DB Corpora – Digital Humanities Uni S

Claus-Michael Schlesinger | Michael Werner Czechowski

26.1.2017 (v1.101)

Contents

Kurzbeschreibung

- Corpusdaten für die computergestützte Textanalyse
- Trennung von plain text und Metadaten
- Abbildung von verschiedenen Zuständen des gleichen Texts
- analyseorientiert
- präzise Vorgaben

Die Datenbank dient der Speicherung von Corpusdaten. Das Grundelement des Gesamtcorpus ist ein Einzeltext. Jeder Text kann erstens eine Reihe von Zuständen haben und ist zweitens durch eine Reihe von Metadaten markiert. Über die Suchfunktion einer Webanwendung können Subcorpora aus dem Gesamtcorpus zusammengestellt werden. Datenbank und Webanwendung dienen ausschließlich der Vorhaltung und Bereitstellung von Texten, es werden keinerlei Analysefunktionen und keine Schnittstellen für bereichsexterne Anwendungen bereitgestellt. Die Datenbank liegt direkt auf dem Compute-Server oder ist von dort aus erreichbar. Daten lassen sich sowohl über die Webanwendung als auch direkt aus der Datenbank auslesen (SQL-Abfragen nur für User, die auf dem Server eingelogged sind),

sodass SQL-Abfragen in eigene Analyseskripte, die auf dem Computeserver ausgeführt werden, integriert werden können.

Usecase: Speichern und Bereitstellen

- Datenhaltung
- Datenoptimierung

Speichern

1. primäre Datenbank
   - Verwaltung von eigenen Corpora
   - eigene Daten werden nicht selbst gespeichert
2. sekundäre Datenbank
   - Austausch (Repository)
   - eigene Daten werden redundant gespeichert

–> DB Corpora für beides nutzbar, **aber:** Neuentwicklung! Stabilität und Integrität der Daten kann *nicht* garantiert werden!

Anforderungen

Anforderungen Textzustände

Textzustand	Beschreibung
raw	ausgangszustand (vor jeder bearbeitung)
normalisiert	ohne Zusatztexte (z.B. Projekt Gutenberg Paratexte)
normalisiert 2	Normalisierungen auf Zeichenebene (ins Kommentarfeld: welche Bearbeitungsschritte, genaue Zustandsbeschreibung des Texts)
pos-tagged	(ins Kommentarfeld: Tool + Version + Datum der Bearbeitung)
annotiert	TEI

Datenfeld zum Text (alle Zustände): Dokumentation der Bearbeitungsschritte (Freitext)

Anforderungen Metadaten

- Werk
  - Titel (mandatory)
  - Erscheinungsjahr (mandatory)
  - Gattung (Lyrik, Epik, Drama, Märchen)
- Autor

  - Name (mandatory)
  - Geburtsjahr
  - Geschlecht
  - GND-Nummer (Gemeinsame Normdatei)

Notwendige Angaben sind *Name, Titel, Erscheinungsjahr*, alle anderen Angaben sind optional.

Vorschlag für Metadatenformat: MODS, Bibtex, Dublin Core (evtl. für die Ausgabe bei Abfragen)

Anforderungen Anwendung

nth = nice to have

- initialer Upload
  - Texte (in allen Zuständen) als plaintext-Datei (.txt)
  - Metadaten
    - über Formularfelder
    - aus Literaturdatenbank bzw. aus formatierten Daten (bibtex/biblatex) *nth*
- Bearbeitung
  - Alle Felder und Uploads müssen in einer separaten Maske bearbeitbar sein (Klärung der Rechteverteilung steht noch aus; darf jeder Nutzer Änderungen einbringen?)
- Löschen
  - Falls ein Dokument fehlerhaft ist oder nicht mehr benötigt wird, so kann jeder Nutzer eine Lösch-Anfrage stellen. Diese wird vom Admin überprüft und ggf. durchgeführt
- Bulk-Upload *(nice-to-have)* (in Ausnahmefällen skriptbasiert möglich, d.h. mit direktem Zugriff auf Datenbank und Speicher, nur mit Administratorrechten)
- ergänzender Upload für bereits vorhandene Einträge (Zugang über frühe Weiche: Bearbeitungsmodus/Suchmodus), es können nur einzelne Texte bearbeitet werden
  - Eingabemaske für Einzeltext mit allen Feldern
  - Anzeige bereits vorhandener Daten (bei bereits vorhandenen Text(zuständen) nur Link zum Volltext (öffnet im neuen Fenster/Lightbox))
  - Wenn Datenfelder überschrieben werden –> Warnung vor dem Überschreiben mit Angabe der Felder, die überschrieben werden
- Gesamtübersicht
  - Alle bisher hochgeladenen Texte anzeigen lassen
  - Soriermöglichkeiten z.B. nach Datum des Uploads etc. *nth*
- Suchen
  - nur Metadaten
  - nur Volltexte *nth*
  - Metadaten und Volltexte *nth*
- Bildschirmausgabe: Ergebnisliste (Suchfeld bleibt sichtbar)
- Download:
  - gesamte Ergebnisliste

- Auswahl aus Ergebnisliste (Clickboxen?)
- Ausgebeformat: zip-Datei mit einzelnen Textfiles und Liste mit bibliografischen Angaben (Bibtex/MODS)

Die Datenbank bietet neben der Webanwendung auch einen direkten low-level-Zugang (SQL-Abfragen, nur Lesen!) für NutzerInnen, die am Server angemeldet sind.

User Stories

Ansicht

ID	Beschreibung
US100	Nutzerin möchte alle hinterlegten Uploads in einer Übersicht ansehen können
US101	Nutzerin möchte einen Corpus im Detail anschauen und verschiedene Zustände vergleichen
US102	Nutzerin möchte zurück zur Übersicht navigieren

Eingabe

ID	Beschreibung
US201	Nutzerin möchte neuen Corpus mit Metadaten hochladen
US202	Nutzerin möchte neuen Autor mit weiteren Informationen anlegen
US203	Nutzerin möchte neue Gattung hinzufügen
US204	Nutzerin möchte weiteren Textzustand zu bestehendem Eintrag hinzufügen

Bearbeitung

Hinweis: *Bereits hochgeladene Texte vom Typ 'raw' können nicht nachträglich verändert werden. Ebenso können Autorennamen und Gattungsbezeichnungen nur angelegt, nicht geändert werden. Die Metadaten für einen einzelnen Eintrag können geändert bzw. neu ausgewählt werden. Änderungen von Autorennamen, Gattungsbezeichnungen und Ersteinträgen sind nur mit Administratorrechten möglich. Einzelne Textzustände können überschrieben werden.*

ID	Beschreibung
US301	Nutzerin möchte Metadaten eines bestehenden Corpuseintrags ändern
US302	Nutzerin möchte Informationen zu bestehendem Autor ändern
US303	Nutzerin möchte bestehende Gattung ändern
US304	Nutzerin möchte einen bestimmten Textzustand verändern

Löschen

Hinweis: *Bei allen Lösch-Anträgen muss die Nutzerin eine Kontaktadressen angeben, sodass sie informiert werden kann, wie weit der Zustand der Bearbeitung ist.*

ID	Beschreibung
US401	Nutzerin möchte Lösch-Auftrag für kompletten Eintrag beantragen
US402	Nutzerin möchte Lösch-Auftrag für bestimmten Autor beantragen
US403	Nutzerin möchte Lösch-Auftrag für bestimmte Gattung beantragen