Praxislabor Digitale Geisteswissenschaften

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Datenbereinigung mit

OpenRefine

03.02.2023

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

OpenRefine – Was ist das?

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- interaktives Werkzeug zum Bearbeiten, Erkunden und Bereinigen großer Mengen von Daten in Tabellenform („A power tool for working with messy data.“)
- hat große Ähnlichkeiten zu einem Tabellenkalkulationsprogramm mit Zeilen und Spalten (z.B. Excel), funktioniert eher wie eine relationale Datenbank.
  (ein „OpenRefine-Projekt“ = eine Tabelle)
- läuft lokal auf dem Rechner, aber im Browser (keine Internetverbindung notwendig)
- Open Source-Software in Java
  (zuvor zwischenzeitlich zu Google gehörig unter dem Namen Google Refine)

Anwendungsmöglichkeiten und Basisfunktionen

- Reinigung und Aufbereitung von Daten
- „Exploration“ von Daten, Aufspüren von Inkonsistenzen oder Fehlern im Datenformat („the big picture of your data“)
- Umstrukturierung und Überführung von Daten in eine andere Form

Anwendungsmöglichkeiten und Basisfunktionen

- wichtige Werkzeuge in OpenRefine: verschiedene Arten von Filtern und Facetten mit vordefinierten Kriterien zum Anpassen der Anzeige

- alle Operationen geschehen i.d.R. nur auf den gewählten / selektierten Daten.

- alle Veränderungen der Daten geschehen auf einer Kopie des Datensets und können leicht wieder rückgängig gemacht werden (“play with your data”)

- Abfolgen von Operationen können gespeichert und dann auf andere Datensätze ebenfalls angewendet werden.

Anwendungsmöglichkeiten – fortgeschrittene Funktionen

Reconcile & Match

– Vergleichen / Angleichen der eigenen Daten anhand von Datenbanken (z.B. Wikidata, GND)

– Anreicherung von Daten (z.B. mit eindeutigen Identifikatoren)

– Verlinkung von Daten

… mit Hilfe von diversen Webservices mit offenen Schnittstellen (APIs) über das Internet

Grundsätzliche Arbeitsweise mit Open Refine

- Daten, die man verändern möchte, herausfiltern,
  dann die selektierten Daten gemeinsam in einer Operation bearbeiten

- Vorgehensweise:

  - Quelldaten einlesen
  - Daten analysieren
  - Daten aufräumen und optimieren
  - Daten anreichern
  - Daten im Zielformat ausgeben

Daten importieren/exportieren aus OpenRefine

- mögliche Dateiformate für den Import (Auswahl): TSV, CSV, Microsoft Excel, JSON, XML

- mögliche Dateiformate für den Export (Auswahl): TSV, CSV, Microsoft Excel, HTML (Tabelle)

Beispiel für ein Importformat

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

dnb-datashop_2019-5-17T9_10_16.csv

```
"subject";"identifier";"type";"creator";"title";"volume";"edition";"publisher";"year";"format";"ISB
"830 Deutsche Literatur";"URN:urn:nbn:de:101:1-2019032815532333331261, URL:http://nbn-resolving.de/
"830 Deutsche Literatur";"ISBN:978-3-8498-1185-3 Broschur : EUR 39.80 (DE), EUR 41.00 (AT), CHF 51.
"610 Medizin, Gesundheit ; 420 Englisch ; 430 Deutsch";"ISBN:978-3-582-76023-4 Broschur : EUR 13.90
"830 Deutsche Literatur ; B Belletristik";"ISBN:978-3-7424-1176-1 : EUR 10.00 (DE) (freier Preis),
"830 Deutsche Literatur";"IDN:1186196173";"Online-Ressource";"Büchner, Georg [Mitwirkender] ; Griem
"830 Deutsche Literatur ; B Belletristik";"IDN:1186199687";"Online-Ressource";"Büchner, Georg [Mitw
"59 Belletristik";"URN:urn:nbn:de:101:1-2019022511054068929226, URL:http://nbn-resolving.de/urn:nbn
"830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040501314604322571, URL:http://
"830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040501320309141140, URL:http://
"830 Deutsche Literatur ; B Belletristik";"ISBN:978-3-947894-94-9 Broschur : EUR 6.90 (DE), EUR 7.1
"49 Theater, Tanz, Film ; 48 Musik ; 59 Belletristik";"URN:urn:nbn:de:101:1-2019020613352145115773,
"59 Belletristik ; 48 Musik";"URN:urn:nbn:de:101:1-2019020421203627151616, URL:http://nbn-resolving
"830 Deutsche Literatur ; 780 Musik";"URN:urn:nbn:de:101:1-2019040912341948360975, URL:http://nbn-r
"320 Politik";"URN:urn:nbn:de:101:1-2019022020314772961194, URL:http://nbn-resolving.de/urn:nbn:de:
"";"URN:urn:nbn:de:101:1-2019050419082680329417, URL:http://nbn-resolving.de/urn:nbn:de:101:1-20190
"830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040412330737993561, URL:http://
"S Schulbücher";"ISBN:978-3-8490-3257-9 Broschur : EUR 9.95 (DE), EUR 10.30 (AT), CHF 10.50 (freier
"830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040412355162390040, URL:http://
"830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019031322202332505205, URL:http://
"";"ISBN:978-3-582-76005-0 Broschur : EUR 16.90 (DE), 3-582-76005-7, IDN:1180147073";"";"";"Arbeits
"";"ISBN:978-3-582-76008-1 Broschur : EUR 18.90 (DE), 3-582-76008-1, IDN:1180146794";"";"";"Arbeits
"830 Deutsche Literatur";"ISBN:978-3-934820-27-2 Broschur : EUR 24.80 (DE), EUR 25.50 (AT), 3-93482
"";"URN:urn:nbn:de:101:1-2019041416095448364279, URL:http://nbn-resolving.de/urn:nbn:de:101:1-20190
"830 Deutsche Literatur";"URN:urn:nbn:de:101:1-2019031016140517141088, URL:http://nbn-resolving.de/
"S Schulbücher";"ISBN:978-3-582-68914-6 Festeinband : EUR 36.90 (DE), 3-582-68914-X, IDN:1175574600
```

Bearbeitungsbeispiele: Formalisieren

Data you have	Desired data
1st January 2014	2014-01-01
01/01/2014	2014-01-01
Jan 1 2014	2014-01-01
2014-01-01	2014-01-01

https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

https://xkcd.com/1179/

Bearbeitungsbeispiele: Vereinheitlichen

Data you have	Desired data
London	London
London]	London
London,]	London
london	London

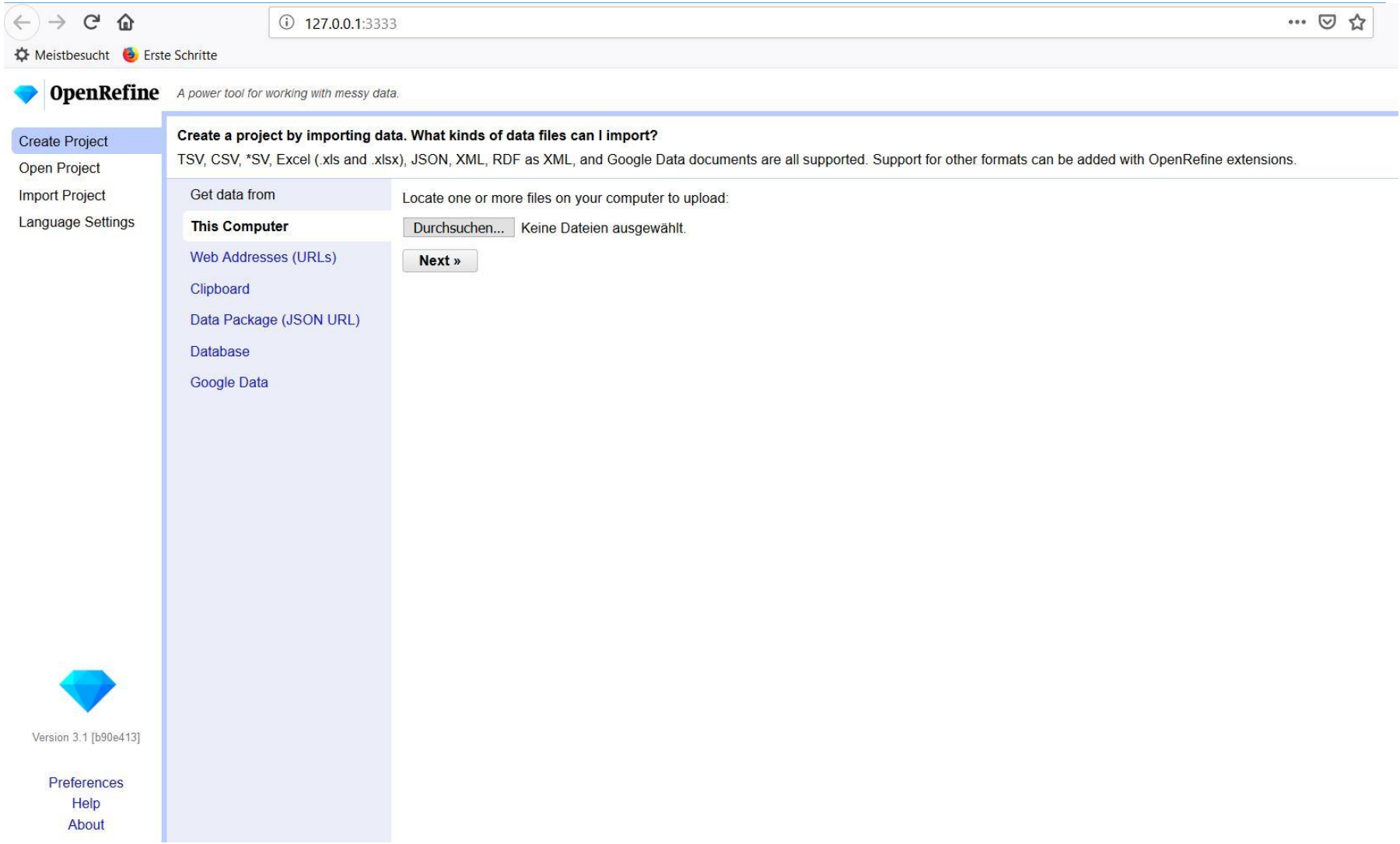
https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

Bearbeitungsbeispiele: Segmentieren

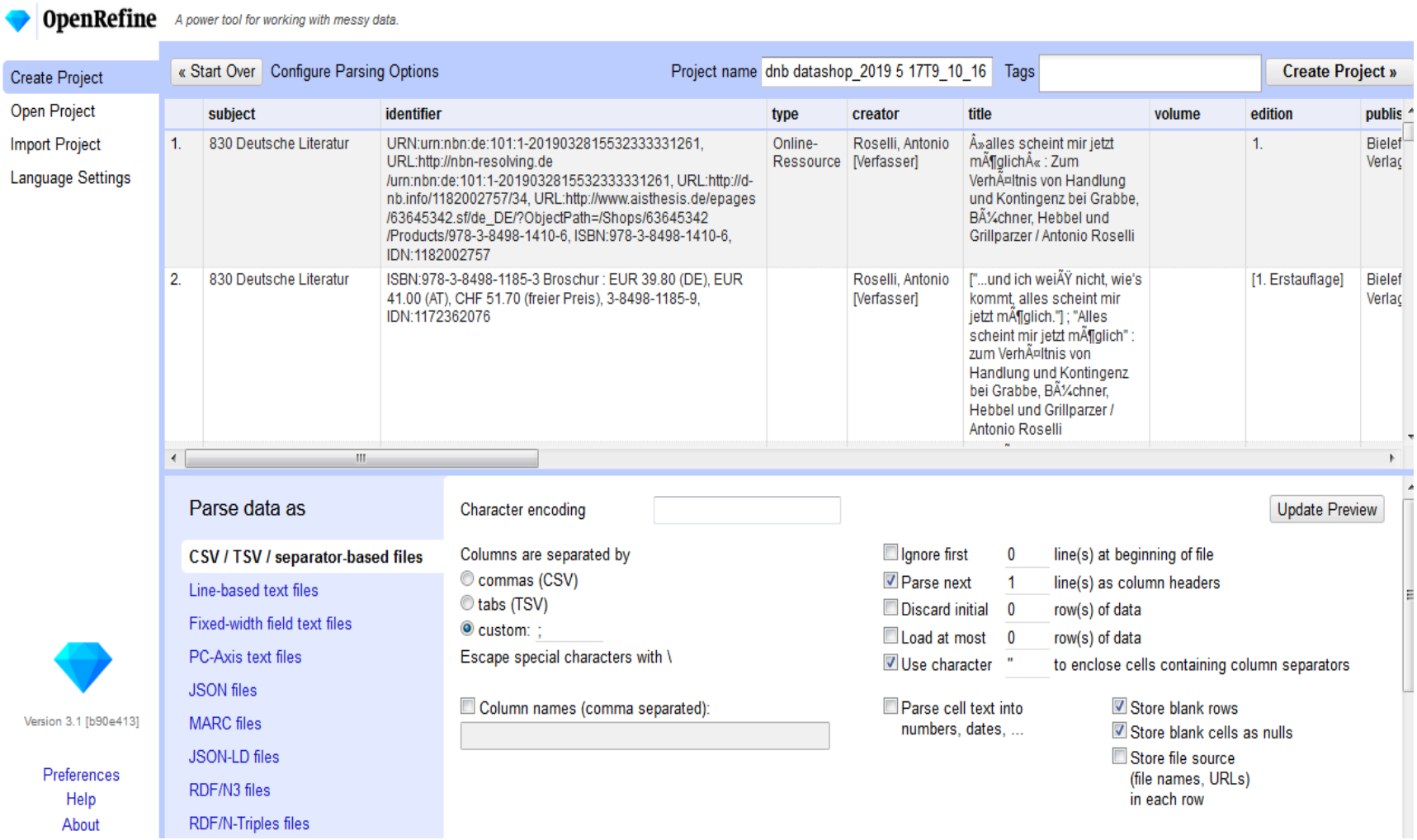
Address in single field	Institution	Library name	Address 1	Address 2	Town/City	Region	Country	Postcode
University of Wales, Llyfrgell Thomas Parry Library, Llanbadarn Fawr, ABERYSTWYTH, Ceredigion, SY23 3AS, United Kingdom	University of Wales	Llyfrgell Thomas Parry Library	Llanbadarn Fawr		Aberystwyth	Ceredigion	United Kingdom	SY23 3AS
University of Aberdeen, Queen Mother Library, Meston Walk, ABERDEEN, AB24 3UE, United Kingdom	University of Abderdeen	Queen Mother Library	Meston Walk		Aberdeen		United Kingdom	AB24 3UE
University of Birmingham, Barnes Library, Medical School, Edgbaston, BIRMINGHAM, West Midlands, B15 2TT, United Kingdom	University of Birmingham	Barnes Library	Medical School	Edgbaston	Birmingham	West Midlands	United Kingdom	B15 2TT
University of Warwick, Library, Gibbett Hill Road, COVENTRY, CV4 7AL, United Kingdom	University of Warwick	Library	Gibbett Hill Road		Coventry		United Kingdom	CV4 7AL

https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

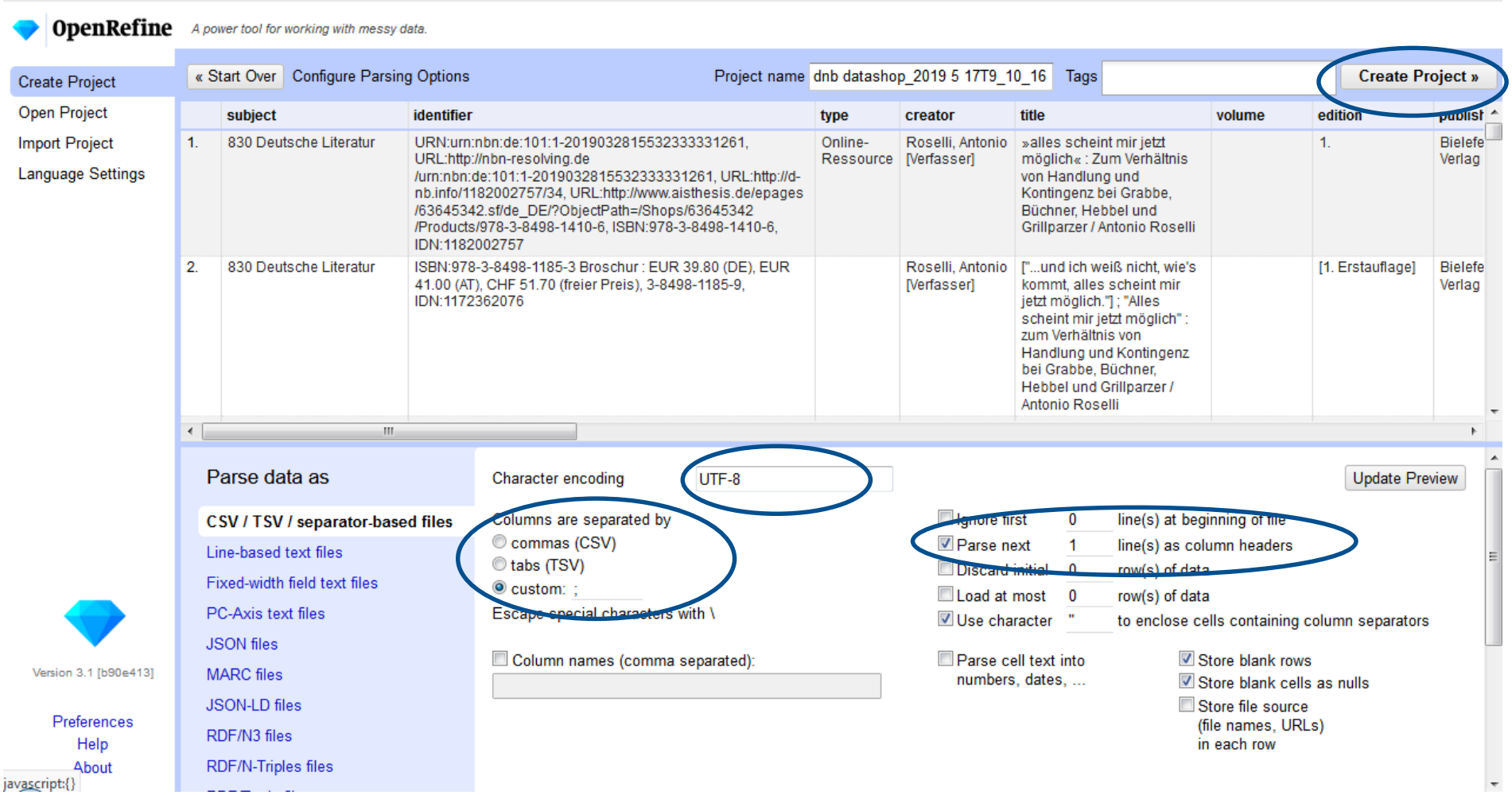
Ein Projekt erstellen…



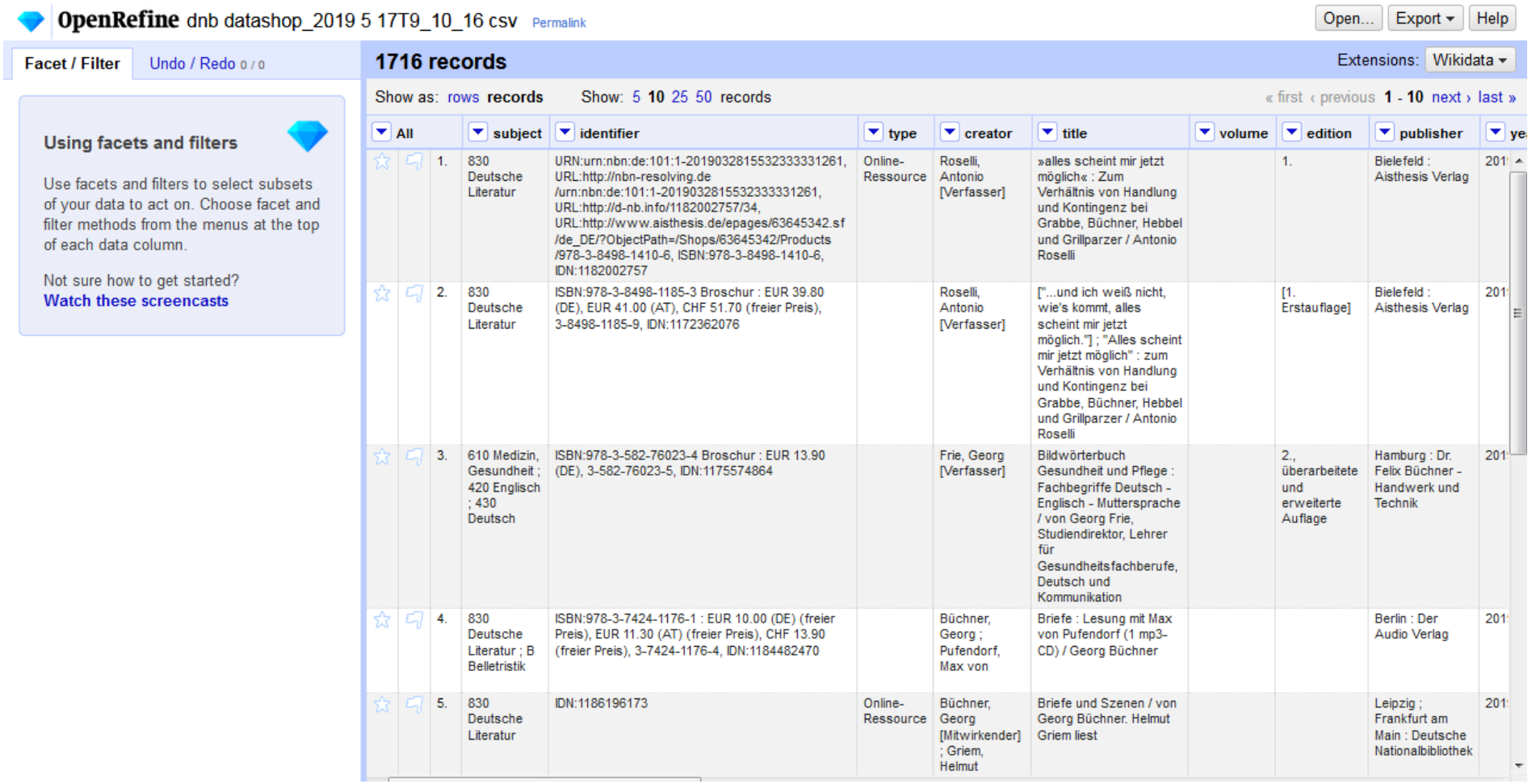
Ein Projekt erstellen…



Ein Projekt erstellen…

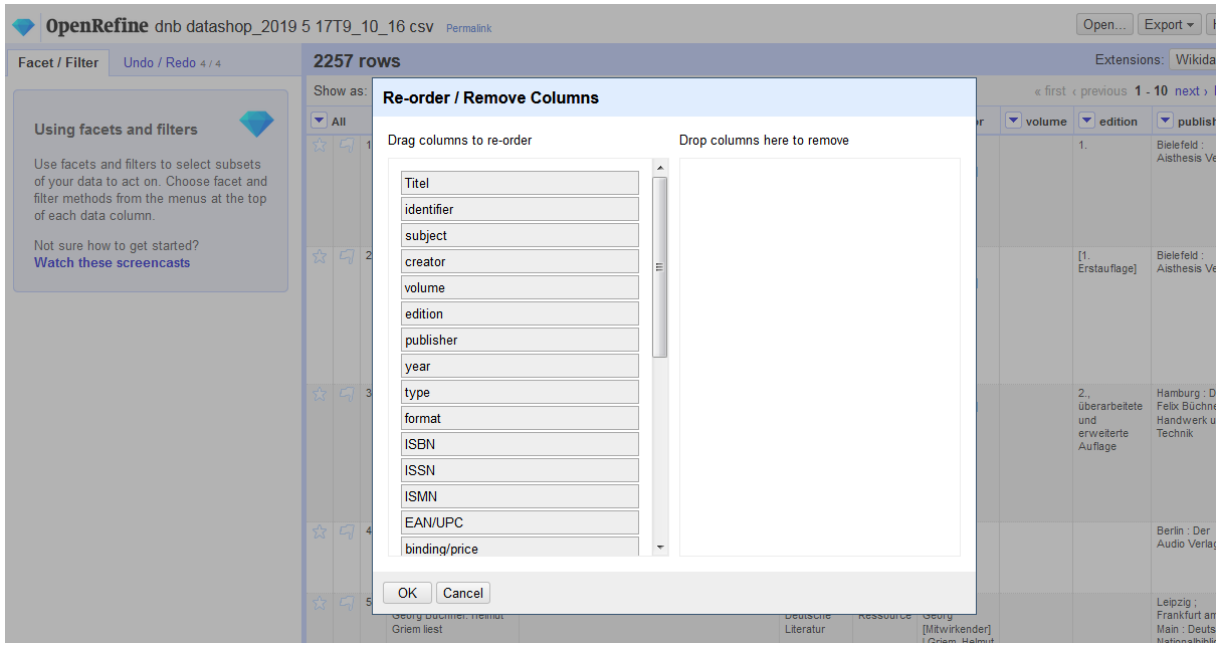


Ein Projekt erstellen…



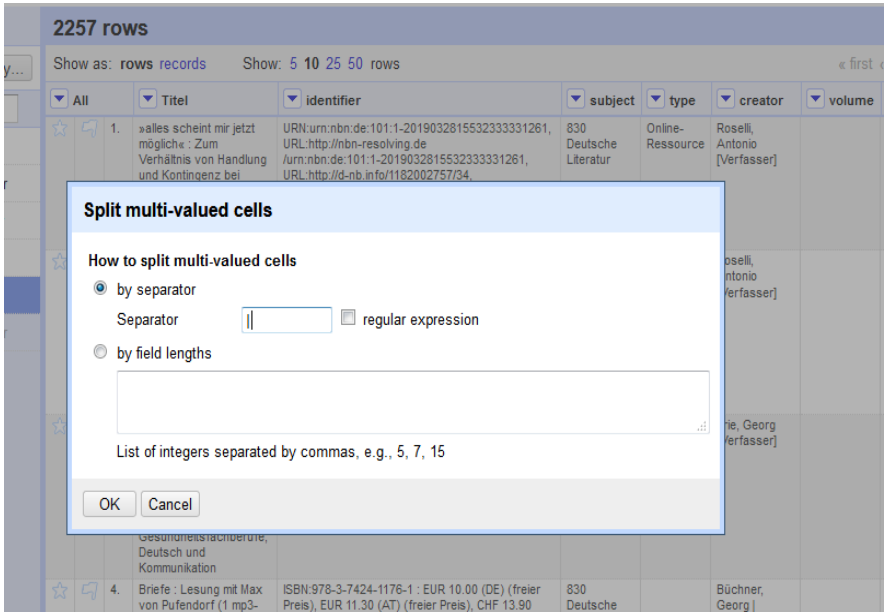
Layout

- Neuordnen von Spalten:
  - Dropdownmenü bei “All”:
    Edit columns → Re-order / Remove Columns



Layout

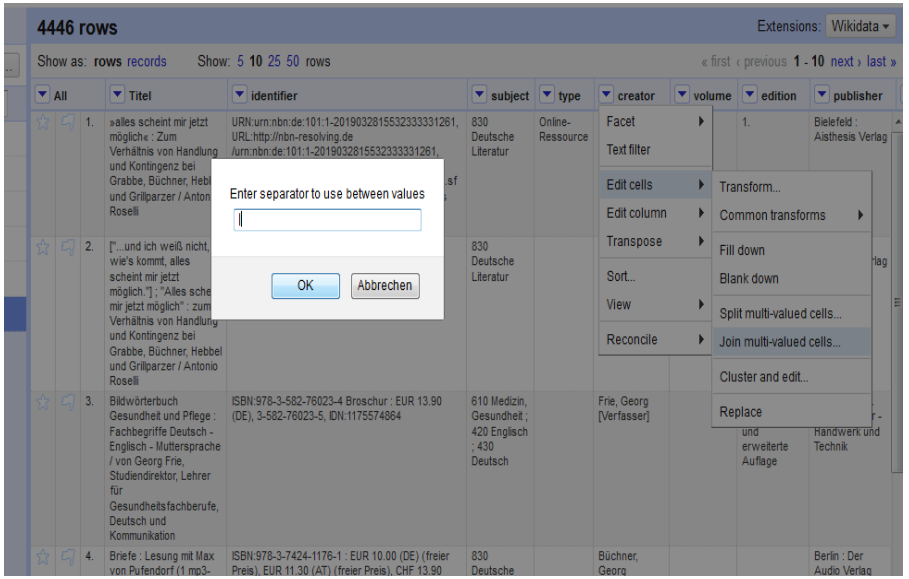
- Splitting Cells:
  - Aufsplitten von z.B. Verfasserangaben, um effektiv mit diesen Arbeiten zu können
  - Edit cells->Split multi-valued cells
- Wechsel rows / record Ansicht



UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Layout

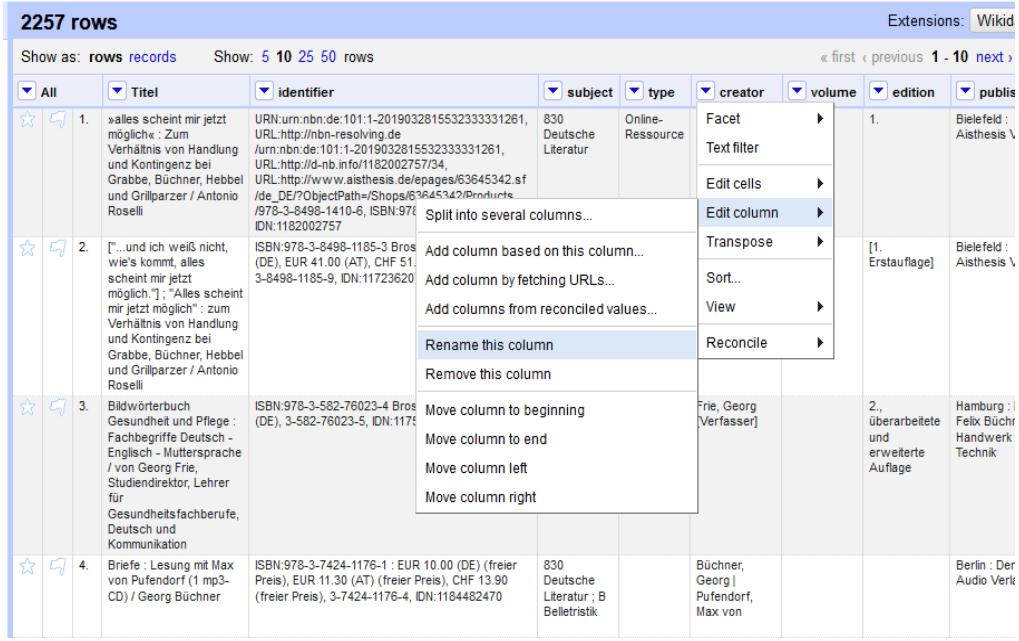
- Joining Cells:
  - Zusammenführen der Verfasserangaben eines Eintrags
  - Edit cells->Join multi-valued cells
  - Separator eintragen (Achtung: Separator sollte nicht bereits in den Daten enthalten sein, z.B. „|“)



UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

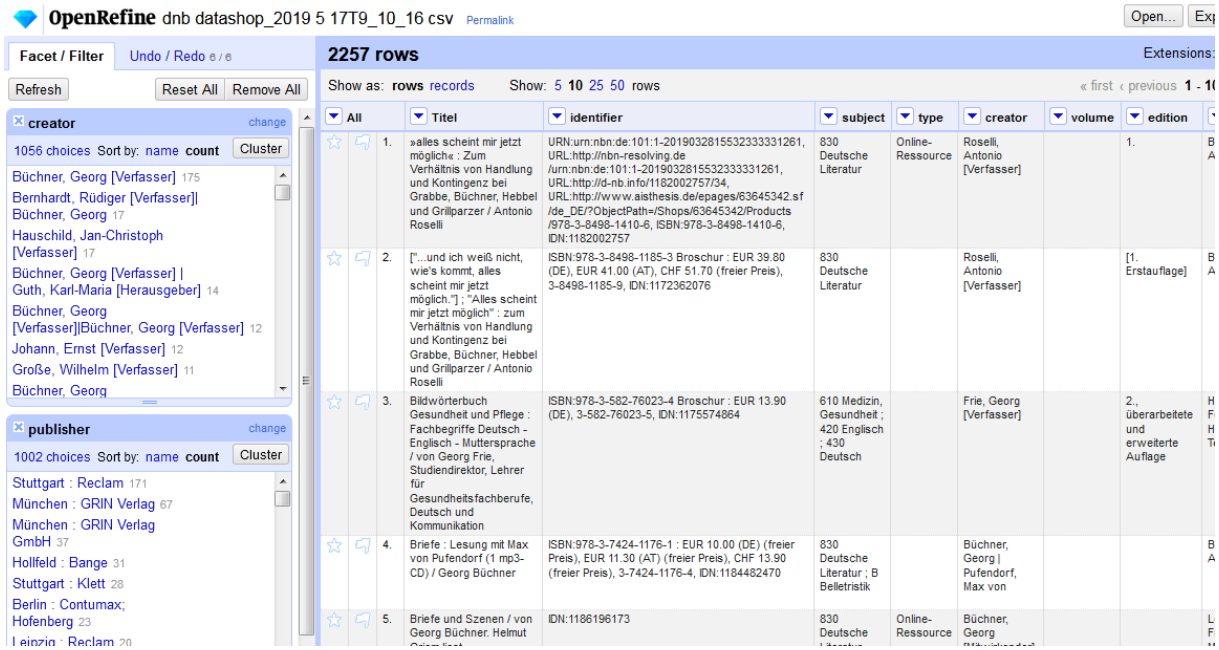
Layout

- Umbenennen von Spalten:
  - Edit column → Rename this column



Facetten und Filter

- Facetten gruppieren die Inhalte einer Spalte
- die am einfachsten zu benutzende Facette ist die Textfacette:
  - Facet → Text Facet
  - Optionen: exclude – include – invert



UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

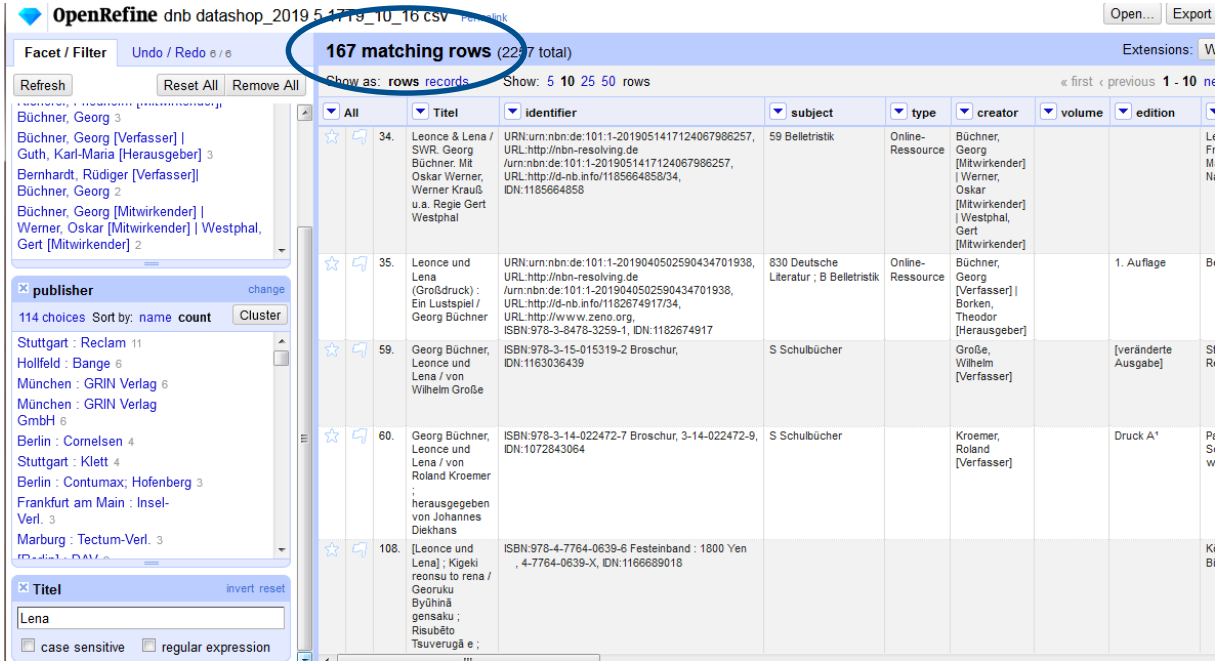
Facetten und Filter

- Customized Facets (Auswahl):
  - Word Facet: segmentiert Text in Wörter und verzeichnet ihre Häufigkeit
  - Text length facet: wertet die Länge des Texts aus und visualisiert diese
  - Facet by Blank: Filtert leere Zellen

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

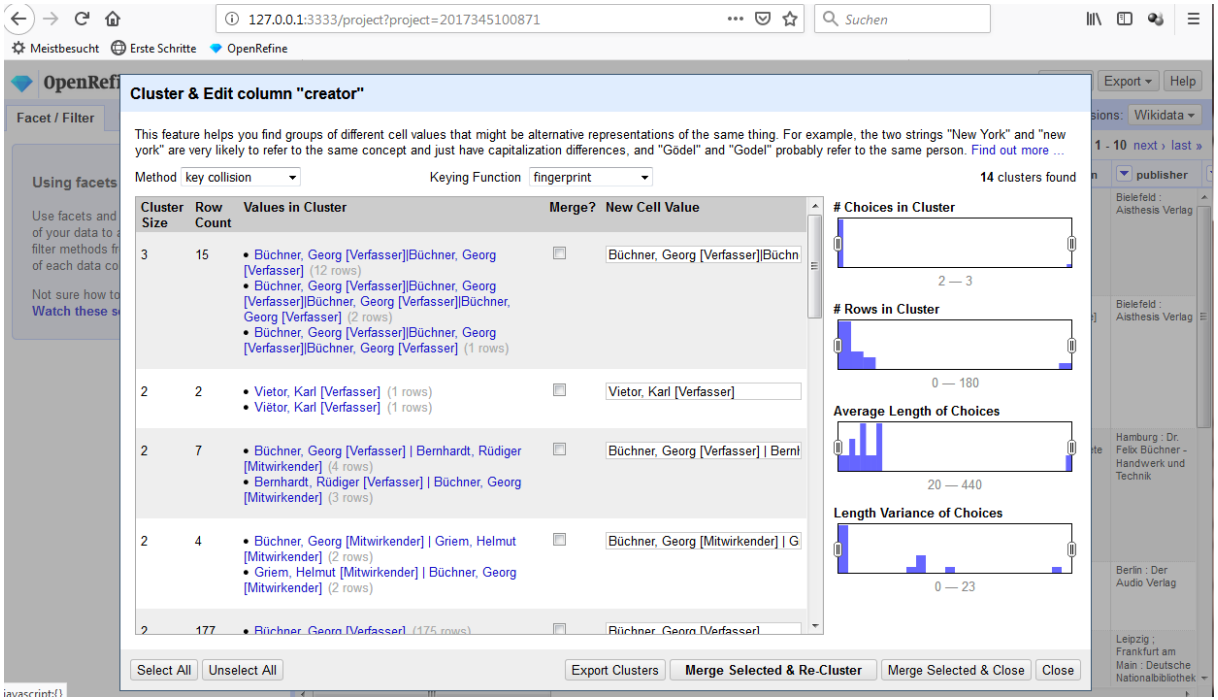
Facetten und Filter

- Text Filter:
  - filtert Werte mit übereinstimmendem String



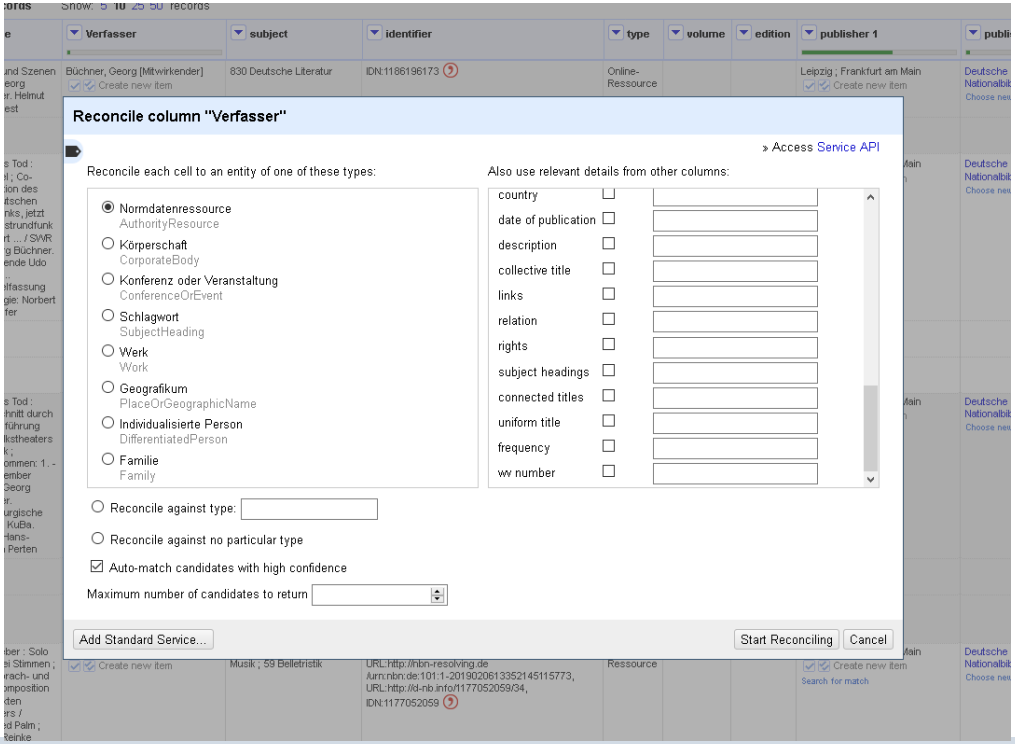
Clustering

- Funktion gruppiert ähnliche, ggf. inkonsistente Einträge und bietet die Möglichkeit einer Zusammenführung
- Sehr nützlich bei Varianten im Bereich von Namen
- Edit → Cluster and edit

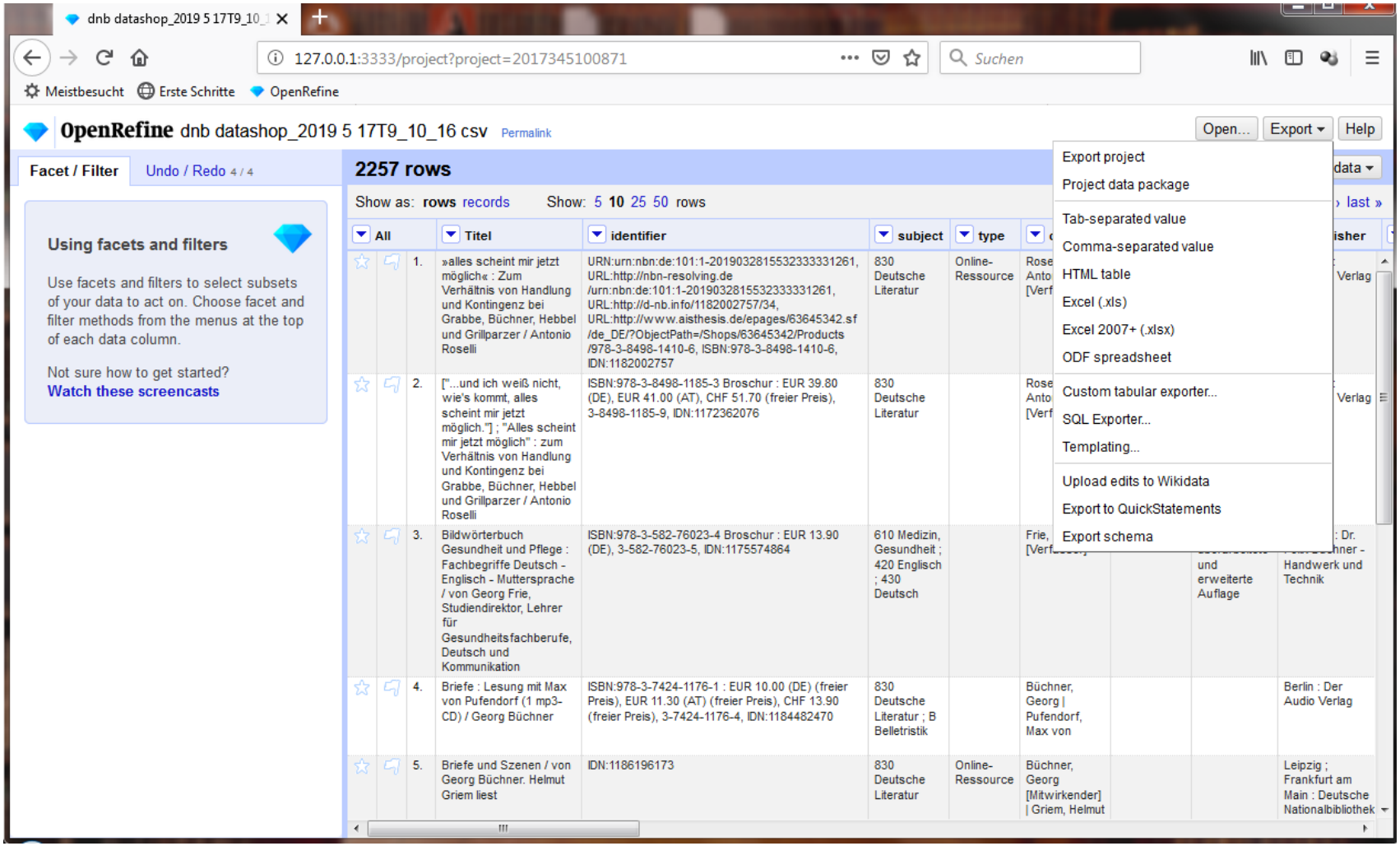


Reconciliation

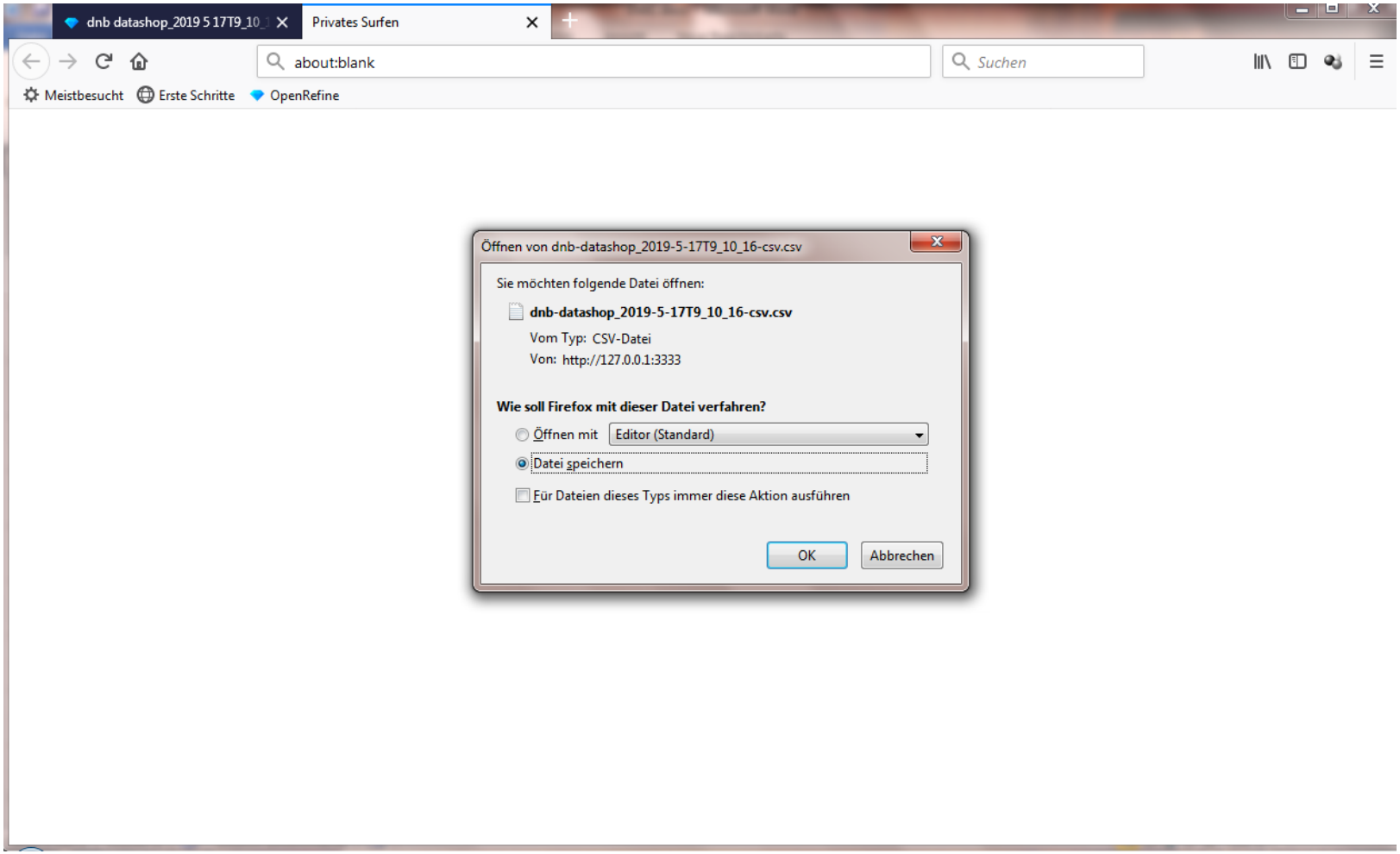
- Z.B. über Reconciliation-Webservice für die GND: https://lobid.org/gnd/reconcile
- Erlaubt den Abgleich und das Aufwerten der eigenen Daten mit den GND-Informationen



Daten speichern / exportieren



Daten speichern / exportieren



Vorbereitung und Hinweise für die Hands-on-Übung

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- Laden Sie bitte die aktuelle Verion von OpenRefine (stable release) herunter, entpacken und installieren Sie die Software auf Ihrem Computer. Sie benötigen den Browser Firefox, 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/setup.html

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

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]

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Danke für Ihre Aufmerksamkeit!

Workshop konzipiert in Anlehnung an "Library Carpentry: OpenRefine Lessons for Librarians."