



Bericht über die Referenzimplementierung der ortsverteilten Infrastruktur in Collections

Deliverable C1.2

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	Christoph Draxler, Jennifer Ecker, Philippe Genêt, Tobias Gradl, Alina Hemmer, Marius Hug, Erik Körner, Daniel Kurzawe, Thorsten Trippel
Projekt	Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur
Bezeichnung	C1.2 Report on the reference implementation for decentralised collections
Förderung	DFG Förderkennzeichen 460033370
Projektlaufzeit	01.10.2021 bis 30.09.2026

Inhaltsverzeichnis

1.	Warum eine dezentrale Infrastruktur?.....	3
2.	Welche Arten von Ressourcen gibt es in den Collections von Text+ und wo sind Besonderheiten? – Beispiele.....	4
2.1	Vielfalt der Ressourcen	4
2.1.1	TEI	5
2.1.2	DTABf.....	5
2.1.3	I5	5
2.1.4	CoNLL-U.....	6
2.1.5	EXMARaLDA.....	6
2.2	Rechtebewehrte Inhalte.....	7
2.2.1	Urheberrecht.....	7
2.2.2	Lizenzrecht	7
2.2.3	Persönlichkeits- und Datenschutzrecht	7
2.3	(Bestehende) Dienste und Services.....	8
2.3.1	BAS	8
2.3.2	DTA	8
2.3.3	Deutsches Referenzkorpus (DeReKo)	9
2.3.4	TextGrid	9
3.	Welchen Zugriff ermöglicht Text+ auf Ressourcen und Dienste?	10
3.1	Nachweissystem(e): Registry	10
3.2	Übergreifende Suche: FCS	11
3.3	Zentrale Zugangsmöglichkeiten: Access und Identity Management (AIM)	12
3.4	Ausblick: Persistierte virtuelle Sammlungen	12
4	Weitere Entwicklungen	12
	Referenzen.....	13

1. Warum eine dezentrale Infrastruktur?

Das NFDI-Konsortium Text+ stellt eine verteilte Infrastruktur für die Archivierung und Bereitstellung von Forschungsdaten zur Verfügung. An der Datendomäne *Collections* beispielsweise sind derzeit elf Datenzentren¹ an verschiedenen Standorten beteiligt. Während der Zugriff auf die jeweiligen Daten in dieser dezentral organisierten Infrastruktur eine Herausforderung darstellt, war und ist der Aufbau eines gemeinsamen „Datensilos“ – in das die Zentren ihre Sammlungen überführen – aus verschiedenen Gründen kein gangbarer Weg.

Schon aus rechtlichen Gründen ist es oft nicht möglich, Datensätze von einer Einrichtung in eine andere zu kopieren oder umzuziehen. Dies gilt insbesondere für „alte“ Datensätze, die von Dritten im Rahmen bilateraler Lizenzvereinbarungen zwischen Datenlieferanten und Einrichtungen, die die Daten anschließend bereitstellen, erworben wurden. Bestehende Verträge und Lizenzen sind in der Regel nicht übertragbar, d. h. sie können nicht von einer anderen juristischen Person genutzt werden, weshalb es nicht ausreichen würde, eine übergeordnete juristische Person für alle Partner, etwa im Rahmen von NFDI e.V. („NFDI nach Satzung“), zu schaffen, um die Nutzung durch alle Partner zu ermöglichen. Hinzu kommt, dass Lizenzgeber, wie z.B. Verlage, möglicherweise nicht mit Änderungen bestehender Vereinbarungen einverstanden wären. Für unsere Partnerin Deutsche Nationalbibliothek (DNB) beispielsweise regelt sogar ein eigenes Gesetz², welche Rechte (und Pflichten) für die archivierten Daten gelten.

Das Portfolio von Text+ enthält viele Bestandsdaten, die einerseits von der wissenschaftlichen Community stark nachgefragt werden, die andererseits aber nur in einer verteilten Forschungsdateninfrastruktur bereitgestellt werden können. Von Dritten zur Verfügung gestellte Datensätze – und bei Text- und Sprachdaten trifft dies auf fast alle Datensätze zu – berühren die Rechte natürlicher und juristischer Personen. Viele gesprochensprachliche Sammlungen bestehen aus Audio- bzw. Videoaufnahmen, die wegen des Persönlichkeits- und Datenschutzes oder aufgrund bilateraler Vereinbarungen nur an dem in der Vereinbarung bezeichneten Datenzentrum gespeichert werden dürfen.

Rechte Dritter berühren dabei Beschränkungen, die durch die Datenschutzgrundverordnung (DSGVO) vorgegeben werden, oder geistiges Eigentum von Verlagen und Autoren, welche der sogenannten Regelschutzfrist unterliegen³. Auch urheberrechtlich geschützte Werke, wie sie etwa in den Beständen der DNB oder des Leibniz-Instituts für Deutsche Sprache (IDS) zahlreich vorhanden sind, dürfen die jeweiligen Institutionen nicht verlassen. Einige Datenbestände dürfen sogar nur in den (physischen) Räumlichkeiten der datenhaltenden Institution eingesehen werden.

Während diese rechtlichen Aspekte einen Hauptgrund darstellen, aufgrund dessen eine Konsolidierung in einer Datenbank oder in einem Computersystem – einschließlich eines Suchindexes – nicht realisiert werden kann, gibt es zahlreiche weitere Gründe.

Die bestehende dezentrale Struktur ist auch historisch zu erklären. Viele der Datenzentren arbeiten bereits seit langer Zeit im Rahmen von Projekten wie CLARIN-D, DARIAH-DE und später CLARIAH-DE zusammen. Die verschiedenen Datenzentren waren als Spezialisten für bestimmte Ressourcentypen in diesen Projekten beteiligt oder haben sich im Laufe auf bestimmte Ressourcentypen oder Inhalte fokussiert und im Hinblick darauf besondere Kompetenzen entwickelt. Sie verfügen daher

¹ Im Einzelnen sind dies: Akademie der Wissenschaften in Hamburg (AdWHH), Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Deutsche Nationalbibliothek (DNB), Leibniz-Institut für Deutsche Sprache (IDS), Bayerisches Archiv für Sprachsignale (BAS) an der Ludwig-Maximilians-Universität München, Universität des Saarlandes (SLUni), Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), Universität Duisburg-Essen (UniDUE), Universität Hamburg (UniHH), Universität zu Köln (UniK) und Universität Tübingen (UniTü)

² vgl. [DNBG - Gesetz über die Deutsche Nationalbibliothek](#)

³ Nach § 64 des UrhG beträgt die Regelschutzfrist in Deutschland 70 Jahre, d.h. konkret, dass ein Werk ab dem 1. Januar, der auf den 70. Todestag des Urhebers folgt, gemeinfrei wird (vgl. [§ 64 UrhG - Einzelnorm](#)).

über unterschiedliche Arten von Forschungsdaten mit unterschiedlichen Datenmodellen und haben jeweils eigene Arbeitsabläufe für die Erstellung, Pflege und Analyse von Daten etabliert. So hat z.B. das Hamburger Zentrum für Sprachkorpora (HZSK) der Universität Hamburg ausgewiesene Expertise im Bereich Soziolinguistik und Mehrsprachigkeit, das Data Center for the Humanities (DCH) der Universität zu Köln ist im Bereich der bedrohten bzw. unterrepräsentierten Sprachen profiliert und sammelt dazu teilweise einzigartige Audiodaten. Mit fremdsprachigen Textkorpora befasst sich das Repositorium der Universität des Saarlandes (UdS) besonders intensiv, während die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) sich mit dem Deutschen Textarchiv auf umfassend annotierte, historische Textsammlungen aus dem deutschen Sprachraum konzentriert. Und die Staats- und Universitätsbibliothek Göttingen (SUB) ist mit dem VD17 und dem VD18 am Aufbau einer vollständigen Sammlung aller im deutschen Sprachraum oder in deutscher Sprache erschienenen Drucke des 17. und 18. Jahrhunderts beteiligt.

Im Verbund ist so ein Netzwerk entstanden, dessen Expertise nahezu alle Bereiche sprach- und textbezogener Daten umfasst. Zudem haben viele Zentren in ihren jeweiligen Spezialgebieten auch große Mengen an einschlägigen Daten erst produziert und anschließend aufwändig kuratiert sowie annotiert. Dementsprechend ist auch die technische Umgebung der jeweiligen Repositorien auf diese Daten ausgerichtet – ein nicht zu unterschätzender Aufwand und eine immense Quelle an Erfahrung und Know-How. Diese Expertise sichert auch die Langzeitverfügbarkeit bzw. -nutzbarkeit.

Ebenfalls historisch gewachsen ist das Vertrauen, das die Daten- und Kompetenzzentren der Datendomäne Collections in den Forschungscommunitys genießen. Neben der technischen und inhaltlichen Expertise gründet dieses Vertrauen hauptsächlich in der Verlässlichkeit und Nachhaltigkeit der Zentren, die in Form von Zertifizierungen – etwa mit dem Core Trust Seal⁴ oder dem nestor-Zertifikat⁵ – verbrieft sind. Insbesondere das Versprechen der Langzeitarchivierung, das mit diesen Zertifikaten einhergeht, ist eine wesentliche Voraussetzung, um als Datenzentrum in die Datendomäne Collections aufgenommen zu werden. Schließlich sind Beständigkeit und Zuverlässigkeit grundlegende Eigenschaften einer jeden Infrastruktur.

Wie wir argumentiert haben, gibt es keine legitime und praktikable Möglichkeit, eine dezentrale Dateninfrastruktur zu vermeiden, wenn man die Bestandsdaten der Partner von Text+ einbezieht. Gleichzeitig gibt es angesichts der Fülle von Daten bei diesen Institutionen eine große Motivation, nach Möglichkeiten zu suchen, den Zugang zu diesen Datensätzen über die verteilten Institutionen hinaus zu ermöglichen. Neben der Heterogenität der Sammlungen und Datenformate selbst besteht somit für Text+ eine wichtige Herausforderung darin, einen zentralen Zugang zu den ortsverteilt gespeicherten Ressourcen zu schaffen. Dadurch können auch neue standortübergreifende Sammlungen aus (urheberrechtsfreien) Daten verschiedener Institutionen, sogenannte virtuelle Sammlungen, gebildet werden.

2. Welche Arten von Ressourcen gibt es in den Collections von Text+ und wo sind Besonderheiten? – Beispiele

2.1 Vielfalt der Ressourcen

Neben rechtlichen Gründen, die die Konsolidierung der Sammlungen an einer zentralen Stelle ausschließen, und der historisch gewachsenen Struktur gibt es weitere zwingende Argumente, die für eine Aufrechterhaltung der ortsverteilten Infrastruktur sprechen. Die folgenden Beschreibungen

⁴ [CoreTrustSeal-AMT](#)

⁵ [nestor-Siegel](#)

illustrieren, welche weiteren Gründe vorliegen.

2.1.1 TEI

Die TEI⁶ (Text Encoding Initiative) stellt ein Framework für angepasste XML-Dokumentgrammatiken zur Verfügung, die strukturelle Gemeinsamkeiten haben. So weisen TEI-Dokumente eine gemeinsame Struktur im Header auf, in denen bestimmte beschreibende Metadaten repräsentiert werden. Der inhaltliche Teil der TEI-Dokumente ist aber sehr variabel und erlaubt es daher, so unterschiedliche Datentypen wie gesprochene Sprache, Merkmalsstrukturen, Wörterbücher oder kritische Apparate zu repräsentieren. Im Bereich der Digital Humanities ist TEI ein sehr weitverbreiteter Standard, wobei alleine auf der Grundlage der Verwendung des Standards noch nicht hinreichend klar ist, welche Variante von TEI verwendet wird. Die TEI stellt aber auch technologische Komponenten bereit, um basierend auf dem Framework die passenden XML-Grammatiken, z. B. als XML-Schema, in RelaxNG⁷ oder auch als DTD (Document Type Definition) zur Verfügung zu stellen. Prozesse, die auf allgemeinen TEI-Formaten aufsetzen, sind daher darauf angewiesen, dass neben den Daten und Dokumentgrammatiken auch die Prozesse soweit adaptiert werden, dass sie zu den Daten passen. Allgemeine TEI-Werkzeuge sind ansonsten sehr ähnlich zu anderen (allgemeinen) XML-Werkzeugen, wie z.B. XML-Parsern. Die Anpassung der Prozesse an spezifische Instantiierungen von TEI bringt dabei implizit auch die Voraussetzung mit sich, die dazu notwendigen Kompetenzen vorzuhalten und die angepassten Prozesse zu pflegen. Eine Konsolidierung ist aufgrund der unterschiedlichen wissenschaftlichen Fragestellungen und Datenmodellierung nur begrenzt möglich. Daher werden verschiedene Systeme und Verarbeitungsprozesse notwendig, wobei nicht davon ausgegangen werden kann, dass diese Prozesse auf den gleichen (Software-)Systemen und Stacks aufsetzen. Aus diesem Grund kann die Verwendung der TEI je nach Einsatzort und -zweck stark variieren und nicht automatisch in andere Systeme überführt werden.

2.1.2. DTABf

In der zweiten Förderphase des von der DFG geförderten Projekts Deutsches Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) wurde auf Grundlage eines Subset des von der TEI entwickelten TEI P5 Standards ein projektspezifischer Auszeichnungsstandard entwickelt, das Basisformat des Deutschen Textarchivs (DTABf). Motivation dafür war die Vermeidung von Variationsspielräumen bei der Textauszeichnung, welche durch die P5-Richtlinien der TEI nicht ausgeschlossen werden können. Da die TEI-Guidelines für sämtliche Bedürfnisse bei der (geisteswissenschaftlichen) Textaufbereitung eine Lösung anbieten möchten, müssen diese entsprechend vielfältig sein. So bedeutet das in einem konkreten Fall, dass es für die Auszeichnung von Personen verschiedene valide Möglichkeiten gibt. Während es gemäß der TEI-Richtlinien gleichbedeutend ist, ob eine Person per `<name type="person">` oder mittels `<persName>` ausgezeichnet wird, ist es ein zentrales Anliegen des DTABf entsprechende Ambiguitäten zu vermeiden, um für größtmögliche Interoperabilität zu sorgen. Im Rahmen der DTA-Annotationsrichtlinien⁸ ist das DTABf ausführlich dokumentiert. Aktive Pflege und Weiterentwicklung erfolgt bis heute durch die DTABf-Steuerungsgruppe an der BBAW, die sich aus Expertinnen und Experten für die TEI-Auszeichnung und -Anpassung mit Verankerung in verschiedenen Communitys zusammensetzt.

2.1.3 I5

Am IDS liegen die Korpora geschriebener Sprache im IDS-Textmodell⁹ als IDS-TEI P5-Format, kurz I5, vor. I5, das durch Customisierung¹⁰ mit dem TEI P5-spezifischen ODD-Mechanismus formal

⁶ [Text Encoding Initiative: TEI](#)

⁷ [RELAX NG](#)

⁸ vgl. [Das DTA-Basisformat](#)

⁹ [IDS-Textmodell | IDS](#)

¹⁰ siehe auch [Lüngen & Sperberg-McQueen, 2012](#).

abgeleitet wurde, ist eine TEI-basierte Dokumentengrammatik auf Grundlage des TEI-Standard P5¹¹. Basierend auf dem I5-Format werden die Korpora im Recherchesystem COSMAS II¹² des IDS durchsuchbar gemacht und dargestellt. Für das zweite Recherchesystem KorAP¹³ wird das I5-Format in das KorAP-XML-Format konvertiert, das Primärdaten, Metadaten und Annotationsebenen für ein Dokument in getrennten XML-Dateien darstellt. Die TEI-Variante I5 wird als Primärformat gebraucht, um die Darstellung in den Recherchesystemen zu gewährleisten und dient zudem als primäres Ingest-Format für das Deutsche Referenzkorpus, über das Daten Dritter in das Korpus integriert werden können.

2.1.4 CoNLL-U

Ausgehend von einer Aufgabe für Forschende während der Conference on Natural Language Learning wurde auf der Konferenz das CoNLL-U-Format entwickelt. Das CoNLL-U-Format ist ein Zugeständnis an die Vielfalt der Formate, die bei Natural Language Processing (NLP)-Aufgaben verwendet werden, und stellt eine Art lingua franca-Datenformat im NLP-Kontext dar. Das Format ist Text-basiert, wobei gleiche Datentypen in Spalten mit delimitierenden Zeichen repräsentiert sind. Die Anzahl der Spalten ist dabei zunächst nicht begrenzt, aber einige Spalten sind in der Regel vorhanden, darunter Token und Lemma. Je nach Anwendungsgebiet kommen weitere Spalten hinzu. Ein Anwendungsgebiet sind z.B. Baumbanken, syntaktisch annotierte Korpora, die meist auf Grundlage von Zeitungstexten erstellt werden. Auch wenn für spezifische Suchanforderungen zum Teil eigene, etwa XML-basierte Formate verwendet werden, erfolgt der Austausch von Baumbanken typischerweise im CoNLL-U Format¹⁴. Der Austausch erfolgt dabei aber nicht "blind", das heißt es ist erforderlich, dass auch die empfangende Seite vor der Verarbeitung mittels Dokumentation die Semantik der Spalten betrachtet. Eine Validierung der Daten erfordert auch spezifische, für die jeweiligen Spalten adaptierte Verfahren und Prozesse.

2.1.5 EXMARaLDA

EXMARaLDA¹⁵ (Extensible Markup Language for Discourse Annotation) wurde ursprünglich im Rahmen eines Teilprojekts des Sonderforschungsbereichs 538 ‚Mehrsprachigkeit‘ an der Universität Hamburg mit dem Ziel entwickelt, transkribierte Sprachkorpora unabhängig von verwendeten Transkriptionssystemen nachhaltig nutzbar und archivierbar zu machen.¹⁶ Bei EXMARaLDA handelt sich um ein System zur computergestützten Erstellung, Verwaltung und Analyse von digitalen Korpora gesprochener Sprache, das den Austausch gesprochensprachlicher Korpora zwischen Forschenden und technologischen Umgebungen erleichtert und eine hohe Kompatibilität mit bestehenden Standards gewährleistet. Es besteht aus Datenmodell, Dateiformaten sowie Software-Tools und hat sich durch kontinuierliche Weiterentwicklung zu einem Standardsystem für Forschende u.a. in den Bereichen Pragmatik, Gesprächsanalyse und Multimodalitätsforschung entwickelt.

Gesprochensprachliche Daten, die in Form von Audio- oder Videoaufnahmen vorliegen, können mithilfe des EXMARaLDA Partitur-Editors¹⁷ transkribiert werden, wobei die Integration digitaler Audio- oder Videodateien vorgesehen ist. In zeitalignierten Transkripten werden Zeitpunkten aus verknüpften Audio-/Videoaufnahmen korrespondierende Stellen in der Transkription zugewiesen. Da die Verknüpfung optional ist, können Transkripte oder Korpora auch ohne für die Transkription verwendete Aufnahmen verwendet werden, sofern diese nicht veröffentlicht werden dürfen. Das EXMARaLDA-System umfasst neben dem Partitur-Editor noch den Corpus-Manager Coma¹⁸, durch den EXMARaLDA-Korpora erzeugt werden können.

¹¹ [P5 Guidelines – TEI](#)

¹² [COSMAS II - Startseite](#)

¹³ [KorAP - Corpus Analysis Platform](#)

¹⁴ vgl. [CoNLL-U Format](#)

¹⁵ [EXMARaLDA](#)

¹⁶ vgl. [Schmidt & Wörner, 2005, 171](#).

¹⁷ [Partitur-Editor – EXMARaLDA](#)

¹⁸ [Corpus-Manager \(Coma\) – EXMARaLDA](#)

Durch die lange Historie und große Nutzendengruppe liegen auch in den Text+ Datenzentren zahlreiche Korpora und Sammlungen in EXMARaLDA-Datenformaten vor.

2.2 Rechtebewehrte Inhalte

Viele Sammlungen in Text+ enthalten rechtebewehrte Inhalte. Dabei handelt es sich einerseits um lizenzrechtlich geschützte Inhalte und andererseits um datenschutzrechtlich geschützte und/oder sensible Inhalte. Beide Arten rechtebewehrter Inhalte sind in der Regel an ein Datenzentrum gebunden und dürfen nur dort archiviert und/oder verwendet werden. Die ortsverteilte Infrastruktur in Text+ ermöglicht es, dass Sammlungen, die rechtebewehrte Inhalte umfassen, mithilfe von Werkzeugen wie der Registry und Federated Content Search (FCS) dennoch an einer zentralen Stelle sichtbar werden, wodurch sich ihre Nutzendengruppe deutlich vergrößern kann. Im Folgenden werden einige Beispiele von Sammlungen in Text+ vorgestellt, die rechtebewehrte Inhalte aufweisen und deswegen zwangsläufig auf eine ortsverteilte Infrastruktur angewiesen sind.

2.2.1 Urheberrecht

Die Sammlung digitaler Objekte der DNB umfasst derzeit über 14 Millionen Objekte und beinhaltet u.a. E-Books, E-Paper, Online-Hochschulschriften, Noten sowie Websites. Die meisten dieser Objekte liegen in den Formaten PDF oder EPUB vor. Sie gehören damit zur Kategorie der „unstrukturierten Texte“, da Strukturinformationen wie Absätze, Satzlängen etc. nicht erfasst sind, selbst wenn ihre Volltexte mit Optical Character Recognition (OCR)-Verfahren durchsuchbar gemacht wurden. Zudem ist der weit überwiegende Teil der Sammlung urheberrechtlich geschützt. Zugänglich gemacht werden darf der gesamte Bestand der DNB – auch in digitaler Form – ausschließlich in den Räumlichkeiten ihrer Standorte in Frankfurt am Main und Leipzig.

2.2.2 Lizenzrecht

Ein anderes Beispiel für rechtlich geschützte Daten ist das Deutsche Referenzkorpus (DeReKo) vom IDS. Das IDS bietet damit Korpora geschriebener Gegenwartssprache an, die über die Recherchesysteme COSMAS II und KorAP abfragbar sind. Die geschriebenen deutschsprachigen Texte aus der Gegenwart und der neueren Vergangenheit beinhalten 55 Milliarden Wörter (Stand 08.03.2023). Zu den Texten gehören Zeitungstexte, belletristische, wissenschaftliche und populärwissenschaftliche Texte sowie weitere Textarten. Das IDS kann Texte in DeReKo durch juristische Vereinbarungen mit Verlagen, Zeitungsredaktionen und Autoren zur Verfügung stellen. Somit sind die Texte urheberrechtlich abgesichert. Zu wissenschaftlichen, nichtkommerziellen Zwecken können alle Korpora IDS-intern und der größte Teil davon weltweit öffentlich durch die genannten Recherchesysteme genutzt werden.

2.2.3 Persönlichkeits- und Datenschutzrecht

Einige Sammlungen in Text+ enthalten hochsensible Inhalte, für die persönlichkeits- oder datenschutzrechtliche Vereinbarungen aus den jeweiligen Erhebungsprojekten vorliegen, die einen Umzug oder das Kopieren der Sammlungen an einen anderen Ort ausschließen. Die für diese Sammlungen geschlossenen Vereinbarungen regeln neben dem Speicherort der Sammlungsinhalte auch die Überwachung der Nutzungsabsicht von Dritten durch explizit benannte Datenverantwortliche. Eine Gruppe von Sammlungen, die nur unter der Voraussetzung strenger datenschutzrechtlicher Vereinbarungen erstellt werden können, sind Sprachdaten, die durch Audio- oder Videoaufzeichnungen von Interaktionen sowie deren anschließender Transkription gewonnen werden. Hierzu zählt beispielsweise das Korpus „Dolmetschen im Krankenhaus (DiK)“¹⁹ der HZSK-

¹⁹ [Dolmetschen Im Krankenhaus \(DiK\) | ZFDM Repository, 2009](#)

Community der Universität Hamburg. Die Sammlung enthält Transkriptionen von Audioaufnahmen von Arzt-Patienten-Kommunikation im Krankenhaus. Dieser besonders sensible Teil der Gesundheitskommunikation kann der Forschungsgemeinschaft nur unter Einhaltung der im Rahmen der Datenerhebung geschlossenen Vereinbarungen zugänglich gemacht werden. Die Übertragung der Daten ist strikt ausgeschlossen, da sie nicht nur Datenschutzvereinbarungen brechen, sondern auch das Vertrauen der datengebenden Personen verletzen würde. Die Beibehaltung der ortsverteilten Infrastruktur in Text+ bietet die Möglichkeit, sensible, aber gleichzeitig für die Forschungsgemeinschaft hochgradig interessante Daten besser auffindbar und zugänglich zu machen, ohne geschlossene Vereinbarungen zu verletzen.

2.3 (Bestehende) Dienste und Services

Die ortsverteilte Infrastruktur von Text+ liegt nicht nur – wie bereits ausführlich dargelegt – in der Ebene der Daten begründet, vielmehr bringen die beteiligten Institutionen auch spezialisierte Dienste und Services in das NFDI-Konsortium ein. Bevor also in Kapitel 3 ein Weg zu den Ressourcen und Diensten in Text+ aufgezeigt werden soll, wird im Folgenden an vier Beispielen exemplarisch beschrieben, welchen Nutzen die forschende Community einerseits von diesen Diensten und Services hat, und warum andererseits diese auf langjährige Expertise gründenden Angebote nicht einfach zentralisiert werden können.

2.3.1 BAS

Das Bayerische Archiv für Sprachsignale (BAS) betreibt ein Repository²⁰ für Sprachdatenbanken. Diese bestehen jeweils aus Audiodaten, mindestens einer orthografischen Transkription sowie in der Regel einem Aussprachewörterbuch. Für viele Sprachdatenbanken sind zudem auch zeit-alignierte Transkriptionen sowie phonemische Segmentationen (Zuordnung von Lauten zur Position im Signal) vorhanden.

Die Sprachdatenbanken sind für Forschung und Technologieentwicklung unter drei Lizenzen zugänglich: 1) freier Zugang, 2) Zugang nur für akademische Nutzer und Nutzerinnen, sowie 3) individueller Zugang nach Abschluss eines Lizenzvertrags. Der Katalog des Repository wird regelmäßig von Suchmaschinen und ähnlichen Diensten gescannt und im Virtual Language Observatory von CLARIN-EU gespiegelt.

Das BAS wurde von Anfang an mit dem Ziel aufgebaut, die in Verbund- und Forschungsprojekten (z.B. VERBMOBIL, SmartKom) gesammelten gesprochensprachlichen Daten dauerhaft zur Verfügung zu halten. Aktuell (Sept. 2023) umfasst das Repository 54 Sprachdatenbanken, und zunehmend kommen weitere, hauptsächlich von Dritten erstellte Sprachdatenbanken hinzu.

Neben dem Repository bietet das BAS auch eine Reihe sprachtechnologischer Webdienste²¹ an. Diese sind – mit Ausnahme der automatischen Spracherkennung – frei verfügbar. Für die automatische Spracherkennung ist eine Authentifizierung als akademische/r Nutzer/in notwendig. Die bekanntesten Webdienste sind 1) MAUS (Munich Automatic Segmentation) zur Segmentierung auf Wort- und Lautebene, 2) Anonymizer zum gezielten Ausblenden von privaten Inhalten aus transkribierten Audiodateien, und 3) WikiSpeech für skriptgesteuerte Sprachaufnahmen über das Internet.

2.3.2 DTA

Das Deutsche Textarchiv (DTA) ist ein aktives Archiv für deutschsprachige, historische Korpora und Sammlungen und umfasst annotierte Volltexttranskriptionen von Drucken, Zeitungen und Zeitschrif-

²⁰ [BAS CLARIN Repository](#)

²¹ [BAS | Web Service Interface](#)

ten sowie handgeschriebene Dokumente verschiedener Gattungen und Textarten vom 16. bis zum frühen 20. Jahrhundert. Die aktuell rund 40 Textsammlungen des DTA sind im Basisformat DTABf²² kodiert und werden in der Regel unter einer CC BY-SA Lizenz zur Nachnutzung bereitgestellt. Ein großer Mehrwert für die forschende Community besteht aber vor allem in der Bereitstellung der Texte innerhalb der DTA-Infrastruktur. So werden alle Texte der Sammlungen mittels CAB (Cascaded Analysis Broker) linguistisch annotiert. Neben einer Transliterierung – *langes s (l)* wird bspw. in ein *rundes s* übersetzt –, einer Zurückführung auf eine phonetische Repräsentation – *TheyI, ThayI* und *TeyI* werden bspw. auf *Teil* zurückgeführt – und einer graphematischen Ersetzung – für jede historische Wortform wird automatisch das ‘ähnlichste’ moderne Wort ermittelt – spielt hier die Lemmatisierung auf eine neuhochdeutsche Grundform mittels der TAGH-Morphologie eine wichtige Rolle. Diese Morphologie ist eine zentrale Komponente in den Softwarediensten des DTA, allerdings lizenzrechtlich so eingeschränkt, dass sie nicht an eine zentrale Infrastruktur übergeben werden kann.

2.3.3 Deutsches Referenzkorpus (DeReKo)

Der Aufbau elektronischer Korpora am IDS begann in den 1960er Jahren und der Umfang der Korpora hat sich seitdem stetig gesteigert. Das Korpus-Archiv bzw. DeReKo-Archiv wird beständig erweitert und bestehendes Korpusmaterial überarbeitet. Veröffentlicht werden die Arbeiten dazu als Releases in den Recherchesystemen COSMAS II und KorAP. Die Releases sind rekonstruierbar²³, um Nachvollziehbarkeit und Replizierbarkeit zu garantieren. KorAP wurde entwickelt, um COSMAS II als primäre Analyseplattform sukzessive abzulösen. Mit Registrierung kann in den beiden Recherchesystemen gesucht werden, wobei in KorAP Ausschnitte frei zugänglicher Daten (z.B. Nutzerdiskussionen von Wikipedia) auch ohne Registrierung eingesehen werden können. Ein Zugriff auf die Daten kann in KorAP auf mehreren Annotationsebenen geschehen²⁴. Dagegen sind in COSMAS II nur ein Teil der Texte aus “Archiv W” morphosyntaktisch annotiert²⁵ (mit unterschiedlichen Taggern und somit auch anderen zugrundeliegenden Tagsets).

2.3.4 TextGrid

TextGrid ist eine Virtuelle Forschungsumgebung für die Geisteswissenschaften, die für die Arbeit mit TEI-codierten Ressourcen optimiert ist und den gesamten Forschungsprozess bis zur Publikation abbildet. Das TextGrid Repository dient als Quellengrundlage sowie als Plattform zum Austausch und zur Publikation. Angeboten wird das TextGrid Repository seit 2012 – zunächst als Verbundprojekt von zehn institutionellen und universitären Partnern, seit 2016 als Teil der DARIAH-DE Forschungsinfrastruktur²⁶. Es bildet inzwischen einen wichtigen Bestandteil der Services von Text+ und wird laufend weiterentwickelt. Das TextGrid Repository bietet technische Schnittstellen über REST und SOAP sowie Client Libraries für Java und XQuery an. Die sogenannten TextGrid Python Clients²⁷ (tgclients) bieten einen einfachen Zugriff auf die TextGrid Repository Services API²⁸. Schulungen und Workshops hierzu sind im Rahmen von Text+ geplant.

Aktuell wird ein neuer Import-Workflow für die Publikation von Forschungsdaten in das TextGrid Repository entwickelt, der es Nutzenden einfach und übersichtlich ermöglichen soll, für TEI-Korpora automatisiert TextGrid-Metadaten zu erstellen, diese (semi-)automatisch anzureichern und nach erfolgreicher Bearbeitung letztlich in TextGrid zu publizieren.

Entlang neu publizierter Korpora wurden verschiedene weitere Funktionalitäten entwickelt und

²² Siehe dazu auch Kapitel 2.1.

²³ vgl. [Ausgaben | IDS](#)

²⁴ vgl. [KorAP-Annotations](#)

²⁵ vgl. [Textorganisation Unter COSMAS II - MECOLB-Minimal Tagset](#)

²⁶ vgl. <https://textgrid.de/web/guest/projekt>

²⁷ [TextGrid Python Clients Documentation](#)

²⁸ [The TextGrid Repository Documentation](#)

getestet, wie projektspezifische Landing-Pages für das TextGrid Repository, sowie neue Möglichkeiten der Facettierung der Suchergebnisse, beispielsweise die Verwendung der bibliothekarischen Basisklassifikation zur Erschließung der Texte oder neue Suchoptionen in Bezug auf die GND-Normdaten. Diese Funktionalitäten stehen jetzt allen TextGrid-Nutzenden zur Verfügung.

Im Sommer 2023 wurde eine Umfrage zu den Kategorien durchgeführt, die Forschende in der Literaturwissenschaft verwenden, um Korpora zusammenzustellen. Die Ergebnisse sollen dazu dienen, verschiedene Ressourcen (insbesondere TextGrid Repository) kritisch zu evaluieren und Anreicherungen zu planen.

3. Welchen Zugriff ermöglicht Text+ auf Ressourcen und Dienste?

3.1 Nachweissystem(e): Registry

In einer verteilten Infrastruktur, die Forschungsdaten nach den FAIR-Prinzipien bereitstellt, kommt denjenigen Werkzeugen eine besondere Bedeutung zu, die Daten auffindbar machen, den Zugang zu den Daten sowie deren die Nutzbarkeit in verschiedenen Umgebungen ermöglichen und die Nachnutzbarkeit sicherstellen.

Für die Auffindbarkeit von Daten sind dabei Nachweissysteme essentiell, über die Nutzende Informationen zu den Daten erhalten. Zu diesen Informationen gehört, wo – also bei welcher Einrichtung – die Daten zu finden sind und wie man Zugriff darauf erhalten kann (also Zugangsbeschränkungen, Zugangsmöglichkeiten, Lizenzbedingungen, Kontaktmöglichkeiten). Das zentrale Einstiegswerkzeug zu den Forschungsdaten in Text+ erfolgt über das Webportal und dort über einen Verzeichnisdienst zu den Forschungsdaten in Text+, der als Registry²⁹ bezeichnet wird. In dieser Registry sind kuratierte Zusammenstellungen von Forschungsdatensätzen der beteiligten Institutionen verzeichnet, so dass Nutzende via Registry Zugriff auf die vollständige Beschreibung der Forschungsdatensätze bei den datenhaltenden Einrichtungen erhalten. Die Registry regelt dabei nicht den Zugang zu den Daten oder sorgt für deren persistente Identifikation, da dies Funktionen sind, die Teil der unterschiedlichen Archiv- und Repositoriumssysteme sind.

Um diese zentrale Hub-Funktion zu den Daten zu ermöglichen, liest die Registry über standardisierte Schnittstellen festgelegte Informationen von den datenhaltenden Einrichtungen ein und präsentiert sie den Nutzenden. Typischerweise werden dazu beschreibende (Meta-)Daten zu den kuratierten Sammlungen über eine OAI-PMH-Schnittstelle oder ähnliche harvest-bare Funktionen zusammengetragen.³⁰ Die autoritativen Referenz-Metadaten zu den kuratierten Datensammlungen verbleiben an der datenhaltenden Einrichtung, d.h. die Verantwortung für die Pflege und Distribution verbleibt dort. Die föderierte Architektur erlaubt es dabei auch, verschiedene Metadatenschemata zu verwenden, etwa abhängig von den Forschungsdatentypen, unterschiedlich stark semantisch angereicherten Metadaten oder auch entsprechend unterschiedlichen Normen z.B. nach ISO 24622-1/24622-2, MARC21, DCAT, etc.

Durch die Verwendung der Schnittstellen ist dabei auch gewährleistet, dass die Infrastruktur offen bleibt. Neue Partner können einfach aufgenommen werden, indem die Adressen zu den Schnittstellen der Partner auch durch die Registry und entsprechende Werkzeuge geharvestet werden.

²⁹ vgl. [Genêt et al., 2023](#)

³⁰ zum Registry-Ingest vgl. [Deliverable IO 3.3a](#).

3.2 Übergreifende Suche: FCS

Neben der Registry ist ein weiteres zentrales Werkzeug die übergreifende Suche. Sie ermöglicht den Zugriff auf Forschungsdaten in einer dezentralen Infrastruktur über eine zentrale Stelle. Die übergreifende Suche wird als föderierte Inhaltssuche (Federated Content Search, FCS) umgesetzt.³¹ Bei der FCS handelt es sich um eine Suchinfrastruktur und Spezifikation, die die Kommunikation zwischen Klienten und Endpunkten über gemeinsame Protokolle, Anfragesprachen und Dateiformate für Rückgaben definiert. Die Suchergebnisse werden für Nutzende aggregiert im Webportal von Text+ bereitgestellt. Da die FCS eine verteilte Inhaltssuche über eine zentrale Schnittstelle erlaubt, verbleibt die Kontrolle über die Sammlungen bei den jeweiligen Datenzentren. Die Datenzentren binden ihre Sammlungen über eigene Endpunkte an und können so weiterhin Umfang, Zugang und Nutzung der Daten steuern. Um die Sammlungen eines Datenzentrums an die FCS anbinden zu können, muss eine Möglichkeit für die Suche auf den jeweiligen Daten existieren, die in den Endpunkt integriert wird.

Die Sammlungen der verschiedenen Datenzentren zeichnen sich durch einen hohen Grad an inhaltlicher und formaler Heterogenität aus, die in der FCS berücksichtigt werden muss. Um den Spezifika der jeweiligen Datenbestände gerecht zu werden, wird auf sogenannten Annotationslayern gesucht, wodurch spezifische Textstrukturen übergreifend identifiziert und individuelle Annotationen einbezogen werden können. Darüber hinaus soll mit der FCS perspektivisch auch über rechtlich geschützte Objekte gesucht werden können. Bei der Implementierung dieser Suche wird sich die Rückgabe der Inhalte dahingehend unterscheiden, dass nur über das Vorhandensein von Suchtreffern informiert wird, aber keine Vorschau auf die Suchtreffer gegeben werden kann. Neben der Möglichkeit zur Volltextsuche stehen Filter zur Verfügung, mit denen Ressourcen ausgewählt werden können, die bestimmten Kriterien entsprechen, oder um anschließend eine Suche auf der Ressourcenauswahl auszuführen. Die Suchfacetten des Filters werden künftig um die Metadaten erweitert, die in der Registry zu jeder Sammlung erhoben werden. Das Metadatenmodell³² wurde in der TA Collection (AG Standardisierung) entwickelt und spiegelt die Vielfalt der Sammlungen wider, ist gleichzeitig aber funktional und abstrahierend aufgebaut. Die Suchfacetten sind ein Beispiel dafür, wie Registry und FCS konzeptionell mehr und mehr verschränkt werden und so unmittelbar ineinandergreifen.

Damit die verschiedenen Datenzentren ihre Sammlungen an die FCS anschließen können, müssen sie eigene Endpunkte entwickeln, die die jeweiligen Spezifika der eigenen Sammlungen berücksichtigen. Für den Standort Hamburg (Akademie der Wissenschaften in Hamburg und Universität Hamburg) wurde der FCS-Endpunkt beispielsweise mithilfe verschiedener Suchportale implementiert. Dieser Endpunkt verarbeitet Anfragen gemäß der SRU/CQL-Spezifikationen, die im Rahmen des CLARIN-Verbundes erarbeitet wurden. Diese werden weitergeleitet an lokale Suchportale, die mithilfe der Plattformen ANNIS³³ oder TSAKorpus³⁴ implementiert wurden, sowie ein webbasiertes Suchportal³⁵ des Langzeitprojektes *Formulae Litterae Chartae*. Die Implementierung erfolgte mithilfe des Python Frameworks FastAPI.

Durch die FCS bleibt die Datenverantwortung bei den jeweiligen Datenzentren. Gleichzeitig ermöglicht sie Nutzenden einen einfachen und schnellen Zugang zu den verschiedenen Sammlungen sowie analytische Vorarbeiten durch Filter- und Rankingmechanismen für die Suchtreffer.

Auch diese Infrastruktur ist offen für die Integration neuer Partner. Durch Einrichtung entsprechender Endpunkt können sie ihre Daten in die FCS integrieren. Zuletzt hinzugekommen ist die umfangreiche Sammlung historischer Zeitungen des Deutschen Zeitungsportals, das die Deutsche Digitale

³¹ Nähere Informationen zur FCS finden sich in Milestone C5.1: <https://zenodo.org/doi/10.5281/zenodo.12770996>.

³² vgl. Registry-Felder in Milestone C1.1: <https://zenodo.org/doi/10.5281/zenodo.12771255>

³³ ANNIS

³⁴ TSAKorpus

³⁵ [Formulae - Litterae - Chartae: Erweiterte Suche](#)

Bibliothek in den Text+ Suchraum einbringt.

3.3 Zentrale Zugangsmöglichkeiten: Access und Identity Management (AIM)

Um in einer verteilten Infrastruktur auf Daten und Systeme zugreifen zu können, die bestimmte Anforderungen an die Nutzenden stellen, ist ein Access und Identity Management (AIM) notwendig. Hierzu gehören etwa Zugriffe auf Daten, die nicht allgemein freigegeben sind, Rechenkapazitäten, die nur für die akademische Gemeinschaft genutzt werden dürfen etc. Standortübergreifende AIM-Systeme ermöglichen es dabei, dass nicht jeder Standort eine eigene Liste von Nutzenden erstellen muss, sondern über Single-Sign-On und ähnliche Verfahren ein Zugriff erfolgen kann, sobald Nutzende von ihrer (vertrauenswürdigen) Heimatinstitution als zugehörig bescheinigt werden. Dazu werden Verfahren wie SAML verwendet. Im akademischen Bereich sind solche Verfahren z.B. für die Fernleihe in den Bibliotheken implementiert worden, so dass Forschende über die Fernleihe auch auf Werke an anderen Bibliotheken Zugriff erhalten können, ohne dass die andere Bibliothek wissen muss, wer sie sind, oder die Identität verifizieren muss.

Im Rahmen von NFDI und Text+ erfolgt die Identifizierung von Nutzenden sowie die Authentifizierung an ihren Heimatinstitutionen über deren Zugangskontrollsysteme, die ortsverteilte Infrastruktur implementiert dann technische Verfahren, die es ermöglichen, dass unter Berücksichtigung von Datenschutzvorgaben auch der Zugang zu Daten und Systemen möglich werden, die an anderen Institutionen beheimatet sind. Text+ setzt hier auf etablierte Verfahren und verwendet die Systeme und Prozesse, die im Rahmen von BASE4NFDI weiterentwickelt werden.

3.4 Ausblick: Persistierte virtuelle Sammlungen

Vor dem Hintergrund, dass die Beschreibungen aller Text+ Forschungsdaten über Werkzeuge wie die Registry zugänglich sind, ist es möglich, dass Nutzende auf ihre Forschungsfragen zugeschnittene Sammlungen zusammenstellen – basierend auf gefundenen Informationen und nach ihrem eigenen Bedarf. Diese Zusammenstellung kann sowohl aus Teilmengen von Daten einer datenhaltenden Einrichtung bestehen, aber auch Daten verschiedener Einrichtungen umfassen. Um auf diese selbst zusammengestellten Datensätze dauerhaft zugreifen zu können, könnten virtuelle Sammlungen gebildet werden, d.h. Zusammenstellungen, die über persistente Identifikatoren auf andere Datensätze verweisen. Solche virtuellen Sammlungen können dabei auch für die Erschließung großer Datensammlungen von einzelnen Partnern sehr hilfreich sein, da sie es ermöglichen, Gesamtbestände gemäß unterschiedlicher Kriterien zu partitionieren.

4 Weitere Entwicklungen

Der zentrale Gegenstand der NFDI und damit auch von Text+ mit der Datendomäne Collections besteht darin, das Forschungsdatenmanagement zu unterstützen und dazu beizutragen, dass Forschungsdaten FAIR verfügbar und zugänglich sind. Daher nehmen die datenhaltenden Institutionen in Text+ auch weitere Daten auf, das Datenportfolio wird kontinuierlich weiterentwickelt und vergrößert. Die Datenzentren im Bereich Collections haben daher jeweils Datenübernahmerichtlinien und Verfahren zur Aufnahme von neuen Daten entwickelt. Diese Aufnahme von Daten wird gesondert im Deliverable C2.2 beschrieben und innerhalb von Collections aufeinander abgestimmt, so dass für Nutzende auch außerhalb von Text+ ein möglichst einheitlicher und verlässlicher Weg zur Datenübernahme durch ein Datenzentrum besteht.

Neben diesem Weg, Daten im Netzwerk von Text+ langfristig bereitzustellen – häufig wird dies als Depositing bezeichnet –, ist der Zugang zu den Daten, Metadaten und deren Vernetzung unerlässlich. Dazu werden z.B. die beschreibenden Metadaten von den Datenzentren über Standardschnitt-

stellen für Kataloge und Webservices bereitgestellt, die Registry und FCS greifen darüber auf diese Daten zu. Es gibt aber auch weitere Entwicklungen, wie diese Informationen bereitgestellt werden können. Ein vielversprechender Ansatz besteht in den Linked Data, in denen mit unterschiedlichen Technologien die Verknüpfung von Daten möglich wird, sowohl von Objektdaten als auch von Metadaten. In distribuierten Umgebungen können so z.B. Daten in RDF über SPARQL-Endpunkte abgerufen werden. Diese Technologie kann dabei auch innerhalb von einzelnen Standorten Verwendung finden. Eine weitere Möglichkeit besteht darin, über JSON-LD Daten bereitzustellen, so dass sie in Knowledge Graphen, etwa im Google Knowledge Graph oder auch in anderen Systemen verwendet werden können. Diese Entwicklungen werden in Text+ beobachtet und gegebenenfalls ebenfalls umgesetzt.

Referenzen

§ 64 UrhG - Einzelnorm. (n.d.). Gesetze im Internet. Retrieved November 22, 2023, from https://www.gesetze-im-internet.de/urhg/_64.html

ANNIS. (n.d.). <https://corpus-tools.org/annis/>

Ausgaben | IDS. (n.d.). IDS Mannheim. Retrieved November 22, 2023, from <https://www.ids-mannheim.de/digspra/kl/projekte/korpora/releases>

BAS CLARIN Repository. (2022, July 25). BAS CLARIN Repository. Retrieved November 22, 2023, from <https://clarin.phonetik.uni-muenchen.de/BASRepository/>

BAS | web service interface. (n.d.). BAS | web service interface. Retrieved November 22, 2023, from <https://clarin.phonetik.uni-muenchen.de/BASWebServices/>

CoNLL-U Format. (n.d.). Universal Dependencies. Retrieved November 22, 2023, from <https://universaldependencies.org/format.html>

CoreTrustSeal-AMT. (n.d.). CoreTrustSeal-AMT. Retrieved November 22, 2023, from <https://amt.coretrustseal.org/certificates>

Corpus-Manager (Coma) – EXMARaLDA. (n.d.). EXMARaLDA. Retrieved November 22, 2023, from <https://exmaralda.org/de/corpus-manager-de/>

COSMAS II - Startseite. (n.d.). IDS Mannheim. Retrieved November 22, 2023, from <https://www2.ids-mannheim.de/cosmas2/>

Das DTA-Basisformat. (n.d.). Deutsches Textarchiv. Retrieved November 22, 2023, from <https://www.deutschestextarchiv.de/doku/basisformat/>

Deliverable IO 3.3a. (n.d.). <https://textplus.sync.academiccloud.de/s/BDw66W7CdM5wNK7>

DNBG - Gesetz über die Deutsche Nationalbibliothek. (2006, June 22). Gesetze im Internet. Retrieved November 22, 2023, from <https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html>

Dolmetschen im Krankenhaus (DiK) | ZFDM Repository. (2009, January 5). ZFDM Repository. Retrieved November 22, 2023, from <https://www.fdr.uni-hamburg.de/record/8308>

EXMARaLDA. (n.d.). EXMARaLDA. Retrieved November 22, 2023, from <https://exmaralda.org/de/>

Formulae - Litterae - Chartae: Erweiterte Suche. (n.d.). in der Formulae – Litterae – Chartae Werkstatt! Retrieved November 22, 2023, from https://werkstatt.formulae.uni-hamburg.de/search/advanced_search

Genêt, P., Gradl, T., Hensen, K., Kudella, C., & Schulz, D. (2023). *F wie Registry - Die Text+ Registry als Hilfsmittel zur Auffindbarkeit von Ressourcen*.

<https://zenodo.org/doi/10.5281/zenodo.8392492>

IDS-Textmodell | IDS. (n.d.). IDS Mannheim. Retrieved November 22, 2023, from <https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell/>

KorAP-Annotations. (n.d.). KorAP. Retrieved November 22, 2023, from <https://korap.ids-mannheim.de/doc/data/annotation>

KorAP - Corpus Analysis Platform. (n.d.). KorAP - Corpus Analysis Platform. Retrieved November 22, 2023, from <https://korap.ids-mannheim.de/>

Lüngen, H., & Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3. <http://journals.openedition.org/jtei/508>

Milestone C1.1. (n.d.). <https://zenodo.org/doi/10.5281/zenodo.12771255>

Milestone C5.1. (n.d.). <https://zenodo.org/doi/10.5281/zenodo.12770996>

nestor-Siegel. (2023, January 23). Nestor. Retrieved November 22, 2023, from https://www.langzeitarchivierung.de/Webs/nestor/DE/Zertifizierung/nestor_Siegel/nestor_siegel_node.html

P5 Guidelines – TEI. (n.d.). Text Encoding Initiative. Retrieved November 22, 2023, from <https://tei-c.org/guidelines/p5/>

Partitur-Editor – EXMARaLDA. (n.d.). EXMARaLDA. Retrieved November 22, 2023, from <https://exmaralda.org/de/partitur-editor-de/>

RELAX NG. (2014, February 25). RELAX NG home page. Retrieved November 22, 2023, from <https://relaxng.org/>

Schmidt, T., & Wörner, K. (2005). Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung*, 6, 171-195. <http://gespraechsforschung-ozs.de/heft2005/px-woerner.pdf>

Text Encoding Initiative: TEI. (n.d.). Text Encoding Initiative: TEI. Retrieved November 22, 2023, from <https://tei-c.org/>

TextGrid Python clients documentation. (2023, June 16). Developing tgclients. Retrieved November 22, 2023, from <https://dariah-de.pages.gwdg.de/textgridrep/textgrid-python-clients/docs/development.html>

The TextGrid Repository Documentation. (2023, June 16). The TextGrid Repository Documentation. Retrieved November 22, 2023, from <https://textgridlab.org/doc/services/index.html>

Textorganisation unter COSMAS II - MECOLB-Minimal Tagset. (n.d.). IDS Mannheim. Retrieved November 22, 2023, from <https://www2.ids-mannheim.de/cosmas2/projekt/referenz/annotationen.html>

Tsakorpus. (n.d.). <https://tsakorpus.readthedocs.io/en/latest>