



# Datenaufbereitung & Qualitätsverbesserung mit OpenRefine

---

Cora Assmann

18.10.2023

# Agenda

---

## 1.) Einführung

- Was ist OpenRefine?
- Gut zu wissen?
- Wofür?

## 2.) Demo: OpenRefine

- Sortieren, Filtern
- Faceten
- Clustern

## 3.) Ausblick

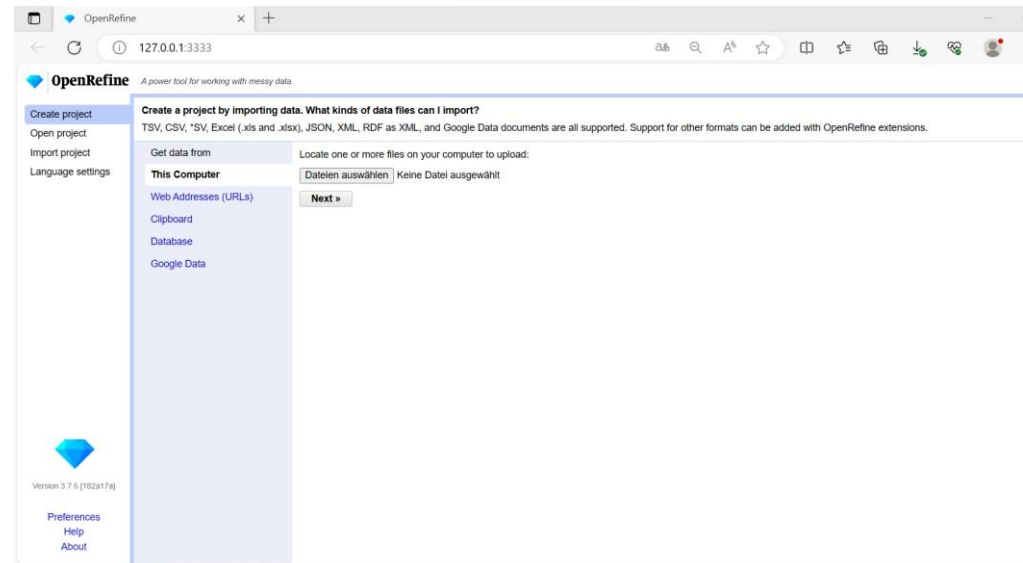
- Transformationen
- Weiterführender Workshop

# Was ist OpenRefine?



- *"ein leistungsstarkes Werkzeug für die Arbeit mit unübersichtlichen/ unsauberen Daten"* (David Huynh)

- *Desktop-Anwendung*



# Was sollte ich noch wissen.....

---



- Keine Internetverbindung erforderlich und keine der Daten oder Befehle werden an einen Remote-Server gesendet!
- Sie ändern NICHT die Original-/Rohdaten.
- Die Projekte werden alle 5 min automatisch gespeichert (Strg+C).
- Dateien werden lokal gespeichert!

# OpenRefine kann helfen ....

---

- einen Überblick über einen Datensatz zu erhalten
- Inkonsistenzen in einem Datensatz aufzulösen z. B. Standardisierung der Datumsformatierung
- Daten in feinere Teile aufzuteilen
- Abgleich lokaler Daten mit anderen Datensätzen
- Datensätze mit Daten aus anderen Quellen anzureichern

# Welche Arten von Dateiformaten kann ich importieren?

---

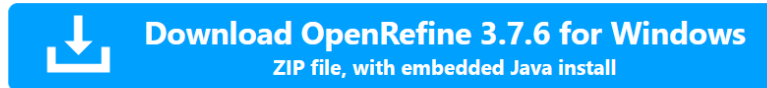
- TSV (tab-separated values)
- CSV (comma-separated values)
- TXT (plain text file without formatting)
- Excel
- JSON (javascript object notation)
- XML (extensible markup language)
- Google Spreadsheet

# Los geht's mit OpenRefine

---

Download für (Windows, Mac OS, Linux) hier  
<https://openrefine.org/download.html>

Anleitung zum Starten: <https://openrefine.org/docs/manual/running>



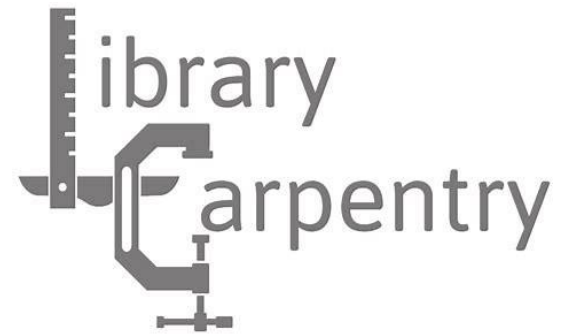
[Privacy notice](#) – [Release notes](#)

Start OpenRefine.exe und öffne <http://127.0.0.1:3333/>

# Demonstrationsdatensatz

---

Demonstrationsdatensatz [doaj-article-sample.csv](#)



Demonstrationsdatensatz stammt vom [Library Carpentry](#) Programm

weitere [Carpentrys Kurse](#) an der Uni-Jena



# Datenimporteinstellungen in OpenRefine

OpenRefine A power tool for working with messy data.

Create project « start over Configure parsing options Project name doaj article sample csv Tags Create project »

Open project

Import project

Language settings

	Title	Authors	DOI	URL	Date	Language	Subjects
1.	The Fisher Thermodynamics of Quasi-Probabilities	Flavia Pennini Angelo Plastino	10.3390/e17127853	<a href="https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412">https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412</a>	01/11/2015	English	Fisher information quasi-probabilities
2.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective	Naveed Aslam Peter C. Wynn	10.3390/agriculture5041172	<a href="https://doaj.org/article/0edc5af6672641c0bd45608812a34f9e">https://doaj.org/article/0edc5af6672641c0bd45608812a34f9e</a>	01/11/2015	English	aflatoxins AFM1 AFB1 milk market (General) S1-972 Agriculture Pakistan

Parse data as

Character encoding UTF-8

Update preview

Disable auto preview

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

Columns are separated by

commas (CSV)

tabs (TSV)

custom ,

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Column names (comma separated)

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Use character " to enclose cells containing column separators

Trim leading & trailing whitespace from strings

Escape special characters with \

Attempt to parse celltext into numbers

Store blank rows

Store blank cells as nulls

Store file source

Store archive file

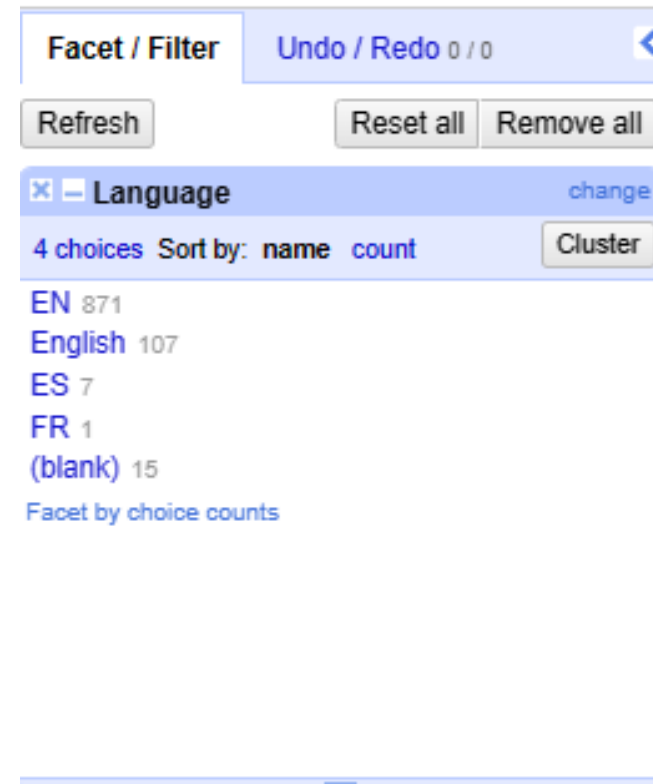
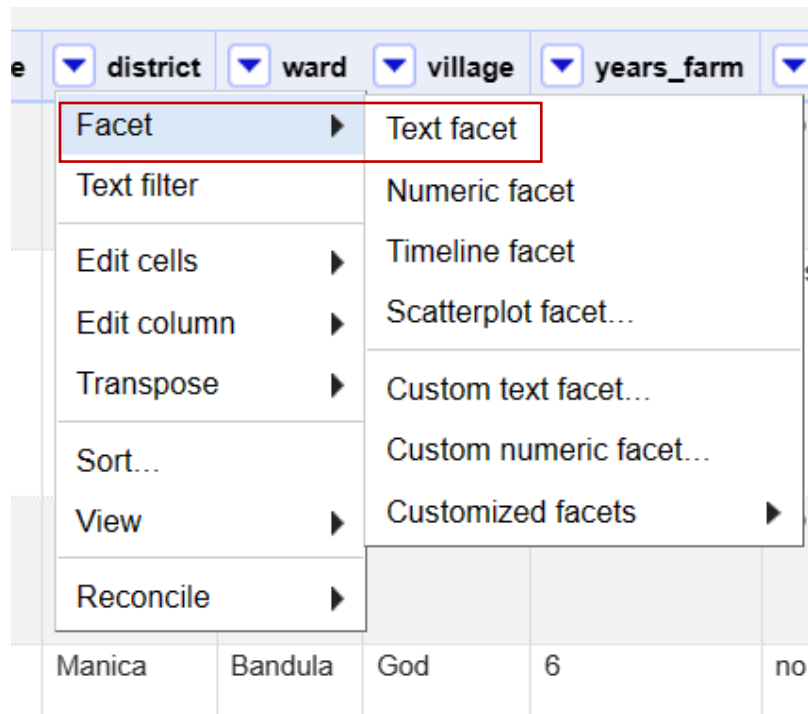
# Überblick über die Daten bekommen

---

- Anzahl der Werte überschaubar (ca.100 Zeilen) -> Facetten und Filter
- Facetten = Gruppierung mit Häufigkeit des Vorkommens von Werten
- Filter= gezielte Suche nach Werten
- Größere Datenmengen oder viele unterschiedliche Werte in einer Spalte vorhanden (>100 Zeilen) -> Clustering
- Clustering = Vorschlag von OpenRefine welche Einträge in einer Spalte das gleiche meinen könnten

# Finden von Inkonsistenzen in den Daten

- Facet -> Text Facet



# Finden von Inkonsistenzen in den Daten

- Facet -> Text Facet

The image shows the OpenRefine interface. On the left, a dropdown menu is open for the 'village' column, with 'Facet' selected. The 'Text facet' option is highlighted with a red box. An arrow points from this option to the 'Publisher' facet results panel on the right.

**Facet Menu Options:**

- Facet (selected)
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

**Facet Results Panel: Publisher**

7 choices Sort by: **name** count Cluster

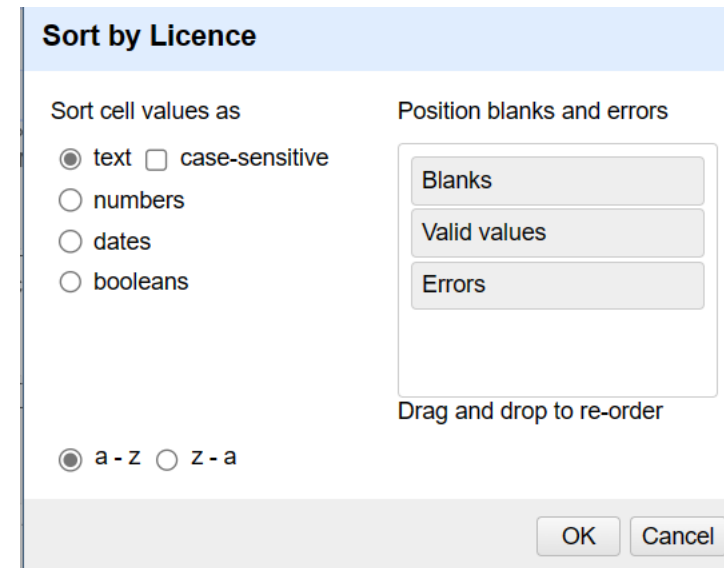
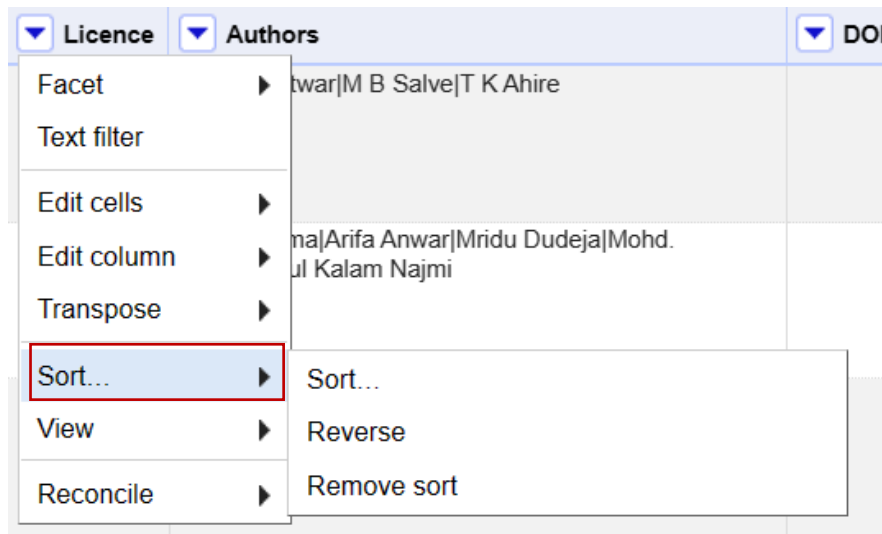
Akshantala Enterprises	13
Aurel Vlaicu University Editing House	17
Consejo Superior de Investigaciones Científicas	11
International Union of Crystallography	858
MDPI AG	3
MDPI AG	93
Society of Pharmaceutical Technocrats	6
Facet by choice counts	

**Data Table:**

district	ward	village	years_farm	
Manica	Bandula	God	6	no

# Daten sortieren /ordnen

- Sort-> Settings

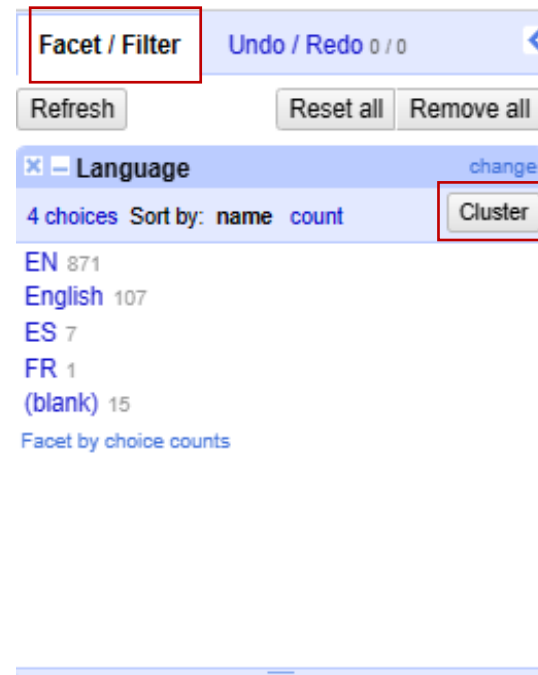


# Clustering

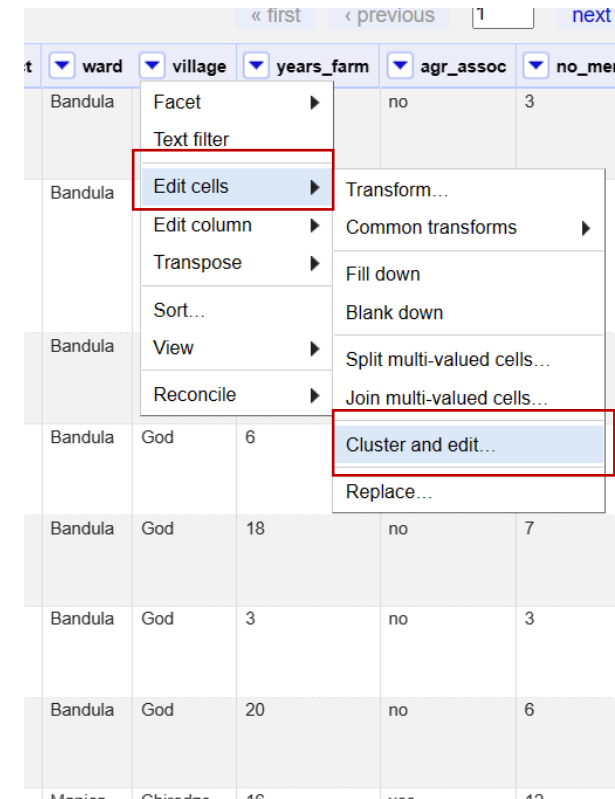
Es gibt zwei Möglichkeiten Cluster aufzurufen:

Spaltenmenü

-> Facet -> Text facet,  
oben rechts den Button  
Cluster anwählen



Spaltenmenü ->  
Edit cells ->  
Cluster and edit





# Clustering Methoden

---

OpenRefine selbst hat auf [Github](#) ausführliche Erläuterungen zu den Clustermethoden publiziert. Hier nur ein kurzer Auszug daraus: (Quelle: [Cluster – histHub](#))

**Key collision:** Unter diesem Begriff werden Clustermethoden gesammelt, die nach dem wesentlichen Teil («key») des Werts suchen. Stimmt dieser überein, schlägt die Methode Cluster vor. Key collision Methoden brauchen wenig Zeit für die Berechnung, auch wenn große Datensätze bearbeitet werden.

**fingerprint** ist die Standardmethode, weil sie wenige falsche Treffer vorschlägt. Sie entfernt Leerzeichen und Satzzeichen, betrachtet alle Buchstaben als Kleinbuchstaben und ignoriert Umlaute. Einzelne Worte werden auseinandergenommen. Die Methode ist nicht ideal für Datensätze mit vielen Umlauten, weil sie diese ignoriert.

## ***ngram-fingerprint***

Diese Methode funktioniert ähnlich, wird aber durch ein [N-Gramm](#) ergänzt. Dadurch werden auch Cluster erkannt, wenn innerhalb des [Strings](#), also der Zeichenkette, Buchstaben vertauscht sind. Die Grösse des N-Gramms kann gewählt werden. Ein großer Wert hat keine Vorteile gegenüber der Standard-fingerprint-Methode. Ein 2-Gramm oder 1-Gramm hingegen wird einige falsche Vorschläge, aber auch neue Matches hervorbringen.

# Clustering Methoden

---

## *metaphone-3*

Diese Methode basiert auf einem phonetischen fingerprint. Metaphone-3 ist spezifisch auf die englische Aussprache zugeschnitten, ist also vor allem bei englischsprachigen Datensätzen anzuwenden.

## *cologne-phonetic*

Wie metaphone-3 ist cologne-phonetic eine phonetische fingerprint Methode. Cologne-phonetic ist auf die deutsche Sprache ausgelegt. Die phonetischen Methoden helfen, Schreibfehler oder Ungleichheiten aufgrund verschiedenerer Rechtschreibungen aufzuspüren.

## *nearest neighbour*

Präziser sind Methoden, die nach der Nächste-Nachbarn-Klassifikation funktionieren. Strings werden verglichen, und wenn die Abweichung einen bestimmten Grenzwert nicht überschreitet, wird ein Cluster vorgeschlagen. Der Nachteil daran ist, dass es schon für kleine Datenmengen sehr viele Berechnungen braucht, was die Methode langsam macht. OpenRefine schaltet deshalb eine key collision davor, die Blöcke mit ähnlichen Einträgen schafft. Es ist möglich, die Größe der Blöcke im Feld «Block» zu bestimmen, wenn eine «Nearest-neighbour»-Methode gewählt wurde. Werte unter drei verlangsamen die Berechnung und bringen selten bessere Ergebnisse. Der Unterschied zwischen den nearest neighbour Methoden liegt in der Art der Berechnung der Differenz zwischen zwei Strings.



# Clustering Methoden

---

## *Levensthein*

Auch Bearbeitungs-Distanz genannt, berechnet diese Methode die Anzahl Bearbeitungsschritte zwischen zwei strings. «Teheran» und «Tehran» ist einen Arbeitsschritt entfernt (einen Buchstaben einfügen), «Buenos Aires» und «Buenosaires» zwei. In OpenRefine kann die Distanz im Feld «Radius» eingestellt werden. Grosse Zahlen liefern hier, vor allem bei kurzen Zeichenketten, übermässig viele Treffer.

## *PPM*

Diese Methode baut auf der Kompression von Textstrings auf. Eine solche Kompression berechnet den Inhalt eines Strings. Entsprechend muss der Inhalt von Zelle A ungefähr dem Inhalt von A+B entsprechen. Diese Methode gibt sehr viele Treffer zurück. Deshalb sollte sie erst nach den anderen Cluster-Methoden genutzt werden. Ausserdem ist sie für längere Einträge pro Zelle präziser als für kürzere.

# Transformationen

- für häufige Transformationen wie z.B. Leerzeichen entfernen, Wechsel zu Groß- oder Kleinschreibung gibt es fertige Funktionen

## Spaltenmenü

-> Edit cells

-> Common transform

-> Collapse consecutive whitespace

The screenshot shows a spreadsheet application with a context menu open over a cell. The menu contains the following options:

- Trim leading and trailing whitespace
- Collapse consecutive whitespace** (highlighted with a red rectangle)
- Unescape HTML entities
- Replace smart quotes with ASCII
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- To null
- To empty string

The 'Transform...' option is also highlighted with a red rectangle, and its sub-menu is open, showing the following options:

- Common transforms** (highlighted with a red rectangle)
- Fill down
- Blank down
- Split multi-valued cells...
- Join multi-valued cells...
- Cluster and edit...
- Replace...

# Ausblick

---

- Komplexere Transformationen z. B.
  - Extrahieren eines bestimmten Datentyps aus einer längeren Textfolge (z. B. Auffinden von ISBNs in einem bibliografischen Zitat)
  - Aufteilung von Daten in einer einzigen Spalte in mehrere Spalten (z. B. Aufteilung einer Adresse in mehrere Teile)
  - Standardisierung des Datenformats in einer Spalte, ohne die Werte zu ändern (z. B. Entfernen von Satzzeichen oder Standardisierung eines Datumsformats)

-> werden mit GREL- General Refine Expression Language  
<https://docs.openrefine.org/manual/grelfunctions> umgesetzt

Vielen Dank für Ihre Aufmerksamkeit.