Fall ebenfalls die Konversion der Daten in das DTABf und die Integration der Ressource in die DTA-Infrastruktur. Da im Gegensatz zur Digitalen Bibliothek neben den Textquellen auch ein Zugriff auf die zugehörigen Bilddaten (Faksimiles) möglich war, war die Voraussetzung für eine Veröffentlichung der Ergebnisse im DTA als Text-Bild-Ansicht gegeben. Die Extraktion der Bilddaten aus externen Quellen und Formaten (PDF-Dateien) war Teil des unten beschriebenen Workflows.

Das Korpus umfasst ausgewählte deutschsprachige Zeitungen aus dem letzten Drittel des 17. Jahrhunderts, die in der Regel periodisch ein- oder zweimal wöchentlich erschienen. Die Ausgaben umfassten meist vier oder acht Seiten und setzten sich im Wesentlichen aus Korrespondenzen aus europäischen Städten zusammen.

Der *makeDTA* genannte Workflow besteht aus einer Sammlung von XSLT-2.0-Stylesheets und verschiedenen Makefiles. Die Transkripte als doc-Dateien werden zunächst ins docx-Format konvertiert, mit Hilfe des Tools `docxtotei` der Text Encoding Initiative nach TEI transformiert und schließlich mit den erwähnten XSLT-Stylesheets in das Basisformat des DTA konvertiert. *makeDTA* ist ein flexibler und modularer Workflow, in welchem quellenspezifische Anpassungen vorgesehen sind. Ein einziger Programmaufruf genügt und setzt den Workflow in Gang. Dabei ist make so programmiert, dass nur fehlende Dateien oder veränderte Dateien neu generiert werden. Die einzelnen Konversions-Schritte müssen nicht manuell in Gang gesetzt werden. Das ist Vor- und Nachteil gegenüber dem oben beschriebenen Konversionsworkflow der „Digitalen Bibliothek“. In diesem ist davon abgesehen worden, die einzelnen Arbeitsschritte in einem Programm zusammenzufassen. Bei diesem hohen Grad der Automatisierung ist allerdings ein händisches Eingreifen innerhalb des Workflows nicht vorgesehen. Für die Konversion des „Lefèvre-Korpus“ bedeutete das konkret, dass die Ausgangsdaten solange korrigiert werden mussten, bis der dann vollautomatische Workflow zum gewünschten Ergebnis führte. Abschließend erfolgte eine finale manuelle Korrektur der Daten in der Umgebung zur Qualitätssicherung des Deutschen Textarchivs (DTAQ).[40]

5.2 Lexikalische Daten

Das im zweiten Beispiel unter Kapitel 4.1 eingesetzte *makeDTA* wurde in einer Anpassung auch für die Kuration und Konversion von lexikalischen Daten verwendet. Die Ausgangsdaten stammen erneut von der Editura GmbH & Co. KG und wurden aus deren Zeno-XML-Format nach DTABf umgewandelt.

[39] CLARIAH-DE (2021a)
[40] Deutsches Textarchiv (2021c)

Die meist mehrbändigen Zeno-Lexika lagen entweder bandweise getrennt, als einzelnes Master-XML oder wiederum in mehrere XML-Dateien aufgesplittet und dabei jeweils mehrere Bände enthaltend vor. Die Metadaten sind in der Regel separat gespeichert, auch hier gibt es jedoch Ausnahmen, d.h. in einzelnen Lexika sind diese abweichend vom sonstigen Vorgehen enthalten.

Ein aufwändiger Arbeitsschritt bestand auch hier v.a. in der Kuration der Ausgangsdaten. Dort wurden Inkonsistenten und durch Kenntnis der vorangegangenen Konversionen erkannte Fehler korrigiert. Die so angepassten Originalquellen wurden per *makeDTA* automatisch nach DTABf konvertiert. Das Ergebnis wurde in die DTAQ-Qualitätssicherungsumgebung eingespielt und kann dort kontrolliert sowie korrigiert werden. Die Kuration umfasste sich wiederholende Arbeitsvorgänge bis das gewünschte Ergebnis erreicht wurde:[41]

- *Herders Conversations-Lexikon*. 5 Bde. Freiburg im Breisgau, 1854–1857.[42]
- Victor von Röll (Hrsg.): *Enzyklopädie des Eisenbahnwesens*. 2. Aufl. 10 Bde. Berlin, Wien, 1912–1923.[43]
- *Dr. Vollmer's Wörterbuch der Mythologie aller Völker*. 3. Aufl. Stuttgart, 1874.[44]
- Wander, Karl Friedrich Wilhelm (Hrsg.): *Deutsches Sprichwörter-Lexikon*. 5 Bde. Leipzig, 1867–1880.[45]

Lexikalische Daten stellen auch eine wichtige Datengrundlage des im Aufbau befindlichen Zentrums für digitale Lexikographie der deutschen Sprache (ZDL) dar. Dieses ist zuständig für die umfassende und wissenschaftlich verlässliche Beschreibung der deutschen Sprache in Gegenwart und Geschichte.[46] Das ZDL bietet Zugriff auf umfangreiche gegenwartssprachliche und historische Wortinformationen. Enzyklopädien sind im Kontext des ZDL v.a. unter einem Vernetzungsgedanken mit anderen Ressourcen interessant. Voraussetzung für eine solche Vernetzung ist hier die Erschließungstiefe unterhalb der Artikelebene. Dabei ist eine explizite Annotation von Stichwörtern, Verweisartikeln oder auch einer Definition des Eintrags zu nutzen, um auf dieser Basis die Elemente mit weiteren Datensätzen zu verlinken.

Das soll im Folgenden an einem konkreten Beispiel verdeutlicht werden. Das folgende Listing zeigt exemplarisch einen Wörterbucheintrag im Zeno-XML:

[41] Da die einzelnen Werke im DTA über die Metadaten in den jeweiligen Werkkontext gestellt sind und also beim Aufruf eines mehrbändigen Werkes jeder Band die zugehörigen anderen Bände verlinkt, ist hier zur besseren Übersicht jeweils nur die URL eines Bandes aufgelistet.

[42] *https://www.deutschestextarchiv.de/book/show/nn_conversationslexikon01_1857*

[43] *http://deutschestextarchiv.de/roell_eisenbahnwesen01_1912*

[44] *https://deutschestextarchiv.de/vollmer_mythologie_1874*

[45] *https://www.deutschestextarchiv.de/wander_sprichwoerterlexikon01_1867*

[46] Vgl. „Zentrum für digitale Lexikographie der deutschen Sprache" (2021). Neben dem ZDL gibt es natürlich auch andere Einrichtungen mit dem Schwerpunkt auf Wörterbüchern oder Enzyklopädien. Bspw. 1) das „CLARIN-Repositorium der Eberhard Karls Universität Tübingen" (2021) (CTS-zertifiziert) mit Schwerpunkt auf lexikalischen und linguistischen Daten und annotierten Korpora, 2) die „Korpora am Institut für Maschinelle Sprachverarbeitung" (2021) an der Universität Stuttgart, 3) die „Corpora der Universität Leipzig" (2021) (Wörterbuch-Ressourcen) oder 4) das „Forschungsdate-Repositorium" (2021) für gegenwartsdeutsche Sprachdaten.

```
<article>
    <lem>Aeaea</lem>
    <text><p>
        <lemfloat>Aeaea</lemfloat>, (<i>Gr. M.</i>)
        <hide name="herk">griechisch</hide>
        Beiname der <link lem="Circe"><ls/></link> Circe von ihrem Aufenthalt,
        der Insel Aea.
        Ulysses gelangte von den furchtbaren Lästrygonen zu ihr.
    </p></text>
</article>
```

Abbildung 4: Zeno-XML-Annotation des Eintrags zu „Aeaea“ in Vollmers *Wörterbuch der Mythologie.* Stuttgart 1874.

Nach der oben beschriebenen automatischen *makeDTA*-Konversion in das DTABf entspricht der Eintrag der folgenden Annotation:

```
<div n="2" type="lexiconEntry">
    <p>
        <hi rendition="#b">Aeaea</hi>, (Gr. M.)
        Beiname der Circe (s. d.) von ihrem Aufenthalt, der Insel Aea.
        Ulysses gelangte von den furchtbaren Lästrygonen zu ihr.
    </p><lb/>
</div>
```

Abbildung 5: DTABf-Wörterbucheintrag nach vollautomatischer *makeDTA*-Konversion.

Für den oben beschriebenen Vernetzungsgedanken mit anderen lexikalischen Ressourcen ist diese Auszeichnung so nicht ausreichend, da außer den typisierten `<div>`-Elementen keine lexikographisch-semantische Annotationen enthalten sind.

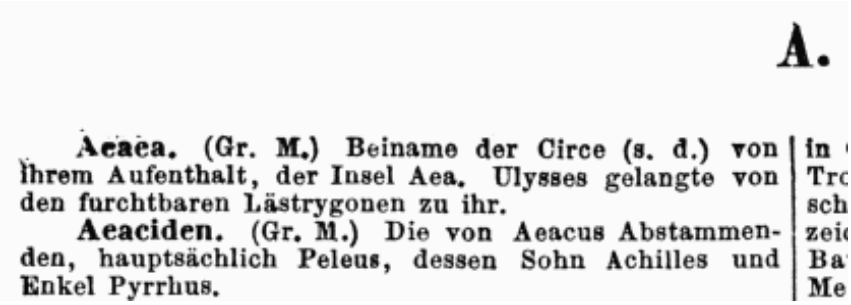
Anders sieht das mit der im November 2020 veröffentlichten Ressource EncycNet aus.[47] Dieses Korpus enthält 22 TEI-kodierte Enzyklopädien, die ursprünglich ebenfalls durch die Editura GmbH & Co. KG digitalisiert wurden. Der bereits bekannte Eintrag zu „Aeaea“ entspricht darin folgendem Schema:

```
<entry xml:id="d3e2047Vollmer-1874" xml:lang="de">
    <form type="lemma">
        <term>Aeaea</term>
    </form>
    <sense xml:id="d3e2047">
        <def>
            <term type="headword">Aeaea</term>, (<hi rend="italic">Gr.M.</hi>)
            <span type="origin">griechisch</span>
            Beiname der <ref type="entry" target="#d3e45093Vollmer-1874"/>
            Circe von ihrem Aufenthalt, der Insel Aea.
            Ulysses gelangte von den furchtbaren Lästrygonen zu ihr.
        </def>
    </sense>
</entry>
```

Abbildung 6: Wörterbucheintrag aus dem EncycNet

[47] Hagen, Ketzan, Jannidis & Witt (2020)

Im Kontext der Auszeichnung von Enzyklopädien ist die Annotation per `<entry>`, `<form>` und `<term>` als best practice zu bezeichnen. Durch die Verwendung von `@xml:id` und `@target` ist bereits eine interne Verweisstruktur angelegt. Bezogen auf die Vorlagen und textgetreue Rekonstruktion als digitalen Volltext ist die gewählte Auszeichnungsstrategie allerdings nicht unproblematisch. Als `/form[@type="lemma"]/term` entspricht der Eintrag dem gängigen Annotationsschema. Da anschließend „Aeaea“ innerhalb von `/sense/def` als Bestandteil der eigenen Definition erneut erscheint und „griechisch“ zweifelsfrei als Metainformation zusätzlich zu den Metainformation „Gr. M.“ als „griechische Mythologie” angefügt wurde – diese zusätzlichen Informationen sind im Druck nicht vorhanden – kommt es hier offensichtlich zu einer Vermischung von Text- und Metadaten. Ein Blick auf das Faksimile (s. Abbildung 7) zeigt das deutlich.



A.

Aeaea. (Gr. M.) Beiname der Circe (s. d.) von ihrem Aufenthalt, der Insel Aea. Ulysses gelangte von den furchtbaren Lästrygonen zu ihr.

Aeaciden. (Gr. M.) Die von Aeacus Abstammenden, hauptsächlich Peleus, dessen Sohn Achilles und Enkel Pyrrhus.

Abbildung 7: Faksimile des Eintrags zu „Aeaea“ in Vollmers *Wörterbuch der Mythologie*. Stuttgart 1874

Das Beispiel macht anschaulich, dass Entscheidungen über die XML-Auszeichnung der Ausgangsdaten nicht nur abhängig vom Repositorium getroffen werden, in welchem die langfristige Archivierung der Daten erfolgen soll, sondern der Kontext der wissenschaftlichen Nachnutzung in die Entscheidungen mit einzubeziehen ist. Eine umfangreiche Vernetzung mit anderen Ressourcen wirkt sich auf die zu wählende Annotationsstrategie aus.

5.3 Editionen

TEI-XML ist der international etablierte Standard für Editionsdaten. Abhängig vom Einzelfall bieten sich für die Integration von Editionsdaten in bestehende Infrastrukturen – ebenso wie für Textdaten – das DTA oder das TextGrid Repository an. Die Integration in das DTA erfordert zwar eine Konversion der Editionsdaten in das DTABf und ist gegebenenfalls mit einem entsprechenden Mehraufwand verbunden. Jedoch bietet die Integration von Editionsdaten in das DTABf den Vorteil, dass die Daten anschließend in ein größeres, homogenes Korpus integrierbar sind, für das bereits etablierte Tools für die Nachnutzung der Daten bereit stehen. Da das DTABf ein TEI-XML-Substandard ist, können alle Editionsdaten, die in das DTA integriert werden, auch in TEI-basierte Infrastrukturen wie dem TextGrid Repository vorgehalten werden.

Für die Konversion von TEI-XML-konformen Editionsdaten in das DTABf hat CLARIAH-DE eine Reihe an Empfehlungen und Hilfsmitteln erarbeitet. Einen allgemeinen Überblick zur Konversion von Editionsdaten in das DTABf beschreibt der Arbeitsbericht „Das DTABf in der Edition“.[48] Einen tiefergehenden Einblick geben die dazugehörigen Fallbeispiele.[49] Außerdem steht mit der Editionsmatrix (EdMa)[50] ein Klassifikationsmodell für digitale Editionen zur Verfügung, das Vergleichbarkeit im sehr heterogenen Feld digitaler Editionen schaffen soll.

Als interaktives Angebot zur Integration von Editionsdaten in die Infrastruktur bietet CLARIAH-DE einen Web-Fragebogen an, der Nutzerinnen und Nutzer dabei unterstützt, digitale Editionen anhand von EdMa zu kategorisieren. Nach erfolgter Kategorisierung erhalten Nutzerinnen und Nutzer erste

[48] Fisseni et al. (2021)
[49] Fisseni, Meiners, Sendler & Sikora (2021); Schulz (2021a); Schulz (2021b)
[50] Schulz, Fisseni & Sendler (2021)

Hinweise und Hilfestellungen für eine Konversion digitaler Editionsdaten in ein Austauschformat wie das DTABf.

Die für die Aufbereitung der Fallbeispiele genutzten Skripte und Programme werden auf GitHub sowie GitLab zur Verfügung gestellt werden. Mit Unterstützung der Skripte lässt sich aufwandsarm und zeitsparend eine Übersicht über Kompatibilitäten und Inkompatibilitäten eines projektspezifischen Editionsschemas mit dem DTABf erstellen. Diese Übersicht bietet einen Anhaltspunkt für eine Aufwandsabschätzung, sollte eine Konversion der Editionsdaten in das DTABf erwünscht sein.

6 Zusammenfassung

Ziel dieser Handreichung war die Beschreibung beispielhafter Workflows zur Integration von Daten aus der Community in bereits bestehende, etablierte und eine nachhaltige Speicherung garantierende Infrastrukturen. Dazu wurde zunächst ein allgemeiner Einblick in den Lebenszyklus von digitalen Daten gegeben und auf die Bedeutung von Standards hingewiesen. Da die Anforderungen je nach Einrichtung aber auch bezogen auf den Datentyp stark variieren, konzentriert sich die Handreichung auf zwei zentrale Einrichtungen aus den Infrastrukturprojekten CLARIN-D und DARIAH-DE: Das Deutsche Textarchiv sowie das TextGrid Repository. Beide Repositorien sind CoreTrustSeal-zertifiziert und bieten eine nachhaltige Archivierung von Daten gemäß der FAIR-Prinzipien.

Aufgrund des strengeren Standards sind die Anforderungen an Datengeber bezüglich der Übernahme ihrer Ergebnisse in die DTA-Infrastruktur vergleichsweise höher. Auch ist die eigenständige Speicherung und Publikation von Daten durch Datengeberinnen und Datengeber im DTA – im Gegensatz zum TextGrid Repository – nicht vorgesehen. Zur Erreichung von Interoperabilität ist die Einhaltung von Standards jedoch ein zentrales Kriterium. Dies gilt vor allem dann, wenn es sich wie beim DTABf um einen Standard handelt, dem es daran gelegen ist, Ambiguitäten im Bereich der Annotation zu vermeiden. Belohnt werden die Mühen zudem mit einer vollautomatischen linguistischen Analyse der Texte, die wiederum die Grundlage für die Verwendung der mächtigen DDC-Suchmaschine, wie auch für in die Infrastruktur eingebundene linguistische Analysetools wie Wortverlaufskurven oder DiaCollo darstellt. Die Verwendung externer Tools wie Voyant oder eine Anbindung an das Language Resource Switchboard gewährleisten sowohl das DTA wie auch das TextGrid Repository, welches mit der sogenannten Regalfunktion zusätzlich die Möglichkeit bietet, individuelle Kollektionen zusammenzustellen.

Die Integration von Daten in bestehende Ressourcen, das haben die Beispiele oben gezeigt, ist kein generischer Prozess. Vielmehr geht es darum, eine je auf die spezifischen Bedürfnisse zugeschnittene Herangehensweise abzustimmen. Den Anfang sollte demnach in jedem Fall die Kontaktaufnahme zu den jeweiligen Betreibern der Repositorien machen:

Das DTA ist erreichbar unter: *redaktion@deutschestextarchiv.de*.

Der Kontakt zum TextGrid Repository lautet: *anfragen@textgrid.de*.

Über weitere im Kontext von CLARIN-D und DARIAH-DE entstandene und langfristig nutzbare Infrastrukturen informiert schließlich auch das CLARIAH-DE Helpdesk: *https://www.clariah.de/support*.

Referenzen und Literatur

Appleby, M., Crane, T., Sanderson, R., Stroop, J. & Warner, S. (Hrsg.). (2021). *IIIF Image API 2.1*. Verfügbar unter: *https://iiif.io/api/image/2.1/*

Aschenbrenner, A. & Meffert, K. (2009). Wissenschaftliche Infrastruktur in den Geisteswissenschaften? Eine Wegbeschreibung. *Jahrbuch für Computerphilologie*, *9*, 9–28.

Bayerisches Archiv für Sprachsignale. (2021). Verfügbar unter: *http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B*

Berlin-Brandenburgische Akademie der Wissenschaften. (2016). D* Corpora. Verfügbar unter: *https://kaskade.dwds.de/dstar/*

Berlin-Brandenburgische Akademie der Wissenschaften. (2021). Das CLARIN-Servicezentrum des Zentrums Sprache an der BBAW. Verfügbar unter: *https://clarin.bbaw.de/de/*

Burnard, L. & Bauman, S. (Hrsg.). (2020). *Guidelines for Electronic Text Encoding and Interchange*. Chicago, New York: TEI P5 Text Encoding Initiative. Verfügbar unter: *https://tei-c.org/guidelines/p5/*

CLARIAH-DE. (2021b). Webseite des Projekts CLARIAH-DE. Verfügbar unter: *https://www.clariah.de/*

CLARIAH-DE. (2021a). DiBiLit-Korpus. Verfügbar unter: *http://www.deutschestextarchiv.de/dibilit*

CLARIN. (2021). Language Resource Switchboard. Verfügbar unter: *https://switchboard.clarin.eu/*

CLARIN-Repositorium der Eberhard Karls Universität Tübingen. (2021). Verfügbar unter: *https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/clarin-center/repository.html*

CoreTrustSeal Homepage. (2021). Verfügbar unter: *https://www.coretrustseal.org/*

Corpora der Universität Leipzig. (2021). Verfügbar unter: *https://corpora.uni-leipzig.de/en?corpusId=deu_news_2020*

Creative Commons Homepage. (2021). Verfügbar unter: *https://creativecommons.org/*

DARIAH-DE. (2019a). *TextGrid Repository Documentation: IIIF Implementation*. Verfügbar unter: *https://textgridlab.org/doc/services/submodules/tg-iiif-metadata/docs_tgrep/index.html*

DARIAH-DE. (2019b). TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften. Verfügbar unter: *https://textgrid.de*

DARIAH-DE. (2019c). *TextGrid Repository Documentation: TextGrid Import*. Verfügbar unter: *https://textgridlab.org/doc/services/submodules/kolibri/kolibri-addon-textgrid-import/docs/index.html*

DARIAH-DE. (2019d). *TextGrid Repository Documentation: TG-digilib*. Verfügbar unter: *https://textgridlab.org/doc/services/submodules/tg-digilib/docs_tgrep/index.html*

DARIAH-DE. (2019e). Voyant in TextGrid. Verfügbar unter: *https://textgridrep.org/docs/voyant*

DARIAH-DE. (2019f). Switchboard. Verfügbar unter: *https://textgridrep.org/docs/switchboard*

DARIAH-DE. (2019g). Regal. Verfügbar unter: *https://textgridrep.org/docs/shelf*

DARIAH-DE. (2019h). Die Digitale Bibliothek bei TextGrid. Verfügbar unter: *https://textgrid.de/de/digitale-bibliothek*

DARIAH-DE Research Data Life Cycle. (2021). Verfügbar unter: *https://de.dariah.eu/web/guest/weiterfuhrende-informationen*

Deutsche Forschungsgemeinschaft. (2016). *DFG-Praxisregeln "Digitalisierung"*. Deutsche Forschungsgemeinschaft, Bonn. Verfügbar unter: *https://www.dfg.de/formulare/12_151/12_151_de.pdf*

Deutsches Textarchiv. (2020a). Dokumentation zum DTA-Basisformat). Verfügbar unter: *http://www.deutschestextarchiv.de/doku/basisformat/*

Deutsches Textarchiv. (2020b). DTA-Basisformat – Richtlinien zur Transkription. Verfügbar unter: *https://www.deutschestextarchiv.de/doku/basisformat/transkription.html*

Deutsches Textarchiv. (2020c). Webseite des Deutschen Textarchivs (DTA). Verfügbar unter: *http://www.deutschestextarchiv.de/*

Deutsches Textarchiv. (2021a). Deutsches Textarchiv – DTAE. Verfügbar unter: *http://www.deutschestextarchiv.de/dtae*

Deutsches Textarchiv. (2021c). Deutsches Textarchiv – Qualitätssicherung. Verfügbar unter: *http://www.deutschestextarchiv.de/dtaq/about*

Deutsches Textarchiv. (2021b). CLARIN-D Metadatenormular zur Aufnahme einzelner Ressourcen. Verfügbar unter: *https://www.deutschestextarchiv.de/dtae/submit/clarin*

DFG-Fachkollegium Sprachwissenschaften. (2015). *Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora.* Deutsche Forschungsgemeinschaft, Bonn.

digilib Community. (2021). *The digilib Scaler API.* Verfügbar unter: *https://robcast.github.io/digilib/scaler-api.html*

ePIC Persistent Identifier Consortium for eResearch. (2016). Persistent Identifiers for eResearch. Verfügbar unter: *https://www.pidconsortium.net/*

Fisseni, B., Meiners, H.-L., Sendler, S. & Sikora, U. (2021). *Evaluationsbericht: Das DTA-Basisformat als Austauschformat für kritische Editionen (Neologie): untersucht und erläutert am Beispiel der* Bibliothek der Neologie. CLARIAH-DE-Arbeitsberichte Nr. 2. CLARIAH-DE. *https://doi.org/10.14618/ids-pub-10498*

Fisseni, B., Sendler, S., Schulz, D., Boenig, M., Meiners, H.-L. & Sikora, U. (2021). *Das DTABf in der Edition: zusammenfassender Evaluationsbericht.* CLARIAH-DE-Arbeitsberichte Nr. 1. CLARIAH-DE. *https://doi.org/10.14618/ids-pub-10496*

Forschungsdate-Repositorium. (2021). Verfügbar unter: *http://repos.ids-mannheim.de/*

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C. & Wiegand, F. (2018). Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN (Germanistische Sprachwissenschaft um 2020). In H. Lobin, R. Schneider & A. Witt (Hrsg.), *Digitale Infrastrukturen für die germanistische Forschung* (Band 6, S. 219–248). Berlin/Boston. Verfügbar unter: DOI: 10.1515/9783110538663-011

Geyken, A. & Gloning, T. (2015). A living text archive of 15th-19th-century German. *Historical corpora: Challenges and perspectives*, *5*, 165–179.

Geyken, A. & Haaf, S. (2018). Integration heterogener historischer Textkorpora in das Deutsche Textarchiv (Osnabrücker Beiträge zur Sprachtheorie). In J. Gessinger, A. Redder & U. Schmitz (Hrsg.), *Korpuslinguistik* (S. 175–192). Duisburg: Universitätsverlag Rhein-Ruhr.

Haaf, S. & Thomas, C. (2016). Die historischen Korpora des Deutschen Textarchivs als Grundlage für Sprachgeschichtliche Forschungen.

Hagen, T., Ketzan, E., Jannidis, F. & Witt, A. (2020). Twenty-two Historical Encyclopedias Encoded in TEI: a New Resource for the Digital Humanities. *https://doi.org/10.5281/zenodo.4159491*

Hamburger Zentrum für Sprachkorpora. (2021). Verfügbar unter: *https://corpora.uni-hamburg.de/hzsk/*

Hedges, M., Neuroth, H., Smith, K. M., Blanke, T., Romary, L., Küster, M. et al. (2013). TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship. *Journal of the Text Encoding Initiative*, (Issue 5). *https://doi.org/10.4000/jtei.774*

ISO/TC 37/SC 4 Language resource management. (2016). *ISO 24624:2016 Language resource management — Transcription of spoken language.* Verfügbar unter: *https://www.iso.org/standard/37338.html*

Korpora am Institut für Maschinelle Sprachverarbeitung. (2021). Verfügbar unter: *https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/*

Neuroth, H., Lohmeier, F. & Smith, K. M. (2011). TextGrid – Virtual Research Environment for the Humanities. *International Journal of Digital Curation*, *6*, 222–231.

Schulz, D. (2021a). *Evaluationsbericht: Das DTA-Basisformat für Manuskripte (DTABf-M) als Austauschformat für kritische Editionen handgeschriebener historischer Texte (Hainhofer): untersucht am Beispiel des Projekts* Philipp Hainhofer: Die Reiseberichte und Sammlungsbeschreibungen 1594–1636. CLARIAH-DE-Arbeitsberichte Nr. 4. CLARIAH-DE. *https://doi.org/10.14618/ids-pub-10500*

Schulz, D. (2021b). *Evaluationsbericht: Das DTA-Basisformat für Manuskripte (DTABf-M) als Austauschformat für Editionen handgeschriebener historischer Texte (Diarium): untersucht am Beispiel des Projekts* Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek. Digitale Edition des Diariums von Herzog August dem Jüngeren, Selbstzeugnis-Repertorium und Forschungsportal. CLARIAH-DE-Arbeitsberichte Nr. 3. CLARIAH-DE. *https://doi.org/10.14618/ids-pub-10499*

Schulz, D., Fisseni, B. & Sendler, S. (2021). *EdMA: eine Matrix zur Kategorisierung digitaler Editionen.* CLARIAH-DE-Arbeitsberichte Nr. 5. CLARIAH-DE. *https://doi.org/10.14618/ids-pub-10501*

Sinclair, Stéfan and Rockwell, Geoffrey. (2016). Voyant Tools. Verfügbar unter: *https://voyant-tools.org/*

Söring, S., Lohmeier, F. & Rodriguez, K. J. (2011, Oktober). 'If You Build It, They Will Come'. Text,. Zugriff am 15.8.2019. Verfügbar unter: *https://ehri-project.eu/content/if-you-build-it-they-will-come*

Thomas, C. & Wiegand, F. (2015). Making great work even better: Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache). In J. Gippert & R. Gehrke (Hrsg.), *Historical corpora: challenges and perspectives* (Band 5, S. 181–196). Tübingen. Verfügbar unter: *https://edoc.bbaw.de/opus4-bbaw/frontdoor/index/index/docId/2005*

Wiegand, F., Thomas, C., Haaf, S., Geyken, A., Jurish, B. & Boenig, M. (2018). Recherchieren, Arbeiten und Publizieren im Deutschen Textarchiv: ein Praxisbericht. *Zeitschrift für germanistische Linguistik*, *46*(1), 147–161. *https://doi.org/doi:10.1515/zgl-2018-0009*

Wissik, T. & Ďurčo, M. (2015). Research data workflows: from research data lifecycle models to institutional solutions (Linköping Electronic Conference Proceedings). In K. De Smedt (Hrsg.), *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland* (S. 94–107). Verfügbar unter: *https://ep.liu.se/ecp/123/008/ecp15123008.pdf*

Zentrum für digitale Lexikographie der deutschen Sprache. (2021). Verfügbar unter: *https://www.zdl.org/*

Alle Links wurden am 23. August 2021 überprüft.