

Protokoll zur Datenbereinigung in Open Refine

Sandra Ziegos

Datenquelle: rhizom.db (SQLite-Datenbank zum Projekt) – daraus:

Tabelle: involved_artists (Datei involved_artists_original.csv – exportiert aus dem DB Browser)

Ziel: Export der in Open Refine bereinigten Daten (Datei involved_artists_cleaned.csv)

Einzelschritte

Nr.	Schritt	Vorgehensweise & Ergebnis
1	Daten importieren	Dateien auswählen > involved_artists_original.csv Importoptionen prüfen: Separator=Komma, Kodierung=UTF-8, leere Zeilen speichern deaktiviert, leere Zellen als null speichern aktiviert Neuer Projektname: involved_artists_cleaned
2	Leere Zellen vereinheitlichen	Transformation: if(isBlank(value), null, value) Text transform on 744 cells in column art_url: grel:if(isBlank(value), null, value) > Alle inhaltlich leeren oder nur aus Leerzeichen bestehenden Zellen wurden in echte Nullwerte (null) umgewandelt.
3	Überflüssige Leerzeichen entfernen	Transformation (global auf alle Textspalten angewendet): Alle > Umwandeln (GREL): value.trim() Text transform on 0 cells in column art_nationality: grel:value.trim() > Keine Änderungen nötig, da keine überflüssigen Leerzeichen vorhanden.
4	Groß-/Kleinschreibung vereinheitlichen	art_nationality > Zellen bearbeiten > Gemeinsame Umwandlungen > In titlecase Text transform on 9 cells in column art_nationality: value.toTitlecase() > Ergebnisse überprüft und Schritt wieder rückgängig gemacht, da z.B. Usa oder Bih als Ergebnis
5	Begriffe gezielt bereinigen	art_area > America (South), America (North) und America (Central) gecheckt (ob nirgendwo South America, North America oder Central America steht): art_area > Facette > Textfacette > alles korrekt
6	Duplikate & Varianten bereinigen	art_name > Facette > Textfacette > Cluster & Bearbeiten > alle Methoden und Keying Funktionen durchprobiert > einen doppelten Eintrag gefunden: Clemens Pliem/Klemens Pliem > zu korrektem Klemens Pliem vereinheitlicht (neuer Zellenwert: Klemens Pliem > Auswahl zusammenführen & neu gruppieren) > Entsprechende Anpassung auch unter art_name_first > Eintrag mit leerer art_area-Zeile entfernt > Ergebnis: nur noch 1 Eintrag "Klemens Pliem" Zwei weitere potenzielle Dubletten gefunden (Nachfragen erforderlich).
7	Jahresangaben in numerische Werte konvertieren	art_year_birth > Zellen bearbeiten > Gemeinsame Umwandlungen > In Nummer Text transform on 130 cells in column art_year_birth: value.toNumber() art_year_death > Zellen bearbeiten > Gemeinsame Umwandlungen > In Nummer Text transform on 20 cells in column art_year_death: value.toNumber() (Hätte man auch gleich beim Importieren machen können.)
8	Datenanreicherung und Linked-Data-Abgleich mit Wikidata	art_name > Abgleichen > Starten Sie den Abgleich... > Dienst: https://wikidata.reconci.link/en/api > Entitätstyp: human (Q5) > errors: 20, matched: 227, none: 496 art_name > Abgleichen > Aktionen > Jede Zelle dem besten Kandidaten zuordnen Match each of 227 cells to its best candidate in column art_name > neue Spalte hinzugefügt: Create new column art_name_uri based on column art_name by filling 227 rows with grel:"https://www.wikidata.org/entity/" + cell.recon.match.id
9	Daten exportieren	Export des bereinigten Datensatzes als involved_artists_cleaned.csv (Export > Kommagetrennter Wert)