

Praxislabor Digitale Geisteswissenschaften

# Hands-On-Übung: Datenbereinigung, -transformation und -analyse mit



# OpenRefine

23.05.2024

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

# Vorbereitung | Hinweise

- Melden Sie sich für den moodle-Kurs [Praxislabor Digitale Geisteswissenschaften](#) an und schreiben Sie sich ein:

**Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities**

Praxislabor Digitale Geisteswissenschaften:  
Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>.

**Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities**

Startseite / Kurse / Verschiedenes / Praxislabor Digitale Geisteswissenschaften

**Allgemeines**

In dieses kollaborative Dokument können Themenvorschläge für die Hands-on-Sessions eingetragen werden. Informationen zur Anmeldung und Kurszeiten unter: <http://www.ub.uni-frankfurt.de/digitalhumanities>

**Einführung in TEI / XML**

Dozentin: Agnes Brauer

Der Workshop führt in die Grundlagen der Textauszeichnung mit TEI ein, einer XML-basierten und sich mittlerweile als De-facto-Standard etablierten Auszeichnungssprache speziell für die Zwecke der Geisteswissenschaften. Nach einer knappen allgemeinen Einführung werden die Teilnehmer anhand einer kleinen Übung die Praxis der Textauszeichnung mit TEI kennenlernen und sich einen ersten Überblick über die Bedeutung und die verschiedenen Module dieser Sprache verschaffen.

Link: <http://www.tei-c.org/>

**Hands-on Übung zur TEI/XML-Einführung**

Dozentin: Agnes Brauer

**Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities**

Praxislabor Digitale Geisteswissenschaften:  
Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>

Trainerin: Agnes Brauer  
Trainerin: Jakob Frohmann

**Selbsteinschreibung (Teilnehmer/in)**

Kein Einschreibeschlüssel notwendig

**EINSCHREIBEN**

# Introduction to OpenRefine

---

- OpenRefine is 'a tool for working with messy data'
- OpenRefine works best with data in a simple tabular format
- OpenRefine can help you split data up into more granular parts
- OpenRefine can help you match local data up to other data sets
- OpenRefine can help you enhance a data set with data from other sources

## Vorbereitung Hands-on-Übung

- Laden Sie bitte die aktuelle Version von OpenRefine herunter, entpacken und installieren Sie die Software auf Ihrem Computer. Sie benötigen außerdem einen Browser, in dem OpenRefine läuft:  
<http://openrefine.org/download.html>
- OpenRefine ist eine Java-Anwendung → es wird eine Java-Laufzeitumgebung benötigt; wählen Sie bitte die Variante „Windows kit with embedded Java“, sollte Java auf Ihrem PC nicht vorhanden sein
- Starten Sie die Anwendung aus dem entpackten Verzeichnis, es öffnet sich eine Shell 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/>.
- Hilfe / Infos zum Setup:  
<https://librarycarpentry.org/lc-open-refine/index.html#getting-ready>

## Heute ...

---

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

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

---

Beispiel: Schlagworte zum Thema „Originalerhalt historischer Bibliotheksbestände“

1. Erstellen Sie ein OpenRefine-Projekt mit Daten der Übungsdatei „originalerhalt.txt“
2. Legen Sie eine einspaltige Tabelle an, in der in jeder Zeile nur ein Schlagwort steht und geben Sie der Spalte einen neuen Namen.
3. Verschaffen Sie sich einen ersten Überblick über die Daten mit Hilfe der Funktion „Text facet“. Welche Schlagworte kommen im Datensatz mehrmals vor?
4. Probieren Sie auf der Grundlage der Facette die Funktion „Cluster“ aus und beseitigen Sie Tippfehler.
5. Vergleichen Sie die Ergebnisse beim Einsatz unterschiedlicher Clustering-Methoden und – Algorithmen.

## Übung 2: „Aufräumen“ am Beispiel von Daten aus einer historischen Sammlung

---

1. Erstellen Sie das OpenRefine-Projekt mit Daten aus dem Dokument „muenzen.txt“  
Vgl. <https://ikmk.smb.museum/object?lang=de&id=18206726>
2. Explorieren Sie die Daten und betreiben Sie etwas Datenbereinigung mithilfe der Funktionen „Text-Facet“ und „Cluster“.
3. Transponieren Sie die Daten in mehrere Spalten (mit *Transpose*). Wie viele Spalten sind sinnvoll?
4. 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. Wikidata)
- Anreicherung von Daten (z.B. mit Identifiern oder geographischen Koordinaten)
- 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.*



## Übung 3: Reconcile & Match am Beispiel von Personendaten

1. Legen Sie ein neues Projekt anhand der Datei „praesidenten.txt“ an
2. Bereinigen Sie die Daten
3. Splitten Sie die Daten mithilfe eines geeigneten Separators
4. Führen Sie eine Reconciliation mit Wikidata durch und „matchen“ Sie die Werte.
5. Fügen Sie neue Spalten auf Basis der Reconciliation hinzu, z.B.:
  1. Beruf
  2. Geburtsort
  3. Koordinaten des Geburtsortes
  4. Geburtsdatum
6. Fügen Sie eine Spalte hinzu, in der Links zu den einzelnen Personendatensätzen in der Gemeinsamen Normdatei im Katalog der Deutschen Nationalbibliothek ausgegeben werden.

Struktur der Links in die GND „<https://d-nb.info/gnd/>“ + „GND-ID“ ([Beispiel](#))

# Tipps & Tricks / Links / Literatur

---

## Blogs

<https://histhub.ch/erste-schritte-mit-openrefine-ein-erstes-projekt/> (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

guide that takes you from data analysis and error fixing to linking your dataset *Ruben Verborgh/Max de Wilde*, Using OpenRefine. The essential OpenRefine t to the Web (Community experience distilled), Birmingham, Mumbai 2013. [[Online-Ressource über UB FFM](#)]

Danke für Ihre Aufmerksamkeit!

Workshop konzipiert in Anlehnung an "[Library Carpentry: OpenRefine Lessons for Librarians.](#)"