Workshop 'Datenbanken' – Ergebnisse DB Corpora

cmx

26.1.2017

Datenbank Corpora

- 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.

Usecases: Speichern und Bereitstellen

Nutzerin A:

Speichern

- 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: Textzustände

- 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

- Titel
    - 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, Dublin Core (evtl. für die Ausgabe bei Abfragen)

Anforderungen Anwendung

- initialer Upload
    - Texte (in allen Zuständen) als plaintext-Datei (.txt)
    - Metadaten
        * über Formularfelder
        * aus Literaturdatenbank bzw. aus formatierten Daten (bibtex/biblatex) *nth*
- Bulk-Upload *(nice-to-have)*
- 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
- 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 auf dem Server eingelogged sind.