

## digis-Workshop 11.04.2018

### Metadaten – Datenbereinigung und -anreicherung mit OpenRefine - Leitfaden -

Referenten: Anna-Lena Nowicki, Heinz-Günther Kuper

#### Projekt anlegen

- Beispieldaten (CSV u.a.) hochladen
- Encoding prüfen/verändern
- Spaltentrenner prüfen/verändern

#### Oberfläche kennenlernen

- OpenRefine-Logo Button (weiteres Projekt anlegen, Projekt erneut öffnen, ...)
- "Open" um parallel in mehreren Projekten zu arbeiten
- "Help" Weiterleitung zu ausführlicher Dokumentation in GitHub !!!!
- "Export" um Daten wieder aus dem System zu bekommen
- "Export" => "Templating" für fortgeschrittene Anwendung, Vertiefung ggf. bei Wikidata-Teil
- Reiter-Wechsel (Facets / History)
- DropDown  
ALL => auf alle Spalten gleichzeitig zugreifen  
=> Spalten entfernen/verschieben: Edit Columns  
Einzelspalten => nur auf ausgewählte Spalte zugreifen

#### DATEN prüfen

- Übersicht zu Spalteninhalt bekommen: Facet => Text Facet  
Bsp. Spalten "Datenbankname", "Objektart", "Andere Objektbezeichnung:"
- Auf Duplikate in einer Spalte prüfen:  
Bsp. Spalte "Objektnummer" => Facet => Customized Facet => Duplicates Facet
- Numerische Werte visualisieren:  
Spalte "Objektnummer" => Facet => Numeric facet
- Datierungen visualisieren (Zeitleiste):  
Spalte "Tagesdatum\_normalisiert" => Facet => Timeline facet
- Prüfe URL auf Erreichbarkeit (wähle Spracheinstellung 'Clojure'):

```
(  
  let [  
    connection (.openConnection (java.net.URL. value))  
    response (.getResponseCode connection)  
  ]  
  (.. connection getInputStream close)  
  response  
)  
=> 200 = OK
```

#### DATEN angleichen

- Führende und abschließende Leerzeichen entfernen  
Spalte "Art der Verpackung:" => Facet => Text Facet  
=> Edit Cells => Common Transforms => Trim leading and trailing whitespaces

- Manuell einzelne Zeichen(-ketten) bearbeiten  
Spalte "Schlagwort:" => Facet => Text Facet  
=> Edit: Punkt entfernen  
Spalte "Farbe(n):" => Facet => Text Facet  
=> Edit: farbig vs. bunt
- Restrukturieren von Feldinhalten  
Spalte "Farbe(n):" => Facet => Text Facet  
=> Edit Cells => Transform => value.split('/').sort().join('/')
- Vergleiche Inhalte zweier Spalten:  
'Marke^Markenname: ' / 'Markenname Vorlageform'  
=> Edit Column => Add Column based on this column  
=> if (value == cells['Marke:^Markenname:'].value, true, false)
- Clustering zur Vereinheitlichung nutzen  
Spalte 'Körperschaftsname\_Vorlageform:'  
=> Text Facet => Cluster  
=> verschiedene Methoden ausprobieren um zum besten Ergebnis zu kommen
- Arbeiten mit regulären Ausdrücken  
Bsp. value.replace(/s+/, " ")

#### **DATEN spaltenübergreifend zusammenstellen**

- Neue Spalte aus Inhalten mehrerer Felder erzeugen:  
Bsp. Maßangaben: => Edit Column => Add column based on this column...  
=> 'Breite: ' + cells['Breite (in cm):'].value + ' cm / Höhe: ' + value + ' cm'
- Maßeinheiten wechseln: Bsp. cm to mm  
=> Edit cells => Transform => value.toNumber() \* 10  
=> Spaltentitel anpassen => Edit Column => Rename this column

#### **DATEN gruppiert ergänzen**

- Text zeilenübergreifend verändern  
Spalte 'Column': => Facet => Text Facet => Edit CC BY => CC BY 4.0
- Neue Spalte basierend auf vorhandener anlegen:  
Spalte 'Column': => Edit Column => Add column based on this column...  
CC BY 4.0 mit URI ergänzen => <https://creativecommons.org/licenses/by/4.0/>
- Spalte umbenennen: Spalte 'Column' => Edit Column => Rename this column
- Spalte 'Technik(en) ' Identifier ergänzen: Druck => AAT URI:  
=> Edit Column => Add column based on this column...  
<http://vocab.getty.edu/aat/300185327>

#### **Fließtext analysieren / bearbeiten**

- Facet => Customized Facet => Word Facet  
über Wortliste / Sätze iterieren: forEach(value.split(" "), v, v)

#### **DATIERUNG bearbeiten**

Zeitspanne (Bsp. 1803 - 1889) aufteilen  
Spalte "Packungs-Datum": => Edit column => split into several columns =>  
Trennzeichen: '-'

Bsp. 01.12.1904 in ISO-8601-konforme Schreibweise umwandeln: => Edit cells  
=> Transform => value.split('.').reverse().join('-')

Datentyp verändern (von string zu date): => Edit cells => Transform => value.toDate()

Datierung um 364 Tage nach hinten verschieben: value.inc(364,"days")

Datierung um 2 Monate vorverlegen: value.inc(-2,"month")

Formatierte Zeichenkette aus Datierung extrahieren:

value.datePart("year") + '-' + value.datePart("month") + '-' + value.datePart("day")

### DATEN erweitern:

- Geoinformationen mit Google Maps abgleichen:  
Spalte 'Ort\*:'  
=> Edit Column => Add column by fetching URLs:  
"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=" +  
escape(value,'url')

Aus JSON-Ergebnis formatierte Adresse extrahieren:

=> Edit Column => Add column based on this column...

with(value.parseJson().results[0], address, address.formatted\_address)

Aus JSON-Ergebnis Koordinaten zur Ortsangabe extrahieren:

=> Edit Column => Add column based on this column...

with(value.parseJson().results[0].geometry.location, pair, pair.lat + ", " + pair.lng)

### RECONCILIATION GND / Wikidata

- URL GND Standard Service:  
<http://refine.codefork.com/reconcile/viafproxy/DNB>

Personennamen müssen vor Reconciliation in einheitliche Form gebracht werden:

<Nachname>, <Vorname>

Spalte 'Urheber^Personenname:':

=> Edit column => Add column based on this column ...

=> value.split("[eigentlich ")[1].replace(']', ").split(' ').reverse().join(', ')

=> Edit Cells => Transform => value.split(" ")[0]

=> Reconcile => Start reconciling ... => "DNB (by way of VIAF)" => "Person"

Zugriff auf Ergebnisliste:

=> Edit column => Add column based on this column ...

=> cell.recon.candidates[index].id

=> Edit column => Add column based on this column ...

cell.recon.candidates[index].name

Manuelle Auswahl treffen und mit gematchten Ergebnissen arbeiten:

=> Edit column => Add column based on this column ...

=> cell.recon.match.id => GND ID

=> Edit column => Add column based on this column ...

cell.recon.match.name => Name und ggf. Lebensdaten der Person

Lebensdaten aus Personenangabe extrahieren:

=> Edit column => Add column based on this column ...

=> cell.recon.match.name.split(', ')[2]

- URL **Wikidata DE** Standard Service:

<https://tools.wmflabs.org/openrefine-wikidata/de/api>

Spalte Marke Markenname: reconcile Wikidata

Zugriff auf Ergebnisse wie bei GND Beispielen