Praxislabor Digitale Geisteswissenschaften

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Datenbereinigung mit OpenRefine: Hands-on Übung

OpenRefine

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

Vorbereitung | Hinweise

- Laden Sie bitte OpenRefine herunter (93MB) und installieren / entpacken Sie die Software auf Ihrem Computer. Sie benötigen den Browser Firefox, in dem OpenRefine läuft (Java-Programm).

http://openrefine.org/download.html
https://github.com/OpenRefine/OpenRefine/releases

- 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/.

- Tragen Sie bitte bei Bedarf Themenvorschläge für die Hands-on Übung in das kollaborative Dokument ein bzw. beachten sie die dortigen Links.

Heute …

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

Einfaches Beispiel zum „Aufräumen“ von Daten

Bitte beachten Sie auch die Links im kollaborative Dokument.

Beispiel: Schlagworte zum Thema „Bestandserhaltung“

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Google Docs-Dokument und 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.

Übung 1: Facettieren und Filtern

Bitte beachten Sie auch die Links im kollaborative Dokument .

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Google Docs-Dokument bzw. mit den vorliegenden bibliografischen Daten: https://raw.githubusercontent.com/LibraryCarpentry/lc-open-refine/gh-pages/data/doaj-article-sample.csv

- Welche Lizenzen werden für die Artikel in diesem Datensatz genutzt? Was ist die häufigste Lizenz in diesem Datensatz? Wie viele Artikel in diesem Datensatz haben keine Lizenz?

- “Facet by blank”-Funktion: Welche Publikationen in diesem Datensatz haben keine DOI in der entsprechenden Spalte eingetragen?

Übung 2: Clustern und „Aufräumen“

Bitte beachten Sie auch die Links im kollaborative Dokument.

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Google Docs-Dokument bzw. mit den vorliegenden Daten zu mittelalterlichen Münzen (gezipte txt-Datei):

https://drive.google.com/file/d/1IOTzVFK9cRJjx1qB-ABc4Cy9OA3UcgM_/view?usp=sharing

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

- Zerlegen sie die Daten in der Spalte in mehrere Spalten, zum Beispiel indem sie einen Doppelpunkt als Separator benutzen.

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

Hinweis: Transformation von Zellen (-inhalten)

Abgesehen von voreingestellten, einfacheren Transformationen („Common transforms“ / „to titlecase“, „to uppercase“, „to lowercase“), können unter „Transform…“ auch umfangreichere Veränderungen an den Daten in einer Zelle vorgenommen werden mit Hilfe von *General Refine Expression Language* (GREL)

Mehr Informationen hier:
https://librarycarpentry.org/lc-open-refine/07-introduction-to-transformations/index.html

Anwendungsmöglichkeiten – fortgeschrittene Funktionen

Reconcile & Match

- Vergleichen / Angleichen der eigenen Daten anhand von Datenbanken (z.B. Wikidata)
- Anreicherung von Daten (z.B. mit eindeutigen Identifikatoren)
- 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.

Anwendungsmöglichkeiten

Reconcile & Match

Gemeinsame Normdatei (GND) der Deutschen Nationalbibliothek via https://lobid.org/gnd

Beispieldaten:

name;beruf;ort
J. Weizenbaum;Informatiker;Berlin
Twain, Mark;Schriftsteller;
Kumar, Lalit;;
Jemand;;

Quelle des Beispiels: http://blog.lobid.org/2018/08/27/openrefine.html

Übung 3: Präsidenten der USA mit GND-Daten anreichern

Bitte beachten Sie auch die Links im kollaborative Dokument .

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Google Docs-Dokument.
- Fügen Sie eine eigene Spalte hinzu, welche nur die Namen der Präsidenten enthält.
- Gleichen Sie die Namen mit der GND ab und wähle jeweils eine Person aus (Link zum Webservice: *https://lobid.org/gnd/reconcile*).
- Ergänzen Sie eine Spalte mit den GND-Nummern.
- Ergänzen Sie eine Spalte mit Links zu den GND-Einträgen

(Linkstruktur: *http://d-nb.info/gnd/* […]).

Tipps & Tricks / Links / Literatur

OpenRefine
http://openrefine.org/

Ressourcen
https://www.wikidata.org/ (Daten aller Art, default in OpenRefine)
https://www.geonames.org/ (Geografika)
http://swb.bsz-bw.de/DB=2.104/ (Personen, Körperschaften, Konferenzen, Geografika, Sachschlagwörter, Werktitel) [GND - Gemeinsame Normdatei der Deutschen Nationalbibliothek, mehr Infos hier)

Visualisierung:
https://geobrowser.de.dariah.eu/
http://hdlab.stanford.edu/palladio/

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]

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Danke für Ihre Aufmerksamkeit!

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

Workshop konzipiert in Anlehnung an "Library Carpentry: OpenRefine Lessons for Librarians." (2016), https://librarycarpentry.org/lc-open-refine/