Praxislabor Digitale Geisteswissenschaften

Datenbereinigung mit

OpenRefine

Agnes Brauer
a.brauer@ub.uni-frankfurt.de

OpenRefine – Was ist das?

- 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)
- 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

Siehe auch: https://www.tanner.de/blog/systeme-technologien/datenanalyse-und-optimierung-mit-openrefine-1/

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

```
1  "subject";"identifier";"type";"creator";"title";"volume";"edition";"publisher";"year";"format";"ISB
2  "830 Deutsche Literatur";"URN:urn:nbn:de:101:1-2019032815532333331261, URL:http://nbn-resolving.de/
3  "830 Deutsche Literatur";"ISBN:978-3-8498-1185-3 Broschur : EUR 39.80 (DE), EUR 41.00 (AT), CHF 51.
4  "610 Medizin, Gesundheit ; 420 Englisch ; 430 Deutsch";"ISBN:978-3-582-76023-4 Broschur : EUR 13.90
5  "830 Deutsche Literatur ; B Belletristik";"ISBN:978-3-7424-1176-1 : EUR 10.00 (DE) (freier Preis),
6  "830 Deutsche Literatur";"IDN:1186196173";"Online-Ressource";"Büchner, Georg [Mitwirkender] ; Griem
7  "830 Deutsche Literatur ; B Belletristik";"IDN:1186199687";"Online-Ressource";"Büchner, Georg [Mitw
8  "59 Belletristik";"URN:urn:nbn:de:101:1-2019022511054068929226, URL:http://nbn-resolving.de/urn:nbn
9  "830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040501314604322571, URL:http://
10 "830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040501320309141140, URL:http://
11 "830 Deutsche Literatur ; B Belletristik";"ISBN:978-3-947894-94-9 Broschur : EUR 6.90 (DE), EUR 7.1
12 "49 Theater, Tanz, Film ; 48 Musik ; 59 Belletristik";"URN:urn:nbn:de:101:1-2019020613352145115773,
13 "59 Belletristik ; 48 Musik";"URN:urn:nbn:de:101:1-2019020421203627151616, URL:http://nbn-resolving
14 "830 Deutsche Literatur ; 780 Musik";"URN:urn:nbn:de:101:1-2019040912341948360975, URL:http://nbn-r
15 "320 Politik";"URN:urn:nbn:de:101:1-2019022020314772961194, URL:http://nbn-resolving.de/urn:nbn:de:
16 "";"URN:urn:nbn:de:101:1-2019050419082680329417, URL:http://nbn-resolving.de/urn:nbn:de:101:1-20190
17 "830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040412330737993561, URL:http://
18 "S Schulbücher";"ISBN:978-3-8490-3257-9 Broschur : EUR 9.95 (DE), EUR 10.30 (AT), CHF 10.50 (freier
19 "830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019040412355162390040, URL:http://
20 "830 Deutsche Literatur ; B Belletristik";"URN:urn:nbn:de:101:1-2019031322202332505205, URL:http://
21 "";"ISBN:978-3-582-76005-0 Broschur : EUR 16.90 (DE), 3-582-76005-7, IDN:1180147073";"";"";"Arbeits
22 "";"ISBN:978-3-582-76008-1 Broschur : EUR 18.90 (DE), 3-582-76008-1, IDN:1180146794";"";"";"Arbeits
23 "830 Deutsche Literatur";"ISBN:978-3-934820-27-2 Broschur : EUR 24.80 (DE), EUR 25.50 (AT), 3-93482
24 "";"URN:urn:nbn:de:101:1-2019041416095448364279, URL:http://nbn-resolving.de/urn:nbn:de:101:1-20190
25 "830 Deutsche Literatur";"URN:urn:nbn:de:101:1-2019031016140517141088, URL:http://nbn-resolving.de/
26 "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

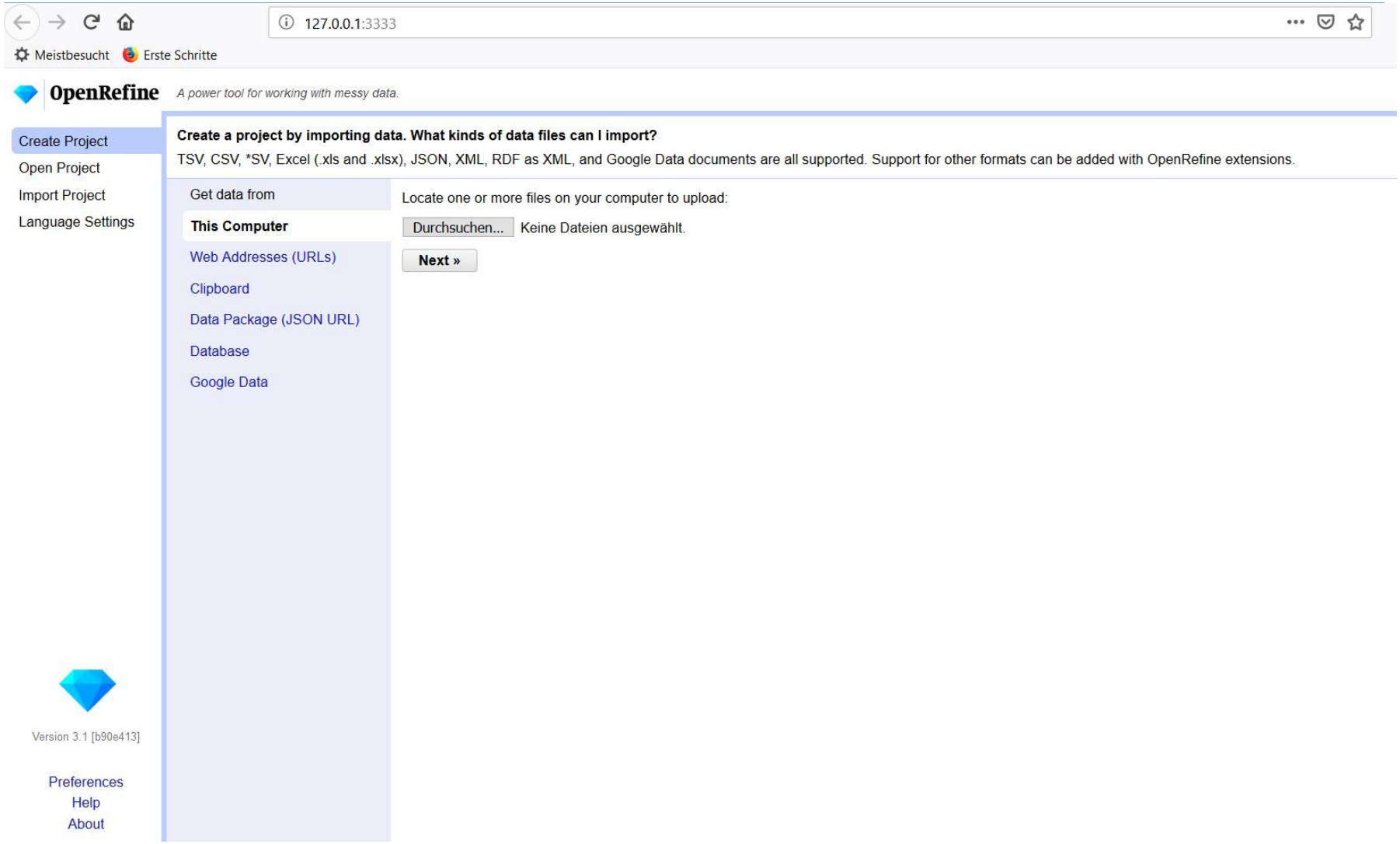
https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

Bearbeitungsbeispiele: Segmentieren

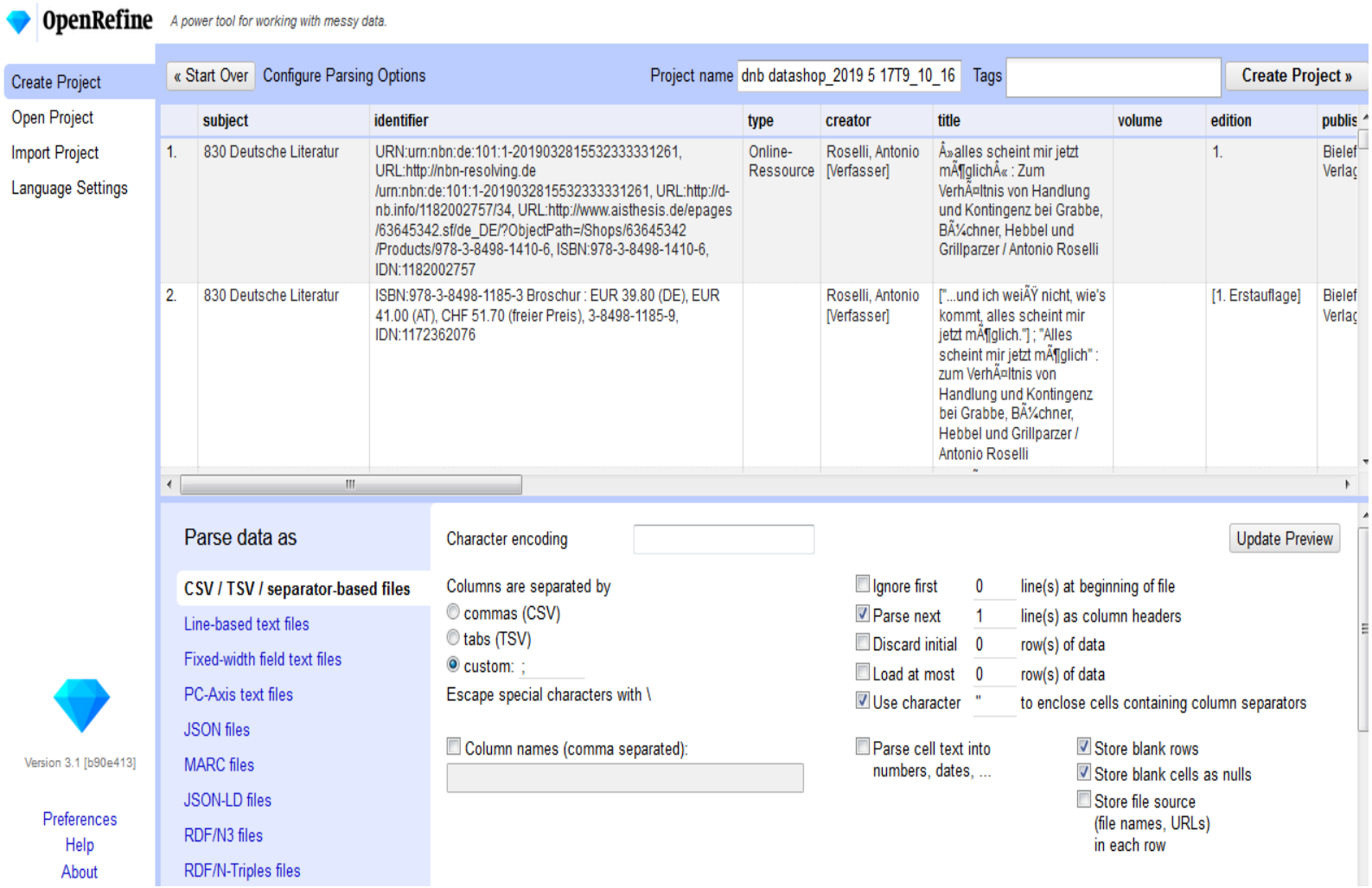
Address in single field	Institution	Library name	Address 1	Address 2	Town/City	Region	Country	Postcode
University of Wales, Llyfrgell Thomas Parry Library, Llanbadarn Fawr, ABERYSTWYTH, Ceredigion, SY23 3AS, United Kingdom	University of Wales	Llyfrgell Thomas Parry Library	Llanbadarn Fawr		Aberystwyth	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…

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

OpenRefine *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- Language Settings

« Start Over | Configure Parsing Options | Project name: dnb datashop_2019 5 17T9_10_16 | Tags | **Create Project »**

	subject	identifier	type	creator	title	volume	edition	publisl
1.	830 Deutsche Literatur	URN:urn:nbn:de:101:1-2019032815532333331261, URL:http://nbn-resolving.de /urn:nbn:de:101:1-2019032815532333331261, URL:http://d-nb.info/1182002757/34, URL:http://www.aisthesis.de/epages /63645342.sf/de_DE/?ObjectPath=/Shops/63645342 /Products/978-3-8498-1410-6, ISBN:978-3-8498-1410-6, IDN:1182002757	Online-Ressource	Roselli, Antonio [Verfasser]	»alles scheint mir jetzt möglich« : Zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli		1.	Bielefe Verlag
2.	830 Deutsche Literatur	ISBN:978-3-8498-1185-3 Broschur : EUR 39.80 (DE), EUR 41.00 (AT), CHF 51.70 (freier Preis), 3-8498-1185-9, IDN:1172362076		Roselli, Antonio [Verfasser]	["...und ich weiß nicht, wie's kommt, alles scheint mir jetzt möglich."] ; "Alles scheint mir jetzt möglich" : zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli		[1. Erstauflage]	Bielefe Verlag

Parse data as

- **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
- RDF/N-Triples files

Character encoding: UTF-8 | Update Preview

Columns are separated by
- commas (CSV)
- tabs (TSV)
- custom: ;

Escape special characters with \

Column names (comma separated):

- Ignore first 0 line(s) at beginning of file
- Parse next 1 line(s) as column headers
- Discard initial 0 row(s) of data
- Load at most 0 row(s) of data
- Use character " to enclose cells containing column separators

Parse cell text into numbers, dates, …

- Store blank rows
- Store blank cells as nulls
- Store file source (file names, URLs) in each row

Version 3.1 [b90e413]

Preferences
Help
About

javascript:{}

Ein Projekt erstellen…

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

OpenRefine dnb datashop_2019 5 17T9_10_16 csv Permalink

Open… Export ▾ He

Facet / Filter | Undo / Redo 0 / 0

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

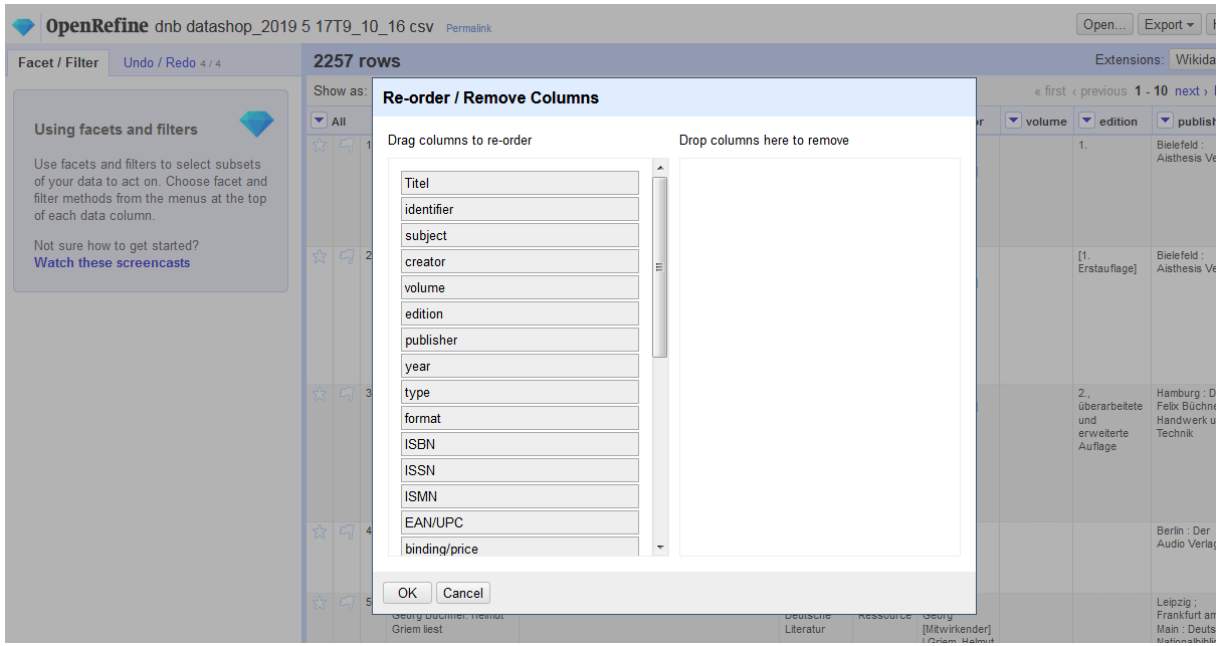
1716 records Extensions: Wikidata

Show as: rows **records** Show: 5 **10** 25 50 records « first ‹ previous **1 - 10** next › las

All		subject	identifier	type	creator	title	volume	edition	publisher	
1.	830 Deutsche Literatur	URN:urn:nbn:de:101:1-2019032815532333331261, URL:http://nbn-resolving.de /urn:nbn:de:101:1-2019032815532333331261, URL:http://d-nb.info/1182002757/34, URL:http://www.aisthesis.de/epages/63645342.sf /de_DE/?ObjectPath=/Shops/63645342/Products /978-3-8498-1410-6, ISBN:978-3-8498-1410-6, IDN:1182002757	Online-Ressource	Roselli, Antonio [Verfasser]	»alles scheint mir jetzt möglich« : Zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli		1.	Bielefeld : Aisthesis Verlag	2
2.	830 Deutsche Literatur	ISBN:978-3-8498-1185-3 Broschur : EUR 39.80 (DE), EUR 41.00 (AT), CHF 51.70 (freier Preis), 3-8498-1185-9, IDN:1172362076		Roselli, Antonio [Verfasser]	["...und ich weiß nicht, wie's kommt, alles scheint mir jetzt möglich."] ; "Alles scheint mir jetzt möglich" : zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli		[1. Erstauflage]	Bielefeld : Aisthesis Verlag	2
3.	610 Medizin, Gesundheit ; 420 Englisch ; 430 Deutsch	ISBN:978-3-582-76023-4 Broschur : EUR 13.90 (DE), 3-582-76023-5, IDN:1175574864		Frie, Georg [Verfasser]	Bildwörterbuch Gesundheit und Pflege : Fachbegriffe Deutsch - Englisch - Muttersprache / von Georg Frie, Studiendirektor, Lehrer für Gesundheitsfachberufe, Deutsch und Kommunikation		2., überarbeitete und erweiterte Auflage	Hamburg : Dr. Felix Büchner - Handwerk und Technik	2
4.	830 Deutsche Literatur ; B Belletristik	ISBN:978-3-7424-1176-1 : EUR 10.00 (DE) (freier Preis), EUR 11.30 (AT) (freier Preis), CHF 13.90 (freier Preis), 3-7424-1176-4, IDN:1184482470		Büchner, Georg ; Pufendorf, Max von	Briefe : Lesung mit Max von Pufendorf (1 mp3-CD) / Georg Büchner			Berlin : Der Audio Verlag	2
5.	830 Deutsche Literatur	IDN:1186196173	Online-Ressource	Büchner, Georg [Mitwirkender] ; Griem, Helmut	Briefe und Szenen / von Georg Büchner. Helmut Griem liest			Leipzig ; Frankfurt am Main : Deutsche Nationalbibliothek	2

Layout

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

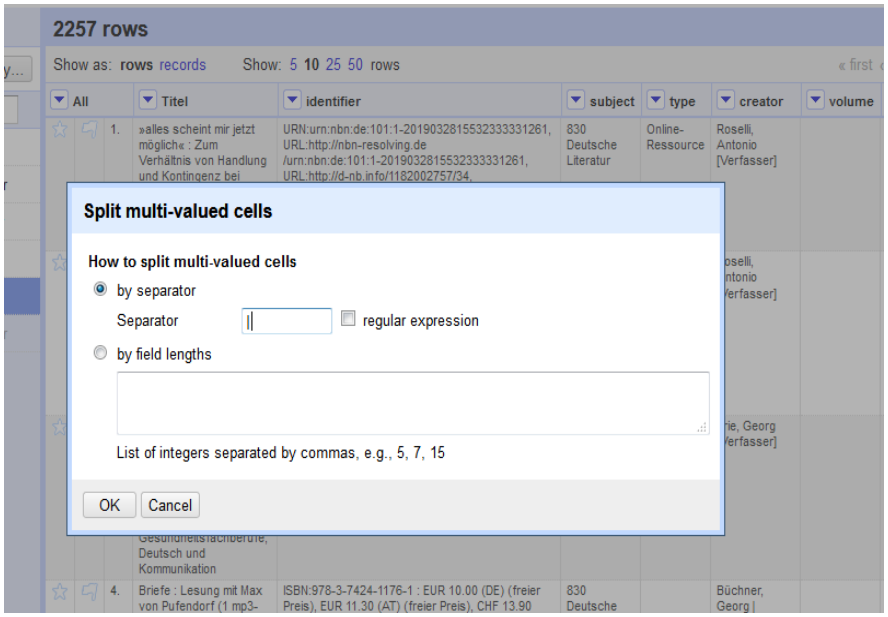


Layout

– Splitting Cells:

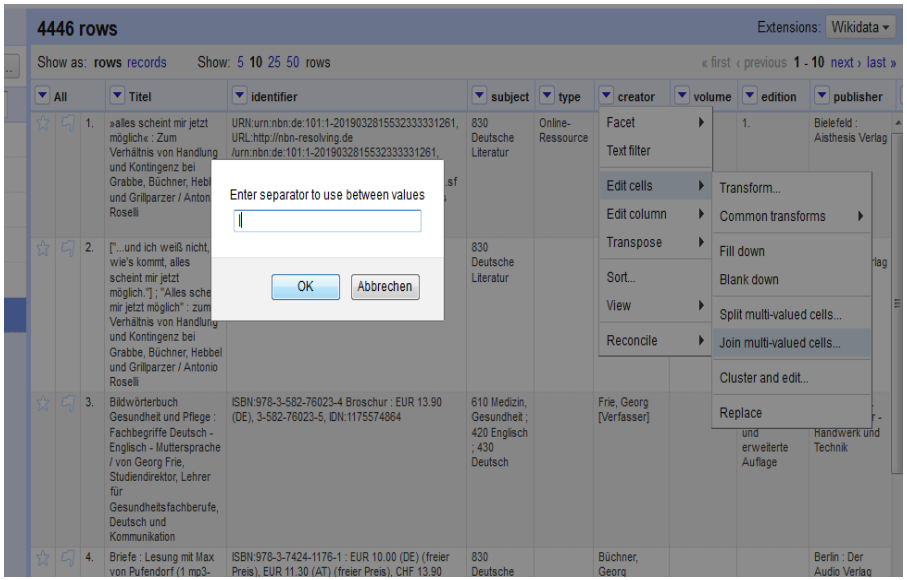
- Aufsplitten von z.B. Verfasserangaben, um effektiv mit diesen Arbeiten zu können
- Edit cells->Split multi-valued cells

– Wechsel rows / record Ansicht



Layout

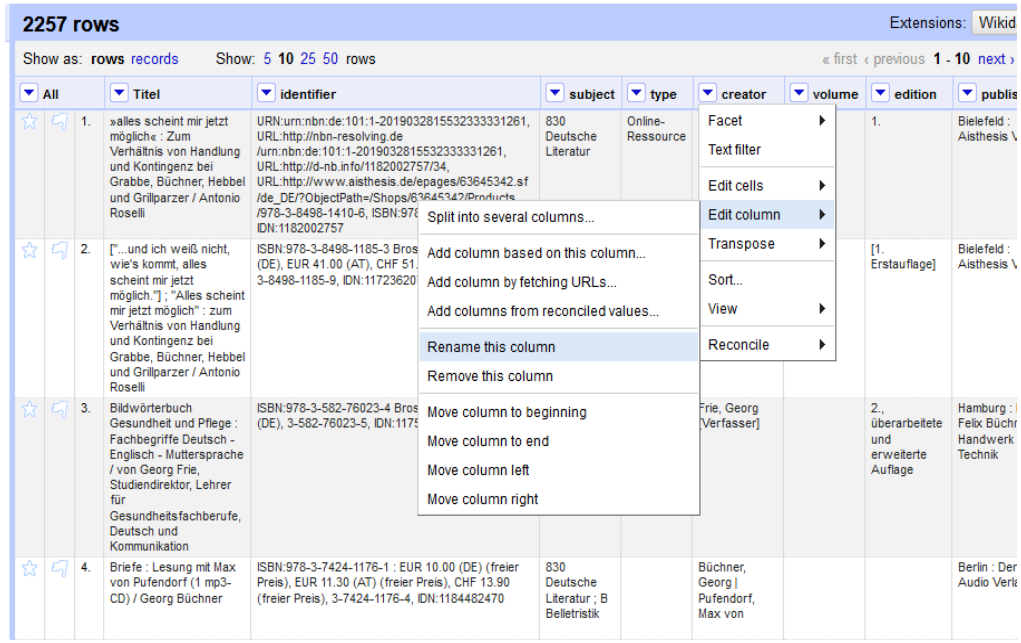
- Joining Cells:
  - Zusammenführen der Verfasserangaben eines Eintrags
  - Edit cells->Join multi-valued cells
  - Separator eintragen (Achtung: Separator sollte nicht bereits in den Daten enthalten sein, z.B. „|“)



UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

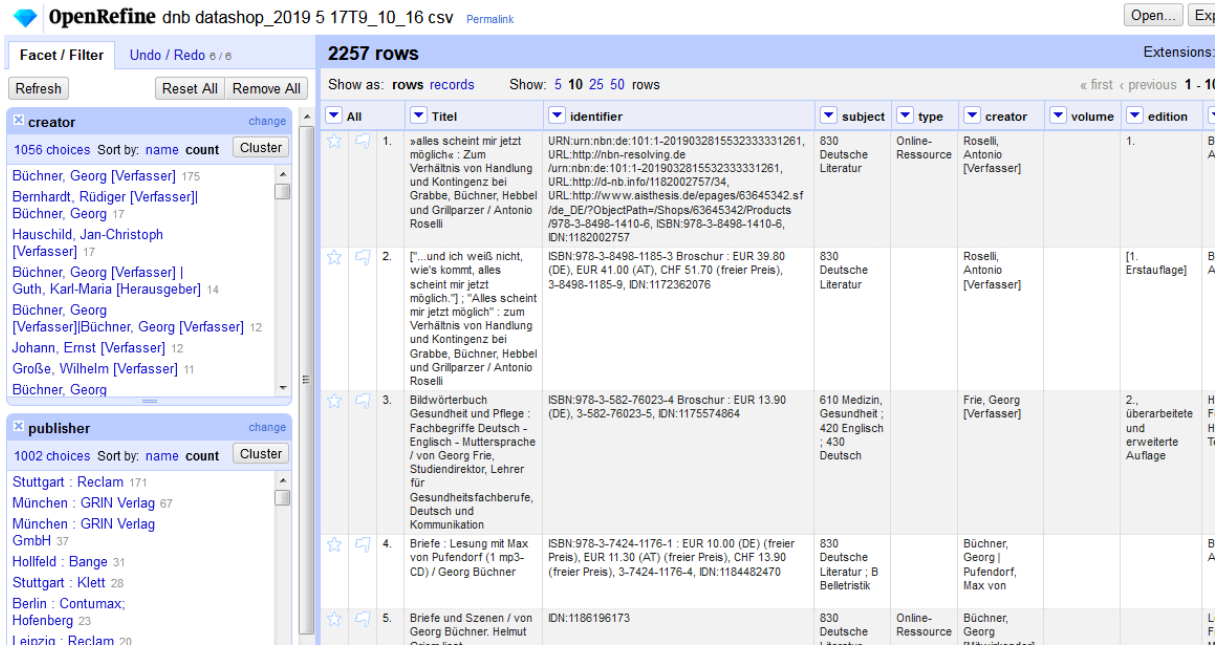
Layout

– 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

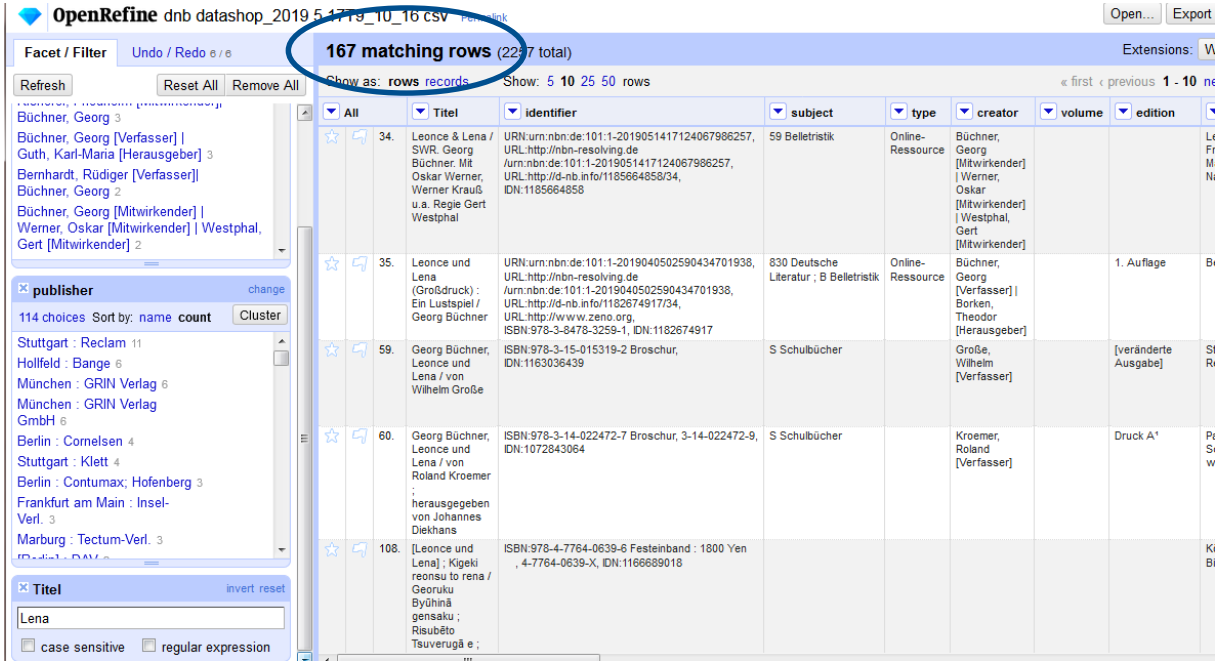


Facetten und Filter

- 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

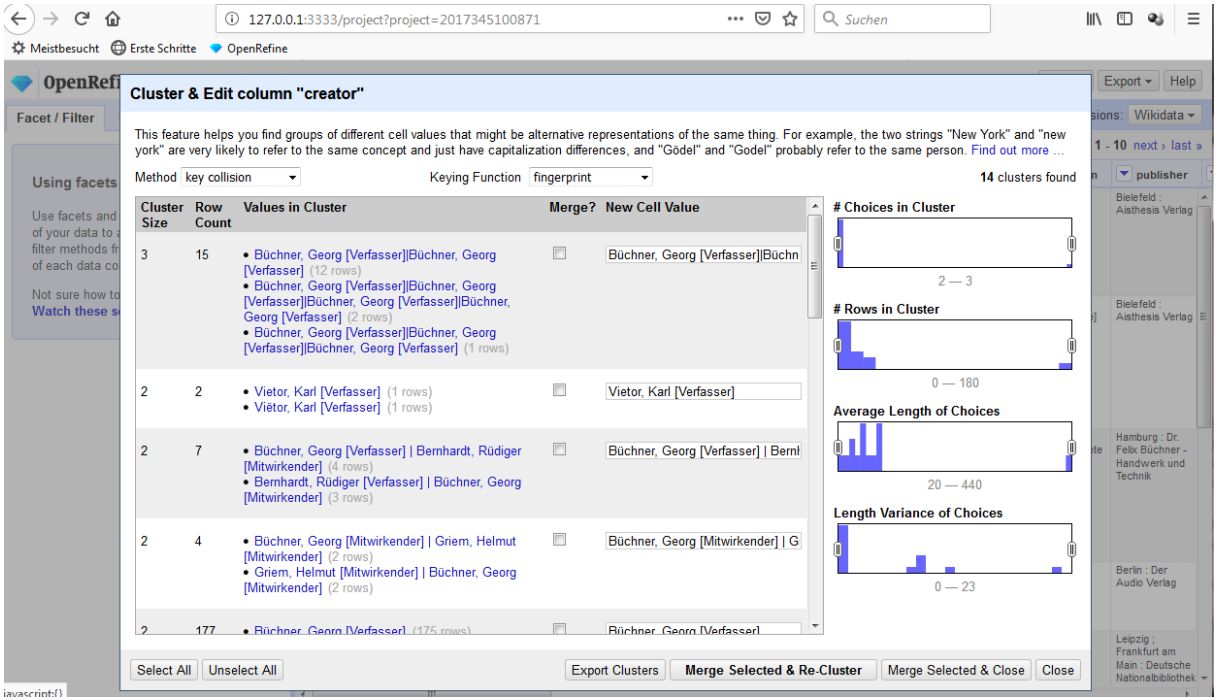
Facetten und Filter

- Text Filter:
  - filtert Werte mit übereinstimmendem String



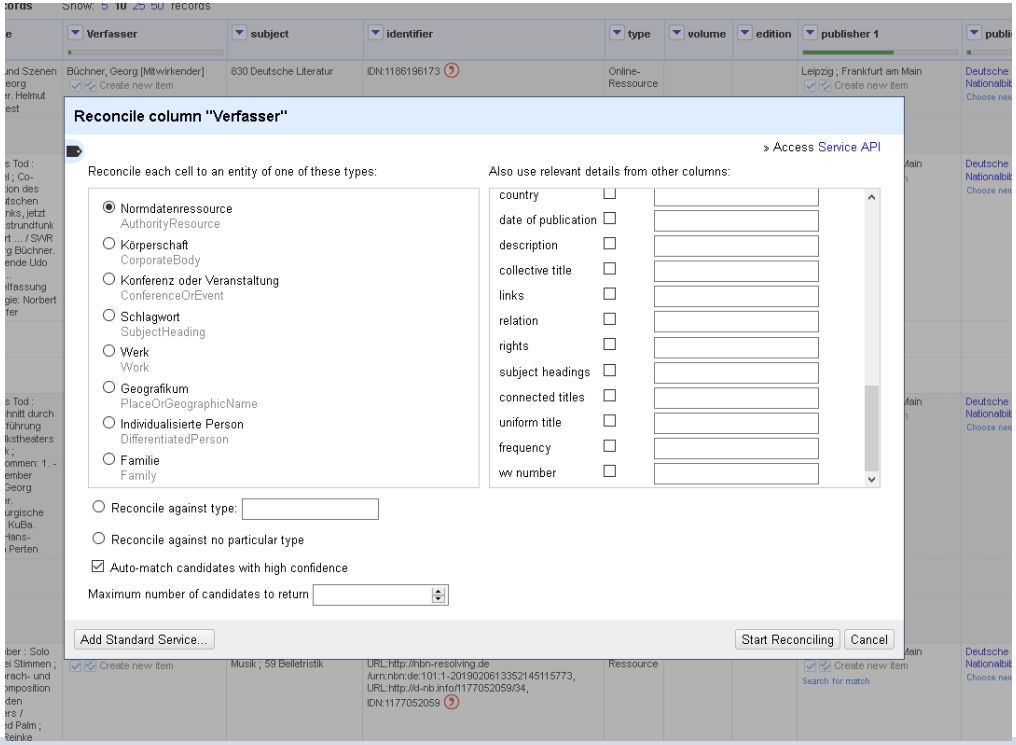
Clustering

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

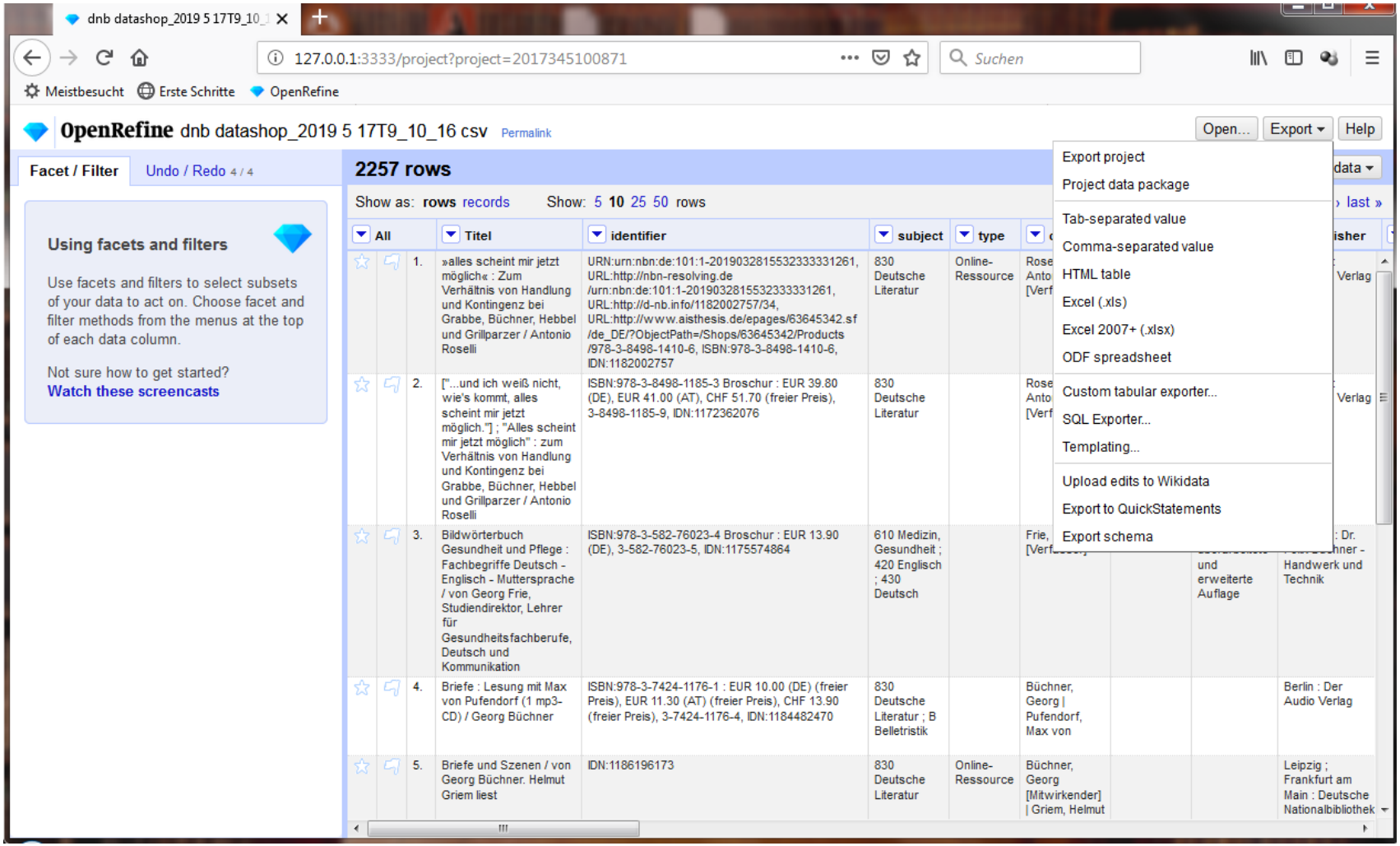


Reconciliation

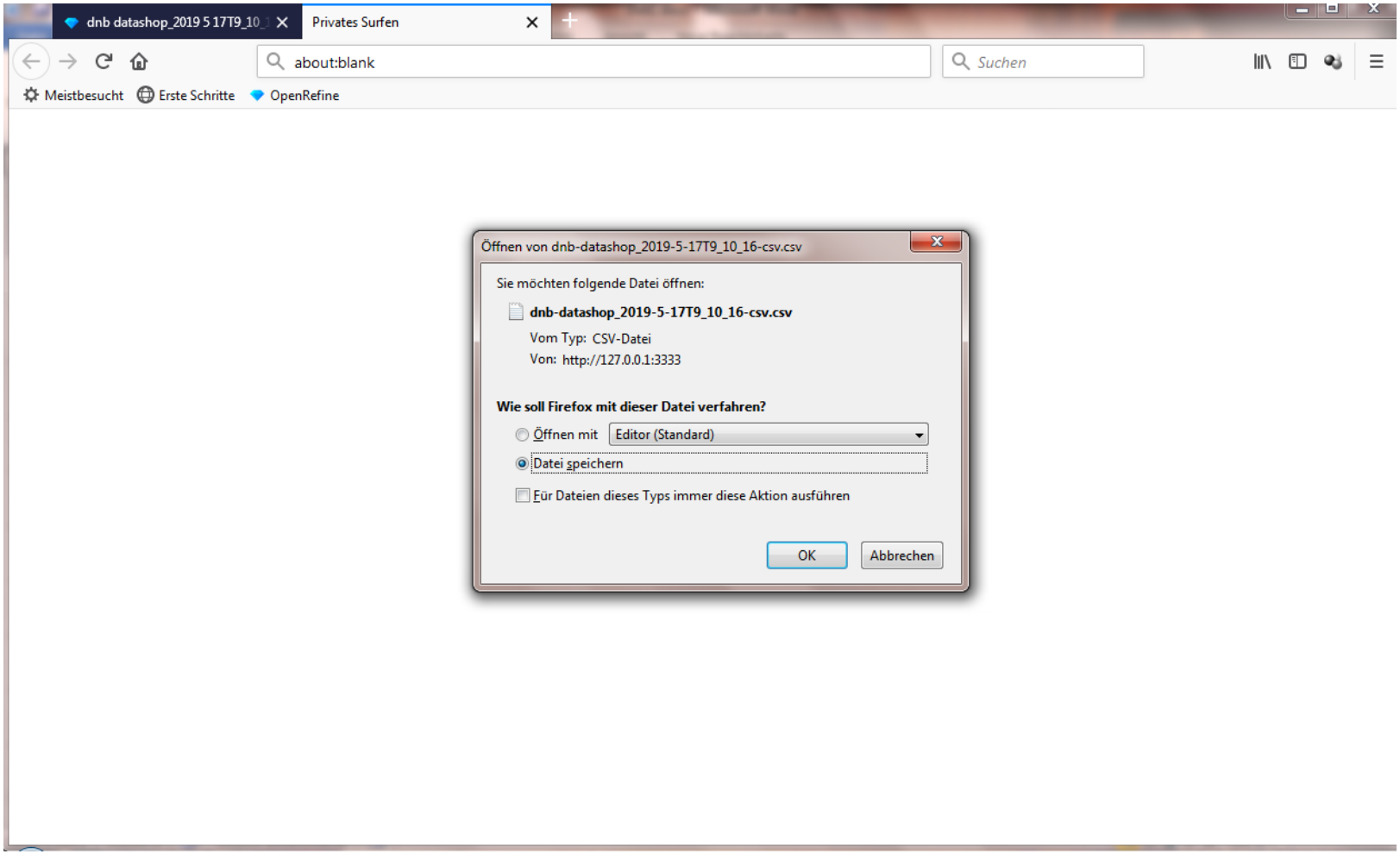
- Z.B. über Reconciliation-Webservice für die GND: https://lobid.org/gnd/reconcile
- Erlaubt den Abgleich und das Aufwerten der eigenen Daten mit den GND-Informationen



Daten speichern / exportieren



Daten speichern / exportieren



Vorbereitung und Hinweise für die Hands-on-Übung

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- Laden Sie bitte OpenRefine (3.4.1) 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 (bitte installieren, falls auf Ihrem Gerät noch nicht vorhanden: http://java.com/ bzw. Open Source Variante:OpenJDK https://github.com/ojdkbuild/ojdkbuild

- 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

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Blogs

https://histhub.ch/cat/net/blog/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 zur Verfügungstellung von Materialien und Daten an Jakob Frohmann (Universitätsbibliothek Johann Christian Senckenberg)

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