Praxislabor Digitale Geisteswissenschaften

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Hands-on-Übung zur Datenbereinigung mit

OpenRefine

14.07.2023

Jakob Frohmann
jakob.frohmann@ub.uni-frankfurt.de

OpenRefine – Was ist das?

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- interaktives Werkzeug zum Bearbeiten, Erkunden und Bereinigen großer Mengen von Daten in Tabellenform („A power tool for working with messy data.“)
- hat große Ähnlichkeiten zu einem Tabellenkalkulationsprogramm mit Zeilen und Spalten (z.B. Excel)
  (ein „OpenRefine-Projekt“ = eine Tabelle)
- läuft lokal auf dem Rechner, aber im Browser (keine Internetverbindung notwendig)
- Open Source-Software in Java
  (zuvor zwischenzeitlich zu Google gehörig unter dem Namen Google Refine)

Vorbereitung Hands-on-Übung

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- Laden Sie bitte die aktuelle Version von OpenRefine herunter, entpacken und installieren Sie die Software auf Ihrem Computer. Sie benötigen außerdem einen Browser, in dem OpenRefine läuft: http://openrefine.org/download.html

- OpenRefine ist eine Java-Anwendung → es wird eine Java-Laufzeitumgebung benötigt; wählen Sie bitte die Variante „Windows kit with embedded Java", sollte Java auf Ihrem PC nicht vorhanden sein

- Starten Sie die Anwendung aus dem entpackten Verzeichnis, es öffnet sich eine Shell 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/index.html#getting-ready

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 „Originalerhalt historischer Bibliotheksbestände“

- Erstellen Sie ein OpenRefine-Projekt mit Daten der Übungsdatei „originalerhalt.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 „praesidenten.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

Tipps & Tricks / Links / Literatur

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Blogs

https://histhub.ch/erste-schritte-mit-openrefine-ein-erstes-projekt/ (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]

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Danke für Ihre Aufmerksamkeit!

Workshop konzipiert in Anlehnung an "Library Carpentry: OpenRefine Lessons for Librarians.“

Vielen Dank für die Zurverfügungstellung von Materialien und Daten an Prof. Dr. Torsten Hiltmann (Humboldt-Universität zu Berlin).