Praxislabor Digitale Geisteswissenschaften

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

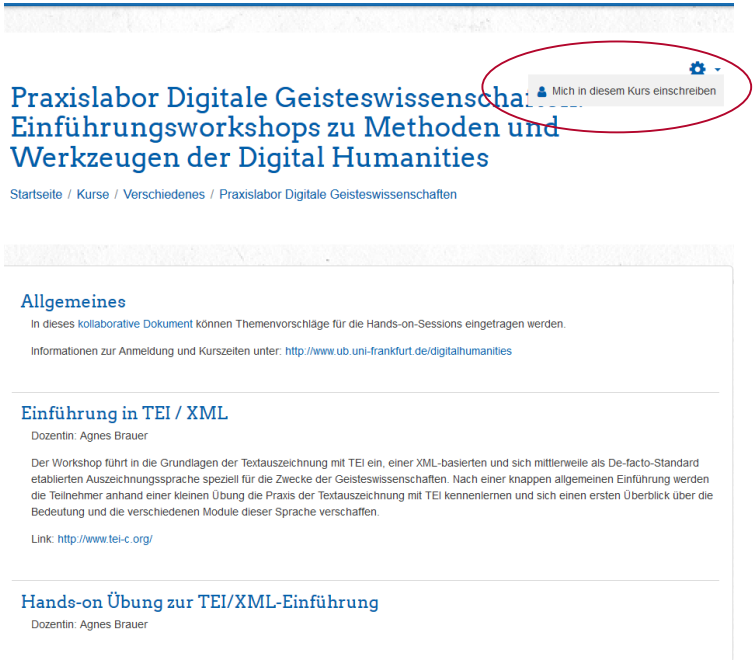
Hands-on-Übung zur Datenbereinigung mit

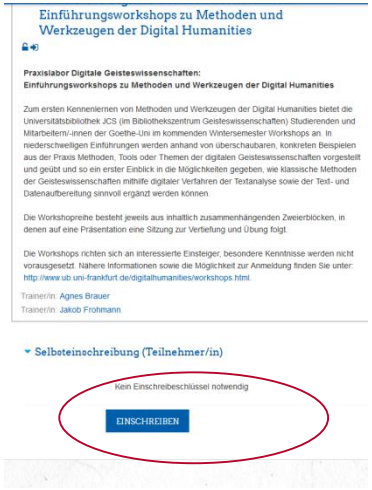
OpenRefine

Agnes Brauer
a.brauer@ub.uni-frankfurt.de

Vorbereitung | Hinweise

- Melden Sie sich für den moodle-Kurs Praxislabor Digitale Geisteswissenschaften an und schreiben Sie sich ein:





Vorbereitung und Hinweise für die Hands-on-Übung

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- Laden Sie bitte OpenRefine (3.1) herunter, entpacken und installieren Sie die Software auf Ihrem Computer. Sie benötigen den Browser Firefox, in dem OpenRefine läuft:
  http://openrefine.org/download.html

- OpenRefine ist eine Java-Anwendung → es wird eine Java-Laufzeitumgebung benötigt (bitte installieren, falls auf Ihrem Gerät noch nicht vorhanden: http://java.com/ bzw. Open Source Variante:OpenJDK https://github.com/ojdkbuild/ojdkbuild

- Starten Sie die Anwendung aus dem entpackten Verzeichnis, es öffnen sich eine Kommandozeile und kurz danach der Browser mit dem geladenen Programm – sollte der Browser nicht starten, benutzen Sie bitte den Link: http://127.0.0.1:3333/.

- Hilfe / Infos zum Setup: https://librarycarpentry.org/lc-open-refine/setup.html

Heute …

- Zellen teilen und (wieder) vereinigen mit Hilfe von Separatoren
- Facetieren, Filtern + Clustern von Daten
- Anreichern eigener Daten aus externen Quellen (Beispiel: Wikidata)

Einfaches Beispiel zum „Aufräumen“ von Daten

Beispiel: Schlagworte zum Thema „Bestandserhaltung“

- Erstellen Sie ein OpenRefine-Projekt mit Daten der Übungsdatei „BE.txt“
- Legen Sie eine einspaltige Tabelle an, in der in jeder Zeile ein Schlagwort steht („Edit Cells“ → „Split multi-valued cells…“ → als Separator ein Leerzeichen wählen)

- Verschaffen Sie sich einen ersten Überblick über die Daten mit Hilfe der Funktion „Text facet“
- Welche Schlagworte kommen im Datensatz mehrmals vor?

- Probieren Sie auf der Grundlage der Facette die Funktion „Cluster“ aus und beseitigen Sie Tippfehler
- Vergleichen Sie die Ergebnisse beim Einsatz unterschiedlicher Clustering-Methoden und -Algorithmen

Übung 2: „Aufräumen“

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Dokument „Muenzen.txt “ Vgl. https://ikmk.smb.museum/object?lang=de&id=18206726

- Explorieren Sie die Daten und betreiben Sie etwas Datenbereinigung mithilfe der Funktionen „Text-Facet“ und „Cluster“

- Transponieren Sie die Daten in mehrere Spalten: Transpose → Transpose cells in rows into columns…
- Wie viele Spalten sind sinnvoll?

- Überlegen Sie sich weitere sinnvolle „Aufräumarbeiten“ und probieren Sie sie einfach aus

Anwendungsmöglichkeiten – fortgeschrittene Funktionen

Reconcile & Match

- Vergleichen / Angleichen der eigenen Daten anhand von Datenbanken (z.B. Wikidata)
- Anreicherung von Daten (z.B. mit Identifiern oder geographischen Koordinaten)
- Verlinkung von Daten

Reconciliation is the process of matching name strings to identifiers of entities in a database like an authority file, Wikidata etc. This is useful whenever you want to merge differing name strings for the same person in your data or when you want to fetch additional data from the target database you are reconciling against.

Reconcile & Match

- Legen Sie ein neues Projekt anhand der Datei „Präsidenten.txt“ an
- Bereinigen Sie die Daten
- Splitten Sie die Daten mithilfe eines geeigneten Separators
- Führen Sie eine Reconciliation mit Wikidata durch
- Fügen Sie neue Spalten auf Basis der Reconciliation hinzu, z.B.:
  - Beruf
  - Geburtsort
  - Koordinaten des Geburtsortes
  - Geburtsdatum

Reconcile & Match

- Bereiten Sie die Datei „Praesidenten2.csv“ für den Dariah-Geobrowser auf: https://geobrowser.de.dariah.eu/index.html

- Aufbau der Tabelle:

	A	B	C	D	E	F	G	H	I
1	Name	Address	Description	Longitude	Latitude	TimeStamp	TimeSpan:begin	TimeSpan:end	GettyID
2									
3									
4									
5									
6									
7									

Tipps & Tricks / Links / Literatur

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Blogs

https://histhub.ch/cat/net/blog/openrefine/ (Blog-Serie zur Arbeit an historischen Daten mit OpenRefine)

http://blog.lobid.org/2018/08/27/openrefine.html (einfache Anreicherung von Daten in OpenRefine mit [Personen-] Daten aus der GND via lobid.org)

Literatur

Ruben Verborgh/Max de *Wilde*, Using OpenRefine. The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web (Community experience distilled), Birmingham, Mumbai 2013. [Online-Ressource über UB FFM]

UB
UNIVERSITÄTS
BIBLIOTHEK
FRANKFURT AM MAIN

Danke für Ihre Aufmerksamkeit!

Vielen Dank für die zur Verfügungstellung von Materialien und Daten an Jakob Frohmann (Universitätsbibliothek Johann Christian Senckenberg) und und Jun.-Prof. Dr. Torsten Hiltmann (Zentrum für Digitale Geschichtswissenschaft, Universität Münster).

Workshop konzipiert in Anlehnung an "Library Carpentry: OpenRefine Lessons for Librarians." (2016)