

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



Deutsche
Forschungsgemeinschaft

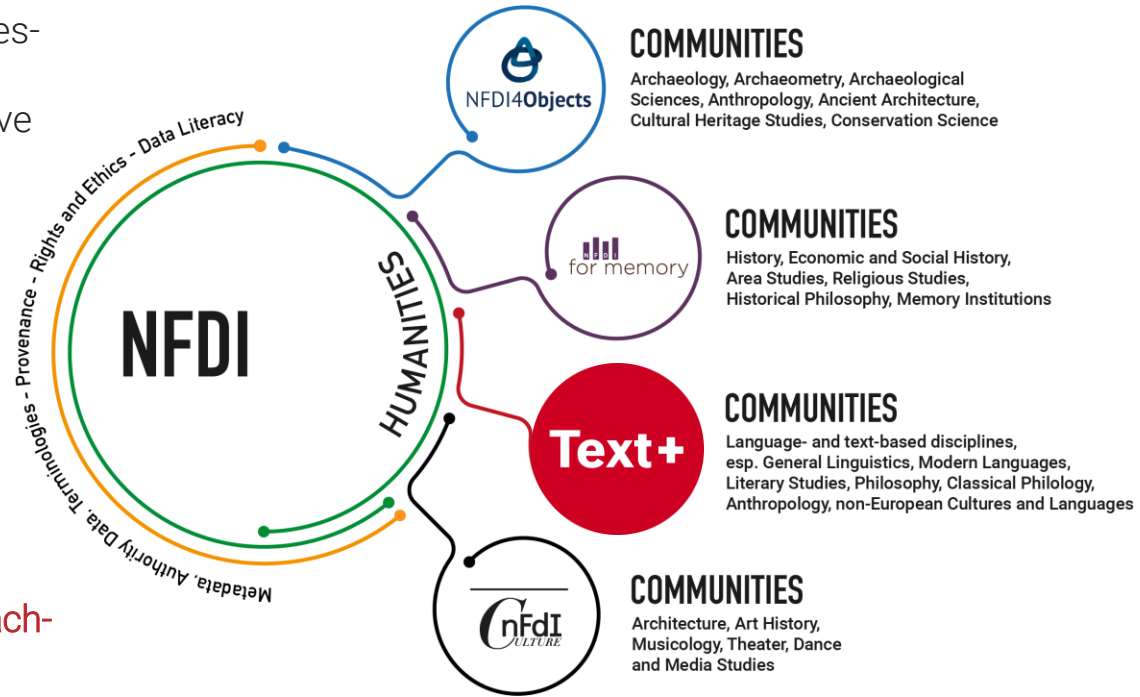
Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche
Forschungsgemeinschaft (DFG) - Projektnummer 460033370

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 Fachcommunities** 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

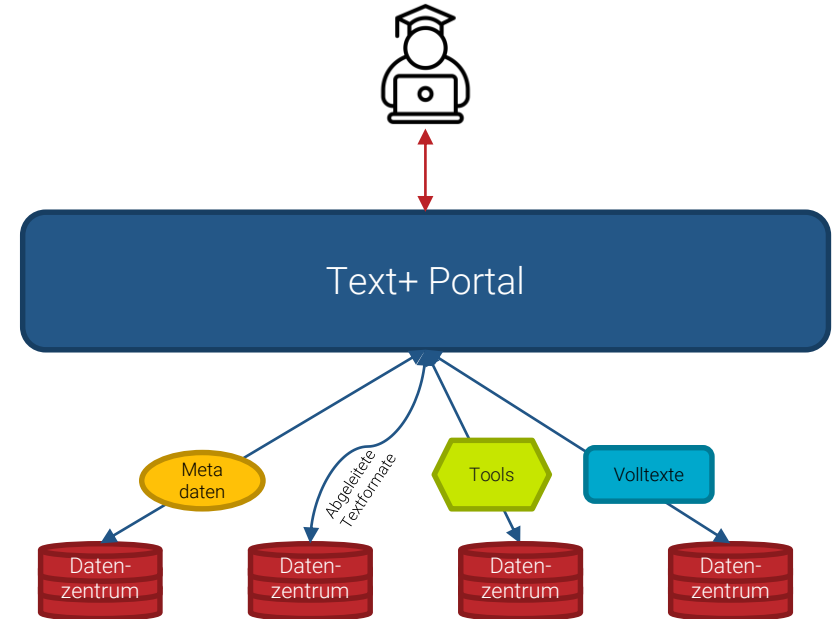
Ortsverteilte Infrastruktur

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



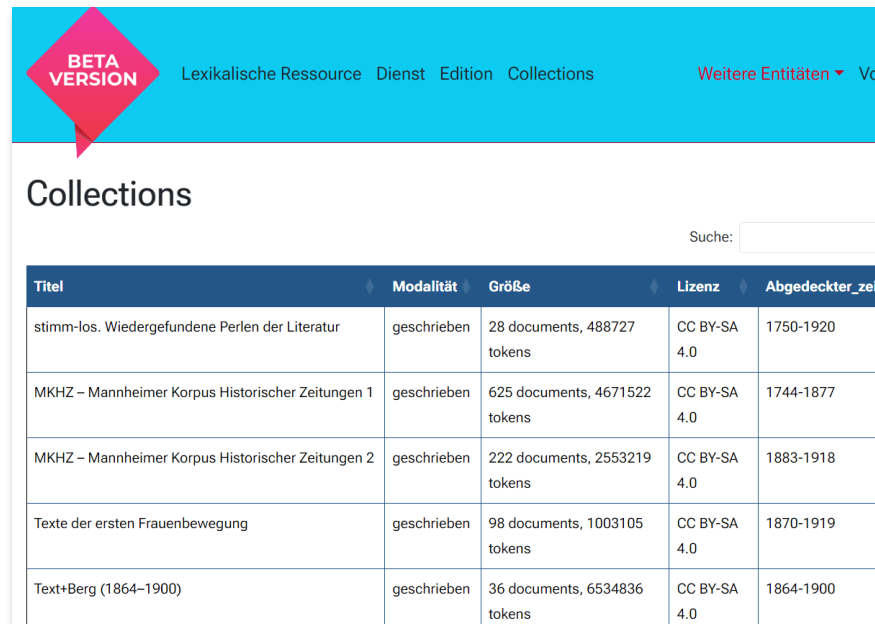
Die Text+ Registry

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



The screenshot shows the Text+ Registry website interface. At the top, there is a blue header with a red diamond containing the text 'BETA VERSION'. To the right of the diamond are navigation links: 'Lexikalische Ressource', 'Dienst', 'Edition', 'Collections', and 'Weitere Entitäten'. Below the header, the word 'Collections' is displayed in large blue font. To the right of 'Collections' is a search bar with the placeholder text 'Suche:'. Below the search bar is a table with the following columns: 'Titel', 'Modalität', 'Größe', 'Lizenz', and 'Abgedeckter Zeitraum'. The table contains five rows of data.

Titel	Modalität	Größe	Lizenz	Abgedeckter Zeitraum
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

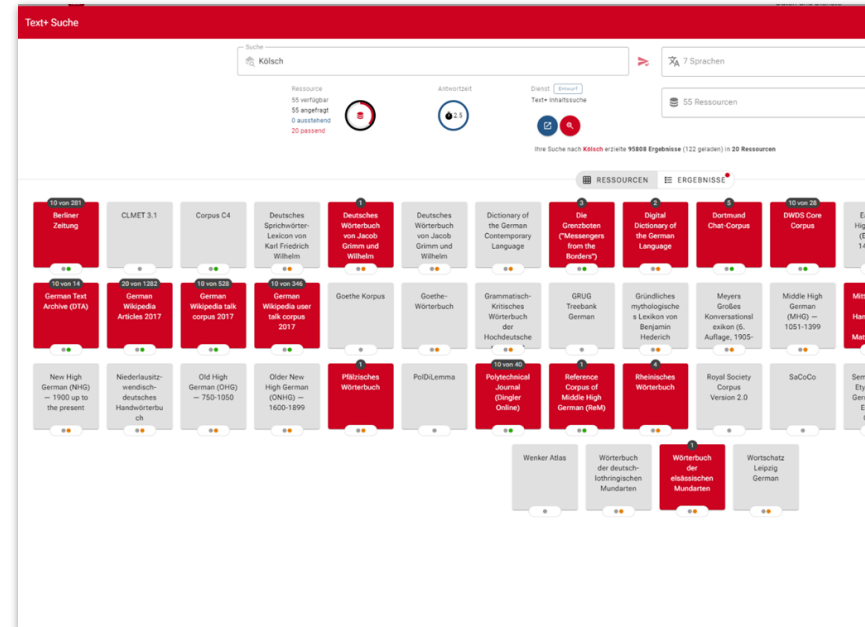
Die föderierte Inhaltssuche (FCS)

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



Daten inhaltlich vernetzen

Auf Metadatenebene:

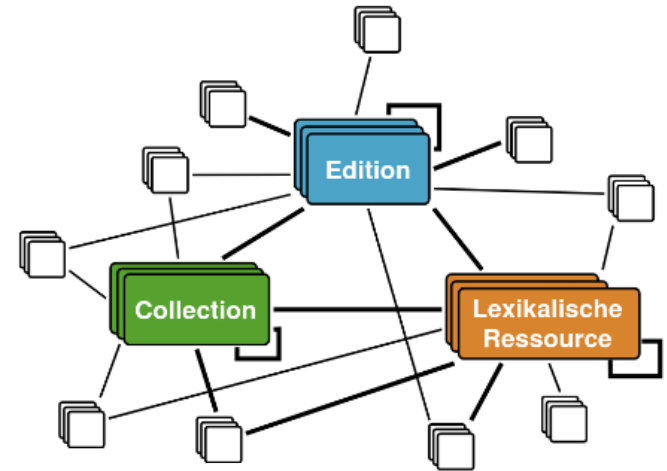
- Beziehungen zwischen Sammlungen (z.B. Teil/Ganzes)
- Bezüge zwischen Ressourcen unterschiedlicher Art
- Kontextualisierung anhand von Entitäten wie Organisationen, Personen, Projekten und Werkzeugen

Dank Interoperabilität:

- Linked Data Techniken, gemeinsame Vokabulare, Klassifikationssysteme 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

 **KorAP**
in einem virtuellen Korpus ▾ mit Poliqarp ▾

WUD17/E96/91039	dir da mal gemacht geht schneller au erklären. Einfach nachmachen. Ei
WUD17/F95/68641	dann auf Bearbeiten gehen? Prinzipiell richtig . Allerdings: Ein neuer Artik
WUD17/B95/39906	g geschummelt, das geht bestimmt eleganter , ich hatte aber keine Lust
WUD17/298/78265	n bsp unnötig, das geht automatisch, vgl. WP:SVZ#Zahlen mit Maßeinl
WUD17/B89/17757	macro ausführen. Es geht rasend schnell . Grüße, -- 09:18, 13. Jan. 2015
WUD17/G83/55224	. Sichter zu werden geht echt schneller als gedacht. Dann werd ich mich
WUD17/B93/89065	kürzt werden. Bitte gehe zukünftig vorsichtiger vor. -- 14:23, 11. Dez.
WUD17/810/37584	l im ANR zu starten geht regelmäßig schief -durch Löschanträge- wenn
WUD17/E89/14860	iese hier (Das Spiel ging gut los ! Wenig Bewegung aber das 1:0 fiel dann

 **KorAP** @DNB
in allen Korpora ▾ mit Poliqarp ▾

DNB12/BWM/34261	uf den Tisch und ging überraschend behände zu dem unter Ordner
DNB12/HHA/54561	ein müsste? »Wir gehen vorsichtig näher an Auxonia heran, Mnemo
DNB12/CAW/19397	meisten Besitzer gingen ausgesprochen hart und grausam mit ihre
DNB12/HHM/14525	r schwitzte, und ging eilig weiter und das Lächeln verließ langsam ε
DNB12/WSG/14665	ny strahlend und ging sichtlich zufrieden zum Ausgang. "Ach ja, zie
DNB12/MKH/98280	die Schulter und ging fröhlich pfeifend seines Weges.« »Juchhuu!«,
DNB12/VEA/34496	chwindigkeit und ging gleichzeitig tiefer . Auf dem Frontbildschirm s
DNB12/RCT/10226	Visitenkarte und ging äußerlich gefasst in die Dunkelheitdes verlas
DNB12/YKA/14953	cl!" Er redete und ging wild gestikulierend auf und ab. Amber hatte

Rechtebewehrte Ressourcen erschließen und bereitstellen



Abgeleitete Textformate:

- Informationsreduktion bis zur Unterschreitung der Schutzwelle 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

TDM unterstützen

Tools zur Verfügung stellen:

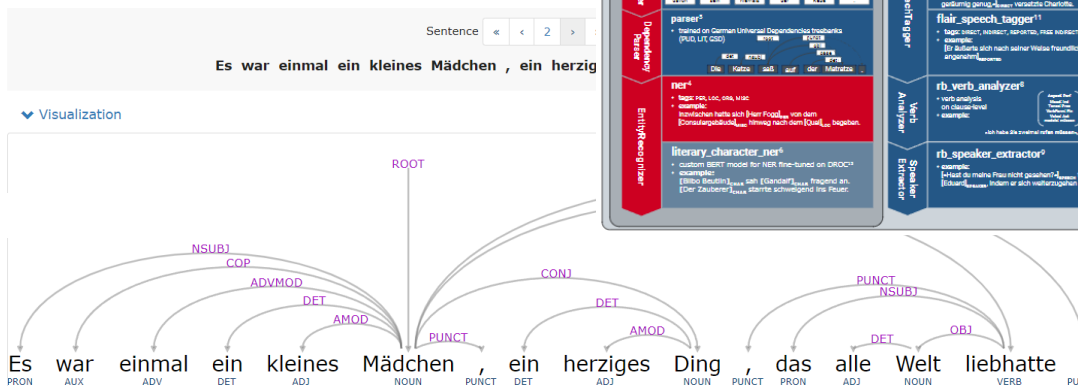
- Mit frei verfügbaren Werkzeugen fördert Text+ die Verarbeitung von Sprach- und Textdaten.
- WebLicht, MONAPipe etc. ermöglichen die gängigen Methoden Text- and-Data-Mining anzuwenden.
- Von der Forschungscommunity entwickelte Komponenten können eingebaut werden.

Sammlungen für TDM optimieren:

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

The screenshot displays the Text+ interface with a sidebar on the left containing tool categories: Input, Tokenizer, Tagging, Morphology, Sentences, Lemmatization, and Dependency. The main area shows a list of tools with their descriptions and examples. On the right, a vertical navigation bar lists tool categories: Normalizer, Slice, Temporal, Classifier, Annotation, Semantic, Speech, Web, and Speaker/Extractor.

- language model**: custom language model for German using Universal Dependencies annotations
- tokenizer⁴**: Naïve-based main steps: 1) raw text to split on whitespace characters, 2) apply tokenizer occupation rules, 3) split off paths, suffix or tags
- tok2vec⁴**: most recent text embeddings: word2vec, word_embeddings
- tagger⁴**: POS tagging
- morphologizer⁴**: Morphological analysis
- sentencizer⁴**: [1] Erstes Kapital [2] Keine nach Rutland und St. Paternan [3] Ich trat meine Hufe nach Rutland-och Hase ab mitan im Winter an, weil ich ganz fertig war, weil Frost und Schnee die Wege ... [4] mich sehen ... [5] Substantiv mitlie
- trainable_lemmatizer⁴**: Lemmatization
- parser⁴**: based on German Universal Dependencies toolkits (PUD, LIT, GSD)
- NER⁴**: Named Entity Recognition
- literary_character_ner⁴**: custom NER model for NER fine-tuned on DRDCH
- neural_normalizer**: normalizes the space 'to' to a given amount of sentences, tokens or characters
- from_start_scifer**: reduces the space 'to' to a given amount of sentences, tokens or characters
- heideftime_temponym_tagger⁷**: handles time and place names
- dependency_clauser⁴**: dependency parsing
- catma_annotation_reader⁹**: reads annotation collection export from CATMA
- germanet_semantic_tagger⁹**: reads CATMA tags and properly maps to spacy objects ('Doc' and 'Token')
- quotation_marks_speech_tagger⁹**: tags speech
- lit_speech_tagger¹¹**: tags speech
- rb_verb_analyzer⁸**: verb analysis
- rb_speaker_extractor⁸**: speaker extraction



Integration von Normdaten im TextGrid-Repository



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



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



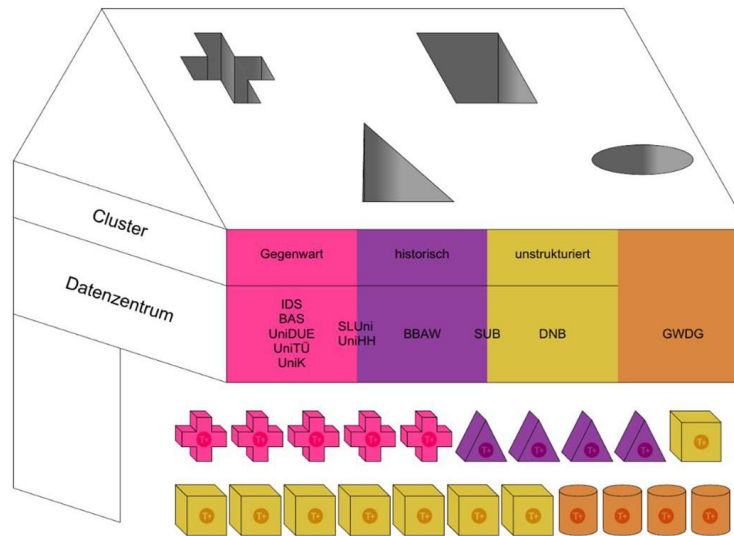
Unstrukturierte Texte:

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



Historische Texte:

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



Andere Daten mit geisteswissenschaftlichem Bezug:

- Projektergebnisse (PDF, TEI, METS...)
- Projektdaten (Text, Bild, Ton, Video)
- Daten aus Nachlässen, Archiven, Bibliotheksbeständen



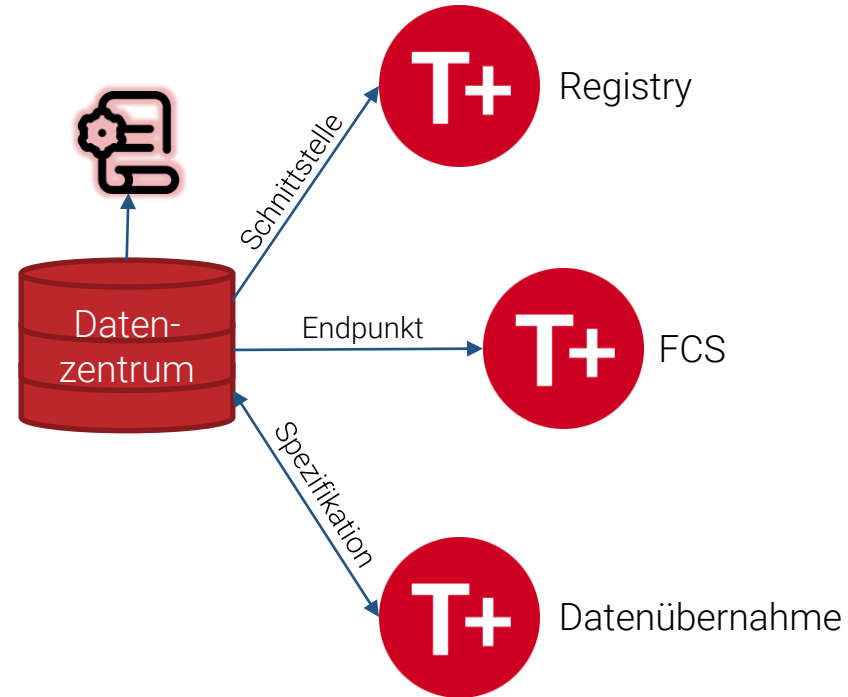
Text+ Datenzentrum werden

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



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



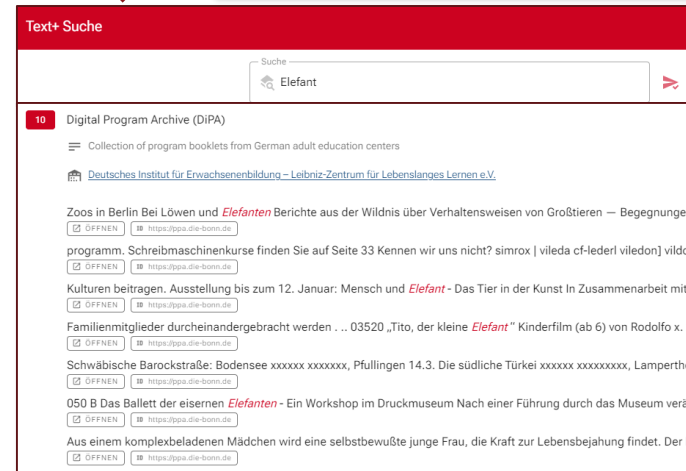
Suchen
Durchsuchen Sie die Volltexte unseres digitalen Archivs

DiPA Digitales Programmarchiv

Suche
Elefant

1
Programm 2. Semester 1972
Volkshochschule Kreis und Stadt Hersfeld
Hersfeld (Hessen), 1972
Enthält 2 Seiten mit mind. einem Treffen:
S. 2b: ... vieler junger Mädchen. 23 „Ich bin ein **Elefant**“, Madame“ Donnerstag, 18. Januar 1973 ... selbstironischen Pop- Titel „Ich bin ein **Elefant**“, Madame“ ist der Primaner Jochen Rull ...
S. 4b: ... Ich - Natalie 18. 1.73 Ich bin ein **Elefant**“, Madame Vorschau auf evtl. kommende ...

2
Programm 1. Semester 1999
Volkshochschule Wiesbaden
Wiesbaden (Hessen), 1999
Enthält 3 Seiten mit mind. einem Treffen:
S. 7: ... Filmehe zum Thema Ich bin ein **Elefant**“, Madame BRD 1968, 100 Min, FSK: ab ...
S. 4b: ... aus Plüsch, z. B. Hasen, Katzen, **Elefanten**“, Pinguine. Vorkenntnisse sind nicht ...
S. 8b: ... einen xxxx ge- 65. setzt, um den **Elefanten**“ aus der Ien Höhe zu betasten. Ein ... stand einfach auf dem Weg. Als ein **Elefant**“ in die Nähe der drei Blinden kam, ... ien an << - Der dritte, der neben dem **Elefanten**“ stand, ffj., nahm dessen Rüssel ih ...



Text+ Suche

Suche
Elefant

10 Digital Program Archive (DiPA)
Collection of program booklets from German adult education centers
Deutsches Institut für Erwachsenenbildung – Leibniz-Zentrum für Lebenslanges Lernen e.V.

Zoos in Berlin Bei Löwen und **Elefanten** Berichte aus der Wildnis über Verhaltensweisen von Großtieren — Begegnung
[OFFNEN] [https://ppa.die-bonn.de]

programm. Schreibmaschinenkurse finden Sie auf Seite 33 Kennen wir uns nicht? simrox | vileda cf-leder| viledon| vild
[OFFNEN] [https://ppa.die-bonn.de]

Kulturen beitragen. Ausstellung bis zum 12. Januar: Mensch und **Elefant** - Das Tier in der Kunst In Zusammenarbeit mit
[OFFNEN] [https://ppa.die-bonn.de]

Familienmitglieder durcheinandergebracht werden ... 03520 „Tito, der kleine **Elefant**“ Kinderfilm (ab 6) von Rodolfo x.
[OFFNEN] [https://ppa.die-bonn.de]

Schwäbische Barockstraße: Bodensee xxxxxxx xxxxxxxx, Pfullingen 14.3. Die südliche Türkei xxxxxxx xxxxxxxx, Lampthe
[OFFNEN] [https://ppa.die-bonn.de]

050 B Das Ballett der eisernen **Elefanten** - Ein Workshop im Druckmuseum Nach einer Führung durch das Museum verä
[OFFNEN] [https://ppa.die-bonn.de]

Aus einem komplexbeladenen Mädchen wird eine selbstbewußte junge Frau, die Kraft zur Lebensbejahung findet. Der
[OFFNEN] [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. –verbände

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