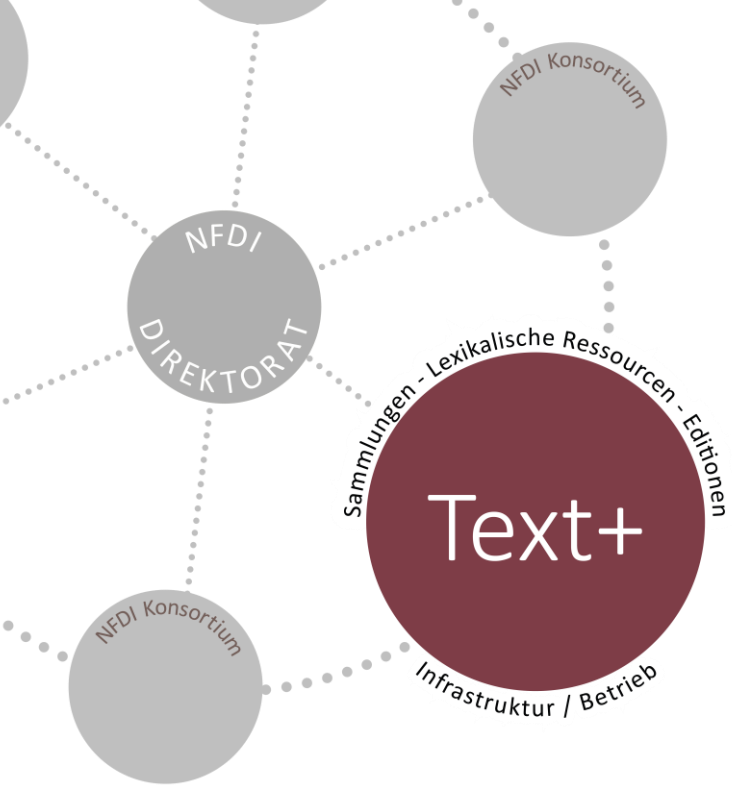
Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 460033370

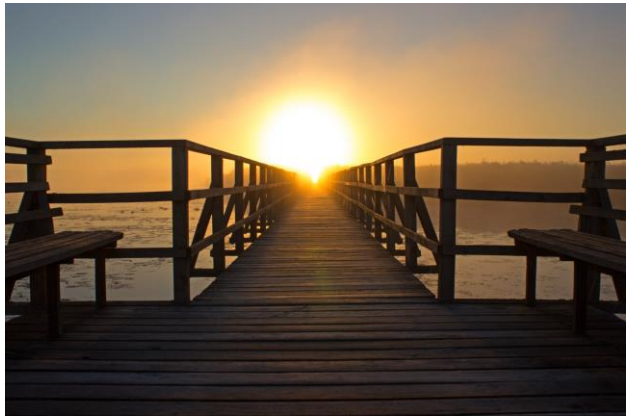


Use Case: Data Depositing in Text+: Strategie und Voraussetzungen

Andreas Witt, Leibniz-Institut für Deutsche Sprache

text-plus.org
office@text-plus.org

NFDI Vision



Daten als gemeinsames Gut für exzellente Forschung, organisiert durch die Wissenschaft in Deutschland.

NFDI Mission

Schritt für Schritt verbessern wir die Nutzungsmöglichkeiten von Daten für Wissenschaft und Gesellschaft. Durch unser Zusammenwirken im NFDI-Verein entsteht eine Dachorganisation für das Forschungsdatenmanagement in allen Wissenschaftszweigen. In Zusammenarbeit mit nationalen und internationalen Partnern schaffen wir die Rahmenbedingungen für rechtskonforme, interoperable und nachhaltige Dateninfrastrukturen, die für Forschende in ihrem Arbeitsalltag gut zugänglich sind. Wir bilden aus, stärken die Kompetenz im Umgang mit Daten und eröffnen neue Berufswege.

Text+ Vision

Die text- und sprachorientierten Geistes- und Sozialwissenschaften nutzen die Möglichkeiten der Digitalisierung in ihrer Forschung umfassend und haben eine gemeinsame Datenkultur.

Text+ Mission

Wir ertüchtigen Text- und Sprachdaten und den Zugang zu solchen Daten – der Durchgriff auf digitale Quellen soll zum Standard werden. Wir stärken die Digital Literacy der Forschenden. Wir decken die Diversität der Forschungsgemeinde ab und bauen auf ihre Partizipation. Wir befördern Interdisziplinarität und Innovation durch Integration von Infrastruktur und Forschung.

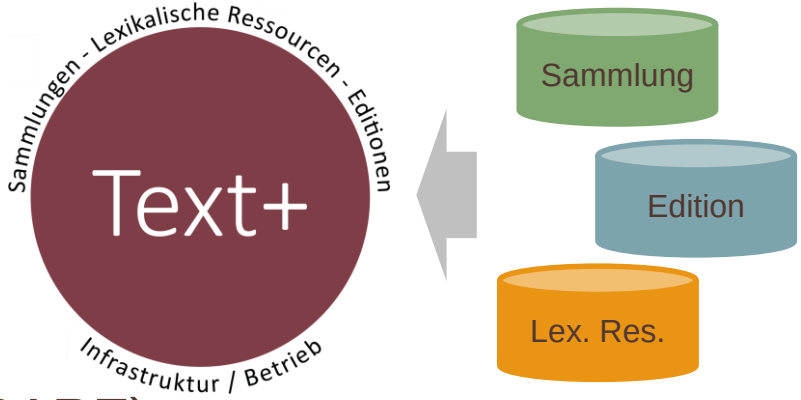
Das Versprechen: Text+ nimmt Daten

» Based on the reference implementation, data centres will expand their portfolio regarding data and services. Clusters are obliged to integrate any data resources and services compliant with their specialisation, the technical and quality criteria set by the Text+ Scientific Board and the respective SCC.

Integration von Daten und Diensten, wenn sie …

- … zur Spezialisierung der Partner passen
- … spezifizierte technischen Anforderungen einhalten
- … spezifizierte Qualitätskriterien erfüllen
- … die Community-basierten Gremien (SCCs/OCC) zustimmen

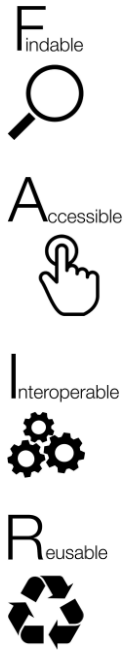
Data Depositing in Text+: Was habe ich davon?



- Nachhaltige Datenhaltung (FAIR/CARE)
- Auffindbarkeit über Kataloge (z. B. VLO, Editionskatalog)
- Nachnutzbarkeit (z. B. LRS, Geo-Browser)

FAIR Prinzipien

- Auffindbarkeit:
  - Daten und Metadaten sollten sowohl von Menschen als auch von Maschinen leicht zu finden sein.
- Zugänglichkeit:
  - Daten und Metadaten sollten verfügbar gemacht und langzeitarchiviert werden, sodass sie leicht von Menschen und Maschinen heruntergeladen und genutzt werden können.
- Interoperabilität:
  - Die Daten sollten derart vorliegen, dass sie mit anderen Datensätzen von Menschen und Maschinen verknüpft werden können.
- Wiederverwendbarkeit:
  - Zur Wiederverwendbarkeit trägt eine Beschreibung der Datensätze über Metadaten bei, sodass sie für weitere Forschungen nachnutzbar und mit anderen Datensätzen vergleichbar sind.



Data Depositing bei und mit Datenzentren

siehe auch: https://www.text-plus.org/forschungsdaten/daten-und-kompetenzzentren/

Cluster in den Sammlungen

- Gegenwartssprache
- Historische Texte
- „Unstrukturierte Texte“

Daten- und Kompetenzzentren

Was sind Daten- und Kompetenzzentren?

- Datenzentren :=
  - Partnereinrichtungen mit Spezialisierung auf bestimmte Arten von Daten, die sie vorhalten
  - Betreiben eine Infrastruktur, die eine langfristige Bereitstellung und Archivierung von Daten ermöglicht
  - Nutzen zur Datenhaltung zertifizierte Repositorien
  - Stellen die Metadaten zu den Daten über Schnittstellen bereit
  - Stellen Schnittstellen zu weiteren Diensten von Text+ zur Verfügung (z. B. zur verteilten Suche)
  - Nehmen Daten gemäß der Spezialisierung auch von Dritten entgegen (mehr dazu später)

- Kompetenzzentren :=
  - Partnereinrichtungen mit speziellen Kenntnissen und Fähigkeiten für bestimmte Arten von Daten
  - Beraten zur Erstellung und Archivierung von bestimmten Daten
  - Können auch an der Entwicklung von Diensten in Text+ beteiligt sein, die auf Schnittstellen aufbauen
  - Benötigen keine Archivinfrastruktur
  - Unterstützen bei der Archivierung an anderen Orten

Welche Zentren gibt es?

- für jede Datendomäne mit unterschiedlichen Schwerpunkten
  - Von „allgemein“ bis „sehr speziell“
  - Born Digital bis Retrodigitalsiert
- Unterschiedliche Sprachen, Epochen, Datenformate

Datenzentren im Bereich „Sammlungen“ von Text+

- IDS
- BBAW
- DNB
- SUB
- Hamburg (Uni und Akademie)
- LMU München
- Uni des Saarlandes
- Uni Duisburg-Essen
- Uni Freiburg (K)
- Uni Köln
- Uni Tübingen
- Uni Würzburg (K)

Datenzentren über alle Datendomänen in Text+

- Sammlungen
  - IDS
  - BBAW
  - DNB
  - SUB
  - Hamburg (Uni und Akademie)
  - LMU München
  - Uni des Saarlandes
  - Uni Duisburg-Essen
  - Uni Freiburg (K)
  - Uni Köln
  - Uni Tübingen
  - Uni Würzburg (K)
- Lexikalische Ressourcen
  - IDS
  - BBAW
  - Sächsische Akademie
  - Uni Köln
  - Uni Trier
  - Uni Tübingen
- Editionen
  - BBAW
  - NRW Akademie
  - SUB
  - Akademie Mainz
  - Darmstadt (TU, HS, UB)
  - HAB
  - Leopoldina
  - Steinheim Institut

Daten- und Kompetenzzentren

- Zertifizierung
- Metadaten/ Metadaten-Harvesting zum Aufbau von Katalogen
- APIs/Interfaces

Zertifizierung der Datenzentren

- Core Trust Seal (CTS)
  - Nachhaltige Infrastruktur und Prozesse
  - International etabliert
  - Bewertung aufgrund von einer Kriterienliste
  - Begutachtung durch ‚Qualified volunteer Reviewers' und Zertifizierung durch das CoreTrustSeal Board
- Nestor e.V.
  - Infrastruktur und Prozesse zur Langzeitarchivierung
  - Basiert auf DIN 31644 „Kriterien für vertrauenswürdige digitale Langzeitarchive"
  - Bewertung aufgrund von auf der DIN 31644 basierenden Kriterienliste
  - Begutachtung Mitglieder der nestor-Arbeitsgruppe ‚nestor AG Zertifizierung '

Metadaten

- Verschiedenste Konventionen und Normen
  - Dublin Core/Dublin Core 15
  - Lightweight Information Describing Objects (LIDO)
  - Marc21
  - ISO 24622-1 und ISO 24622-2 (CMDI)
  - TEI-Header
  - ...
- Unterschiedliche Serialisierungen
  - XML
  - JSON/JSON-LD
  - RDF (N-Tubles, Turtle, XML-RDF, JSON-LD)

- „Öffentliche“ Information über Forschungsdaten
  - Für Kataloge und Nachweissysteme
  - Bereitgestellt über Schnittstellen
  - Durchsuchbarkeit über Suchmaschinen
  - Ermöglicht Zitation/Persistente Identifikation

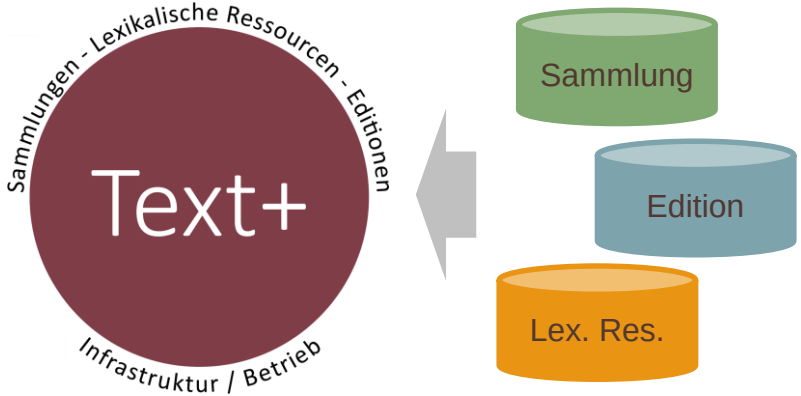
APIs/Interfaces

- Auslieferung von Metadaten:
  - OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting)
    - REST-Interface-Spezifizierung
    - Erlaubt neben Dublin Core auch andere Metadatenschemata
  - SPARQL (SPARQL Protocol And RDF Query Language)
    - Insbesondere für Linked Data Anwendungen
    - Erfordert Linked Data Darstellung der Metadaten im Backend
- Zugriff auf Forschungsdaten in einer ortsverteilten Infrastruktur
  - Föderierte Inhaltssuche (FCS)
  - Basiert auf Webstandards
- Services aus Repositorien: Zugriff auf Erschließungswerkzeuge

Wege ins Text+ Universum

Nutzung eines spezifischen Zentrums

- Daten an ein Datenzentrum geben
- Formate, Anforderungen absprechen
- Ggf. Untersützung bei der Anpassung, Aufbereitung erhalten
- jährliche Ausschreibung zur Förderung von Kooperations-Projekten



Selbst Text+ Zentrum werden

- Die Daten nicht an Text+ „übergeben“
- Daten selbst hosten und verwalten
- Selbst ein Datenzentrum in Text+ betreiben

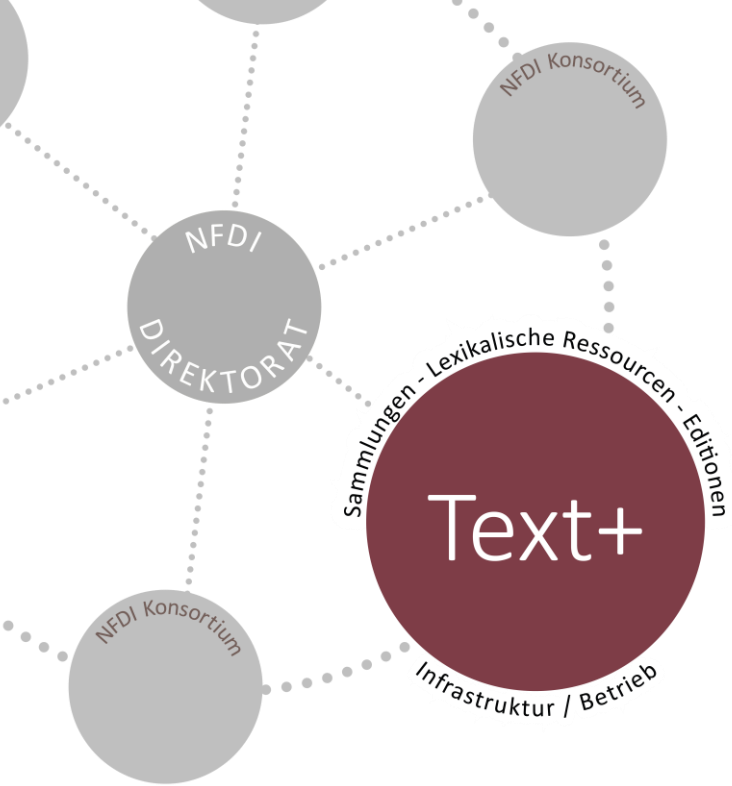
Nutzung der „generischen“ Optionen

- Text+ bietet das DARIAH Repository als "catch-all" Repository für Dateneigentümer, die die clusterspezifischen Kriterien nicht erfüllen

- Text+ bietet ein CTS- und nestor-zertifiziertes Langzeitarchiv, das von der DNB betrieben und von der GWDG entwickelt und gepflegt wird.

Vorstellung der Spezialisierungen der Datenzentren...

» ... Ab kurz nach 10 Uhr



Vielen Dank für Ihre Aufmerksamkeit!

Andreas Witt, Leibniz-Institut für Deutsche Sprache

Mit Material von:
Andreas Henrich, Otto-Friedrich-Universität Bamberg
Christoph Draxler, Bayerische Archiv für Sprachsignale
Alexander Geyken, Berlin-Brandenburgische Akademie der Wissenschaften
Marius Hug, Berlin-Brandenburgische Akademie der Wissenschaften
Christoph Kudella, SUB Göttingen
Peter Leinen, Deutsche Nationalbibliothek
Philipp Wieder, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
office@text-plus.org

Text+ ist ein Konsortium der bundesweiten Initiative zum Aufbau einer nationalen Forschungsdateninfrastruktur (NFDI) und wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 460033370

nfdi Nationale Forschungsdaten Infrastruktur

Gefördert durch DFG Deutsche Forschungsgemeinschaft