Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Konzept und prototypische Implementierung beispielhafter Ressourcen

Milestone C1.1

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	0.1
Redaktion	01.09.2023
Redaktionsteam	Philippe Genêt, Marius Hug, Peter Leinen, Thorsten Trippel
Projekt	Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur
Bezeichnung	C1.1 Concept and prototypical implementation of representative resources
Förderung	DFG Förderkennzeichen 460033370
Projektlaufzeit	01.10.2021 bis 30.09.2026

Inhalt

1. Task Area Collections

An der Task Area Collections sind 19 Institutionen beteiligt, die umfangreiche Sammlungen sprach- und textbasierter Daten in Text+ einbringen. Es handelt sich dabei um Sammlungen geschriebener, gesprochener oder gebärdeter Sprache und Texte sowie sprach- und textbezogene Experimental- oder Messdaten, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden. Dazu gehören: Textsammlungen (z.B. von literarischen Texten, Sachtexten, Zeitungs- und Zeitschriftentexten, Interviews, Inschriften, Handschriften, Drucken), mono- und multimodale Aufnahmen z.B. von spontaner und formaler Sprache (z.B. von Reden, Dialogen, Nachrichten, Interviews, Interaktion im Alltag), Sensordaten (z.B. EEG, Eyetracking, Artikulographie), Befragungen, Reaktionszeiten etc.

Die Heterogenität der Sammlungen spiegelt sich in den Daten wider: Die Texte, Ton-, Bild- oder Bewegtbildaufnahmen sowie zahlreiche andere Daten liegen samt ihrer jeweiligen Metadaten in verschiedensten Formaten vor. Der Grad der Erschließung und Annotation variiert von Institution zu Institution und von Sammlung zu Sammlung. Ebenso bestehen unterschiedliche Stufen der Zugänglichkeit – frei verfügbare, gemeinfreie Inhalte stehen neben Dokumenten, die nur nachweislich akademischem Publikum zugänglich gemacht werden können; urheberrechtlich geschütztes Material ist genauso darunter wie Bestände, die (auch digital) nur in bestimmten Lesesälen einsehbar sind.

2. Konzept für die Integration von Sammlungen in die Text+ Infrastruktur

Ziel der Text+ Infrastruktur im Hinblick auf die Datendomäne Collections ist es, die Sammlungen der beteiligten Partnerinstitutionen über ein zentrales Webportal auffindbar und durchsuchbar zu machen, ohne die Sammlungen dabei in ein zentrales Datensilo zu überführen, sondern sie ortsverteilt in ihren jeweiligen Heimatrepositorien zu belassen und auf unterschiedlichen Ebenen zu vernetzen. Zwei Werkzeuge spielen dabei Schlüsselrollen: Die Text+ *Registry* dient als Verzeichnisinstrument für Editionen, lexikalische Ressourcen oder eben Sammlungen der Auffindbarkeit dieser Ressourcen und beinhaltet zugleich die entsprechenden Zugangsinformationen. Die *Federated Content Search (FCS)*[1] ist die Suchinfrastruktur, mit der die Volltexte der Ressourcen in Text+ zentral durchsucht werden[2] können. Die Task Area Collections hat für beide Instrumente Konzepte entwickelt, um die heterogenen Sammlungen der Datendomäne einheitlich zu integrieren.

a. Registry

Gemeinsam mit Tobias Gradl von der Universität Bamberg, der die Text+ Registry maßgeblich entwickelt, hat sich die AG Reference Implementation & Portfolio Development im Rahmen eines Workshops Mitte Januar 2023 auf ein Set von Metadaten geeinigt, mit dem Sammlungen beschrieben werden (vgl. Anhang A). Dieses Datenmodell, das für bestimmte Felder auch kontrollierte Vokabulare vorsieht, bildet die Basis für die Entwicklung des Registry-Prototypen, in den die von den verschiedenen Datenzentren identifizierten Beispiel-Sammlungen integriert

[1] vgl. Eckart, T. et al. (2023), Schonefeld, O. et al. (2014) und Stehouwer, H. et al. (2012)

[2] zu den verwendeten Standards vgl. Morgan, E.L. (2004) und OASIS (2013)

werden. Anhand des Prototypen wird es möglich sein, das Datenmodell nötigenfalls iterativ anzupassen und weiterzuentwickeln.

b. Federated Content Seach

Im Rahmen des Milestones C5.1 „Software Requirements for collections-specific aspects of the Federated Content Search“[3] hat die Task Area Collections ihre spezifischen Anforderungen an die FCS formuliert und die Besonderheiten insbesondere im Hinblick auf rechtebewehrte Inhalte, Suche auf Annotations-Layern, spezifische Austauschformate sowie Konfiguration, Testing und Entwicklung von Endpunkten hervorgehoben. Auch die Nutzung der Metadatenfelder aus der Registry als mögliche Such- und Filterfacetten und damit die Verschränkung der beiden Instrumente wurde darin expliziert.

3. Beispielhafte Ressourcen

Für die Implementierung des Registry-Prototypen haben die Datenzentren der Datendomäne Collections einzelne beispielhafte Sammlungen ausgewählt[4] und Beschreibungen zur Verfügung gestellt. Schon diese erste Auswahl bildet die große Bandbreite der in Collections versammelten Datentypen ab. Die Auswahl weiterer zu integrierender Sammlungen erfolgt nach Fertigstellung des Registry-Prototypen. Dann werden einige der Datenzentren auch eine automatisierte Bereitstellung der Sammlungsbeschreibungen erproben, die perspektivisch regelmäßig von der Registry geharvested werden sollen.

4. Prototypische Implementierung

Nach dem Relaunch des Text+ Webportals sollen die Prototypen der Registry[5] und der FCS dort integriert werden. So entsteht für Nutzende des Webportals ein einheitlicher, zentraler Zugang nicht nur zu den Daten der Task Area Collections, sondern auch zu den Ressourcen der anderen Datendomänen Editions und Lexical Resources. Dies stellt ein wichtiges Element der ortsverteilten Infrastruktur von Text+ dar.

a. Registry

Die Registry gibt den Datenzentren für die Sammlungsbeschreibungen kein einheitliches Datenformat vor. Dank des Data Modeling Environments (DME)[6] kann die Registry beliebige Datenformate – XML, csv, yml etc. – problemlos ingestieren, wenn ein individuelles Mapping der Datenfelder einmal erstellt wurde. Dies erlaubt den Datenzentren, das Ausgabeformat selbst zu wählen.

Anders als ihr Prototyp soll die produktive Version der Registry die Sammlungsbeschreibungen der Datenzentren regelmäßig harvesten, idealerweise mit Hilfe von OAI-PMH-Schnittstellen, die in zahlreichen Datenzentren ohnehin vorhanden sind.

---

[3] https://zenodo.org/doi/10.5281/zenodo.12770996

[4] Eine Liste der für den Prototyp der Registry ausgewählten Sammlungen findet sich in Anhang B

[5] vgl. Eckart, T. (2021) und Eckart, T.; Gradl, T. (2017)

[6] vgl. Henrich, A.; Gradl, T. (2021) und Gradl, T.; Henrich, A. (2016)

Die Datenhoheit liegt auch bei den Sammlungsbeschreibungen weiterhin ganz bei den Datenzentren. Sie können in der Beschreibung auch festlegen, wie der Zugang zur jeweiligen Sammlung geschieht. Inhalte, die durch Urheber- und/oder Persönlichkeitsrechte geschützt sind, können durch die datenhaltende Institution entsprechend kenntlich gemacht und die Zugangsbedingungen beschrieben werden – etwa „nur im Lesesaal“ oder „nur für Forschende“.

b. Federated Content Seach

Für die FCS ist hingegen eine einheitliche Infrastruktur notwendig. Damit der Aggregator der FCS Suchanfragen an die verschiedenen Datenzentren verteilen kann, müssen dort jeweils Endpunkte entwickelt werden. Diese „übersetzen“ die Anfrage des Aggregators in eine Suchanfrage, die die lokale Suchinfrastruktur verarbeiten kann, und senden das Ergebnis zurück an den Aggregator. Die Entwicklung des Endpunkts obliegt den datenhaltenden Institutionen.

Rechtebewehrte Inhalte sollen von der Suche nicht ausgenommen werden. Auch hier sollen Treffer gemeldet , aber nicht direkt angezeigt werden (etwa als *keyword in context*). Stattdessen werden die Zugangsmöglichkeiten zu den Dokumenten mit Suchtreffern analog zu den Angaben in der Registry beschrieben.

5. Ausblick

Ein im Herbst 2023 anstehender Bericht wird die Umsetzung der prototypischen Implementierung darlegen.

Im Anschluss an die Integration der Beispiel-Ressourcen und der Aufnahme der Registry und der FCS ins Text+ Webportal werden verschiedene Elemente der Umsetzung evaluiert und ggf. iterativ angepasst – so zum Beispiel das Registry-Datenmodell oder die Anzeige der Suchergebnisse der FCS.

Im nächsten Schritt wird die Automatisierung des Registry-Workflows vorangetrieben – ebenfalls in Iterationen – und weitere Ressourcen identifiziert, die in die Text+ Infrastruktur eingebracht werden.

Literatur

Eckart, T.; Gradl, T. (2017): „Working towards a Metadata Federation of CLARIN and DARIAH-DE“. Utrecht. doi: 10.5281/zenodo.1173604.

Eckart, T.; Gradl, T.; Jegan, R. u. a. (2021): „CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services“. Bamberg: Otto-Friedrich-Universität.

Eckart, T.; Herold, A.; Körner, E.; Wiegand, F. (2023): „A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN“. In „Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography“. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o.

Gradl, T.; Henrich, A. (2016): „Die DARIAH-DE-Föderationsarchitektur : Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen“. Berlin; New York, NY: de Gruyter Saur. doi: 10.1515/bfp-2016-0027.

Henrich, A.; Gradl, T. (2021): „Integration von Forschungsdaten: Wie können Forschungsinfrastrukturen helfen?“. Berlin; Boston: De Gruyter. doi: 10.1515/9783110538915-039.

Morgan, E.L. (2004): „An Introduction to the Search/Retrieve URL Service (SRU)“. Ariadne, 40. URL: http://www.ariadne.ac.uk/issue/40/morgan.

OASIS (2013): „searchRetrieve: Part 0. Organization for the Advancement of Structured Information Standards“. URL: http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html.

Schonefeld, O.; Eckart, T.; Kisler, T.; Draxler, C.; Zimmer, K.; Ďurčo, M.; Panchenko, Y.; Hedeland, H; Blessing, A.; Shkaravska, O. (2014): „CLARIN Federated Content Search (CLARIN-FCS) – Core Specification“. URL: https://www.clarin.eu/content/federated-content-search-core-specification.

Stehouwer, H.; Durco, M.; Auer, E.; Broeder, D. (2012): „Federated Search: Towards a Common Search Infrastructure“. In: „Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)“. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3255–3259. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf.

Anhang A

Registry-Felder Collections

konsolidierte Version (Stand: 20.03.2023)

Legende:
inhaltliche Angaben
technische Angaben
organisatorische Angaben
nicht für Registry-Prototyp

Abgleich der Felder mit
VLO: https://textplus.sync.academiccloud.de/f/435106
schema.org: https://textplus.sync.academiccloud.de/f/435103
DCDDM: https://textplus.sync.academiccloud.de/f/419182

Bezeichnung	Pflicht/ opt.	kontr. Vokabular	Mehrfachangaben möglich?	Bemerkungen
Titel/Name der Sammlung	Pflicht	nein	nein	
Beschreibung	Pflicht	nein	nein	hier: inhaltliche Beschreibung, technische Beschreibung s.u. „technische Dokumentation“
Größenbeschreibung	Pflicht	nein	nein	Größe und Ausmaß wird beschrieben, bei wachsenden Sammlungen: ungefähre Angabe.
Lizenz	Pflicht	ja	ja	default-Eintrag anbieten (restriktivste Option)
Modalität	Pflicht	ja	ja	Angabe: geschrieben, gesprochen, gebärdet etc.
Sprache	Pflicht	ja	ja	Angabe "n.a." oder "unbekannt" möglich; Vokabular ISO 639-3 und/oder GND
Datentypisierung	Pflicht	ja (+Freitext)	ja	Orientierung an VLO-Facette "Format" PLUS weitere individuelle Angaben möglich
Erstellungsdatum	eines der 3 Daten ist verpflichtend	ja	nein	vgl. DC creation date, Format-Vorgabe
Veröffentlichungsdatum		ja	nein	vgl. DC publication date, Format-Vorgabe
abgedeckter Zeitraum		ja	nein	vgl. DC temporal coverage, Format-Vorgabe
Volltext verfügbar?	Pflicht	ja	nein	Angabe: ja/nein
Annotationslayer vorhanden?	Pflicht	ja	nein	Angabe: ja/nein
Welche Annotationslayer sind vorhanden?	optional	ja (+Freitext)	ja	Vokabular orientiert an FCS-Vokabular PLUS weitere individuelle Angaben möglich
Annotationen manuell oder automatisch erstellt?	optional	ja	nein	Angabe: manuell/automatisch

Kollektionstypen	optional	ja	ja	vgl. Liste in MWW, VLO-Facette "Collection", ggf. auch GND-basiert
Genre	optional	nein	ja	Ggf. gedoppelt mit „Kollek-tionstypen", GND als Basis?
fachliche Zuordnung	optional	ja	ja	Vokabular: Basisklassifikation (BK)
Schlagworte	optional	nein	ja	Referenz: GND, VIAF o.ä.
Periodizität	optional	ja	nein	Vorzusehen für dynamische Sammlungen (nicht unbedingt in Prototyp)
Status im Datenlebenszyklus	optional	ja	nein	z.B. abgeschlossen, wachsend, veraltet etc. --> für Prototyp nicht relevant (Versionierung)
PID	Pflicht	nein	nein	mehrere Standards zulassen, PID zeigt auf Sammlung selbst
Zugangsinformation	Pflicht	nein	ja	technische Zugangsdaten Endpunkt (FCS, OAI, API) samt Attribut zum Finden der Sammlung, kann auch sein "Lesesaal in der DNB
Hierarchie	Pflicht	nein	ja	Informationen, ob die Sammlung Teil einer anderen Sammlung ist oder andere Sammlungen beinhaltet
Bezug zu anderen Sammlungen	optional	nein	ja	Auch institutionsübergreifend
Liste der Dateien/Datenströme	optional	nein	ja	ggf. Livelink zu Heimatorganisa-tion (Versionierung muss angedacht werden für dynamisch wachsende Sammlungen)
technische Dokumentation	optional	nein	ja	Achtung: Dopplungen mit anderen Feldern vermeiden!
Datenverantwortliche Person/Institution	Pflicht	nein	nein	i.S.v. datenhaltende Person/Institution GND-Einträge als Basis?
Ansprechperson für Datenzugang	Pflicht	nein	ja	technischer Kontakt
Förderer	optional	ja	ja	Vokabular mit den üblichen Förderern PLUS Freitextfeld, GND als Basis?
Titel des Projekts	optional	nein	nein	Freitext
Förderer-ID	optional	nein	ja	Freitext

Anhang B

Liste der Sammlungen, die für den Prototypen der Text+ Registry ausgewählt wurden

Datenzentrum	Name der Sammlung	Link zur Sammlung
Akademie der Wissenschaften Hamburg	Dolganisch-Korpus	https://inel.corpora.uni-hamburg.de/portal/corpora/dolgan/
Akademie der Wissenschaften Hamburg	Deutsche Gebärdensprache (DGS)-Korpus	https://www.sign-lang.uni-hamburg.de/dgs-korpus/
Deutsche Nationalbibliothek	Freie Online-Hochschulschriften	https://www.dnb.de/dissonline
Deutsche Nationalbibliothek	Exilpresse des Deutschen Exilarchivs	https://www.dnb.de/exilpressedigital
Berlin-Brandenburgische Akademie der Wissenschaften	Deutsches Textarchiv Kernkorpus	https://www.deutschestextarchiv.de/
Leibniz-Institut für Deutsche Sprache	Deutsches ReferenzKorpus DeReKo	https://www.ids-mannheim.de/digspra/kl/projekte/korpora/verfuegbarkeit/
Data Center for the Humanities, Köln	Zaghawa-Wagi-Korpus	https://lac.uni-koeln.de/collection/11341/00-0000-0000-0000-1AC6-9
Hamburger Zentrum für Sprachkorpora	Korpus "Dolmetschen im Krankenhaus (DiK)"	https://www.fdr.uni-hamburg.de/record/8308
Hamburger Zentrum für Sprachkorpora	Korpus "Hamburg Adult Bilingual Language (HABLA)"	https://www.fdr.uni-hamburg.de/record/1351
Hamburger Zentrum für Sprachkorpora	Korpus "The Hamburg MapTask Corpus (HAMATAC)"	https://www.fdr.uni-hamburg.de/record/1481
Hamburger Zentrum für Sprachkorpora	Korpus "TraCES"	https://www.fdr.uni-hamburg.de/record/8334
Universität des Saarlandes	CLARIND-UdS Language Resource Repository	https://fedora.clarin-d.uni-saarland.de/index.de.html
Niedersächsische Staats- und Universitätsbibliothek Göttingen	Textgrid-Sammlung "Digitale Bibliothek"	https://textgridlab.org/1.0/tgoaipmh/oai
Niedersächsische Staats- und Universitätsbibliothek Göttingen	Textgrid-Sammlung "European Literary Text Collection (ELTeC)"	https://textgridlab.org/1.0/tgoaipmh/oai
Niedersächsische Staats- und Universitätsbibliothek Göttingen	Textgrid-Sammlung "Achitrave"	https://textgridlab.org/1.0/tgoaipmh/oai
Niedersächsische Staats- und Universitätsbibliothek Göttingen	Textgrid-Sammlung "CoNSSA: Corpus of Novels of the Spanish Silver Age (version 2.0.0)"	https://textgridlab.org/1.0/tgoaipmh/oai
Niedersächsische Staats- und Universitätsbibliothek Göttingen	Textgrid-Sammlung "Neologie"	https://textgridlab.org/1.0/tgoaipmh/oai