

Protokoll zur Datenbereinigung in Open Refine

Sandra Ziagos

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 > <i>involved_artists_original.csv</i> Importoptionen prüfen: Separator=Komma, Kodierung=UTF-8, leere Zeilen speichern deaktiviert, leere Zellen als null speichern aktiviert Neuer Projektname: <i>involved_artists_cleaned</i>
2	Leere Zellen vereinheitlichen	Transformation: <code>if(isBlank(value), null, value)</code> Text transform on 744 cells in column art_url: <code>grel:if(isBlank(value), null, value)</code> > 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): <code>value.trim()</code> Text transform on 0 cells in column art_nationality: <code>grel:value.trim()</code> > Keine Änderungen nötig, da keine überflüssigen Leerzeichen vorhanden.
4	Groß-/Kleinschreibung vereinheitlichen	<i>art_nationality</i> > Zellen bearbeiten > Gemeinsame Umwandlungen > In titlecase Text transform on 9 cells in column art_nationality: <code>value.toTitlecase()</code> > Ergebnisse überprüft und Schritt wieder rückgängig gemacht, da z.B. <i>Usa</i> oder <i>Bih</i> als Ergebnis
5	Begriffe gezielt bereinigen	<i>art_area</i> > America (South), America (North) und America (Central) gecheckt (ob nirgendwo South America, North America oder Central America steht): <i>art_area</i> > Facette > Textfacette > alles korrekt
6	Duplikate & Varianten bereinigen	<i>art_name</i> > 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 <i>art_name_first</i> > Eintrag mit leerer <i>art_area</i> -Zeile entfernt > Ergebnis: nur noch 1 Eintrag "Klemens Pliem" Zwei weitere potenzielle Dubletten gefunden (Nachfragen erforderlich).
7	Jahresangaben in numerische Werte konvertieren	<i>art_year_birth</i> > Zellen bearbeiten > Gemeinsame Umwandlungen > In Nummer Text transform on 130 cells in column art_year_birth: <code>value.toNumber()</code> <i>art_year_death</i> > Zellen bearbeiten > Gemeinsame Umwandlungen > In Nummer Text transform on 20 cells in column art_year_death: <code>value.toNumber()</code> (Hätte man auch gleich beim Importieren machen können.)
8	Datenanreicherung und Linked-Data-Abgleich mit Wikidata	<i>art_name</i> > Abgleichen > Starten Sie den Abgleich... > Dienst: https://wikidata.reconci.link/en/api > Entitätstyp: human (Q5) > errors: 20, matched: 227, none: 496 <i>art_name</i> > 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 <code>grel:"https://www.wikidata.org/entity/" + cell.recon.match.id</code>
9	Daten exportieren	Export des bereinigten Datensatzes als <i>involved_artists_cleaned.csv</i> (Export > Kommagetrennter Wert)