

# Datenbereinigung mit OpenRefine: Hands-on Übungen



# OpenRefine

Jakob Frohmann

[j.frohmann@ub.uni-frankfurt.de](mailto:j.frohmann@ub.uni-frankfurt.de)

## Heute ...

---

Wiederholungen und Übungen zu den Themen der letzten Woche

- Grundsätzliches: Ein Projekt erstellen, bearbeiten, wieder exportieren.
- Einfache Operationen: Tabellen umstrukturieren, Zellen teilen und (wieder) vereinigen mit Hilfe von Separatoren
- Facetieren, Filtern, Clustern von Daten:  
Überblick über die Daten gewinnen, Fehler aufspüren, Datensets zur weiteren Bearbeitung isolieren

Außerdem: Anreichern eigener Daten aus externen Quellen (Beispiel: GND)

## Vorbereitung | Hinweise

---

- Laden Sie bitte **OpenRefine Version 3.4.1** herunter (93MB) und installieren / entpacken Sie die Software auf Ihrem Computer. Sie benötigen einen Browser sowie eine Java-Umgebung zum Ausführen des Programms. Bitte installieren Sie Java separat oder nutzen Sie den Download mit eingebettetem Java (nur für Windows).

<http://openrefine.org/download.html>

<https://github.com/OpenRefine/OpenRefine/releases>

<https://openjdk.java.net/>

- 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 / Interesse Themenvorschläge für die Hands-on Übung in das [kollaborative Dokument](#) ein.

## Link zu Übungsdaten

---

<https://docs.google.com/document/d/1tAb4vkFP-rMCXITdMYT3XDRbHzmIJ1HdvFby-Semp8/edit?usp=sharing>

## Übung 1: Einfaches Beispiel zum „Aufräumen“ von Daten

---

Bitte beachten Sie auch die Links im [kollaborative Dokument](#).

Beispiel: Schlagworte zum Thema „Kulturgut“

- Erstellen Sie das OpenRefine-Projekt mit Daten aus dem verlinkten 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. Beseitigen Sie Tippfehler und gleichen Sie Schlagwörter an, wo es Ihnen sinnvoll erscheint.

## Übung 2: Facettieren und Filtern

---

Bitte beachten Sie auch die Links im [kollaborative Dokument](#) .

- Erstellen Sie das OpenRefine-Projekt 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?
- Teilen Sie die Spalte „Authors“ so auf, dass jede\*n Autor\*in eines Titels in einer eigenen Spalte steht.
- “Facet by blank”-Funktion: Welche Publikationen in diesem Datensatz haben keine DOI in der entsprechenden Spalte eingetragen?



Münzkabinett  
Staatliche Museen zu Berlin

PATENSCHAFT NACHRICHTEN KONTAKT ÜBER UNS DE | EN

Interaktiver Katalog des Münzkabinetts

START SUCHE KARTE

Erkunden Sie eine der  
größten numismatischen  
Sammlungen der Welt

z. B. Gold Antike Architektur



[ERWEITERTE SUCHE](#)



Schlagwortsuche



Münzstätten und  
Fundorte



Nachrichten

IKMK überschreitet die 33.000



© Münzkabinett

Münzkabinett  
Staatliche Museen zu Berlin



## Aachen: Stadt

1374

Ausstellung im Bode-Museum  
Raum 242, BM-054/24 Deutsches Reich (Norden). Silber  
13.-15. Jh.

**Vorderseite** KAROLVS MAG-NVS INPERAT. [Karolus Magnus Imperator]. Brustbild Kaiser Karls des Großen mit Krone, Zepter und Reichsapfel über Adlerschild.

**Rückseite** XC VINCIT - XC REGNA - AN DNI M- CCCLXXIII / MON-ETA - IVNC-HEIT. [Christus Vincit Christus Regnat Anno Domini MCCCLXXIII Moneta Jungheit]. Langes Kreuz umgeben von doppeltem Schriftkreis.

**Dargestellte/r** Karl der Große (768-814), König der Franken, seit 800 Kaiser



**Münzstand** Stadt

**Datierung** 1374  
Spätmittelalter

**Nominal** Groschen

**Material** Silber

**Herstellung** geprägt

**Gewicht** 2,27 g



1104 Suchergebnisse

100



**Aachen: Stadt**  
Jungheit (Aachen) (Deutschland,  
Rheinland)  
Groschen, 1374  
18206308



**Aachen: Stadt**  
Aachen (Deutschland, Rheinland)  
1/4 Groschen (Piéfort), 1498  
18243445



**Aachen: Stadt**  
Aachen (Deutschland, Rheinland)  
1/4 Groschen, 1496  
18243447



**Alpen: Herrschaft**  
(Deutschland, Rheinland)  
Pfennig, 1344-1351  
18244632



**Alpen: Herrschaft**  
(Deutschland, Rheinland)  
Heller, 1386-1401  
18244633



**Alpen: Herrschaft**  
(Deutschland, Rheinland)  
Heller, 1386-1401  
18244634



**Alpen: Herrschaft**  
(Deutschland, Rheinland)  
Weißpfennig, nach 1436 bzw. 1463  
18244635



**Alpen: Herrschaft**  
(Deutschland, Rheinland)  
Schilling, 1422-1465  
18244643

## Übung 3: Clustern und „Aufräumen“

---

Die Daten wurden dem Interaktiven Katalog des Münzkabinetts der Staatlichen Museen zu Berlin entnommen (<https://ikmk.smb.museum/>). Es sind Suchergebnisse einer Schlagwortsuche mit den Stichworten "Mittelalter", "Deutschland" und "Heraldik / Wappen".

- Erstellen Sie das OpenRefine-Projekt mit Daten mit den Daten zu mittelalterlichen Münzen. Wie lassen sich die Daten in eine sinnvolle Tabellenform bringen?
- Explorieren Sie die Daten etwas und betreiben Sie etwas Datenbereinigung mithilfe der Funktionen „Text-Facet“ und „Cluster“
- Zerlegen sie die Daten einer 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!

## Anwendungsmöglichkeiten – fortgeschrittene Funktionen

---

### Reconcile & Match

- Vergleichen / Angleichen der eigenen Daten anhand von Datenbanken (z.B. GND oder 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.*

Quelle des Zitats und des folgenden Beispiels: <http://blog.lobid.org/2018/08/27/openrefine.html>

## 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;;*

## Übung 4: Präsidenten der USA mit GND-Daten anreichern

---

Bitte beachten Sie die Links im [kollaborative Dokument](#) .

- Erstellen Sie das OpenRefine-Projekt mit Daten zu den Präsidenten der USA 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

---

## Blogs

<https://histhub.ch/histhub-lab-tutorials-zu-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](#)]

## Danke für Ihre Aufmerksamkeit!

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

Workshop konzipiert in Anlehnung an "Library Carpentry: OpenRefine"

<https://librarycarpentry.org/lc-open-refine/>

(Licensed under CC-BY 4.0 2016–2020 by [Library Carpentry](#))

Alle Links in der Präsentation zuletzt abgerufen am 09.11.2020.