Sammlungen in Text+

Text+-Partner an folgenden Einrichtungen: Akademie der Wissenschaften in Hamburg, Albert-Ludwigs-Universität Freiburg, Bayerisches Archiv für Sprachsignale/LMU München, Berlin-Brandenburgische Akademie der Wissenschaften, Deutsche Nationalbibliothek, Eberhard Karls Universität Tübingen, Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, Leibniz-Institut für Deutsche Sprache, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Universität Duisburg-Essen, Universität Hamburg, Universität des Saarlandes, Universität Würzburg, Universität zu Köln

Datendomäne

Sprach- und textbasierte Sammlungen umfassen Sammlungen geschriebener, gesprochener oder gebärdeter Sprache und Texte sowie sprach- und textbezogene Experimental- oder Messdaten, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden:

- **Textsammlungen** (literarische Texte, Sachtexte, Zeitungs- und Zeitschriftentexte, Interviews, Inschriften, Handschriften, Drucke etc.)
- **mono- und multimodale Aufnahmen**
  z. B. von spontaner und formaler Sprache (Reden, Dialoge, Nachrichten, Interviews, Interaktion im Alltag etc.)
- **Sensordaten**
  (EEG, Eyetracking, Artikulographie etc.)
- **Befragungen, Reaktionszeiten**

Arbeitspakete

Referenz-Implementierung

- Konzept und prototypische Umsetzung der Integration von repräsentativen Ressourcen
- Betrieb und Weiterentwicklung von Repositorien der Datenzentren
- Zertifizierung der Repositorien
- Beispielhafte Integration der repräsentativen Bestände der Cluster

Portfolio-Entwicklung

- Kooperationsprojekte
- Integration neuer Daten und Datenzentren
- Bereitstellung von Referenzdaten/Bestandsdaten
- Bereitstellung, Anpassung und Weiterentwicklung von Tools für die Arbeit mit Text+-Daten

Standardisierung

- Erarbeitung von Best Practices für Sammlungen
- Datentypen
- Zusammenarbeit mit Standardisierungsorganisationen

Community Aktivitäten

Für Forschende zu allen Aspekten der Erstellung und Verwendung von Daten, anderen Ressourcen sowie im Bereich des Datenmanagements:

- Informationen zu rechtlichen und ethischen Aspekten
- Disseminationsaktivitäten zur Nachnutzung von Daten
- Weitere Kurse und Trainingsaktivitäten
- Individueller Support (Helpdesk)

Software Services

- Anbindung Federated Content Search (FCS) und Registry
- Linked (Open) Data (LOD)
- Weitere Schnittstellen

Inhaltliche Clusterung

Historische Texte

- Dokumente von Beginn der Aufzeichnungen bis in die erste Hälfte des 20. Jahrhunderts
  - Historische Sprachstufen
  - Historische Daten anderer Sprachen
  - Lateinische/nicht-lateinische Schrift
- Erfahrung mit der Datenaufbereitung
  - Anreicherung der Metadaten
  - Orientierung an etablierten Standards, z. B. DTABf
  - Normalisierung historischer Schreibweisen des Deutschen
  - Konvertierung von OCR-ten Texten zu erschlossenen Texten

Gegenwartsbezogene Sprachdaten

- Datenbereiche aus
  - Gesprochene Sprache
  - Geschriebene Sprache
  - Andere Modalitäten
  - Mono- oder Multilingual
  - Spezialkorpora (z. B. zu Parlamentsdebatten oder bedrohten Sprachen)
  - Experimentaldaten (z. B. aus psycho- oder neurolinguistischen Experimenten)
  - Abgeleitete Datenformate (z. B. N-Gramm-Listen)
- Auch für außereuropäische Sprachen
- Grundlage für maschinelles Lernen, Sprachverarbeitung
- Vorhandenes Portfolio an Werkzeugen zur Tiefenerschließung

Unstrukturierte Texte

- Fokus auf
  - Textuelle Sprachdaten ohne Annotation
  - Semi-strukturierte OCR-Texte aus Digitalisierungsprozessen, z. B. von Zeitungen
  - Bilder mit bibliographischen und strukturellen Metadaten (METS/MODS)
- Beispiele: Dokumente der
  - Deutschen Digitalen Bibliothek (DDB)
  - Arbeitsgemeinschaft Sammlung Deutscher Drucke (AG SDD)
  - SUB Göttingen
- Ergänzung der Metadaten, Standardisierungsbemühungen im Rahmen von DINI-KIM (Kompetenzzentrum Interoperable Metadaten)
- Weiterverarbeitung von Texten zu Plain-Text-Dokumenten (Grundlage Tiefenerschließung)

Beiträge zu übergreifenden Fragestellungen

Task-Area-übergreifend

- Linked Data
- Helpdesk
- Registry und föderierte Inhaltssuche (FCS)
- Blog
- Reference Implementation

Konsortiumsübergreifend (einschlägige NFDI-weite Aktivitäten)

- Metadatenstandards
- Linked Data
- NFDI-Sektionen Metadaten, Training und Education, Basisdienste, ELSA (ethische, rechtliche und soziale Aspekte) und perspektivisch NFDI & Industrie

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