

# Datenbereinigung, -transformation und -analyse mit



# OpenRefine

17.06.2025  
18:00-19:30 Uhr

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

# 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

## OpenRefine – Was ist das?

---

- Open Source-Software in Java  
(zuvor zwischenzeitlich zu Google gehörig unter dem Namen Google Refine)
- läuft lokal auf dem Rechner, aber im Browser (keine Internetverbindung notwendig)
- 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)

## Anwendungsmöglichkeiten – 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 – 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 (Vorschlag)

---

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

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

## Daten importieren/exportieren in OpenRefine

---

Mögliche Dateiformate für den Import und Export (Auswahl):

CSV, Microsoft Excel, JSON, XML, HTML

## Beispiel für ein Importformat

```

dnb-datashop_2019-5-17T9_10_16.csv x
1 "subject";"identifizier";"type";"creator";"title";"volume";"edition";"publisher";"year";"format";"ISBN
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.html>

### PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

**2013-02-27**

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13  
 20130227 2013.02.27 27.02.13 27-02-13  
 27.2.13 2013. II. 27. 27<sup>1</sup>/<sub>2</sub>-13 2013.158904109  
 MMXIII-II-XXVII MMXIII <sup>LXVII</sup><sub>CCLXV</sub> 1330300800  
 ((3+3)×(111+1)-1)×3/3-1/3<sup>3</sup> 2013  
 10/11011/1101 02/27/20/13 0 1 2 3 4 5 6 7 8 

<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.html>

## Bearbeitungsbeispiele: Segmentieren

Address in single field	Institution	Library name	Address		Town/City	Region	Country	Postcode
			Address 1	Address 2				
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 Aberdeen	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.html>

# Ein Projekt erstellen...

← → ↻ 🏠 127.0.0.1:3333 ... 🛡️ ☆

⚙️ Meistbesucht 🌈 Erste Schritte

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

Create Project  
Open Project  
Import Project  
Language Settings

**Create a project by importing data. What kinds of data files can I import?**  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

**This Computer**  Keine Dateien ausgewählt.

Web Addresses (URLs)  
Clipboard  
Data Package (JSON URL)  
Database  
Google Data

  
Version 3.1 [b90e413]  
Preferences  
Help  
About

# Ein Projekt erstellen...

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

« Start Over    Configure Parsing Options    Project name     Tags     **Create Project »**

	subject	identifier	type	creator	title	volume	edition	publist
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 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]	Bielefeld 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

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

# Ein Projekt erstellen...

 **OpenRefine** dnb datashop\_2019 5 17T9\_10\_16 csv [Permalink](#)

Open... Export Help

Facet / Filter Undo / Redo 0 / 0

1716 records

Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 records

« first « previous 1 - 10 next » last »

## 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](#)

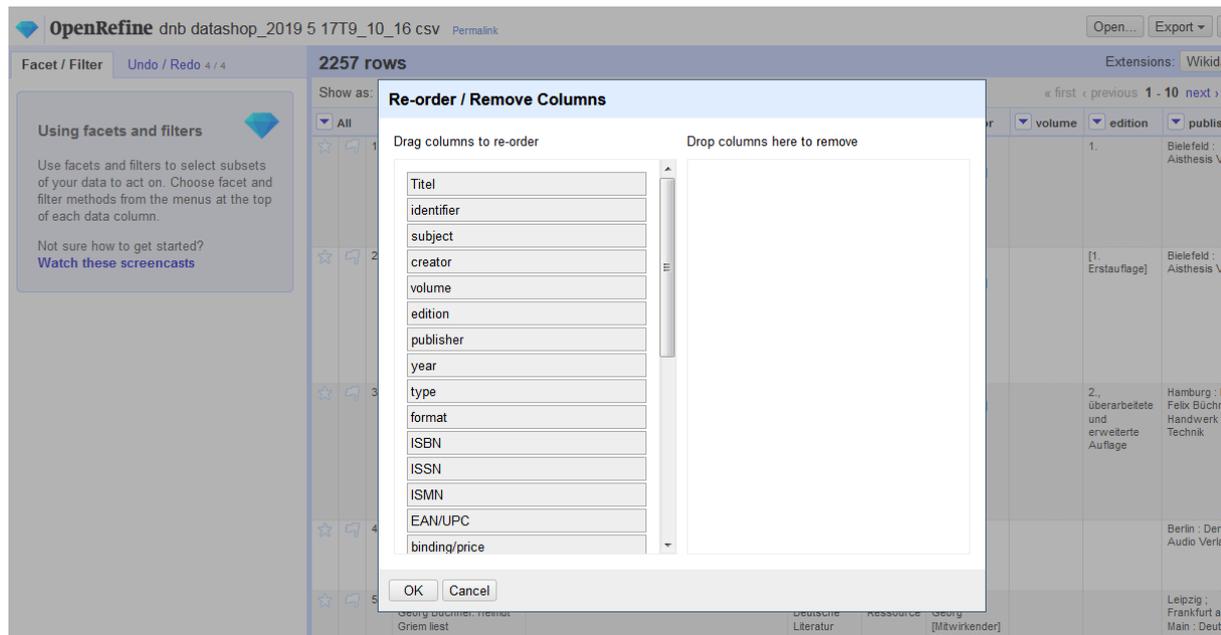
All	subject	identifier	type	creator	title	volume	edition	publisher	year
☆	1. Deutsche Literatur	URN:urn:nbn:de:101:1-201903281553233331261, URL:http://nbn-resolving.de /urn:nbn:de:101:1-201903281553233331261, URL:http://d-nb.info/1182002757/34, URL:http://www.aisthesis.de/epages/63645342.sf /de_DE/?ObjectPath=/Shops/63645342/Products /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	201
☆	2. 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	201
☆	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	201
☆	4. Deutsche Literatur ; 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	201
☆	5. 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	201

## Mit den Daten arbeiten...

### Layout

#### Neuordnen von Spalten:

- Dropdownmenü bei "All":  
Edit columns → Re-order / Remove Columns



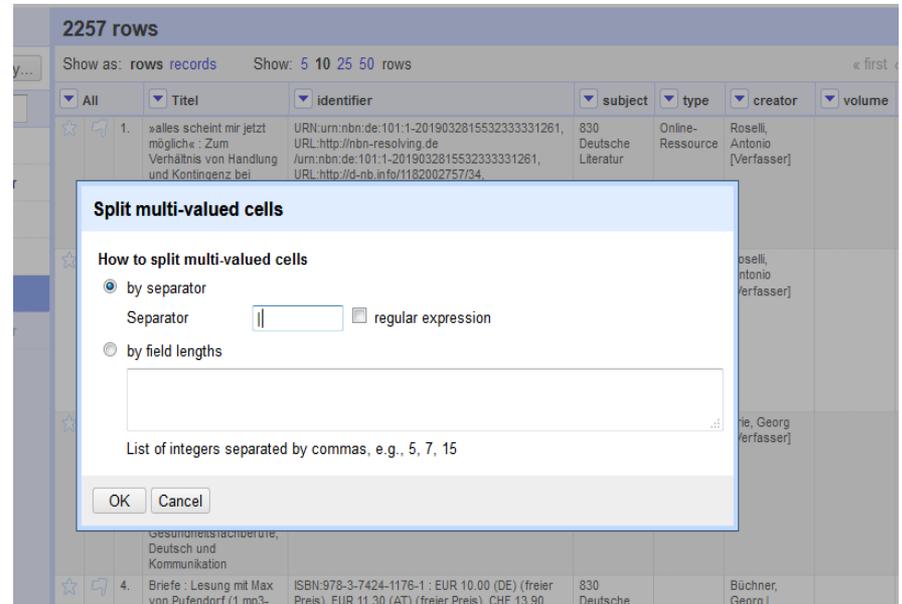
## Mit den Daten arbeiten...

### 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



2257 rows

Show as: rows records Show: 5 10 25 50 rows

All	Titel	identifier	subject	type	creator	volume
1.	»alles scheint mir jetzt mögliche : Zum Verhältnis von Handlung und Kontinenz bei	URN:urn:nbn:de:101:1-201903281553233331261, URL:http://nbn-resolving.de /urn:nbn:de:101:1-201903281553233331261, URL:http://d-nb.info/1182002757/34	830 Deutsche Literatur	Online-Ressource	Roselli, Antonio [Verfasser]	
4.	Briefe : Lesung mit Max von Pufendorf (1 mp3-	ISBN-978-3-7424-1176-1 : EUR 10.00 (DE) (freier Preis), EUR 11.30 (AT) (freier Preis), CHF 13.90	830 Deutsche		Büchner, Georg [Verfasser]	

**Split multi-valued cells**

How to split multi-valued cells

by separator

Separator   regular expression

by field lengths

List of integers separated by commas, e.g., 5, 7, 15

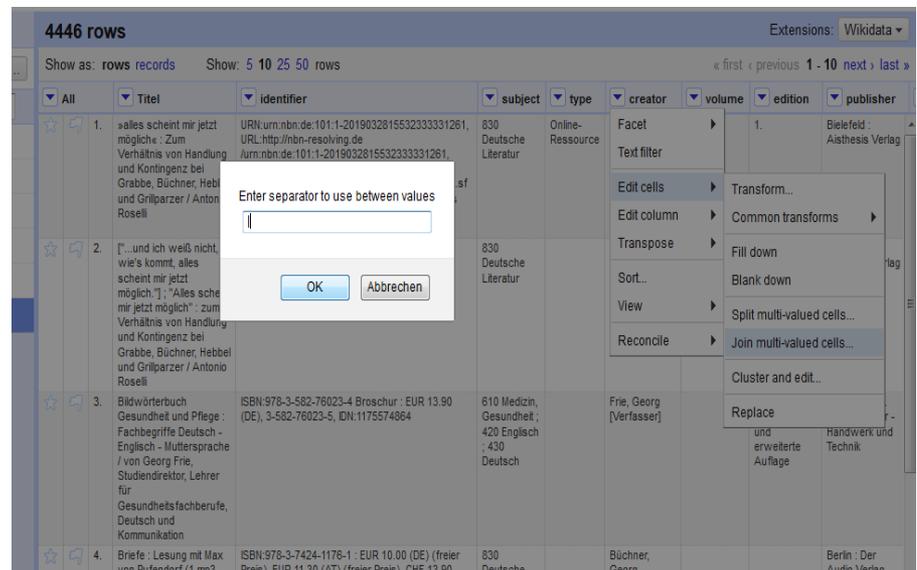
OK Cancel

# Mit den Daten arbeiten...

## 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. „|“)



The screenshot displays a data management interface with a table containing 4446 rows. A context menu is open over a cell, and the 'Join multi-valued cells...' option is selected. A dialog box is shown, asking for a separator to use between values, with the character '|' entered in the input field. The table columns include 'All', 'Titel', 'identifier', 'subject', 'type', 'creator', 'volume', 'edition', and 'publisher'. The first row shows a title about 'Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Anton Roselli' with a URL in the identifier column. The second row shows a title about '„...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'. The third row shows a title 'Bildwörterbuch Gesundheit und Pflege: Fachbegriffe Deutsch - Englisch - Muttersprache / von Georg Frie, Studiendirektor, Lehrer für Gesundheitsfachberufe, Deutsch und Kommunikation'. The fourth row shows a title 'Briefe: Lesung mit Max von Pufendorf (1.mq3-'. The dialog box has 'OK' and 'Abbrechen' buttons.

# Mit den Daten arbeiten...

## Layout

### Umbenennen von Spalten:

- Edit column → Rename this column

2257 rows										Extensions: Wikid	
Show as: rows records		Show: 5 10 25 50 rows			« first < previous 1 - 10 next »						
All	Titel	identifizier	subject	type	creator	volume	edition	publis			
☆	1.	»alles scheint mir jetzt möglich« : Zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli	URN:urn:nbn:de:101:1-201903281553233331261, URL:http://nbn-resolving.de/urn:nbn:de:101:1-201903281553233331261, URL:http://id-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, ISBN-978-1182002757	830 Deutsche Literatur	Online-Ressource	Facet	1.	Bielefeld : Aisthesis V			
☆	2.	[...] 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	ISBN-978-3-8498-1185-3 Bros (DE), EUR 41.00 (AT), CHF 51.3-8498-1185-9, IDN:11723620				[1. Erstauflage]	Bielefeld : Aisthesis V			
☆	3.	Bildwörterbuch Gesundheit und Pflege : Fachbegriffe Deutsch - Englisch - Muttersprache / von Georg Frie, Studiendirektor, Lehrer für Gesundheitsfachberufe, Deutsch und Kommunikation	ISBN-978-3-582-76023-4 Bros (DE), 3-582-76023-5, IDN:11723620			Frie, Georg [Verfasser]	2., überarbeitete und erweiterte Auflage	Hamburg : Felix Büchrr Handwerk Technik			
☆	4.	Briefe : Lesung mit Max von Pufendorf (1 mp3-CD) / Georg Büchner	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	830 Deutsche Literatur ; Belletristik		Büchner, Georg   Pufendorf, Max von		Berlin : Der Audio Verle			

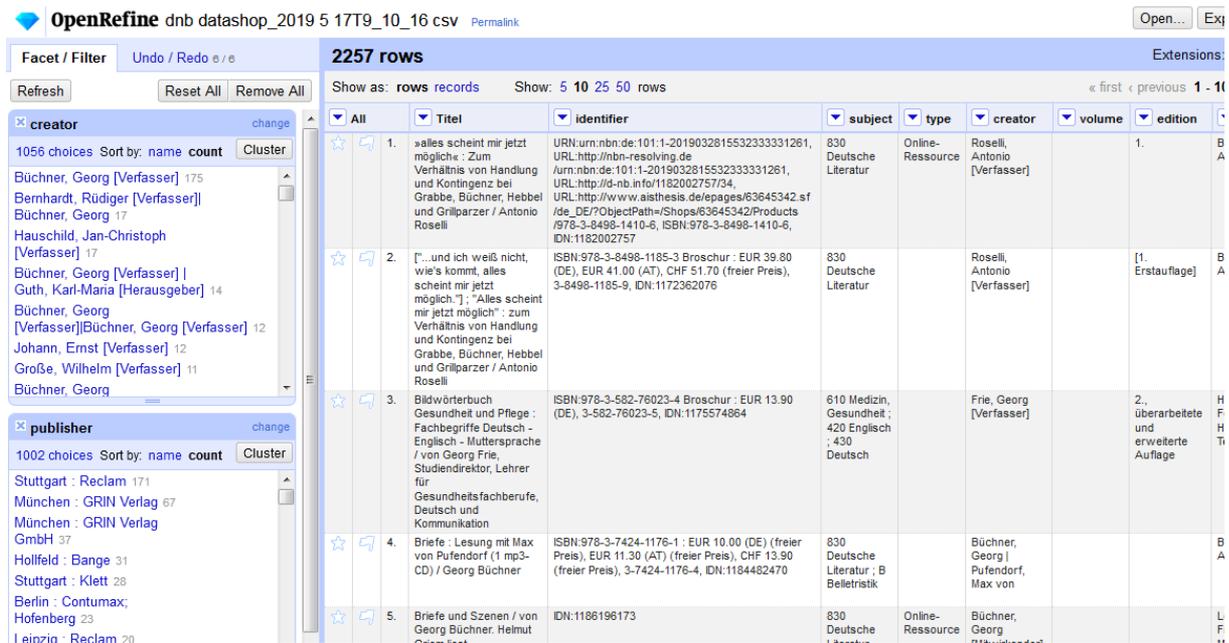
# Mit den Daten arbeiten...

## 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



**OpenRefine** dnb datashop\_2019.5.17T9\_10\_16.csv Permalink

Facet / Filter Undo / Redo 8 / 8

2257 rows Extensions:

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 10

	All	Titel	identifizier	subject	type	creator	volume	edition
1.	»alles scheint mir jetzt möglich«. ; Zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli	URN:nbn:de:101:1-201903281553233331261, URL:http://nbn-resolving.org/urn:nbn:de:101:1-201903281553233331261, 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	830 Deutsche Literatur	Online-Ressource	Roselli, Antonio [Verfasser]		1.	B A
2.	[...]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	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	830 Deutsche Literatur		Roselli, Antonio [Verfasser]		[1. Erstauflage]	B A
3.	Bildwörterbuch Gesundheit und Pflege : Fachbegriffe Deutsch - Englisch - Muttersprache / von Georg Frie, Studiendirektor, Lehrer für Gesundheitsfachberufe, Deutsch und Kommunikation	ISBN-978-3-582-76023-4 Broschur : EUR 13,90 (DE), 3-582-76023-5, IDN:1175574864	610 Medizin, Gesundheit ; 420 Englisch ; 430 Deutsch		Frie, Georg [Verfasser]		2., überarbeitete und erweiterte Auflage	H F T
4.	Briefe : Lesung mit Max von Pufendorf (1 mp3-CD) / Georg Büchner	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	830 Deutsche Literatur ; B Belletristik		Büchner, Georg   Pufendorf, Max von			B A
5.	Briefe und Szenen / von Georg Büchner, Helmut	IDN:1186196173	830 Deutsche Literatur	Online-Ressource	Büchner, Georg			L F

**creator** 1056 choices Sort by: name count Cluster

- Büchner, Georg [Verfasser] 175
- Bernhardt, Rüdiger [Verfasser]
- Büchner, Georg 17
- Hauschild, Jan-Christoph [Verfasser] 17
- Büchner, Georg [Verfasser] | Guth, Karl-Maria [Herausgeber] 14
- Büchner, Georg [Verfasser]|Büchner, Georg [Verfasser] 12
- Johann, Ernst [Verfasser] 12
- Große, Wilhelm [Verfasser] 11
- Büchner, Georg

**publisher** 1002 choices Sort by: name count Cluster

- Stuttgart : Reclam 171
- München : GRIN Verlag 67
- München : GRIN Verlag GmbH 37
- Hollfeld : Bange 31
- Stuttgart : Klett 28
- Berlin : Contumax; Hofenberg 23
- Leinzio : Reclam 20

## Mit den Daten arbeiten...

---

### Facetten und Filter

#### Customized Facets (Auswahl):

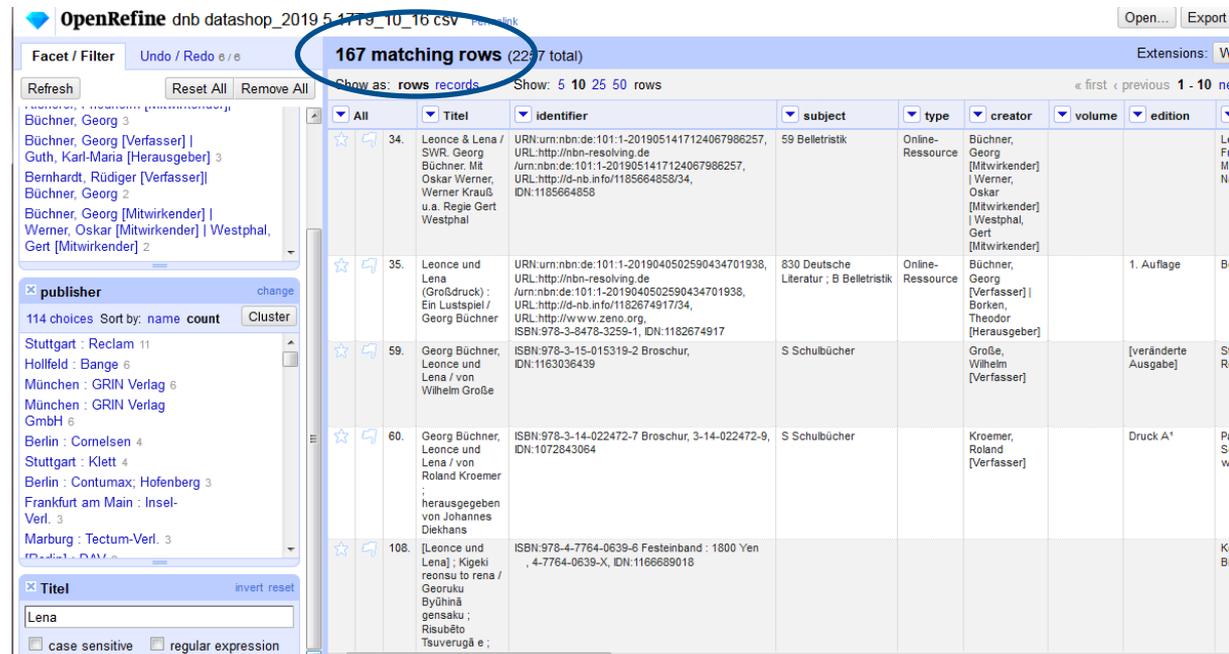
- **Word Facet:** segmentiert Text in Wörter und verzeichnet ihre Häufigkeit
- **Text length facet:** wertet die Länge des Texts aus und visualisiert diese
- **Facet by Blank:** Filtert leere Zellen

# Mit den Daten arbeiten...

## Facetten und Filter

### Text Filter:

- filtert Werte mit übereinstimmendem String



**OpenRefine** dnb datashop\_2019\_5\_17\_19\_10\_16.csv

Facet / Filter Undo / Redo 0 / 0 **167 matching rows** (2297 total) Extensions: W

Refresh Reset All Remove All Show as: rows records Show: 5 10 25 50 rows

All	Titel	identifizier	subject	type	creator	volume	edition
34.	Leonce & Lena / SWR, Georg Büchner, Mit Oskar Werner, Werner Krauß u.a. Regie Gert Westphal	URN:nbn:de:101:1-2019051417124067986257, URL:http://nbn-resolving.de /urn:nbn:de:101:1-2019051417124067986257, URL:http://d-nb.info/1185664858/34, IDN:1185664858	59 Belletristik	Online-Ressource	Büchner, Georg [Mitwirkender]   Werner, Oskar [Mitwirkender]   Westphal, Gert [Mitwirkender]		
35.	Leonce und Lena (Großdruck) : Ein Lustspiel / Georg Büchner	URN:nbn:de:101:1-2019040502590434701938, URL:http://nbn-resolving.de /urn:nbn:de:101:1-2019040502590434701938, URL:http://d-nb.info/1182674917/34, ISBN:978-3-8478-3259-1, IDN:1182674917	830 Deutsche Literatur ; B Belletristik	Online-Ressource	Büchner, Georg [Verfasser]   Borken, Theodor [Herausgeber]		1. Auflage
59.	Georg Büchner, Leonce und Lena / von Wilhelm Große	ISBN:978-3-15-015319-2 Broschur, IDN:1163036439	S Schulbücher		Große, Wilhelm [Verfasser]		[veränderte Ausgabe]
60.	Georg Büchner, Leonce und Lena / von Roland Kroemer ; herausgegeben von Johannes Diekhans	ISBN:978-3-14-022472-7 Broschur, 3-14-022472-9, IDN:1072843064	S Schulbücher		Kroemer, Roland [Verfasser]		Druck A*
108.	[Leonce und Lena] ; Kigeki reonsu to rena / Georuku Byūhinā gensaku ; Risubēto Tsuverugā e ;	ISBN:978-4-7764-0639-6 Festeinband : 1800 Yen , 4-7764-0639-X, IDN:1186689018					

**publisher** change 114 choices Sort by: name count Cluster

- Stuttgart : Reclam 11
- Hollfeld : Bange 6
- München : GRIN Verlag 6
- München : GRIN Verlag GmbH 6
- Berlin : Cornelsen 4
- Stuttgart : Klett 4
- Berlin : Contumax; Hofenberg 3
- Frankfurt am Main : Insel-Verl. 3
- Marburg : Tectum-Verl. 3

**Titel** invert reset

Lena

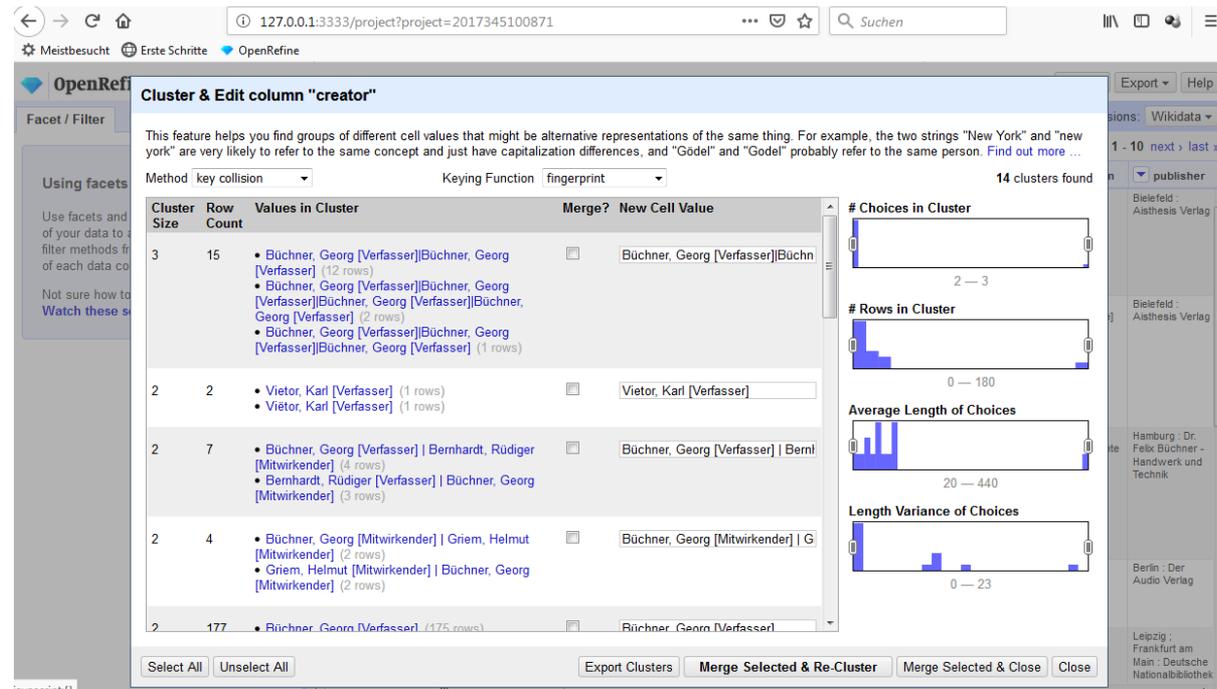
case sensitive  regular expression

# Mit den Daten arbeiten...

## 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



**Cluster & Edit column "creator"**

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Godel" and "Godel" probably refer to the same person. [Find out more ...](#)

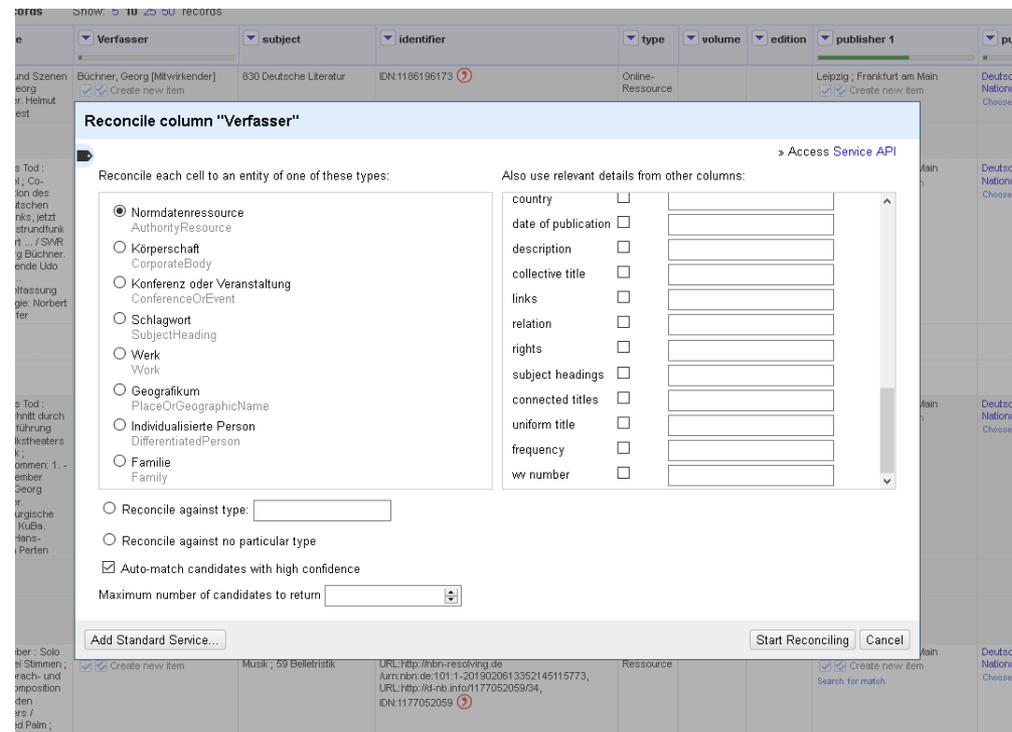
Method: key collision    Keying Function: fingerprint    14 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	15	<ul style="list-style-type: none"> <li>Büchner, Georg [Verfasser] Büchner, Georg [Verfasser] (12 rows)</li> <li>Büchner, Georg [Verfasser] Büchner, Georg [Verfasser] Büchner, Georg [Verfasser] Büchner, Georg [Verfasser] (2 rows)</li> <li>Büchner, Georg [Verfasser] Büchner, Georg [Verfasser] (1 rows)</li> </ul>	<input type="checkbox"/>	Büchner, Georg [Verfasser] Büchn
2	2	<ul style="list-style-type: none"> <li>Viotor, Karl [Verfasser] (1 rows)</li> <li>Viotor, Karl [Verfasser] (1 rows)</li> </ul>	<input type="checkbox"/>	Viotor, Karl [Verfasser]
2	7	<ul style="list-style-type: none"> <li>Büchner, Georg [Verfasser]   Bernhardt, Rüdiger [Mitwirkender] (4 rows)</li> <li>Bernhardt, Rüdiger [Verfasser]   Büchner, Georg [Mitwirkender] (3 rows)</li> </ul>	<input type="checkbox"/>	Büchner, Georg [Verfasser]   Bernl
2	4	<ul style="list-style-type: none"> <li>Büchner, Georg [Mitwirkender]   Griem, Helmut [Mitwirkender] (2 rows)</li> <li>Griem, Helmut [Mitwirkender]   Büchner, Georg [Mitwirkender] (2 rows)</li> </ul>	<input type="checkbox"/>	Büchner, Georg [Mitwirkender]   G
2	177	<ul style="list-style-type: none"> <li>Büchner, Georg [Verfasser] (175 rows)</li> </ul>	<input type="checkbox"/>	Büchner, Georg [Verfasser]

# Mit den Daten arbeiten...

## Reconciliation

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



The screenshot shows a data table with columns for 'Verfasser', 'subject', 'identifizier', 'type', 'volume', 'edition', and 'publisher 1'. A dialog box titled 'Reconcile column "Verfasser"' is open, allowing users to map data from the 'Verfasser' column to various entity types. The dialog includes a list of entity types with radio buttons, a section for 'Also use relevant details from other columns' with checkboxes and input fields, and options for reconciliation settings like 'Auto-match candidates with high confidence' and 'Maximum number of candidates to return'.

**Reconcile column "Verfasser"**

Reconcile each cell to an entity of one of these types:

- Normdatenressource  
AuthorityResource
- Körperschaft  
CorporateBody
- Konferenz oder Veranstaltung  
ConferenceOrEvent
- Schlagwort  
SubjectHeading
- Werk  
Work
- Geografikum  
PlaceOrGeographicName
- Individualisierte Person  
DifferentiatedPerson
- Familie  
Family

Also use relevant details from other columns:

- country
- date of publication
- description
- collective title
- links
- relation
- rights
- subject headings
- connected titles
- uniform title
- frequency
- ww number

Reconcile against type:   
 Reconcile against no particular type  
 Auto-match candidates with high confidence  
 Maximum number of candidates to return:

# Daten speichern / exportieren

dnb datashop\_2019 5 17T9\_10\_16 csv

127.0.0.1:3333/project?project=2017345100871

Suchen

Meistbesucht Erste Schritte OpenRefine

OpenRefine dnb datashop\_2019 5 17T9\_10\_16 csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 4 / 4

2257 rows

Show as: rows records Show: 5 10 25 50 rows

**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](#)

All	Titel	identifier	subject	type	author	publisher
☆	1. »alles scheint mir jetzt möglich« : Zum Verhältnis von Handlung und Kontingenz bei Grabbe, Büchner, Hebbel und Grillparzer / Antonio Roselli	URN:urn:nbn:de:101:1-201903281553233331261, URL:http://nbn-resolving.de/urn:nbn:de:101:1-201903281553233331261, 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	830 Deutsche Literatur	Online-Ressource	Rose Antonio [Verf.]	Verlag
☆	2. [...]...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	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	830 Deutsche Literatur		Rose Antonio [Verf.]	Verlag
☆	3. Bildwörterbuch Gesundheit und Pflege : Fachbegriffe Deutsch - Englisch - Muttersprache / von Georg Frie, Studiendirektor, Lehrer für Gesundheitsfachberufe, Deutsch und Kommunikation	ISBN:978-3-582-76023-4 Broschur : EUR 13.90 (DE), 3-582-76023-5, IDN:1175574864	610 Medizin, Gesundheit ; 420 Englisch ; 430 Deutsch		Frie, Georg [Verf.]	Handwerk und Technik
☆	4. Briefe : Lesung mit Max von Pufendorf (1 mp3-CD) / Georg Büchner	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	830 Deutsche Literatur ; B Belletristik		Büchner, Georg   Pufendorf, Max von	Berlin : Der Audio Verlag
☆	5. Briefe und Szenen / von Georg Büchner. Helmut Griem liest	IDN:1186196173	830 Deutsche Literatur	Online-Ressource	Büchner, Georg [Mitwirkender]   Griem, Helmut	Leipzig ; Frankfurt am Main : Deutsche Nationalbibliothek

Export project

Project data package

Tab-separated value

Comma-separated value

HTML table

Excel (.xls)

Excel 2007+ (.xlsx)

ODF spreadsheet

Custom tabular exporter...

SQL Exporter...

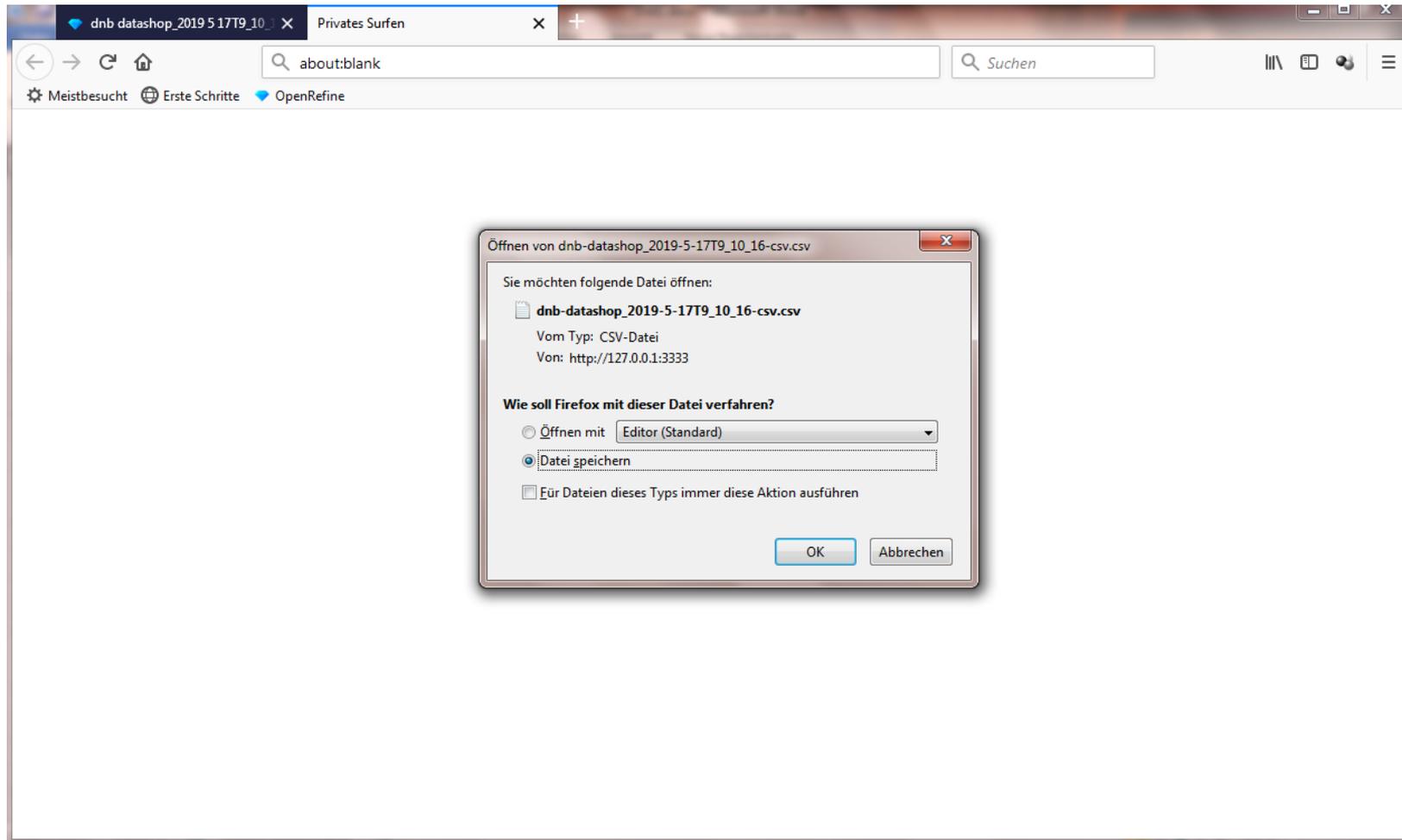
Templating...

Upload edits to Wikidata

Export to QuickStatements

Export schema

## Daten speichern / exportieren



# Hands-on-Übungen mit Praxisbeispielen

Grundfunktionen von OpenRefine selbst ausprobieren, z.B.:

- einen Datensatz erkunden und „aufräumen“
- Zellen teilen und (wieder) vereinigen
- Facetieren / Filtern / Clustern von Daten usw.

Anreichern eigener Daten aus externen Quellen

Data you have	Date of Birth from VIAF (Virtual International Authority File)	Date of Death from VIAF (Virtual International Authority File)
Braddon, M. E. (Mary Elizabeth)	1835	1915
Rossetti, William Michael	1829	1919
Prest, Thomas Peckett	1810	1879

## Hinweise zum Selbst-Ausprobieren

---

- 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>

## Einfache Übungen ...

---

- 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](#))

# Danke für Ihre Aufmerksamkeit!

---

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

- <http://blog.lobid.org/2018/08/27/openrefine.html>  
(einfache Anreicherung von Daten in OpenRefine mit [Personen-] Daten aus der GND via lobid.org)
- *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](#)]