Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Leitlinie für das Integrieren von Daten in Text+/NFDI, kollektionsspezifische Version

Milestone C2.2

Das vorliegende Dokument wurde im Rahmen des Konsortiums Text+ im Kontext der Arbeit des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. verfasst. NFDI wird von der Bundesrepublik Deutschland und den 16 Bundesländern finanziert, und das Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370. Die Autor:innen bedanken sich für die Förderung sowie Unterstützung. Ein Dank geht außerdem an alle Einrichtungen und Akteur:innen, die sich für den Verein und dessen Ziele engagieren.

This document was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Version	1.0
Redaktion	
Redaktionsteam	Florian Barth, Andreas Blätte, José Calvo Tello, Anke Debbeler, Christoph Draxler, Stefan Fischer, Marcel Fladrich, Stefan E. Funk, Philippe Genêt, Mathias Göbel, Alina Hemmer, Marius Hug, Jörg Knappen, Marie-Pauline Krielke, Daniel Kurzawe, Timm Lehmberg, Peter Leinen, Elisabeth Mollenhauer, Felix Rau, Sara Saleh, Florian Schiel, Elke Teich, Thorsten Trippel, Ubbo Veentjer, Lukas Weimer, Antonina Werthmann, Rebecca Wilm, Andreas Witt, Stine Ziegler, Claus Zinn
Projekt	Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur
Bezeichnung	C2.2 Guidelines for integrating data into Text+/NFDI, collection-specific version
Förderung	DFG Förderkennzeichen 460033370
Projektlaufzeit	01.10.2021 bis 30.09.2026

Inhalt

1 Text+ und NFDI

Die Nationale Forschungsdateninfrastruktur (NFDI)[1] sieht Daten als gemeinsames Gut für exzellente Forschung, organisiert durch die Wissenschaft in Deutschland. Die Mission der NFDI ist es, die Nutzungsmöglichkeiten von Daten für Wissenschaft und Gesellschaft Schritt für Schritt zu verbessern. Ziel des Zusammenwirkens im NFDI-Verein ist das Entstehen einer Dachorganisation für das Forschungsdatenmanagement in allen Wissenschaftszweigen, die in Zusammenarbeit mit nationalen und internationalen Partnern die Rahmenbedingungen für rechtskonforme, interoperable und nachhaltige Dateninfrastrukturen schafft, die für Forschende in ihrem Arbeitsalltag gut zugänglich sind. Die NFDI möchte ausbilden, die Kompetenz im Umgang mit Daten stärken und neue Berufswege eröffnen.

Die Vision des NFDI-Konsortiums Text+[2] ist es, dass die text- und sprachorientierten Geistes- und Sozialwissenschaften die Möglichkeiten der Digitalisierung in ihrer Forschung umfassend nutzen und sie eine gemeinsame Datenkultur haben. Seine Mission ist es, Text- und Sprachdaten sowie den Zugang dazu zu ertüchtigen. Dadurch trägt Text+ dazu bei, dass der Durchgriff auf digitale Quellen zum Standard wird. Text+ stärkt die Digital Literacy der Forschenden, wobei die Diversität der Forschungsgemeinde abgedeckt und auf ihre Partizipation gebaut wird. Durch die Integration von Infrastruktur und Forschung werden Interdisziplinarität und Innovation gefördert.

Um seiner Mission und somit auch der Mission der NFDI gerecht zu werden, nimmt Text+ neben bereits an den beteiligten Datenzentren vorhandenen Daten auch zusätzliche Ressourcen in das Text+ Portfolio auf. Dieses Versprechen ist bereits im Projektantrag verankert: “Based on the reference implementation, data centres will expand their portfolio regarding data and services. Clusters are obliged to integrate any data resources and services compliant with their specialisation, the technical and quality criteria set by the Text+ Scientific Board and the respective SCC” (Hinrichs et al., 2022). Daten und Dienste können also dann in Text+ integriert werden, wenn sie zur Spezialisierung der Partner passen, sie bestimmte technische Anforderungen und Qualitätskriterien erfüllen und/oder die Community-basierten Gremien, also die Scientific Coordination Committees (SCCs) und das Infrastructure/Operations Coordination Committee (OCC), zustimmen.

2 Bereitstellung von Daten über Text+

Daten über Text+ verfügbar zu machen, trägt zu einer nachhaltigen Datenhaltung bei. Hier spielen die FAIR-Prinzipien (Wilkinson et al., 2016) – Findability, Accessibility, Interoperability und Reusability – eine wichtige Rolle.

Findability (Auffindbarkeit) bedeutet, dass Daten und Metadaten sowohl von Menschen als auch von Maschinen leicht zu finden sein sollen. Zu diesem Zweck wird ihnen ein persistenter Identifikator zugewiesen, der in den (aussagekräftigen, „reichen“) Metadaten hinterlegt wird. Außerdem werden Daten und Metadaten über durchsuchbare Verzeichnisse und spezialisierte Suchmaschinen auffindbar gemacht. Im Text+-Kontext sind hier vor allem die Text+ Registry, welche sich noch im Aufbau befindet, und die Föderierte Inhaltssuche (Federated Content Search, kurz FCS), welche aktuell als Entwurf[3] zur Verfügung steht, zu nennen.

Accessibility (Zugänglichkeit) heißt, dass Daten und Metadaten verfügbar gemacht und langzeitarchiviert werden sollen, sodass sie leicht von Menschen und Maschinen heruntergeladen und genutzt werden können. Mittels ihres Identifikators können Daten und Metadaten über standardisierte Kommunikationsprotokolle abgerufen werden. Wo nötig, werden Authentifizierungs- und Autorisierungsverfahren eingebunden, die im Rahmen eines Identity and Access Managements (IAM) auch NFDI-weit diskutiert und bereitgestellt werden. Stehen die Daten selbst nicht zur

[1] https://www.nfdi.de/

[2] https://www.text-plus.org/

[3] https://fcs.text-plus.org/

Verfügung, etwa aus urheber-, persönlichkeits- oder datenschutzrechtlichen Gründen, werden zumindest die entsprechenden Metadaten angeboten.

Mit Interoperability (Interoperabilität) ist gemeint, dass die Daten derart vorliegen sollen, dass sie von Menschen und Maschinen mit anderen Datensätzen verknüpft werden können. Zu einer erhöhten Interoperabilität trägt es bei, wenn eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache für die Wissensrepräsentation genutzt wird und kontrollierte Vokabulare verwendet werden, die ebenfalls den FAIR-Prinzipien entsprechen. Zudem sollten Daten und Metadaten qualifizierte Verweise auf andere Daten und Metadaten enthalten, zum Beispiel in Form von Verknüpfungen zu Normdaten (etwa zur Gemeinsamen Normdatei).

Zur Reusability (Wiederverwendbarkeit) trägt eine Beschreibung der Datensätze über Metadaten bei, sodass sie für weitere Forschungen nachnutzbar und mit anderen Datensätzen vergleichbar sind. Dabei sollen die enthaltenen Attribute präzise und relevant sein. Die Angabe einer Nutzungs-lizenz und detaillierter Provenienz-Informationen ist unverzichtbar, genau wie die Einhaltung von fachgebietsrelevanten Community-Standards. Mit Wiederverwendbarkeit steht auch die Möglichkeit in Verbindung, Daten mithilfe weiterer Tools zu verarbeiten und zu analysieren, etwa über CLARINs Language Resource Switchboard (LRS)[4] oder den DARIAH-DE Geo-Browser[5].

Auch die CARE-Prinzipien (Carroll et al., 2020) – Collective Benefit (Kollektiver Nutzen), Authority to Control (Kontrolle über die Daten), Responsibility (Verantwortung) und Ethics (Ethik) – sind von Bedeutung, wenn es um nachhaltige Datenhaltung geht. Ursprünglich aus der Anthropologie stammend werden hier zu reinen Datenverwaltungskriterien weitere Kriterien angegeben, die auf einen Interessenausgleich zwischen den Datengebenden und der wissenschaftlichen Nachnutzung ausgerichtet sind.

3 Daten- und Kompetenzzentren in Text+

Die von Text+ adressierten Forschungsdaten werden in drei Datendomänen eingliedert, die sich (neben *Infrastructure/Operations* und *Administration*) als sogenannte Task Areas in Text+ wiederfinden: *Collections*, *Lexical Resources* und *Editions.* Diese sind jeweils weiter in feinkörnigere thematische Cluster unterteilt. Im vorliegenden Dokument liegt der Fokus auf der Datendomäne *Collections*, welche aus den Clustern *Contemporary Language*, *Historical Texts* und *Unstructured Text* besteht. An jedem Cluster ist mindestens ein Datenzentrum beteiligt. Typischerweise wirken mehrere Daten- und Kompetenzzentren zusammen.

Mit „Datenzentren“ werden in Text+ Partnereinrichtungen bezeichnet, die sich auf bestimmte Arten von Daten spezialisiert haben und über entsprechende eigene Daten sowie eine Infrastruktur verfügen, die eine langfristige Bereitstellung und Archivierung von Daten ermöglicht. Dabei nutzen sie zur Datenhaltung zertifizierte Repositorien. Datenzentren stellen Metadaten zu den Daten über Schnittstellen bereit und stellen Schnittstellen zu weiteren Diensten von Text+ zur Verfügung, zum Beispiel zur verteilten Suche. Außerdem nehmen Datenzentren Daten von Dritten entgegen, sofern diese ihrer Spezialisierung entsprechen und gewisse Anforderungen erfüllt sind.

Auch Kompetenzzentren sind Partnereinrichtungen, die spezielle Kenntnisse und Fähigkeiten in Bezug auf bestimmte Arten von Daten mitbringen. Zu deren Erstellung und Archivierung bringen sie sich jedoch vor allem in beratender Form ein, was nicht ausschließt, dass Kompetenzzentren durchaus auch an der Entwicklung von Diensten in Text+ beteiligt sein können, die auf Schnittstellen aufbauen. Im Gegensatz zu Datenzentren benötigen Kompetenzzentren keine eigene Archivinfrastruktur. Stattdessen leisten sie Unterstützung bei der Archivierung an anderen Orten.

Die Schwerpunkte der einzelnen Daten- und Kompetenzzentren sehen ganz unterschiedlich aus. So

[4] https://switchboard.clarin.eu/

[5] https://de.dariah.eu/geobrowser

kann der Fokus zum Beispiel auf bestimmten Sprachen, Epochen oder Datenformaten liegen. Einzelne Zentren können eine große Bandbreite unterschiedlicher Arten von Daten abdecken oder sich aber auf ganz spezielle Arten von Daten konzentrieren. Auch darin, ob sie eher mit born-digital oder mit retrodigitalisierten Daten arbeiten (oder beidem), können sich Daten- und Kompetenzzentren unterscheiden.

An der Task Area *Collections* sind folgende Datenzentren beteiligt:

- Akademie der Wissenschaften in Hamburg (AdWHH)
- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
- Deutsche Nationalbibliothek (DNB)
- Eberhard Karls Universität Tübingen (UniTÜ)
- Leibniz-Institut für Deutsche Sprache (IDS)
- Ludwig-Maximilians-Universität München (LMU)
- Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
- Universität des Saarlandes (UdS)
- Universität Duisburg-Essen (UniDUE)
- Universität Hamburg (UniHH)
- Universität zu Köln (UniK)

Als Kompetenzzentren wirken folgende Institutionen mit:

- Albert-Ludwigs-Universität Freiburg (UniFR)
- Julius-Maximilians-Universität Würzburg (UniWÜ)

Die von den Datenzentren zur Verfügung gestellten Repositorien sind durch das CoreTrustSeal (CTS)[6] oder das nestor-Siegel[7] zertifiziert, befinden sich im Prozess der Rezertifizierung oder streben eine Zertifizierung an.

Das Core Trust Seal ist ein international etabliertes Siegel zur Auszeichnung von nachhaltiger Infrastruktur. Bei Antrag auf Zertifizierung eines Repositoriums wird dieses durch qualifizierte Mitglieder des sogenannten *Assembly of Reviewers* begutachtet. Der Bewertung liegt eine feste Kriterienliste[8] zugrunde. Gelten die Voraussetzungen als erfüllt, erfolgt die Zertifizierung durch das sogenannte *Core Trust Seal Board*. Das nestor-Siegel für vertrauenswürdige digitale Langzeitarchive wird auf Basis der DIN 31644 "Kriterien für vertrauenswürdige digitale Langzeitarchive" vergeben. Die Begutachtung erfolgt durch Mitglieder der *nestor-AG Zertifizierung*.

Die Datenzentren stellen Metadaten als öffentliche Informationen zu den vorhandenen Forschungsdaten zur Verfügung. Die Bereitstellung erfolgt über Schnittstellen, um die Integration in Kataloge und Nachweissysteme zu erleichtern und die Metadaten so über Suchmaschinen durchsuchbar zu machen. Die Verfügbarmachung entsprechender Metadaten ermöglicht auch eine Zitation beziehungsweise eine persistente Identifikation der Daten.

Die zur Verfügung gestellten Metadaten entsprechen gängigen Konventionen und Normen wie Dublin Core[9], DataCite[10], Lightweight Information Describing Objects (LIDO)[11], MARC (machine-readable cataloguing) 21, ISO 24622-1[12] und ISO 24622-2[13] (die zusammen die Component Metadata

[6] https://www.coretrustseal.org/

[7] https://www.langzeitarchivierung.de/Webs/nestor/DE/Zertifizierung/nestor_Siegel/siegel.html

[8] https://www.coretrustseal.org/why-certification/requirements/

[9] https://www.dublincore.org

[10] https://schema.datacite.org/

[11] https://cidoc.mini.icom.museum/working-groups/lido/lido-overview/

[12] https://www.iso.org/standard/37336.html

[13] https://www.iso.org/standard/64579.html

Infrastructure, kurz CMDI[14], ausmachen) oder TEI-Headern, wie sie in den Guidelines der Text Encoding Initiative (TEI)[15] beschrieben werden. Dabei werden unterschiedliche Serialisierungen wie Extensible Markup Language (XML)[16], JavaScript Object Notation (JSON)[17] oder Resource Description Format (RDF) genutzt, wobei es für letztere wiederum unterschiedliche Serialisierungen wie N-Triples[18], Terse RDF Triple Language (Turtle)[19], RDF/XML[20] oder JavaScript Object Notation for Linked Data (JSON-LD)[21] gibt. Zu diesem Thema ist aktuell das Deliverable C3.2 "Einschlägige Normen für Sammlungen" in Arbeit.

Die Metadaten werden in der Regel von den Datenzentren bereits über Schnittstellen ausgeliefert. Vorgesehen ist zunächst OAI-PMH[22]; für RDF-basierte Repräsentationen bieten sich daneben SPARQL[23]-Schnittstellen an. OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) basiert auf XML und Representational State Transfer (REST). Als kleinsten gemeinsamen Nenner sind Datengebende dazu angehalten, Dublin-Core-Metadaten anzugeben; zusätzlich sind aber auch weitere Metadatenschemata erlaubt – in den letzten Jahren hat sich DataCite als interdisziplinäres Format besonders etabliert. SPARQL (SPARQL Protocol and RDF Query Language) ist insbesondere für Anwendungen im Bereich Linked Data konzipiert. Zumindest im Backend ist dementsprechend eine Darstellung der Metadaten als Linked Data erforderlich.

Die auf Webstandards basierende Text+ FCS ermöglicht (analog zur CLARIN-FCS[24]) den Zugriff auf Daten in einer ortsverteilten Infrastruktur. Dieses Werkzeug dient als niederschwelliger Zugang zu den Daten neben den spezialisierten Erschließungswerkzeugen, die bei den meisten datenhaltenden Institutionen auch für die jeweiligen dort vorgehaltenen Daten zur Verfügung stehen.

4 Wege ins Text+ Universum

Es gibt verschiedene Wege, eigene Daten in die Text+ Infrastruktur zu integrieren. Eine Möglichkeit ist es, die Daten von einem existierenden Datenzentrum aufnehmen zu lassen. Über den Text+ Helpdesk ist es möglich, sich dazu beraten zu lassen, welches Datenzentrum für die gegebenen Daten in Frage kommen könnte. Es muss gewährleistet sein, dass die Anforderungen des jeweiligen Datenzentrums erfüllt sind, und Datenzentrum und Datenhaltende müssen sich darüber abstimmen, unter welchen Bedingungen eine Übernahme der Daten stattfinden kann. Unter Umständen ist es möglich, durch das Datenzentrum Unterstützung bei der Anpassung und Aufbereitung der Daten zu erhalten.

Ist eine Übergabe der Daten an ein existierendes Datenzentrum nicht gewünscht oder möglich, besteht für Datenhaltende die Möglichkeit, selbst ein Datenzentrum aufzubauen und durch die Schnittstellen und eine Zertifizierung Teil des Netzwerks von Text+ zu werden. In diesem Fall würden die Daten nachhaltig eigenständig gehostet, gepflegt und verwaltet.

Ein Mechanismus, über den Mittel für die Aufbereitung von Daten und deren Integration in Text+ zur Verfügung gestellt werden, besteht in den Kooperationsprojekten von Text+. Im Rahmen der jährlichen Ausschreibung zur Förderung von Kooperationsprojekten können diejenigen, die bereits Forschungsdaten haben, sich auf eine finanzielle Förderung bewerben. Die Bewerbung steht sowohl

[14] https://www.clarin.eu/content/component-metadata

[15] https://tei-c.org/

[16] https://www.w3.org/TR/xml/

[17] https://www.json.org

[18] https://www.w3.org/TR/n-triples/

[19] https://www.w3.org/TR/turtle/

[20] https://www.w3.org/TR/rdf-syntax-grammar/

[21] https://www.w3.org/TR/json-ld11/

[22] https://www.openarchives.org/pmh/

[23] https://www.openarchives.org/pmh/

[24] https://contentsearch.clarin.eu/

Projekten offen, die das Ziel haben, vorhandene Daten so aufzubereiten, dass sie in ein bestehendes Datenzentrum integriert werden können, als auch externen Datenzentren, die mit Schnittstellen in die technische Infrastruktur von Text+ eingebunden werden möchten.

Wenn keine andere Lösung in Frage kommt bzw. die clusterspezifischen Kriterien nicht erfüllt sind, besteht die Möglichkeit, Daten zur Bitstream Preservation abzuladen. Hierzu bietet Text+ ein System an, das von der GWDG entwickelt, betrieben und gepflegt wird. Dort können Daten ganz unabhängig von ihrem Format hinterlegt werden. Viele Vorteile, die mit einer nachhaltigen Datennut-zung zusammenhängen, fallen bei dieser Option jedoch weg. Zum Beispiel kann eine Integration in weitere Analysewerkzeuge und inhaltsbasierter Suchfunktionen nicht garantiert werden.

5 Beschreibung der Text+ Datenzentren

a Akademie der Wissenschaften in Hamburg (AdWHH)

Seit ihrer Gründung im Jahr 2004 fördert die AdWHH[25] die interdisziplinäre Forschung zu gesellschaftlich bedeutsamen Zukunftsfragen und grundlegenden wissenschaftlichen Problemen. Darüber hinaus koordiniert die AdWHH derzeit sechs Langzeitvorhaben[26] im Rahmen des Akademienprogramms (das wiederum von der *Union der deutschen Akademien der Wissenschaften*[27] koordiniert wird), die jeweils einen starken Fokus auf die digitale Erschließung und Analyse einzigartigen und vielfältigen Sprachmaterials legen. Als prominentes Beispiel ist das Projekt DGS-Korpus[28] zu nennen, das die umfassende Sammlung von Gebärdensprachdaten und deren Zusammenstellung in Form des öffentlichen DGS-Korpus zum Ziel hat.

Um eine solide Grundlage für die langfristige Verfügbarkeit vielfältiger sprachlicher Ressourcen für weltweite Forschungsgemeinschaften und die interessierte Öffentlichkeit zu schaffen, kooperiert die AdWHH mit dem Zentrum für nachhaltiges Forschungsdatenmanagement (ZFDM)[29]. Als zentrale Betriebseinheit an der Universität Hamburg stellt das ZFDM unter anderem eine lokale technische Infrastruktur (einschließlich eines Forschungsdatenrepositoriums[30]) zur Verfügung. Zudem betreiben die AdWHH und das ZFDM eine gemeinsame Infrastruktur zur Containervirtualisierung. Für die Nutzung des Forschungsdatenrepositoriums gibt es eine Datenübernahmevereinbarung[31], die über die technischen und rechtlichen Rahmenbedingungen aufklärt. Mit dem ZFDM kooperiert auch das an der Universität Hamburg angesiedelte Hamburger Zentrum für Sprachkorpora (HZSK), sodass sich mit der AdWHH und dem HZSK zwei Text+ Datenzentren eine gemeinsame Dienste-Infrastruktur teilen, die etwa Webinterfaces, OAI-Provider, FCS-Endpoints und Elasticsearch umfasst.

Die AdWHH ist spezialisiert auf Manuskripte, Gebärdensprache, historische Enzyklopädien (insbesondere griechisch-byzantinisch sowie klassisches Tamil), indigene Sprachen und frühmittelalterliche Vorlagen für Urkunden und Briefe. Ein weiterer Fokus sind Interdisziplinarität und linguistische Diversität. Die Daten können dabei gesprochen, gebärdet oder Manuskripte sein. Nach Absprache nimmt die AdWHH externe Daten an, die zu dieser Spezialisierung passen. Als Formate werden TEI/ISO und XML und nach Absprache ggf. auch weitere Formate akzeptiert. Ansprechperson ist Timm Lehmberg (timm.lehmberg@awhamburg.de).

[25] https://www.awhamburg.de/

[26] https://www.awhamburg.de/forschung/langzeitvorhaben.html

[27] https://www.akademienunion.de/

[28] http://www.dgs-korpus.de/

[29] https://www.fdm.uni-hamburg.de/

[30] https://www.fdr.uni-hamburg.de/

[31] https://doi.org/10.25592/uhhfdm.1

b Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

Für die BBAW[32] ist das Deutsche Textarchiv (DTA)[33] als Collections-Datenzentrum in Text+ vertreten. Wie Hug (2023) beschreibt, handelt es sich um ein Archiv für deutschsprachige, historische Korpora und Sammlungen, das am Zentrum Sprache[34] der BBAW angesiedelt ist und annotierte Volltexttranskriptionen von Drucken, Zeitungen und Zeitschriften sowie handgeschriebene Dokumente verschiedener Gattungen und Textarten umfasst. Für die Transkriptionen bietet das DTA mit dem DTABf[35] – einem Subset der TEI-P5-Guidelines – ein etabliertes Basisformat an.

Das DTA stellt aktuell etwa 40 Textsammlungen als Forschungsdaten zur Nachnutzung bereit. Im Zentrum steht das DTA-Kernkorpus[36], das mit rund 1.500 Werken die Grundlage für ein Referenzkorpus des Neuhochdeutschen vom 16. bis zum frühen 20. Jahrhundert darstellt. Projekte, die hochwertige Transkriptionen anfertigen, ein nachnutzbares Textformat verwenden, Metadaten bereitstellen und Lizenzfragen geklärt haben, finden im DTA eine etablierte Infrastruktur zur Bereitstellung ihrer Forschungsdaten. Das DTA berät zu allen Belangen, angefangen von Verfahren der Transkription, über die Annotation bis hin zur Dissemination der Forschungsdaten.

Das DTA ist dabei auf geschriebene deutschsprachige Texte spezialisiert, die ungefähr im Zeitraum zwischen 1600 und 1920 entstanden sind und eine offene Lizenz haben. Daten werden in den Formaten DTABf und TEI akzeptiert. Nach Absprache können unter Umständen auch Daten im XML, DOCX- oder EPUB-Format entgegengenommen werden. Ansprechpersonen sind Marius Hug (marius.hug@bbaw.de) und die Redaktion des DTA (redaktion@deutschestextarchiv.de).

Das DTA ist eng mit dem Digitalen Wörterbuch der deutschen Sprache (DWDS)[37] verbunden und innerhalb einer gemeinsamen Infrastruktur zugänglich. Daraus ergibt sich ein Korpusbestand, der mehr als 500 Jahre umfasst, sowie die Möglichkeit, die vom DTA bereitgestellten Forschungsdaten über die Korpustools des DWDS – z.B. die elaborierte DDC-Suchmaschine oder DiaCollo zur diachronen Untersuchung von Wortverbindungen – zu explorieren und zu analysieren. Das Zentrum Sprache der BBAW dient als Kompetenzzentrum für historische Texte und Daten sowie für Formatspezifikationen und Standardisierungsaktivitäten in den internationalen Fachcommunitys und wurde mit seinem CLARIN-Servicezentrum im September 2023 für weitere drei Jahre vom CoreTrustSeal Standards and Certification Board als vertrauenswürdiges Datenrepositorium zertifiziert.

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

c Deutsche Nationalbibliothek (DNB)

Die DNB[38] hat den Auftrag, so legt es das Gesetz über die DNB[39] fest, alle seit 1913 in Deutschland erschienenen Publikationen und Tonträger sowie Werke, die in deutscher Sprache erstellt wurden oder einen Bezug zu Deutschland haben, zu sammeln, zu dokumentieren und zu archivieren.

Zum Sammelgebiet der DNB gehören neben physischen Publikationen und Tonträgern auch soge-

[32] https://www.bbaw.de/

[33] https://www.deutschestextarchiv.de/

[34] https://www.bbaw.de/forschung/zentren/zentrum-sprache

[35] https://www.deutschestextarchiv.de/doku/basisformat/

[36] https://www.dwds.de/d/korpora/dtak

[37] https://www.dwds.de/

[38] https://www.dnb.de/

[39] vgl. https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html

nannte Netzpublikationen, zu denen zum Beispiel E-Books, E-Paper, Online-Hochschulschriften, Noten und Websites zählen. Von den 46.229.317 Objekten, die 2022 zum Gesamtbestand der DNB gehörten, waren 17.511.831 Monografien, 8.561.512 Zeitschriften/Zeitungen und 12.291.159 Netzpublikationen.

Die Inhaltsverzeichnisse aller Monografien des Verlagsbuchhandels werden seit 2008 digitalisiert und über den Katalog und über die Datendienste der DNB zur Verfügung gestellt. Sie sind im Volltext durchsuchbar. Welche Nutzungs- oder Zugriffsrechte für jede einzelne Publikation gelten, entscheiden die abliefernden Rechteinhabenden bzw. legt das geltende Urheberrecht fest.

Die DNB ist spezialisiert auf in Deutschland erschienene Publikationen und Tonträger sowie Werke, die in deutscher Sprache erstellt wurden oder einen Bezug zu Deutschland haben. Sie nimmt sowohl geschriebene als auch gesprochene Daten an, sofern sie durch den Sammelauftrag[40] der DNB abgedeckt sind, was im Einzelfall besprochen werden muss. Die akzeptierten Datenformate[41] sind ONIX for Books, XMetaDissPlus, NISO JATS, Crossref und DDEX. Ansprechpersonen sind Peter Leinen (p.leinen@dnb.de) und Philippe Genêt (p.genet@dnb.de).

Der Zugang zu den meisten Objekten in den Beständen der DNB ist aus urheberrechtlichen Gründen beschränkt und vielfach nur den DNB-Lesesälen möglich. Die DNB versucht jedoch, die Umsetzung von Forschungsprojekten verschiedenster Disziplinen so flexibel wie möglich zu erleichtern, und unterstützt Projekte gerne bei der Korpusbildung. Ferner bietet die DNB Beratung zu DH-Arbeitsplätzen, DH-Stipendien, DH-Calls und dem DNB Lab[42] an.

d Eberhard Karls Universität Tübingen (UniTÜ)

Die Datenressourcen der UniTÜ umfassen Korpora für gesprochene Sprache und geschriebene Texte, die auf verschiedenen Ebenen der linguistischen Annotation von Morphologie, Syntax und Semantik annotiert werden. Dazu kommen lexikalische Ressourcen, die eng mit anderen lexikalischen und textuellen Ressourcen, die in Text+ vertreten sind, verbunden sind. Die Korpora und lexikalischen Ressourcen sind für die datengetriebene Forschung sowohl in der theoretischen als auch in der Computerlinguistik unverzichtbar. Die Annotationen umfassen verschiedene grammatikalische Rahmen und halten sich an die in der Gemeinschaft weit verbreiteten Kodierungsstandards sowie an die Kodierungsstandards der International Standards Organization (ISO)[43]. Diese Ressourcen sind im TALAR[44]-Datenrepository verzeichnet, das CTS-zertifiziert wurde und standardisierte Protokolle für den Daten-Ingest von externen Datenressourcen verwendet.

Das Tübinger Data and Competence Centre beherbergt eine Sammlung von weit verbreiteten syntaktisch annotierten Korpora, die so genannten Tübinger Baumbanken für Deutsch, Englisch und Japanisch. Darüber hinaus enthält das Tübinger Archiv für Sprachressourcen (*Tübingen Archive of Language Resources,* TALAR) eine große Anzahl extern entwickelter Treebanks im Rahmen der Universal Dependencies[45]. Alle sprachlich annotierten Korpora der UniTÜ können mit der Webanwendung Tübingen Annotated Data Retrieval Application (TüNDRA)[46] (Martens, 2013) durchsucht und visualisiert werden, einige sind auch über die CLARIN Federated Content Search[47] zugänglich. Zusätzlich zu den sprachlich annotierten Korpora bietet die UniTÜ Datendienste in Form von Vektorraum-Wortdarstellungen und zugehörigen Softwaretools an. Darüber hinaus bietet sie

[40] https://www.dnb.de/sammelauftrag

[41] https://www.dnb.de/DE/Professionell/Sammeln/Unkoerperliche_Medienwerke/unkoerperliche_medienwerke_node.html#doc210120bodyText9

[42] https://www.dnb.de/dnblab/

[43] https://www.iso.org

[44] https://talar.sfb833.uni-tuebingen.de/

[45] https://universaldependencies.org/

[46] https://weblicht.sfs.uni-tuebingen.de/Tundra/

[47] https://contentsearch.clarin.eu/

Softwaredienste für die inkrementelle Annotation externer Textkorpora über die virtuelle Forschungsumgebung WebLicht[48] (Hinrichs et al., 2010) an. WebLicht ermöglicht u.a. die automatische Anreicherung von Textkorpora mit einer Named-Entity-Erkennung auf der Basis von Deep-Learning-Tools und kann somit als Werkzeug für die automatische Anreicherung unstrukturierter Daten und die anschließende Verknüpfung mit Normdaten sowie verknüpften offenen Daten genutzt werden.

Laut Trippel und Zinn (2023) liegt die Spezialisierung von TALAR auf Baumbanken/Treebanks, Wortnetzen und Word Embeddings. Es werden sowohl geschriebene als gesprochene Daten aufgenommen. Als Datenformate werden das GermaNet XML-Format, CoNLL-U und Word Embeddings akzeptiert; nach Absprache kommen auch andere Formate in Frage. Ansprechpersonen sind Thorsten Trippel (thorsten.trippel@uni-tuebingen.de) und Claus Zinn (claus.zinn@uni-tuebingen.de).

Qualitätsgesicherte Daten, deren Datentypen unterstützt werden, werden initial nach dem BagIT-Standard übermittelt. Eine Hilfeleistung kann das Werkzeug Bagman[49] (Zinn, 2022) bieten. Voraussetzung ist der Abschluss eines Datenüberlassungsvertrages; ein Muster[50] steht online zur Verfügung. Bei einer offenen Lizenz sollte diese CC-BY 4.0 oder höher betragen. Metadaten sollten im CMDI-Format vorliegen; aktuell werden die Profile CourseProfile, ExperimentProfile, LexicalResourceProfile, ResourceBundle, SpeechCorpusProfile, TextCorpusProfile, ToolProfile und WebLicht-WebService unterstützt.

e Leibniz-Institut für Deutsche Sprache (IDS)

Das IDS[51] in Mannheim besitzt die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit. Es handelt sich um die zentrale außeruniversitäre Einrichtung zur Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch und der neueren Geschichte. Zusammen mit 91 außeruniversitären Forschungs- und Serviceeinrichtungen gehört das IDS zur Leibniz-Gemein-schaft, einer der vier großen Forschungsorganisationen in Deutschland.

Für seinen Auftrag, die sprachliche Vielfalt, Struktur und Verwendung der deutschen Sprache zu dokumentieren, zu archivieren und zu erforschen, hat das IDS die wichtigsten Sammlungen des Gegenwartsdeutschen aufgebaut. Im Bereich der Schriftsprache enthält das Deutsche Referenzkorpus (DeReKo)[52] 46,9 Milliarden Wörter aus vielen verschiedenen Gattungen, darunter Zeitungen, wissenschaftliche Texte und Werke der Belletristik, aber auch aus der computervermittelten Kommunikation aus Chats und Usenet sowie Wikipedia. Im Bereich der gesprochenen Sprache bietet das Archiv für Gesprochenes Deutsch (AGD)[53] 46 Korpora mit mehr als 4000 Stunden Audio- und audiovisuellen Aufnahmen an, die z.B. Ressourcen zu Dialekten oder „umgangssprachlichen“ Variationen sowie zur Sprache von Auswanderern nach Israel und deutschsprachigen Minderheiten in Namibia oder Russland sowie z.B. das Wendekorpus zur deutschen Wiedervereinigung oder das GeWiss-Korpus der akademischen Rede enthalten. Das FOLK-Korpus[54] (Forschungs- und Lehrkorpus Gesprochenes Deutsch) bietet eine stratifizierte Auswahl einer großen Vielfalt an gesprochenem Deutsch in natürlichen Interaktionen.

Einige weitere Ressourcen, die insbesondere zur Langzeitarchivierung am IDS vorhanden sind, sind die gesprochensprachliche *Bonner Längsschnittstudie des Alterns* (BOLSA)[55], das schriftliche Lerner-

[48] https://weblicht.sfs.uni-tuebingen.de/

[49] https://weblicht.sfs.uni-tuebingen.de/bagman/

[50] https://uni-tuebingen.de/de/134320

[51] https://www.ids-mannheim.de/

[52] https://www1.ids-mannheim.de/kl/projekte/korpora/

[53] http://agd.ids-mannheim.de/index_en.shtml

[54] http://agd.ids-mannheim.de/folk.shtml

[55] https://repos.ids-mannheim.de/corpora/BOLSA/cmdi/BLSA.cmdi

korpus *Deutsch im Studium: Lernerkorpus* (DISKO)[56] und das multimodale, wissenschaftssprachliche Vorlesungskorpus *Mitschreiben in Vorlesungen: Ein multimodales Lehr-Lernkorpus* (MIKO)[57], welches neben geschriebener Sprache auch Audio- und Videoaufnahmen enthält.

Das IDS entwickelt Werkzeuge und Schnittstellen zur Abfrage und Analyse der Korpora: Für gesprochene Korpora ist die Datenbank für Gesprochenes Deutsch (DGD)[58] die zentrale Schnittstelle mit rund 12.000 registrierten Nutzenden. Für schriftliche Korpora wird die Korpusanalyseplattform KorAP[59] verwendet. KorAP ist für große, mehrfach annotierte Korpora und komplexe Suchmechanismen optimiert und unterstützt mehrere Abfragesprachen, im Frühjahr 2023 gab es über 54.000 registrierte Nutzende. Was lexikalische Daten betrifft, stellt das IDS das Online-Wortschatz-Informationssystem Deutsch (OWID)[60] zur Verfügung – ein Wörterbuchportal, das den Zugang zu verschiedenen am IDS erarbeiteten Wörterbüchern ermöglicht.

Im Hinblick auf die Übernahme von weiteren Daten ist das IDS auf das Neuhochdeutsche und sowohl auf geschriebene als auch gesprochene Sprache spezialisiert. Ein besonderes Interesse gilt linguistisch annotierten Korpora. Ein weiterer Fokus liegt auf Deutsch in nicht-primär deutschsprachigen Ländern. Bezüglich anderer Daten, die ungefähr zum Profil des IDS passen könnten, ist ebenfalls eine Absprache möglich.

Textuelle Daten werden am IDS im I5-Format[61] gespeichert, welches auf dem TEI-Standard P5[62] basiert. Daten, die in diesem Format vorliegen, werden daher präferiert. Was gesprochene Sprache betrifft, erfolgt eine Transkription nach ISO 24624:2016. Das Ergebnis sind XML-Dokumente, die den TEI-Guidelines entsprechen. Zudem liegen die Signaldateien in gängigen Formaten wie Waveform Audio File Format (WAV) vor. Bezüglich weiterer möglicher Dateiformate wird um Absprache gebeten. Ansprechpersonen sind Andreas Witt (witt@ids-mannheim.de), Thorsten Trippel (trippel@ids-mannheim.de) und Antonina Werthmann (werthmann@ids-mannheim.de).

Über die Möglichkeit der Datenübernahme wird anhand der detaillierten Datenübernahmerichtlinien[63] des IDS entschieden. Wichtig ist eine gesicherte Qualität der Daten, welche in unterstützten Datentypen vorliegen sollten. Unverzichtbar ist auch, dass die Rechte geklärt sind. Bei einer offenen Lizenz sollte diese CC-BY 4.0 oder höher sein. Metadaten sollten ISO 24622-1/-2 entsprechen, d.h. im CMDI-Format vorliegen. Weichen die Gegebenheiten von diesen Idealen ab, ist eine Datenübernahme nach Absprache möglich. Kommt es zu einer Einigung zwischen Datengebenden und Datennehmenden, ist ein Datenüberlassungsvertrag zu schließen.

Nach der Klärung der Formate, Lizenzen und dem Abschluss eines Datenübernahmevertrags erfolgt der Ingest ins Repositorium, welcher Voraussetzung für die Langzeitarchivierung ist. Den Daten wird bei diesem Schritt eine PID vergeben. Am Ende des Prozesses stehen neue Zugangsmöglichkeiten zu den Daten. So werden diese über Forschungsdatensuchmaschinen wie die CLARIN-VLO oder die Text+ Registry auffindbar gemacht. Welche weiteren Zugangsmöglichkeiten konkret geschaffen werden können, hängt von Faktoren wie den rechtlichen Möglichkeiten/Lizenzen, Datentypen, Formaten und Umfängen der Daten ab. Gegebenenfalls ist ein Zugang über die Text+ FCS oder IDS-basierte Suchwerkzeuge denkbar. In bestimmten Fällen kann auch ein Rohdatenzugang ermöglicht werden. Sind die nötigen Voraussetzungen erfüllt, können die Daten mithilfe von autorisierten Werkzeugen weiterverarbeitet werden.

---

[56] http://repos.ids-mannheim.de/corpora/DISKO/cmdi/object000000.cmdi

[57] http://repos.ids-mannheim.de/corpora/MIKO/cmdi/object000000.cmdi

[58] https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome

[59] https://korap.ids-mannheim.de/

[59] https://www.ids-mannheim.de/

[60] https://www.owid.de/

[61] https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell

[62] https://tei-c.org/guidelines/p5/

[63] https://doi.org/10.14618/ids-pub-8791

f Ludwig-Maximilians-Universität München (LMU)

Als öffentliche Einrichtung an der LMU repräsentiert das Bayerische Archiv für Sprachsignale (BAS)[64] diese in Text+. Das BAS wurde 1995 mit dem Ziel gegründet, gesprochensprachliche Daten und Sprachverarbeitungsdienste sowohl für die Technologieentwicklung als auch für die Grundlagenforschung zur Verfügung zu stellen. Die vom BAS bereitgestellten Ressourcen umfassen: ein Repository für Sprachdatenbanken, webbasierte Dienste für die Sprachverarbeitung, eigenständige Tools zur Datensammlung und -analyse sowie Beratung, Ausbildung und Unterstützung im Bereich Datenmanagement.

Das Repository enthält derzeit mehr als 40 Sammlungen von Sprachdaten in mehreren Sprachen (Deutsch, Englisch, Japanisch, Italienisch usw.) aus akademischen und industriellen Kooperationsprojekten (z.B. Verbmobil[65], SmartKom[66]) bzw. von Dritten (z.B. Deutsche Telekom, Uni Tübingen, Uni Zürich, MPI Magdeburg). Die Webdienste bieten akademischen Nutzenden einen einfachen Zugang zu mehr als 15 komplexen Sprachverarbeitungsprozessen, z.B. WebMAUS[67] zur Alignierung von Audio und Text auf Wort- und Lautebene, ASR[68] für die KI-basierte automatische Spracherkennung durch akademische und kommerzielle externe Anbieter, Anonymizer[69] zu automatischen Anonymisieren von Sprachaufnahmen durch das Ersetzen von Wörtern im Signal durch ein Rauschen. Ein besonderer Dienst ist COALA[70], der auf der Basis vordefinierter Tabellen automatisch die für das Repository notwendigen Metadaten generiert und somit den Import neuer Korpora erheblich erleichtert. Zu den Tools für die Datensammlung und -analyse gehören WikiSpeech[71] zur webbasierten skriptgesteuerten Sprachaufnahme, Octra[72] zur effizienten orthografischen Transkription gesprochener Sprache, und EMU WebApp[73] zur Einbindung phonetischer Analysen in das Statistikpaket R.

Das BAS ist mit dem CoreTrustSeal zertifiziert. Das Repository wird regelmäßig von Index-Diensten gescannt und unterliegt einem Verfügbarkeitsmonitoring durch CLARIN Europa. Alle Dienste sind sowohl über ein grafisches Interface als auch per API erreichbar.

Die Spezialisierung des BAS liegt, wie Draxler und Schiel (2023) es beschreiben, auf gesprochener Sprache, für welche sowohl ein Repository als auch sprachtechnologische Webdienste zur Verfügung stehen. Neben Sprachaufnahmen in Formaten wie WAV, MP4, AIF oder als raw audio file werden auch zeitalignierte Annotationen zu diesen Aufnahmen angenommen. Für diese kommen Formate wie Textdateien, BPF, AnnotJSON, TextGrid, TEI oder CSV in Frage. Es wird um Absprache mit Christoph Draxler (draxler@phonetik.uni-muenchen.de) und/oder Florian Schiel (schiel@phonetik.uni-muenchen.de) gebeten.

g Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)

Wie in Barth et al. (2023) beschrieben, betreibt die SUB zwei unterschiedliche Repositorien, die für Text+ relevant sind, nämlich das DARIAH-DE Repository und das TextGrid Repository.

Das DARIAH-DE Repository[74] ist ein digitales Langzeitarchiv für geistes- und kulturwissenschaftliche

[64] https://www.bas.uni-muenchen.de/Bas/BasHomedeu.html

[65] https://www.phonetik.uni-muenchen.de/forschung/Verbmobil/Verbmobildeu.html

[66] https://www.dfki.de/web/forschung/projekte-publikationen/projekt/smartkom/

[67] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSGeneral

[68] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/ASR

[69] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Anonymizer

[70] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Coala

[71] hhttps://webapp.phonetik.uni-muenchen.de/wikispeech/

[72] https://clarin.phonetik.uni-muenchen.de/apps/octra

[73] https://ips-lmu.github.io/EMU-webApp/

[74] https://de.dariah.eu/en/repository

Forschungsdaten, das seit 2017 von der SUB betrieben wird. Jedem archivierten Objekt wird eine DataCite DOI zugewiesen, über die es dauerhaft referenzier-, zitier- und verfügbar ist. Über den DARIAH-DE Publikator[75] ist es möglich, Daten vorzubereiten, zu verwalten und zu importieren.

Das DARIAH-DE Repository ist für generische Forschungsdaten und Kollektionen ideal geeignet. Bezüglich der Modalität der angenommenen Daten gibt es keine Einschränkungen. Darüber hinaus werden alle Datenformate akzeptiert, wobei besonders für die Langzeitarchivierung und Nachnutzung geeignete Datenformate empfohlen werden. Für mehr Informationen können sich Interessierte an info@de.dariah.eu wenden.

Das TextGrid Repository[76] ist ein Langzeitarchiv für geisteswissenschaftliche Forschungsdaten, das seit 2011 von der SUB betrieben wird. Es liefert einen umfangreichen, durchsuch- und nachnutzbaren Bestand an Texten und Bildern. An den Grundsätzen von Open Access und den FAIR-Prinzipien orientiert, wurde das TextGrid Repository 2020 mit dem CoreTrustSeal versehen. Für Forschende bietet das TextGrid Repository eine nachhaltige, dauerhafte und sichere Möglichkeit zur zitierfähigen Publikation ihrer Forschungsdaten und zur verständlichen Beschreibung derselben durch erforderliche Metadaten. Als PIDs werden ePIC Handles vergeben.

Im Mission Statement[77] des TextGrid Repository sind weitere Informationen zum Thema Nachhaltigkeit, FAIR und Open Access zu finden. Der Bestand wurde mit dem Erwerb der Digitalen Bibliothek aufgebaut und entwickelt sich fortwährend auf der Basis der TextGrid Community. Durch zahlreiche Editionsprojekte, die in der virtuellen Forschungsumgebung des TextGrid Laboratory[78] entstehen, sind sowohl Manuskripte (Bilder) als auch Transkriptionen (XML/TEI-kodierte Textdaten) vorhanden (z.B. die Bibliothek der Neologie[79] oder auch das Projekt zur deutsch-französischen Reisekorrespondenz ARCHITRAVE[80]). Das TextGrid Repository bietet anpassbare Projektbereiche und projektspezifische Facetten, was die umfangreichen Suchfunktionalitäten des Repositoriums betrifft. Ein Export der Daten in gängige Formate ist möglich.

Das TextGrid Repository ist also auf XML/TEI-kodierte Texte und Bilder spezialisiert und für digitale Editionen besonders geeignet. Als Datenformate werden XML (TEI), unformatierter Text (Unicode) und Rastergrafiken (bis TIFF 6.0) akzeptiert. Auch was das TextGrid Repository betrifft, ist info@de.dariah.eu die richtige Adresse für An- und Nachfragen.

Es wurden ebenfalls neue Korpora in TextGrid hinzugefügt: Die European Literary Text Collection (ELTeC) ist eine Sammlung mehrerer Korpora in verschiedenen europäischen Sprachen, die von der COST Action Distant Reading zusammengestellt wurde. Die Veröffentlichung im TextGrid Repositorium verbessert den FAIR-Status dieser Korpora und erhöht die Mehrsprachigkeit des Repositoriums. Entlang dieser Koprora wurden verschiedene neue Funktionalitäten entwickelt und getestet, darunter eine Überarbeitung des Import-Workflows, projektspezifischer Seiten und der Facettierung, die Verwendung der bibliothekarischen Basisklassifikation zur Erschließung der Texte oder neue Suchoptionen in Bezug auf die GND-Normdaten. Diese und weitere Entwicklungen wurden auf mehreren einschlägigen Tagungen vorgestellt, z.B. auf dem Hispanistentag, FORGE oder der internationalen Tagung Digital Humanities zusammen mit zwei Verantwortlichen von ELTeC. Es wird weiter daran gearbeitet, sowohl die neue Features zu dokumentieren, weiter zu entwickeln, und in Workshops mit der Community zu testen, als auch den neuen Import-Workflow zu etablieren und weitere Korpora zu veröffentlichen.

[75] https://repository.de.dariah.eu/publikator/

[76] https://textgridrep.org/

[77] https://textgridrep.org/docs/mission-statement

[78] https://de.dariah.eu/web/guest/textgridlab

[79] https://bdn-edition.de/index.html

[80] https://architrave.eu/index.html?lang=de

h Universität des Saarlandes (UdS)

Als Daten- und Kompetenzzentrum ist die UdS spezialisiert auf Korpora geschriebener Sprache mit Schwerpunkt auf fremdsprachlichen und multilingualen Korpora, zum Beispiel Übersetzungs- und Dolmetschkorpora. Einen weiteren Schwerpunkt bilden registerspezifische Korpora, welche verschiedene Aspekte sprachlicher Variation sowohl bezüglich Disziplinen und Gattungen (Wissenschaft, Belletristik) als auch bezüglich der Modalität (gesprochen, geschrieben) abbilden. In kleinerem Umfang werden auch generierte Daten, zum Beispiel Daten aus maschineller Übersetzung oder Sprachmodelle, archiviert. Das Datenzentrum ist ein CLARIN-Zentrum Typ B und mit dem CoreTrustSeal zertifiziert. Mehr als 100 Datenressourcen sind bereits im Repositorium der UdS archiviert. Die Ressourcen sind über das Virtual Language Observatory auffindbar und eine Auswahl der archivierten Korpora ist zudem lokal über einen CQPweb-Server[81] und projektübergreifend über die Federated Content Search durchsuchbar. Hiervon sind im Zusammenhang mit Text+ zwei diachrone Korpora für das Englische hervorzuheben: das Royal Society Corpus (RSC)[82] und das Old Bailey Corpus (OBC)[83].

Das RSC beinhaltet wissenschaftliche Publikationen aus den Jahren 1665 bis 1920, die in den Transactions und Proceedings der Royal Society of London veröffentlicht wurden. Das Korpus wurde umfangreich auf Text-, Satz- und Tokenenbene annotiert und umfasst 78,6 Millionen Tokens.

Das OBC dokumentiert gesprochenes Englisch aus zwei Jahrhunderten (1720 bis 1913) und basiert auf Verhandlungsprotokollen des zentralen Strafgerichtshofs in London. Die Texte des OBC umfassen 24,4 Millionen Tokens und wurden mit soziobiografischen und pragmatischen Annotationen versehen. Aufgrund ihrer freien Lizenz, Größe und breiten Nutzung in der Forschung sind diese Datenressourcen für eine Übernahme in Text+ besonders relevant. Weiterhin enthält das Repositorium der UdS Übersetzungskorpora, darunter EuroParl-UdS[84] und EPIC-UdS[85], sowie eine Reihe slawischer Ressourcen.

Das Repositorium der UdS, CLARIND-UdS[86], ist in Knappen et al. (2023) und im Blogeintrag von Fischer et al. (2023) näher beschrieben, dort wird auch auf Zugangswege für interessierte Datengeber:innen eingegangen. Das Repositorium ist seit 2013 in Betrieb und seither zertifiziert, zunächst mit dem Data Seal of Approval (einem der beiden Zertifikate, die später in das CoreTrustSeal (CTS) eingeflossen sind), in der Zwischenzeit durch das CTS. Die letzte Zertifizierung des Repositoriums hat 2019 durch das CoreTrustSeal stattgefunden; die Rezertifizierung steht kurz vor dem Abschluss. Zudem ist CLARIND-UdS in relevanten nationalen und internationalen Verzeichnissen von Forschungsdatenrepositorien aufgelistet. Hierzu gehören der Eintrag in RIsources[87] bei der Deutschen Forschungsgemeinschaft und der Eintrag in re3data,[88] der wichtigsten internationalen Registratur für Repositorien. Eine Liste der akzeptierten Datenformate ist auf der Webseite des Repositoriums zu finden.[89] Als Metadatenformate werden Dublin Core und CMDI verwendet. Ansprechpersonen der UdS sind Elke Teich (e.teich@mx.uni-saarland.de) und Jörg Knappen (j.knappen@mx.uni-saarland.de).

[81] https://corpora.clarin-d.uni-saarland.de/cqpweb/

[82] http://hdl.handle.net/21.11119/0000-0004-8E37-F

[83] http://hdl.handle.net/11858/00-246C-0000-0023-8CFB-2

[84] http://fedora.clarin-d.uni-saarland.de/europarl-uds/

[85] http://fedora.clarin-d.uni-saarland.de/epic-uds/

[86] https://fedora.clarin-d.uni-saarland.de/

[87] https://risources.dfg.de/detail/RI_00435_de.html

[88] https://www.re3data.org/repository/r3d100010384

[89] https://fedora.clarin-d.uni-saarland.de/ressources/AcceptedFormats.en.pdf

i Universität Duisburg-Essen (UniDUE)

Das PolMine-Projekt[90] der Universität Duisburg-Essen[91] ist auf Sammlungen parlamentarischer Textdaten spezialisiert. Das Ziel ist es, der Forschung neue Horizonte zu erweitern, indem digital verfügbare politisch relevante Texte erschlossen und als Korpora verfügbar gemacht werden. Als Kompetenzzentrum für parlamentarische Sprachdaten stellt das Projekt Expertise und Werkzeuge zur Erstellung von Korpora auf Basis von Protokollen des politischen Diskurses bereit und leistet Unterstützung bei entsprechenden Vorhaben.

Die wichtigste Ressource des Projekts ist das GermaParl-Korpus[92], eine digitale Sammlung der Plenarprotokolle des Deutschen Bundestags. GermaParl v2.0.0 wurde am 23. Mai 2023 veröffentlicht und umfasst 19 Legislaturperioden, 72 Jahre (von 1949 bis 2021), 4.341 Protokolle und 273 Millionen Tokens.

Die Aufbereitung von GermaParl erfolgt durch einen reproduzierbaren Workflow, bei dem eine eigens entwickelte Toolchain zum Einsatz kommt. Die Rohdaten, welche zunächst in den Formaten PDF, XML oder TXT vorliegen, werden in einem ersten Schritt mit trickypdf[93] vorverarbeitet, um sie grundlegend zu bereinigen. Hierbei werden unter anderem Kolumnentitel, Seitenzahlen, zweispaltige Formatierung, Inhaltsverzeichnisse und Anhänge entfernt. Das Ausgabeformat ist plain text. Mithilfe von frappp erfolgt als nächstes eine Umwandlung zu XML. Hierbei wird die Debattenstruktur (Sprecher:innenwechsel, Zwischenrufe etc.) aus unstrukturiertem Text rekonstruiert. Im Anschluss werden externe Daten zur Konsolidierung und Anreicherung hinzugezogen. Die vollständigen Namen der Mitglieder des Bundestages und des Präsidiums werden den Stammdaten des Bundestages entnommen. Für die vollständigen Namen anderer Sprecher:innen und Parteiaffiliationen wird Wikipedia herangezogen. Der nächste Schritt ist die linguistische Annotation mithilfe von bignlp[94]. Stanford CoreNLP (Manning et al., 2014) wird verwendet, um den Text in Sätze und Tokens zu segmentieren sowie POS-Tagging mit Universal POS und Named Entity Recognition durchzuführen. Für Lemmatisierung und POS-Tagging mit dem Stuttgart-Tübingen-Tagset (STTS) kommt TreeTagger (Schmid, 1994; 1995) zum Einsatz. Schließlich werden die linguistisch annotierten Daten in die Corpus Workbench (CWB)[95] importiert.

GermaParl steht als XML und linguistisch annotiert und indiziert im CWB-Datenformat frei unter der Lizenz CC-BY-SA über Zenodo zur Verfügung[96]. Die Toolchain für die Reproduktion der Daten und für andere Aufbereitungsprojekte wird über GitHub-Repositorien bereitgestellt. Die wachsende Nutzungscommunity wird aktiv in die Weiterentwicklung der Ressourcen eingebunden, um sie nutzungsfreundlicher zu gestalten und die Qualität der Daten zu erhöhen.

Die Ressource soll kontinuierlich gepflegt, qualitativ verbessert und aktualisiert werden. In GermaParl v3 soll die Interoperabilität des Korpus durch die Adaption des ParlaMint-TEI-XML-Standards verbessert werden. Durch die Evaluation und Nutzung alternativer NLP-Pipelines (insbesondere MONAPipe[97]) sollen die Annotationsschichten in Zukunft erweitert werden. Ein weiterer Plan ist die Nutzung des Korpus als Beispiel- und Referenzdatensatz für die im KonsortSWD-Measure “Linking Textual Data”[98] entwickelten Tools und Workflows.

Das PolMine-Projekt ist eine treibende Kraft für textbasierte Forschung in der Politikwissenschaft zu

[90] https://polmine.github.io/

[91] https://www.uni-due.de/

[92] https://polmine.github.io/GermaParl/

[93] https://github.com/PolMine/trickypdf

[94] https://github.com/PolMine/bignlp

[95] https://cwb.sourceforge.io/

[96] https://zenodo.org/record/7949074

[97] Dönicke et al. (2022); https://gitlab.gwdg.de/text-plus-collections/mona-pipe

[98] https://www.konsortswd.de/konsortswd/das-konsortium/services/linking-textual-data/

Policy und Politik. Durch die linguistische Annotation sind die Sammlungen auch für sprachwissenschaftliche und zeitgeschichtliche Forschung relevant. Einen Einstieg in die sozialwissenschaftliche Arbeit mit Textdaten bieten die Online-Foliensätze „Using Corpora in Social Science Research“[99] sowie die Webinar-Reihe „Cookin’ with GermaParl“.

Ergänzend zu den Sammlungen bietet das Projekt zugehörige Software-Tools an. Das polmineR-Paket, das in der statistischen Programmiersprache R implementiert und über das Comprehensive R Archive Network (CRAN)[100] verfügbar ist, gewährleistet, dass eine Umgebung für die Analyse von Parlamentsdebatten funktional und vollständig interoperabel ist und dass qualitative und quantitative Analyseschritte interaktiv kombiniert werden können. Zudem ist das Projekt in einer sich entwickelnden mehrsprachigen Forschungsgemeinschaft zur parlamentarischen Sprache aktiv.

Parlametarische Materialien sind gemeinfrei, dadurch besteht die Möglichkeit der Nutzung von offenen Verzeichnisdiensten wie GitHub und Zenodo. PolMine ist als CLARIN-Zentrum des Typs C registriert.

Als Text+ Datenzentrum ist PolMine auf die Aufbereitung, Bereitstellung, Analyse und Verknüpfung (im Sinne von Data Linkage) von parlamentarischen Sprachdaten spezialisiert. Es handelt sich um konzeptionell schriftliche Daten, die in schriftlicher Form vorliegen. Daten werden in den Formaten PDF, TXT, XML und CWB entgegengenommen. Ansprechpersonen sind Andreas Blätte (andreas.blaette@uni-due.de) und Stine Ziegler (stine.ziegler@uni-due.de).

j Universität Hamburg (UniHH)

Die Universität Hamburg ist Text+ mit langjähriger Erfahrung aus dem Hamburger Zentrum für Sprachkorpora (HZSK) beteiligt. Dabei bietet die Universität Hamburg mit der Community HZSK innerhalb des Repositorium des Zentrums für nachhaltiges Forschungsdatenmanagement (FDR) eine institutionelle Basis, um die nachhaltige Nutzbarkeit sprachwissenschaftlicher Primärforschungsdaten über zeitlich befristete Forschungsprojekte hinaus zu gewährleisten. Als ein Zusammenschluss von Mitgliedern verschiedener Fakultäten und Institutionen der Universität Hamburg unterstützte das HZSK die Konsistenz und Koordination computergestützter empirischer Forschung und Lehre der Sprachwissenschaft sowie der an die Universität Hamburg angegliederten Nachbardisziplinen über die Projektlaufzeiten hinaus. Fragen der empirischen Forschung sowie des Forschungsdatenmanagements werden mittlerweile für Studierende sprachwissenschaftlicher Studiengänge im Rahmen fachwissenschaftlicher Lehrveranstaltungen des Arbeitsbereichs Deutsch als Fremd- und Zweitsprache sowie im inhaltlichen Schwerpunkt „Digitale Linguistik“ (Androutsopoulos, Zinsmeister) behandelt.

Das Repositorium der Universität Hamburg, welches auf der Plattform Invenio basiert und durch das Zentrum für nachhaltiges Forschungsdatenmanagement (ZFDM) betrieben wird, beherbergt in der HZSK Community mehr als 50 Korpora, die mehrheitlich dem thematischen Bereich der mehrsprachigen mündlichen und schriftlichen Daten sowie Daten aus weniger verbreiteten oder gefährdeten Sprachen angehören. Neben einer Vielzahl von (Kinder-)Spracherwerbskorpora und anderen Korpora, die sich auf einzelne Aspekte der Mehrsprachigkeit konzentrieren, werden weitere hochrelevante Themen zu den gesellschaftlichen Aspekten der Mehrsprachigkeit abgedeckt, z.B. durch die Korpora Dolmetschen im Krankenhaus (DiK) und die Community Interpreting Database (ComInDat). Das verwendete Metadatenschema ist CMDI und als PIDs werden den Daten DOIs zugewiesen. Die technischen und rechtlichen Rahmenbedingungen für die Nutzung des Repositoriums sind in einem Dokument festgehalten worden. Mit der AdWHH, welche ebenfalls mit dem ZFDM kooperiert, teilt sich das HZSK eine gemeinsame Dienste-Infrastruktur, zu der etwa Webinterfaces, OAI-Provider, FCS-Endpoints und Elasticsearch gehören.

[99] https://polmine.github.io/UCSSR/#1

[100] https://cran.r-project.org/

Die Universität Hamburg ist vor allem auf Korpora gesprochener Sprache und Mehrsprachigkeit spezialisiert. Externe Daten nimmt es in den Formaten TEI/ISO und XML nach Absprache entgegen. Ansprechpersonen sind Kristin Bührig, Marcel Fladrich und Alina Hemmer (corpora@uni-hamburg.de).

k Universität zu Köln (UniK)

Das Data Center for the Humanities (DCH) ist eine zentrale Einrichtung der Philosophischen Fakultät der Universität zu Köln und berät und unterstützt Wissenschaftler:innen an der Universität und darüber hinaus bei Fragen der dauerhaften Sicherung, Verfügbarkeit und Präsentation von Daten und Ergebnissen der geisteswissenschaftlichen Forschung. In Abstimmung mit lokalen Partnern ergänzt das DCH das Forschungsdatenmanagement (FDM) an der Fakultät mit einem auf die Geisteswissenschaften zugeschnittenen Profil.

Im Rahmen seiner Möglichkeiten bietet das DCH verschiedene Datenpublikations- und Archivierungsservices für geisteswissenschaftliche Forschungsdaten: Dazu gehören die Datenpublikation über das Language Archive Cologne (LAC) oder über Zenodo, sowie die Datenarchivierung durch den DCH Archiving Service. Darüber hinaus unterstützt das DCH aktiv projektspezifische Datenpublikationsprozesse in Fachrepositorien.

Das Language Archive Cologne (LAC) ist ein CTS-zertifiziertes Fachrepositorium zur Publikation von audiovisuellen Sprachdaten. Es handelt sich ebenfalls um ein CLARIN-Zentrum sowie einen Teil des Digital Endangered Languages and Musics Archives Network (DELAMAN)[101]. Die Spezialisierung des LAC liegt auf Sprachdokumentation, Oral Literature und außereuropäischen Sprachen. Als Datenformate werden gängige Audio- und Video-Formate sowie Community-Standards für Annotation akzeptiert. Genauere Informationen dazu gibt es in den User Guides[102], in denen eine Format Whitelist[103] sowie ein Metadata Template[104] bereitgestellt werden. Das LAC steht auch für Projektbegleitung zur Vefügung und berät in diesem Zusammenhang zu Datenkuration und Korpusaufbau. Auch begleitetes Data Depositing sowie Korpus- und Metadaten-Validierung sind Teil dieses Angebots. Das LAC speichert und versioniert Daten als OCFL-Datenobjekte. Als PIDs werden für alle Korpuskomponenten und Dateien Handles und für ganze Sammlungen/Korpora DOIs vergeben. Für zugangsbeschränkte Daten ist eine Zugangskontrolle eingebaut. Ansprechpersonen sind Felix Rau und das gesamte LAC-Team, welches unter lac-helpdesk@uni-koeln.de zu erreichen ist.

Der DCH Data Publishing Support (via Zenodo) ermöglicht die Publikation großer Sammlungen von Einzeldatensätzen auf Zenodo. Der Schwerpunkt liegt hier auf Paperpublikation und Sammlungen von Einzeldatensätzen aus dem geisteswissenschaftlichen Bereich. Als Datenformate werden Communty-Standards akzeptiert. Der DCH Publishing Support ist nicht nur in die DCH-Services, sondern auch in den Verband Digital Humanities im deutschsprachigen Raum (DHd) eingebunden. Es werden projektbegleitende Beratung zur Daten- und Metadatenkuration, begleitetes Data Depositing sowie Korpus- und Metadatenvalidierung angeboten. Ansprechpersonen sind Patrick Helling und das gesamte DCH-Team, welches unter info-dch@uni-koeln.de zu erreichen ist.

Der DCH Archiving Service bietet eine einfache und niederschwellige Datenarchivierung zur Sicherung und zum Nachweis von Projektdaten und -ergebnissen aus dem geisteswissenschaftlichen Bereich. Die Daten werden als BagIt-Datenobjekte gespeichert; optional ist ein Packaging als FAIR Digital Object-kompatibles RO-Crate möglich. Das Data Depositing kann einmalig oder inkrementell sein. Der Service umfasst optional eine Metadaten-Publikation, DOI-Registrierung und sichere Verschlüsselung der gespeicherten Daten. Der Schwerpunkt liegt bei der projektbegleitenden

[101] https://www.delaman.org/

[102] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides

[103] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist

[104] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/metadata-template

oder ergebnissichernden Datensicherung, nicht auf der Schaffung von Zugriffsmöglichkeiten. Der Service ist also vor allem für Cold und Dark Archiving geeignet. Als Datenformate werden Community-Standards akzeptiert; eine Beratung hierzu kann in Anspruch genommen werden. Ansprechpersonen sind Felix Rau und das Team des DCH; die Kontaktadresse lautet dch-datensicherung@uni-koeln.de

6 Übersicht der Datenzentren

Datenzentrum	**Spezialisierung**	**Akzeptierte Daten**
Akademie der Wissenschaften in Hamburg (AdWHH)	- Korpora gesprochener und gebärdeter Sprache (Schwerpunkt auf Sprachvariation und Sprachdiversität) - Manuskripte - historische Enzyklopädien (insbes. griechisch-byzantinisch) - nordeurasische indigene Sprachen - frühmittelalterliche Vorlagen für Urkunden und Briefe - Interdisziplinarität und linguistische Diversität	TEI/ISO und XML, weitere Formate nach Absprache
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)	- historische (ca. 1600–1920) deutschsprachige Texte - hochwertige Transkriptionen - weiterverarbeitbares, strukturiertes Textformat - saubere Metadaten, sowohl auf Objekt- wie auch auf Sammlungsebene - offene Lizenz - Dokumentation	DTABf, TEI-XML, nach Abstimmung ggf. DOCX oder EPUB
Deutsche Nationalbibliothek (DNB)	- in Deutschland erschienene Publikationen und Tonträger sowie Werke, die in deutscher Sprache erstellt wurden oder einen Bezug zu Deutschland haben - literarische Texte - Sachtexte - Zeitungs- und Zeitschriftentexte - Normdaten - Wörterbücher - Enzyklopädien - (gedruckte) Editionen - monomodale Sprachaufnahmen (Hörbücher)	ONIX for Books, XMetaDissPlus, MARCXML, NISO JATS, Crossref, DDEX (siehe auch hier[105])

[105] https://www.dnb.de/DE/Professionell/Sammeln/Unkoerperliche_Medienwerke/unkoerperliche_medienwerke_node.html#doc210120bodyText9doc210120bodyText9

Eberhard Karls Universität Tübingen (UniTÜ)	- Korpora gesprochener und geschriebener Sprache, morphologisch, syntaktisch und semantisch annotiert - Baumbanken/Treebanks - Word Embeddings	Wortnetze in GermaNet-artigem XML, Baumbanken in CoNLL-U, Korpora nach TEI (nach Absprache), Word Embeddings, weitere Formate nach Absprache
Leibniz-Institut für Deutsche Sprache (IDS)	- neuhochdeutsche Referenzkorpora (gesprochen, geschrieben) - linguistisch annotierte Korpora - Deutsch in nicht primär deutschsprachigen Ländern	TEI-XML (I5), Standard-Formate für gesprochene Sprache, Transkription nach nach ISO/TEI (ISO 24624:2016), weitere Formate nach Absprache
Ludwig-Maximilians-Universität München (LMU)	- gesprochensprachliche Daten	Metadaten in CMDI; WAV, MP4, AIF, raw; Textdateien, BPF, AnnotJSON, TextGrid, TEI oder CSV (für weitere Formate s.[106])
Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)	- Digitalisate - Reisezeitschriften - digitale indische Sprachdaten - DARIAH-DE Repository: geistes- und kulturwissen-schaftliche Forschungsdaten und Kollektionen aller Typen - TextGrid Repository: XML/TEI-kodierte Texte und Bilder	TIFF, JPG, PDF, TEI-XML
Universität des Saarlandes (UdS)	- annotierte Korpora geschriebener Sprache, insbesondere fremdsprachliche und multilinguale Sprachkorpora (parallele und vergleichbare Korpora) und registerspezifische Korpora	vrt (Open Corpus Workbench, CQP), siehe Liste[107] der akzeptierten Formate
Universität Duisburg-Essen (UniDUE)	- gesprochensprachliche Daten mit politischem Bezug	CWB, TEI-XML
Universität Hamburg (UniHH)	- (annotierte) Korpora gesprochener Sprache - (annotierte) Korpora mit mehrsprachigen Daten	TEI/ISO und XML, weitere Formate nach Absprache
Universität zu Köln (UniK)	- audiovisuelle Daten mit Schwerpunkt auf Sprachaufnahmen - Spezialisierung auf kleine und bedrohte Sprachen - Oral Literature - Sprachdokumentation	siehe Metadata Template[108], Format Whitelist[109] und Basic Language Archive Metadata Bundle (BLAM) profile[110]

[106] https://www.bas.uni-muenchen.de/forschung/Bas/BasFormatsdeu.html

[107] https://fedora.clarin-d.uni-saarland.de/ressources/AcceptedFormats.en.pdf

[108] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/metadata-template

[109] https://dch.phil-fak.uni-koeln.de/bestaende/language-archive-cologne/user-guides/format-whitelist

[110] https://catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu%3Acr1%3Ap_1475136016193&registrySpace=public

7 Literatur

Barth, F., Calvo Tello, J., Funk, S. E., Göbel, M., Kurzawe, D., Veentjer, U., & Weimer, L. (2023). Die SUB Göttingen als Datenzentrum innerhalb der Task Area Collections im NFDI-Konsortium Text+. Zenodo. https://doi.org/10.5281/zenodo.8108827

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, *19*(43), 1-12. https://doi.org/10.5334/dsj-2020-043

Dönicke, T., Barth, F., Varachkina, H., & Sporleder, C. (2022). MONAPipe: Modes of narration and attribution pipeline for German computational literary studies and language analysis in spaCy. *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022),* 8-15. https://aclanthology.org/2022.konvens-1.2

Draxler, C., & Schiel, F. (2023). Wohin mit... Sprachdaten? Bayerisches Archiv für Sprachsignale Repository. Zenodo. https://doi.org/10.5281/zenodo.8108803

Fischer, S., Knappen, J., Krielke, M.-P. & Teich, E.: Partner-Parade #01: CLARIND-UdS. Datenzentrum an der Universität des Saarlandes, in: Text+ Blog, 25.09.2023, https://textplus.hypotheses.org/7021.

Hinrichs, E., Geyken, A., Leinen, P., Speer, A., Stein, R., Blumtritt, J., Borek, L., Eckart, T., Engelberg, S., Grötschel, M., Henrich, A., Heyer, G., Horstmann, W., Jefferies, N., Kudella, C., Lobin, H., Müller-Spitzer, C., Neuber, F., Neuefeind, C., ... Witt, A. (2022). Text+: Language- and text-based research data infrastructure. Zenodo. https://doi.org/10.5281/zenodo.6452002

Hinrichs, E., Hinrichs, M., & Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations,* 25-29. https://aclanthology.org/P10-4005

Hug, M. (2023). Historische Textsammlungen im Deutschen Textarchiv (DTA). Zenodo. https://doi.org/10.5281/zenodo.8108758

Knappen, J., Fischer, S., Krielke, M.-P., & Teich, E. (2023). CLARIND-UdS: Repositorium für Sprachressourcen an der Universität des Saarlandes. Zenodo. https://doi.org/10.5281/zenodo.8108813

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 55-60. https://doi.org/10.3115/v1/P14-5010

Martens, S. (2013). TüNDRA: A web application for treebank search and visualization. *Proceedings of The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, 133-144. http://bultreebank.org/wp-content/uploads/2017/06/TLT12Proceedings-compressed.pdf#page=139

Rau, F., Debbeler, A., Saleh, S., & Mollenhauer, E. (2023). Datenpublikation & Archivierung am Data Center for the Humanities. Zenodo. https://doi.org/10.5281/zenodo.8108772

Schmid, H. (1994): Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Schmid, H. (1995): Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf

Trippel, T., & Zinn, C. (2023). Datenübernahme durch TALAR, das Tübinger Archive of Language Resources. Zenodo. https://doi.org/10.5281/zenodo.8108786

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Witt, A., Werthmann, A., & Trippel, T. (2023). Datenübernahme am Text+ Datenzentrum des Leibniz-Instituts für Deutsche Sprache, Mannheim. Zenodo. https://doi.org/10.5281/zenodo.8108796

Zinn, C. (2022). Bagman - A tool that supports researchers archiving their data. *Selected Papers from the CLARIN Annual Conference 2021,* 181-189. https://doi.org/10.3384/ecp18916