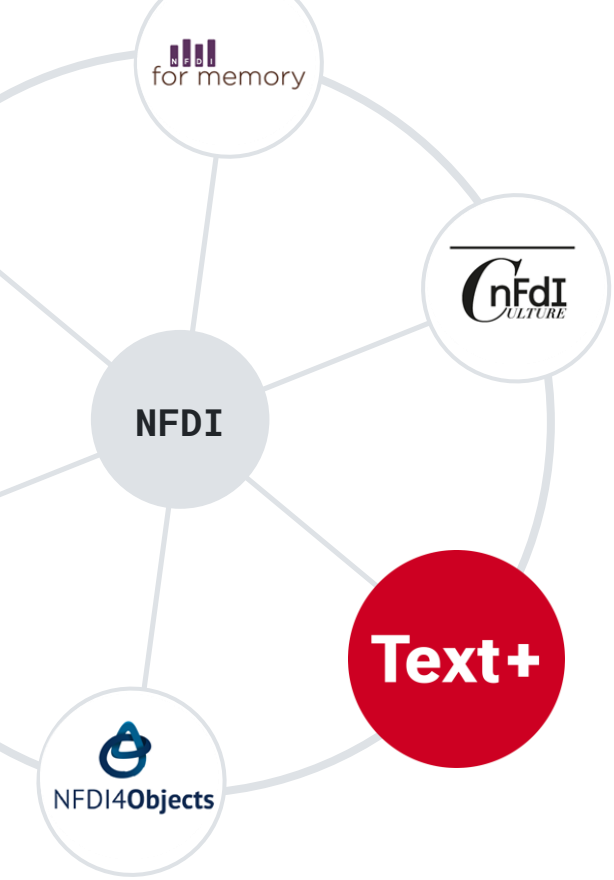


Sammlungen vernetzen und für die Forschung nutzbar machen – eine Aufgabe (auch) für Bibliotheken

Philippe Genêt und Peter Leinen, Deutsche Nationalbibliothek
José Calvo Tello und Lukas Weimer, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 460033370

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

DEUTSCHE NATIONAL BIBLIOTHEK

SUB NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

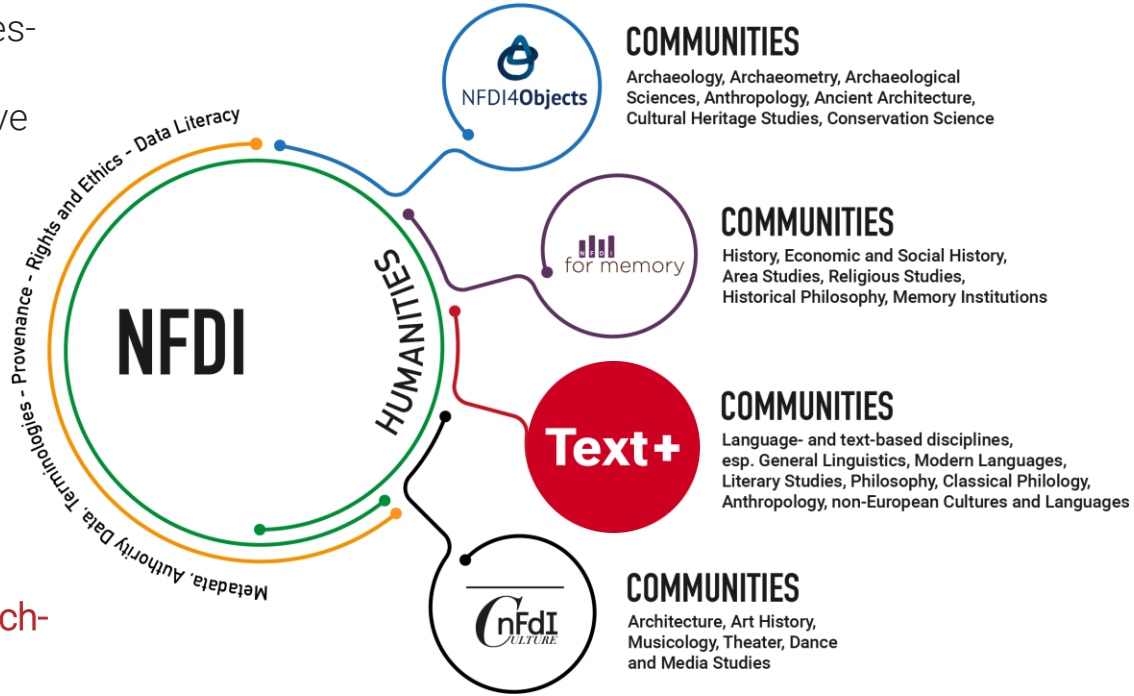
AKADEMIE WISSENSCHAFTEN UND KÜNSTE NORDRHEIN-WESTFALEN

Text+: Ein Konsortium der NFDI

- Das zweite von insgesamt vier geistes- und kulturwissenschaftlichen Konsortien der bundesweiten Initiative zum Aufbau einer nationalen Forschungsdateninfrastruktur (NFDI) Förderperiode I: 10/21 - 09/26

Fokus:

- Aufbau einer auf die Bedarfe der Fachcommunitys zugeschnittenen Forschungsdateninfrastruktur
- Langfristige Erhaltung und breite wissenschaftliche Nutzung von Sprach- und Textdaten



Ziele von Text+

Ortsverteilte Infrastruktur

- Zentraler Zugriff über Text+ Portal
- Gemeinsamer Suchraum für dezentral vorliegende Ressourcen
- Vernetzung und Kontextualisierung von Forschungsdaten (z.B. mit Normdaten)
- Unterstützung von Text-and-Data-Mining (TDM)

FAIRification von Forschungsdaten

- Findable: für andere auffindbar
- Accessible: für andere zugänglich
- Interoperable: für andere mit gängigen Werkzeugen be-/ verarbeitbar
- Reusable: für andere wiederverwendbar in anderen Forschungskontexten

„One stop shop“ für sprach- und textbasierte Ressourcen

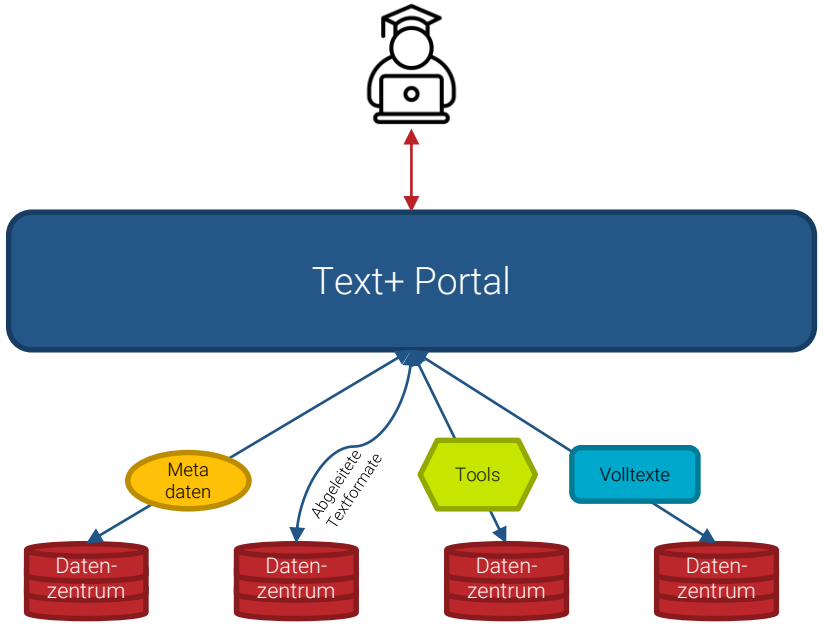
Ortsverteilte Infrastruktur

Gemeinsamer Suchraum

- Suche nach Ressourcen über Metadaten
  → Text+ Registry
- Semantische Suche in Volltexten
  → Federated Content Search
- Erschließung rechtebewehrter Inhalte
  → Abgeleitete Textformate und Authentifizierungs- und Autorisierungs-Infrastruktur (AAI)

Nutzung der Daten

- Bereitstellung von Tools
- Aufbereitung/Vorprozessierung der Ressourcen



Icon by freepik

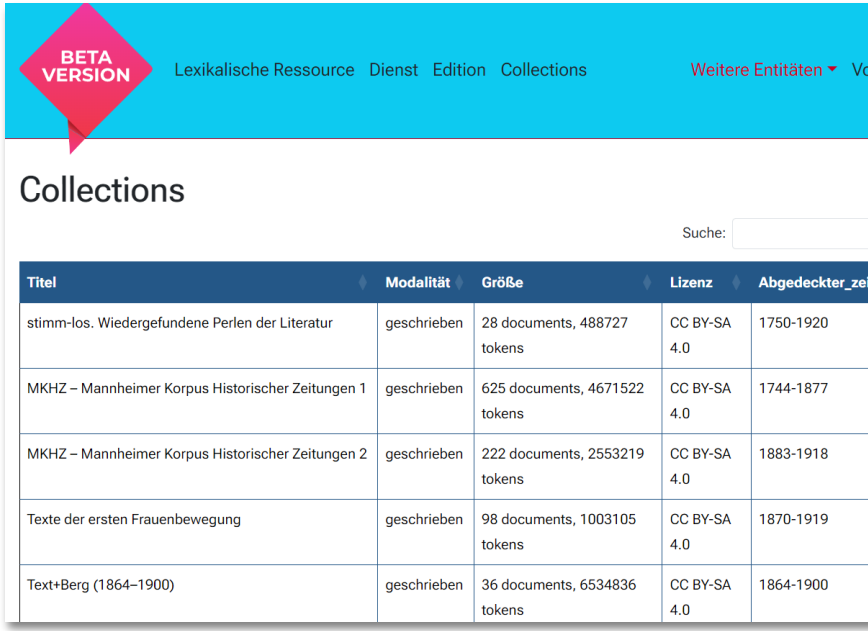
Die Text+ Registry

Suche nach Ressourcen über Metadaten

- Verzeichnissystem für Sammlungen, lexikalische Ressourcen, Editionen und Dienste
- Erhöhte Sichtbarkeit und Auffindbarkeit der eigenen Sammlungen
- Zentraler Zugangspunkt zu Ressourcen
- Schnittstellen zur Anschlussfähigkeit an bestehende Systeme (intern/extern) und übergreifende Strukturen (DFG, NFDI, EOSC etc.)

Das Plus der Text+ Registry

- Umfangreiches Datenmodell für komplexe Suchen
- Gute Bedienbarkeit durch wenige einfache Pflichtfelder
- Offen für neue Ressourcen



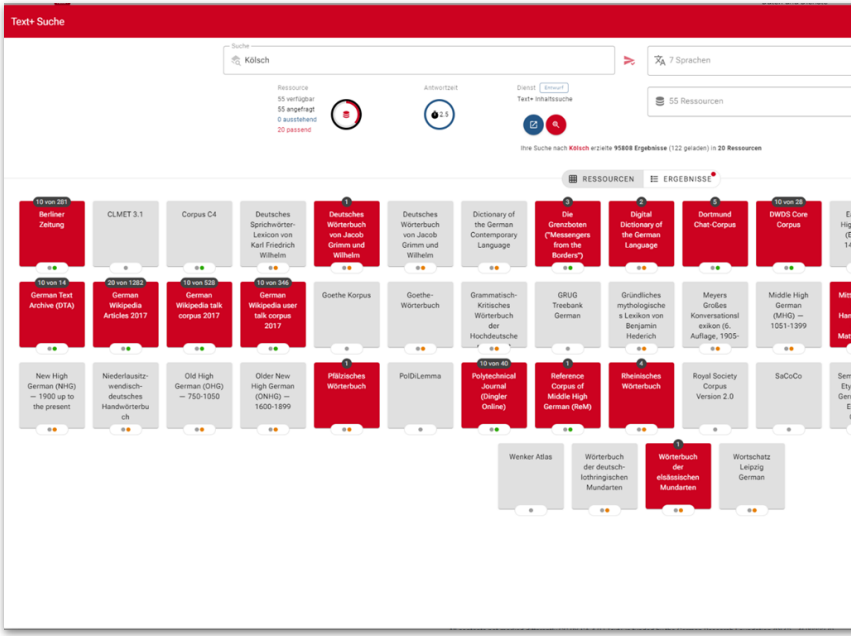
Titel	Modalität	Größe	Lizenz	Abgedeckter_ze
stimm-los. Wiedergefundene Perlen der Literatur	geschrieben	28 documents, 488727 tokens	CC BY-SA 4.0	1750-1920
MKHZ – Mannheimer Korpus Historischer Zeitungen 1	geschrieben	625 documents, 4671522 tokens	CC BY-SA 4.0	1744-1877
MKHZ – Mannheimer Korpus Historischer Zeitungen 2	geschrieben	222 documents, 2553219 tokens	CC BY-SA 4.0	1883-1918
Texte der ersten Frauenbewegung	geschrieben	98 documents, 1003105 tokens	CC BY-SA 4.0	1870-1919
Text+Berg (1864–1900)	geschrieben	36 documents, 6534836 tokens	CC BY-SA 4.0	1864-1900

Quelle: https://c105-100.cloud.gwdg.de/registry/, Genêt, P., Gradl, T., Hensen, K., Kudella, C., & Schulz, D. (2023). F wie Registry - Die Text+ Registry als Hilfsmittel zur Auffindbarkeit von Ressourcen. FORGE 2023 - Anything Goes?! Forschungsdaten in den Geisteswissenschaften - kritisch betrachtet (FORGE2023), Tübingen, Deutschland. Zenodo. https://doi.org/10.5281/zenodo.8392493.

Die föderierte Inhaltssuche (FCS)

Suche nach Inhalten in den Ressourcen von Text+

- dezentrale Volltextsuche
- Anschluss über individuell zu programmierende Endpunkte
- Enge Verzahnung mit Registry (Metadaten = Suchfacetten)
- Zwei Suchvarianten: „Simple Search“ & „Advanced Search“

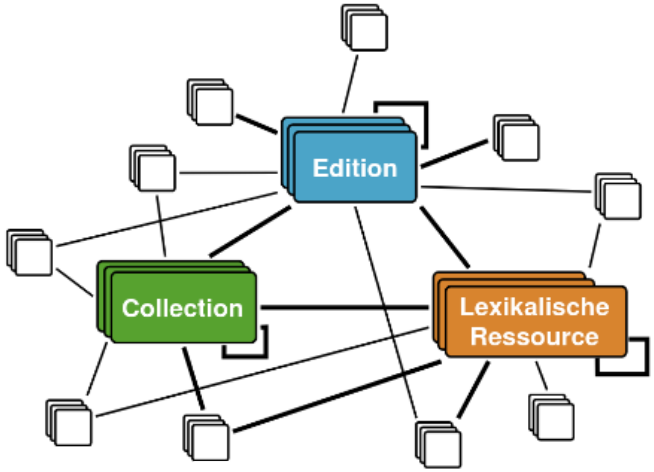


Das Plus der Text+ FCS

- Referenzierbarkeit über PIDs
- Schnittstellen
- Eingeschränkte Trefferanzeige, z.B. für urheber-, lizenz-, datenschutzrechtlich relevante Ressourcen möglich

Quelle: https://www.text-plus.org, Herold, A., & Wiegand, F. (2022). Wissen über Wörter – Die Datendomäne »Lexical Resources«. Text+ Plenary 2022 (TextPlusPlenary), Mannheim. Zenodo. https://doi.org/10.5281/zenodo.7252068.

Daten inhaltlich vernetzen



Auf Metadatenebene:

- Beziehungen zwischen Sammlungen (z.B. Teil/Ganzes)
- Bezüge zwischen Ressourcen unterschiedlicher Art
- Kontextualisierung anhand von Entitäten wie Organisationen, Personen, Projekten und Werkzeugen

Dank Interoperabilität:

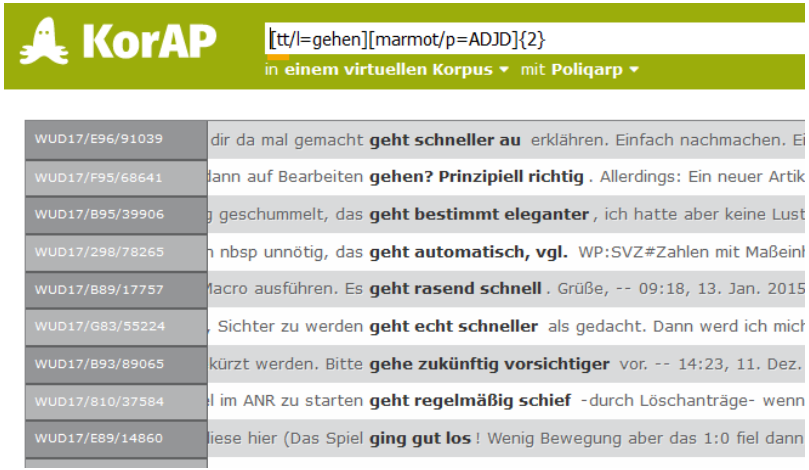
- Linked Data Techniken, gemeinsame Vokabulare, Klassifikations-systeme und Normdaten
- Schnittmengen in Ressourcenbeschreibungen für die übergreifende Betrachtung von Ressourcen

Mit Daten anderer:

- Schnittstellen machen die Daten in Text+ anschlussfähig an andere Infrastrukturangebote (NFDI, EOSC etc.) und Wissensgraphen

Beispiel KorAP@DNB

- Kooperation zwischen DNB und IDS innerhalb von Text+
- Prototypische Implementierung einer übergreifenden Suche auf linguistischen Annotationen, Metadaten und im Text (KorAP = Korpus-Analyseplattform) über Bestände der beiden Institutionen hinweg
- Vorarbeit: Automatische Extraktion und Annotation von Volltexten aus ePUBs der DNB



Rechtebewehrte Ressourcen erschließen und bereitstellen

Abgeleitete Textformate:

- Informationsreduktion bis zur Unterschreitung der Schutz-schwelle des Urheberrechts
- z.B.: Bag-of-words, N-Gramme, Maskierung/Randomisierung von Zeichen, Worten oder Sätzen etc.

Gestaltung von Trefferlisten bei der Volltextsuche:

- z.B. „Treffer gefunden, Ressource nicht online verfügbar“
- oder Beschränkung auf wenige Zeichen/Worte vor/nach dem gesuchten Begriff

Zugangsmanagement:

- AAI zur Identifikation akademischer Nutzender

Icons by freepik

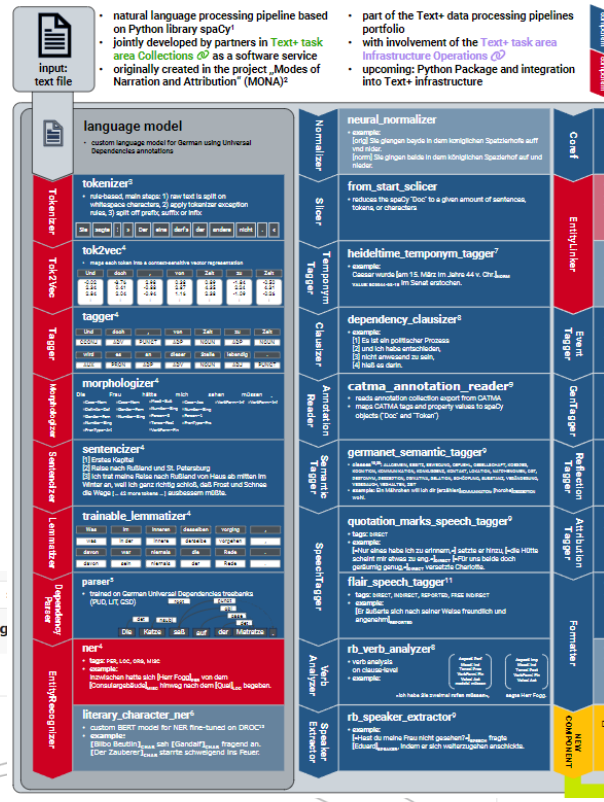
TDM unterstützen

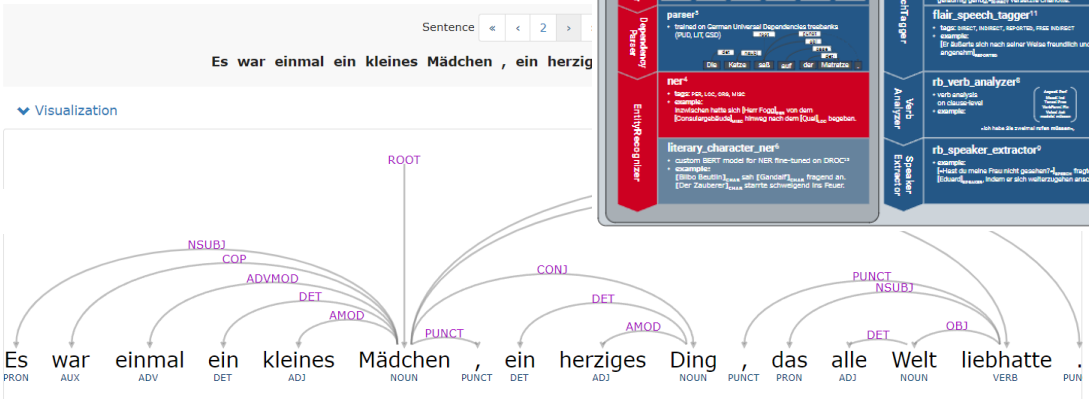
Tools zur Verfügung stellen:

- Mit frei verfügbaren Werkzeugen fördert Text+ die Verarbeitung von Sprach- und Textdaten.
- WebLicht, MONAPipe etc. ermöglichen die gängigen Methoden Text-and-Data-Mining anzuwenden.
- Von der Forschungscommunity entwickelte Komponenten können eingebaut werden.

Sammlungen für TDM optimieren:

- Tools können auch von den Institutionen in Text+ genutzt werden, um ihre Ressourcen für TDM-Verfahren zu optimieren





Quellen: https://weblicht.sfs.uni-tuebingen.de/weblicht/, https://doi.org/10.5281/zenodo.8424925.

Integration von Normdaten im TextGrid-Repository

- TEI-spezifisches Repositorium
- Unterstützung beim Import mit neuem Workflow
- Ressourcen miteinander verknüpfen
- Eindeutige Identifizierung von Entitäten (Personen, Orte, Begriffe)

GND

Gemeinsame Normdatei (GND)

- Identifikation von Autor*innen
- +++ NEU +++ NEU +++ NEU +++
  - Nach Autor*innen mit der GND-ID suchen
  - Verwendung von Sachbegriffen (v.a. Gattungen) für die Erschließung von Texten
  - Integration von Daten aus der GND in das Portal (experimentell)

Basisklassifikation (BK) im Projekt ELTeC

- Texte werden als Primärquelle und als Teil einer Sprache/Sprachfamilie eindeutig beschrieben.
- Basisklassifikation als mittelgroßes System für die Erschließung von Forschungsdaten und elektronischen Publikationen besser geeignet als andere.

Übernahme von Sammlungen/Forschungsdaten

Gegenwartssprachliche Daten:

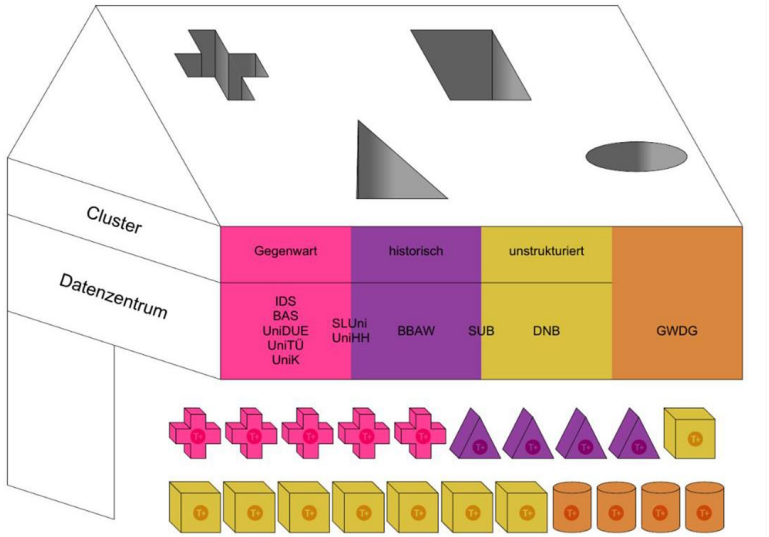
- soziolinguistische Interviewdaten.
- Dokumentation des Sprachwandels,
- Anwendungsorientierte Sprachdaten

Unstrukturierte Texte:

- Online-Hochschulschriften
- E-Journals
- E-Books

Historische Texte:

- Novellen aus dem 19. Jahrhundert
- Briefkorpus (1745–1872)
- Pesttraktate (1473–1700)



Andere Daten mit geisteswissen-schaftlichem Bezug:

- Projektergebnisse (PDF, TEI, METS…)
- Projektdaten (Text, Bild, Ton, Video)
- Daten aus Nachlässen, Archiven, Biblio-theksbeständen

Quelle: Draxler, C., Genêt, P., Hug, M., Lehmberg, T., & Trippel, T. (2023). Wohin damit? Wir geben Ihren Daten ein Zuhause. Text+ Plenary 2023: Connecting People and Data, SUB Göttingen. Zenodo. https://doi.org/10.5281/zenodo.10034001.

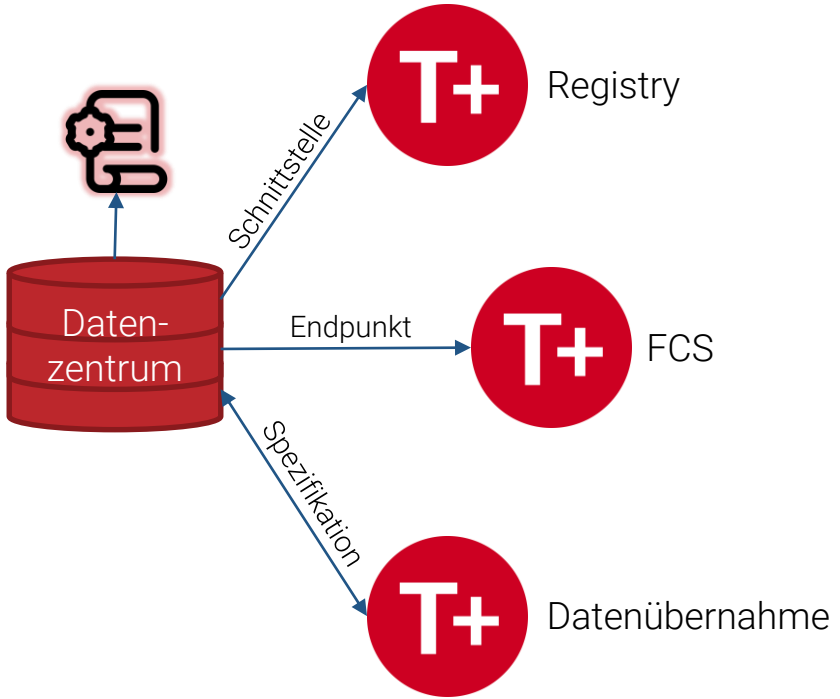
Text+ Datenzentrum werden

Eigene Bestände in Text+ integrieren:

- Sammlungsbeschreibungen erstellen und per Schnittstelle zum Harvesting durch die Text+ Registry bereitstellen
- Anschluss an die Federated Content Search durch Entwicklung eines eigenen FCS-Endpunkts

Daten Dritter aufnehmen:

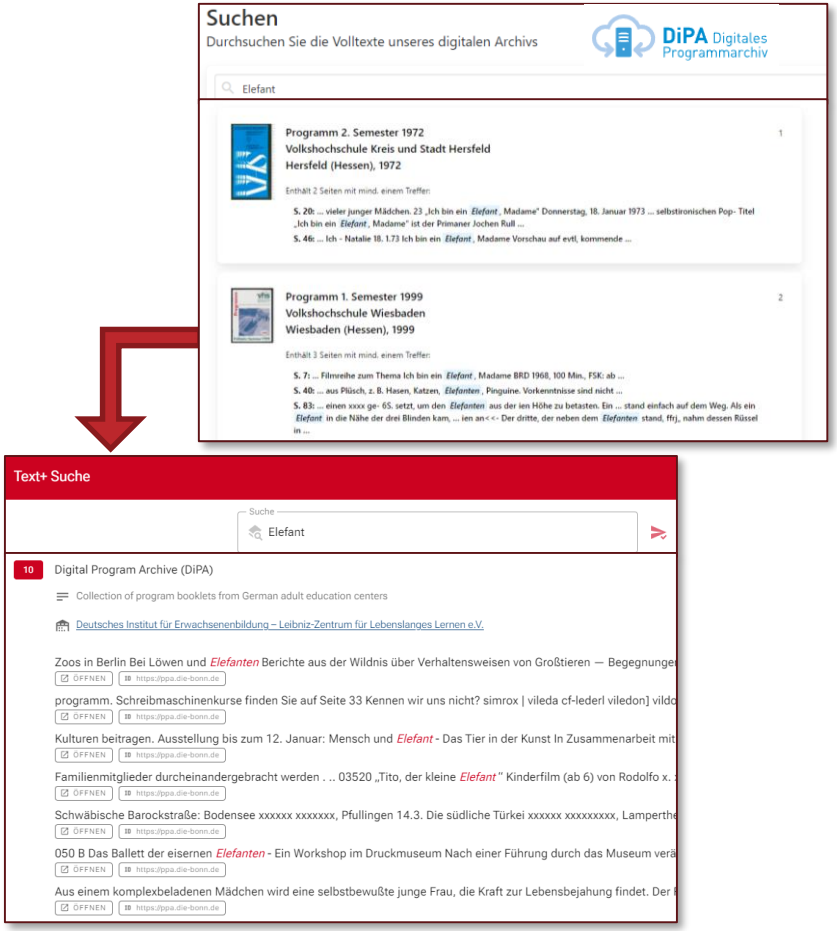
- ggf. spezifizieren, welche Daten von Ihrem Datenzentrum aufgenommen werden können



Icon by freepik

Beispiel DiPA+

- Durch Text+ gefördertes Kooperationsprojekt des Deutschen Instituts für Erwachsenenbildung (DIE)
- **Ziel:** Integration des Volkshochschul-Programmarchivs (ab 1950) in die Text+ Infrastruktur
- **Herausforderung:** Daten können aus rechtlichen Gründen nicht an ein anderes Datenzentrum überführt werden.
- **Lösung:** Das DIE wird selbst zum Datenzentrum
  - Eintragung der Sammlung in die Text+ Registry
  - Anbindung der Programm-Volltexte an die FCS



Quellen: https://text-plus.org/vernetzung/kooperationsprojekte/, https://www.die-bonn.de/id/42004/about/html, https://ppa.die-bonn.de/#/.

Text+: Die sprach- und textbasierte Forschungsdateninfrastruktur

9 Akademien

14 Universitäten

5 Bibliotheken und Archive

7 Stiftungen und Zentren

27 Fachverbände bzw. –verbünde

11 Fachinformationsdienste

11 Daten- und Kompetenzzentren:

- BBAW: Deutsches Textarchiv
- Deutsche Nationalbibliothek
- GWDG: Text+ Langzeitarchiv
- Leibniz Institut für Deutsche Sprache
- LMU: Bayerisches Archiv für Sprachsignale Repository
- SUB Göttingen: DARIAH DE Repository; TextGrid Repository
- UdS: Repositorium für Sprachressourcen
- Uni Duisburg-Essen: Kompetenzzentrum für parlamentarische Sprachdaten
- Uni Hamburg: Hamburger Zentrum für Sprachkorpora, Akademie der Wissenschaften in Hamburg
- Uni Tübingen: Tübingen Archive of Language Resources
- Uni Köln: Data Center for the Humanities Köln

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 46003337