OpenRefine

Kathi Woitas, Digital Scholarship Services
Universitätsbibliothek Bern

DOI: 10.5281/zenodo.5776098

Thadguidry, CC0, via Wikimedia Commons

Ziele – Ablauf

Teil 1 – Vorstellung des Tools

Dauer: ca. 1h

Show & Tell

Lernziele:

Die Teilnehmenden kennen den **Funktionsumfang** von OpenRefine und können über einen **möglichen Einsatz** entscheiden.

Teil 2 – Hands-On-Workshop

Dauer: ca. 3h (inkl. Pause)

Lernziele:

Teilnehmende wissen, wie Daten **laden, sortieren, filtern, bereinigen**, **transformieren** und **exportieren**. Sie können zudem Daten mit externen Reconciliation-Services **anreichern** und in der **Bearbeitungshistorie** vor und zurück navigieren.

UNIVERSITÄT
BERN

OpenRefine

Teil 1: Vorstellung des Tools

Kathi Woitas, Digital Scholarship Services
Universitätsbibliothek Bern

Was ist OpenRefine?

- als Google Refine entwickelt und unterstützt von Google 2010-13
- danach: ***OpenRefine***
- **Open-Source**-Software (BSD 3), verfügbar auf Github
- läuft **lokal** auf Computer (ebenso eigene Daten)
- Oberfläche **im Browser:** 127.0.0.1:3333 oder localhost:3333

- **Datenbearbeitungstool**: Daten einlesen, manipulieren bzw. bereinigen, anreichern, exportieren
- beherrscht **viele Datenformate** für In- und Output
- viel **per Menüs/Klicks** machbar, aber auch Code einsetzbar für komplexe Aufgaben

Was bietet OpenRefine?

- **Excel:** Datenbereinigung, Transformationen, Syntax relativ schwierig
- **Data Processing** mit R, Python: Code-Kenntnisse nötig
- **OpenRefine** als “goldener Mittelweg” bez. Funktionalität und Einfachheit

autom. Bearbeitungshistorie:

- einfaches Rückgängigmachen jeden Schrittes: exploratives Vorgehen möglich
- autom. Aufzeichnung in JSON-Notation: zum Wiederholen auf neue Daten und/oder Teilen
- internes Speicherformat „OpenRefine-Projekt“ enthält auch Historie

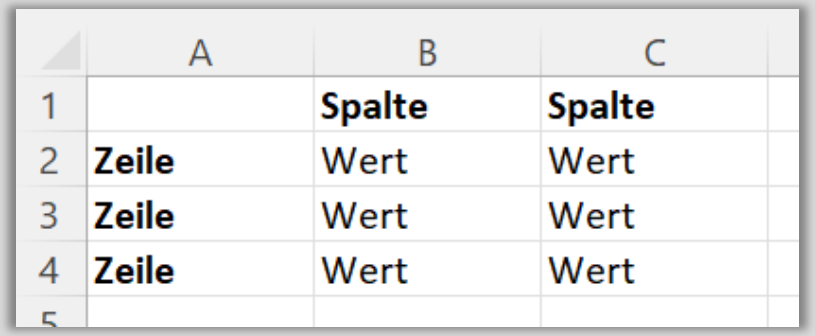
Was kann ich damit machen? Daten laden

UNIVERSITÄT BERN

- Vielzahl von Datenformaten:

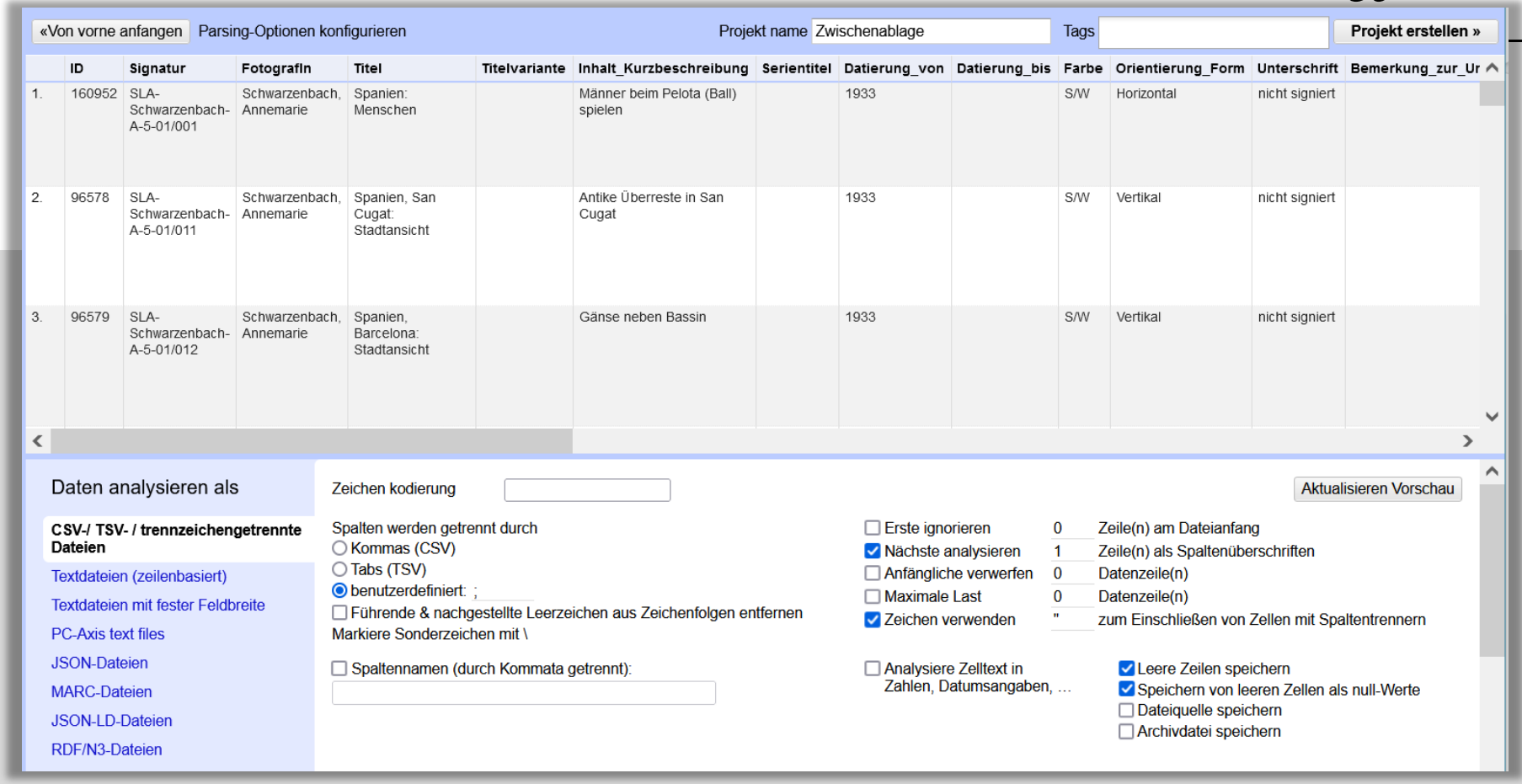
TSV, CSV, *SV, Excel, XML, RDF (in XML), JSON, Google Data Dokumente

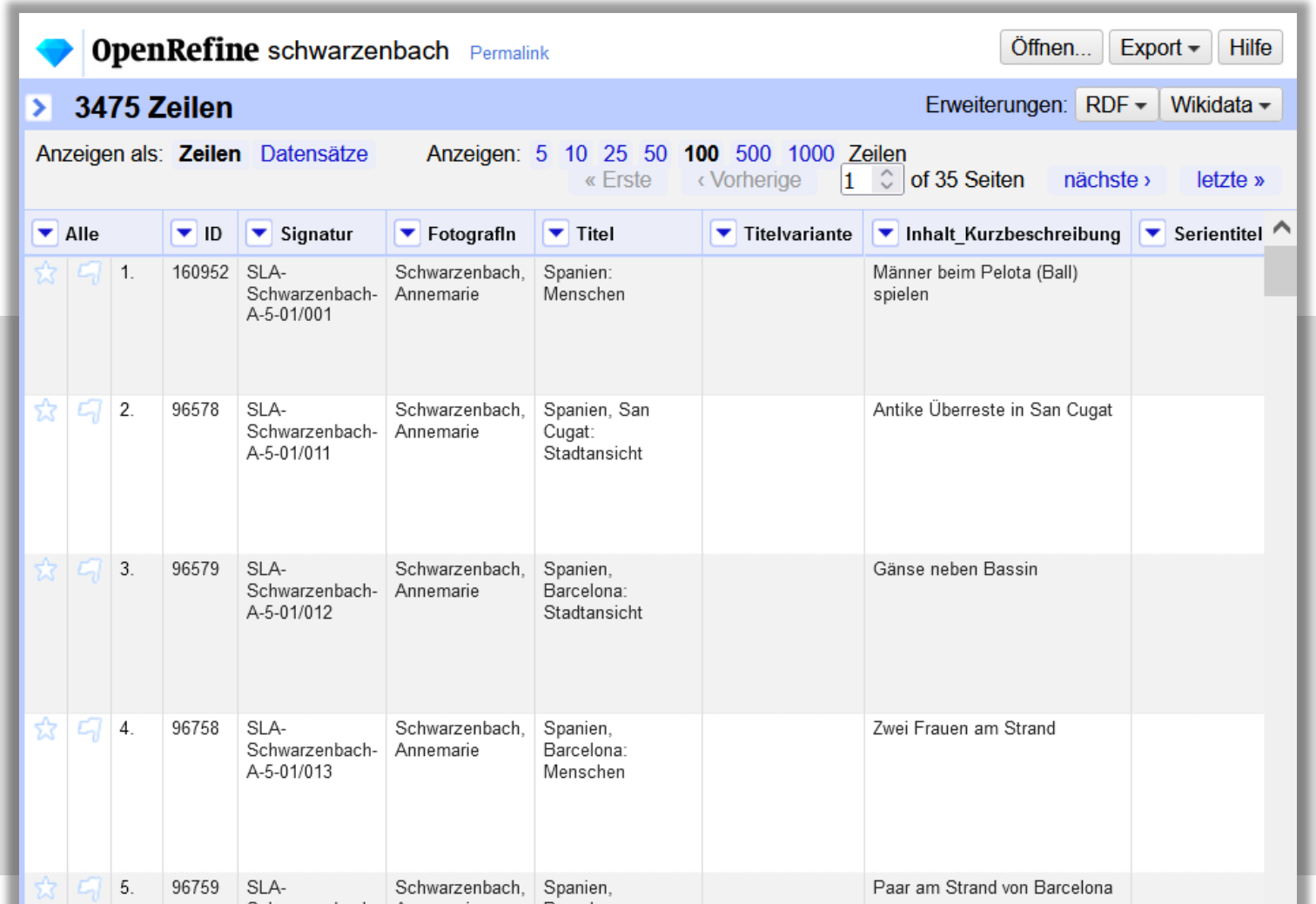
- von URL, per Copy&Paste
- diverse **Optionen** zum Datenimport **mit Vorschau**



	A	B	C
1		**Spalte**	**Spalte**
2	**Zeile**	Wert	Wert
3	**Zeile**	Wert	Wert
4	**Zeile**	Wert	Wert
5			

Wie sieht das aus? localhost:3333/


«Von vorne anfangen
Parsing-Optionen konfigurieren
Projekt name Zwischenablage
Tags
Projekt erstellen »
ID Signatur Fotografin Titel Titelvariante Inhalt_Kurzbeschreibung Serientitel Datierung_von Datierung_bis Farbe Orientierung_Form Unterschrift Bemerkung_zur_Ur
1. 160952 SLA-Schwarzenbach-A-5-01/001 Schwarzenbach, Annemarie Spanien: Menschen Männer beim Pelota (Ball) spielen 1933 S/W Horizontal nicht signiert
2. 96578 SLA-Schwarzenbach-A-5-01/011 Schwarzenbach, Annemarie Spanien, San Cugat: Stadtansicht Antike Überreste in San Cugat 1933 S/W Vertikal nicht signiert
3. 96579 SLA-Schwarzenbach-A-5-01/012 Schwarzenbach, Annemarie Spanien, Barcelona: Stadtansicht Gänse neben Bassin 1933 S/W Vertikal nicht signiert
Daten analysieren als
CSV-/ TSV- / trennzeichengetrennte Dateien
Textdateien (zeilenbasiert)
Textdateien mit fester Feldbreite
PC-Axis text files
JSON-Dateien
MARC-Dateien
JSON-LD-Dateien
RDF/N3-Dateien
Zeichen kodierung
Aktualisieren Vorschau
Spalten werden getrennt durch
Kommas (CSV)
Tabs (TSV)
benutzerdefiniert: ;
Führende & nachgestellte Leerzeichen aus Zeichenfolgen entfernen
Markiere Sonderzeichen mit \
Spaltennamen (durch Kommata getrennt):
Erste ignorieren 0 Zeile(n) am Dateianfang
Nächste analysieren 1 Zeile(n) als Spaltenüberschriften
Anfängliche verwerfen 0 Datenzeile(n)
Maximale Last 0 Datenzeile(n)
Zeichen verwenden " zum Einschließen von Zellen mit Spaltentrennern
Analysiere Zelltext in Zahlen, Datumsangaben, …
Leere Zeilen speichern
Speichern von leeren Zellen als null-Werte
Dateiquelle speichern
Archivdatei speichern



OpenRefine schwarzenbach Permalink

Öffnen... Export Hilfe

3475 Zeilen

Erweiterungen: RDF Wikidata

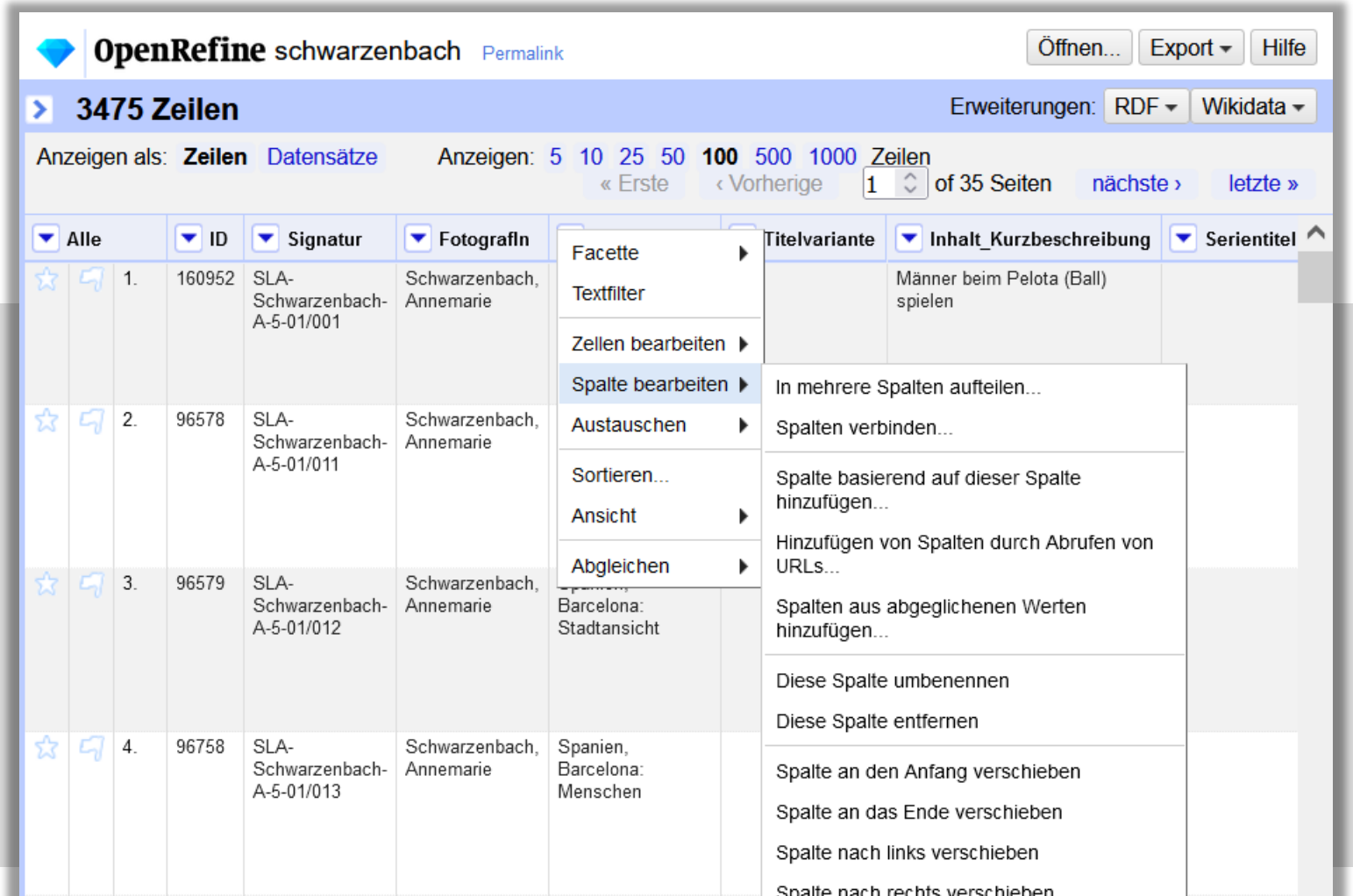
Anzeigen als: Zeilen Datensätze

Anzeigen: 5 10 25 50 100 500 1000 Zeilen

« Erste ‹ Vorherige 1 of 35 Seiten nächste › letzte »

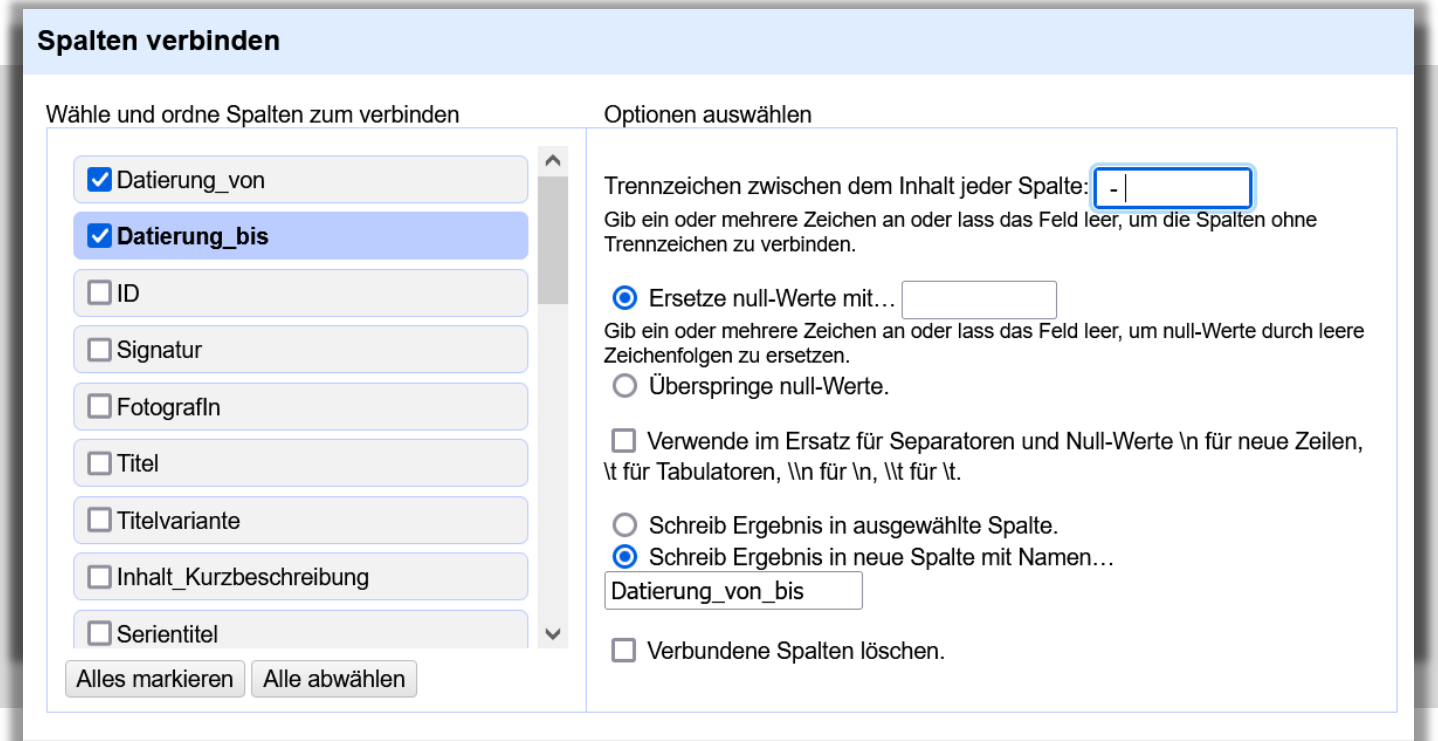
Alle		ID	Signatur	Fotografln	Titel	Titelvariante	Inhalt_Kurzbeschreibung	Serientitel
1.	160952	SLA-Schwarzenbach-A-5-01/001	Schwarzenbach, Annemarie	Spanien: Menschen		Männer beim Pelota (Ball) spielen	
2.	96578	SLA-Schwarzenbach-A-5-01/011	Schwarzenbach, Annemarie	Spanien, San Cugat: Stadtansicht		Antike Überreste in San Cugat	
3.	96579	SLA-Schwarzenbach-A-5-01/012	Schwarzenbach, Annemarie	Spanien, Barcelona: Stadtansicht		Gänse neben Bassin	
4.	96758	SLA-Schwarzenbach-A-5-01/013	Schwarzenbach, Annemarie	Spanien, Barcelona: Menschen		Zwei Frauen am Strand	
5.	96759	SLA-	Schwarzenbach,	Spanien,		Paar am Strand von Barcelona	

UNIVERSITÄT BERN



UNIVERSITÄT BERN

Was kann ich damit machen?
Spalten aufteilen und verbinden



Was kann ich damit machen?
Facetten vs. Filter

Facetten

automatische Facetten nach **allen Ausprägungen der Werte** in der Spalte

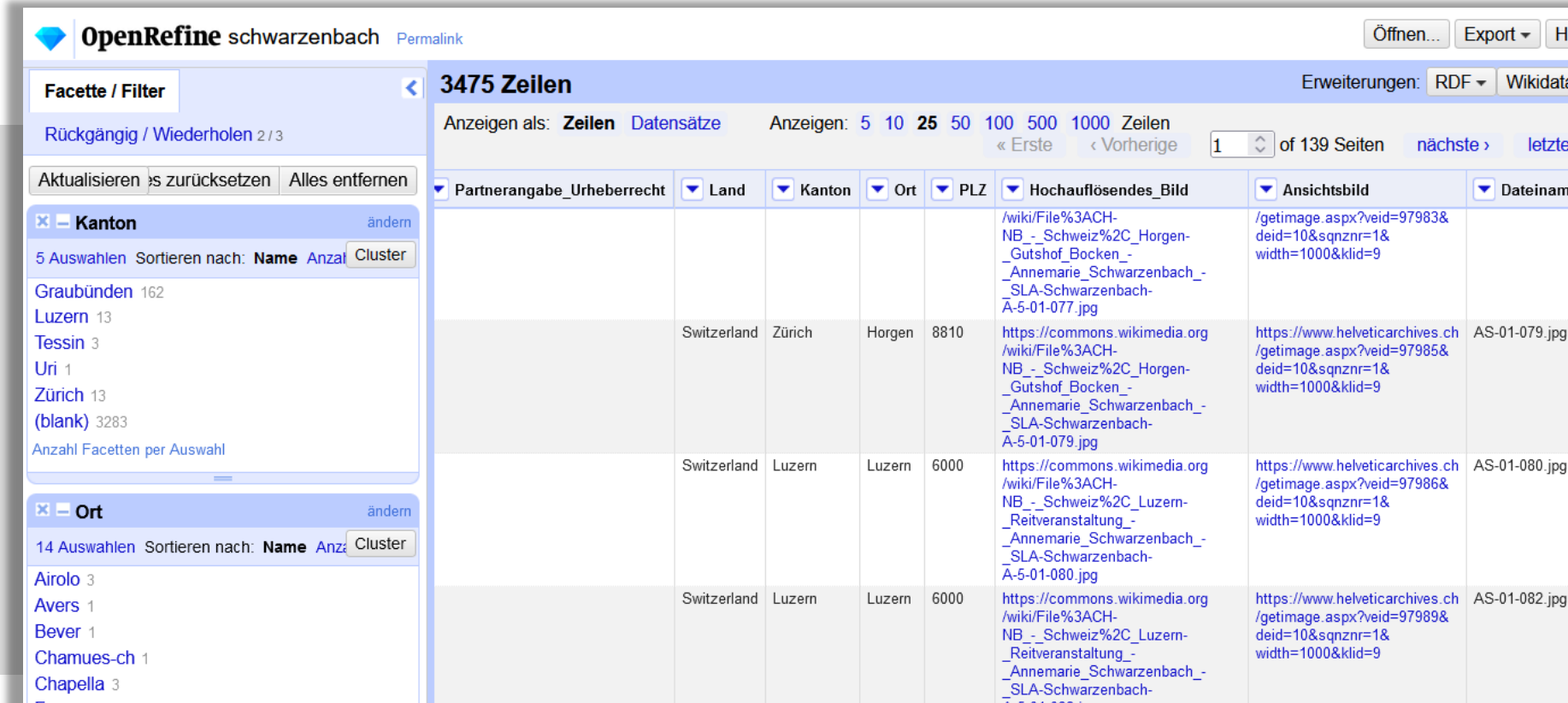
- (bekannt durch Katalog)
- guter Datenüberblick
- Basis: immer gesamter Feldinhalt

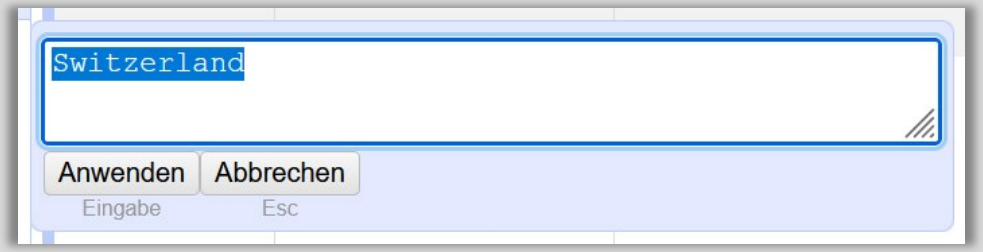
Wofür?

- **Erkennen von Unregelmässigkeiten/Fehlern**!
- geeignet für eher einfache und nicht übermässig viele verschiedene Ausprägungen

Was kann ich damit machen? Facettieren

UNIVERSITÄT BERN





Was kann ich damit machen?

Facetten-Werte auswählen und bearbeiten

Was kann ich damit machen?
Facetten vs. Filter

Filter

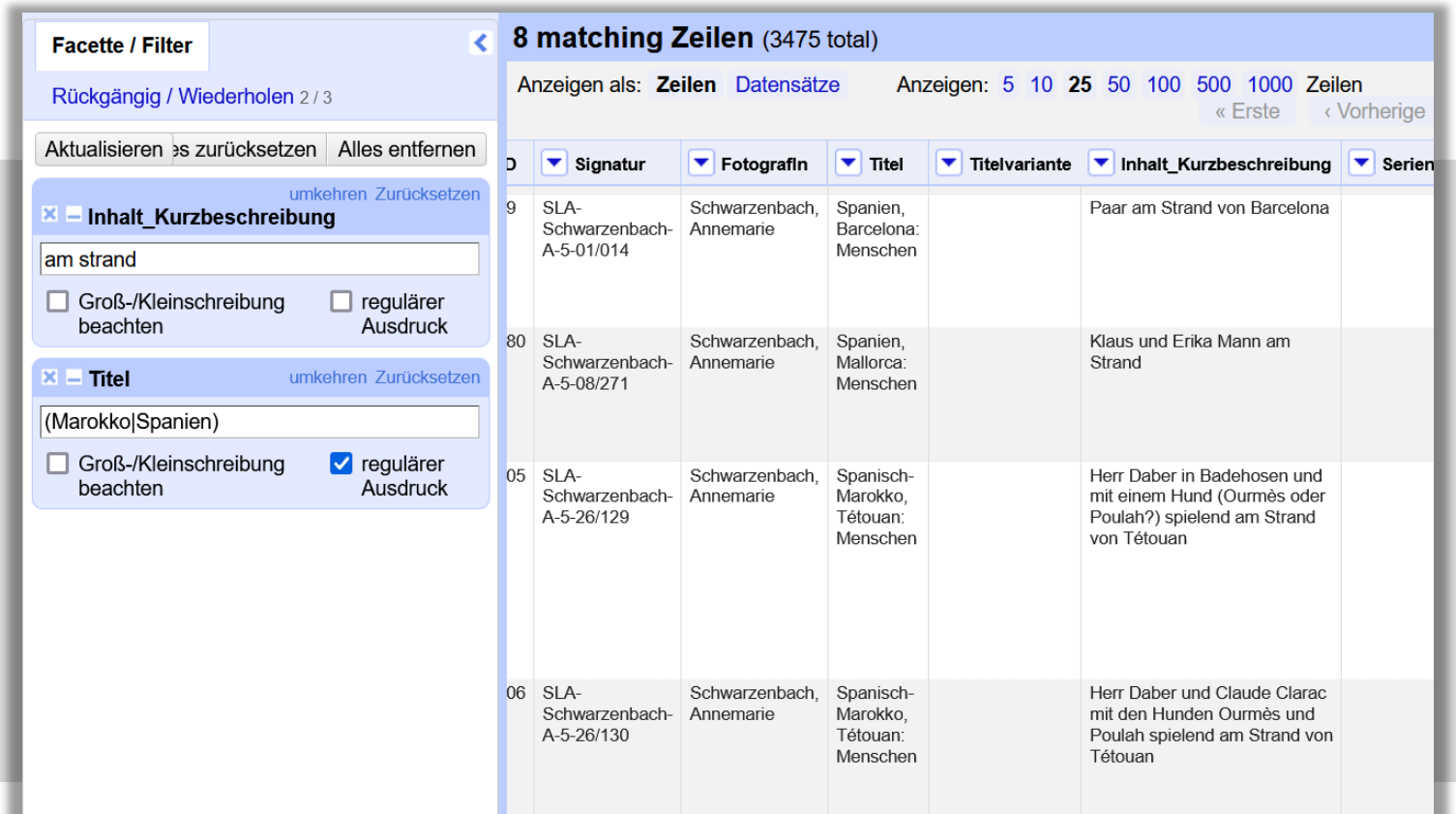
automatische Suche nach (auch Mehr-Wort-)**Strings** oder regulären Ausdrücken (Regex) in der Spalte

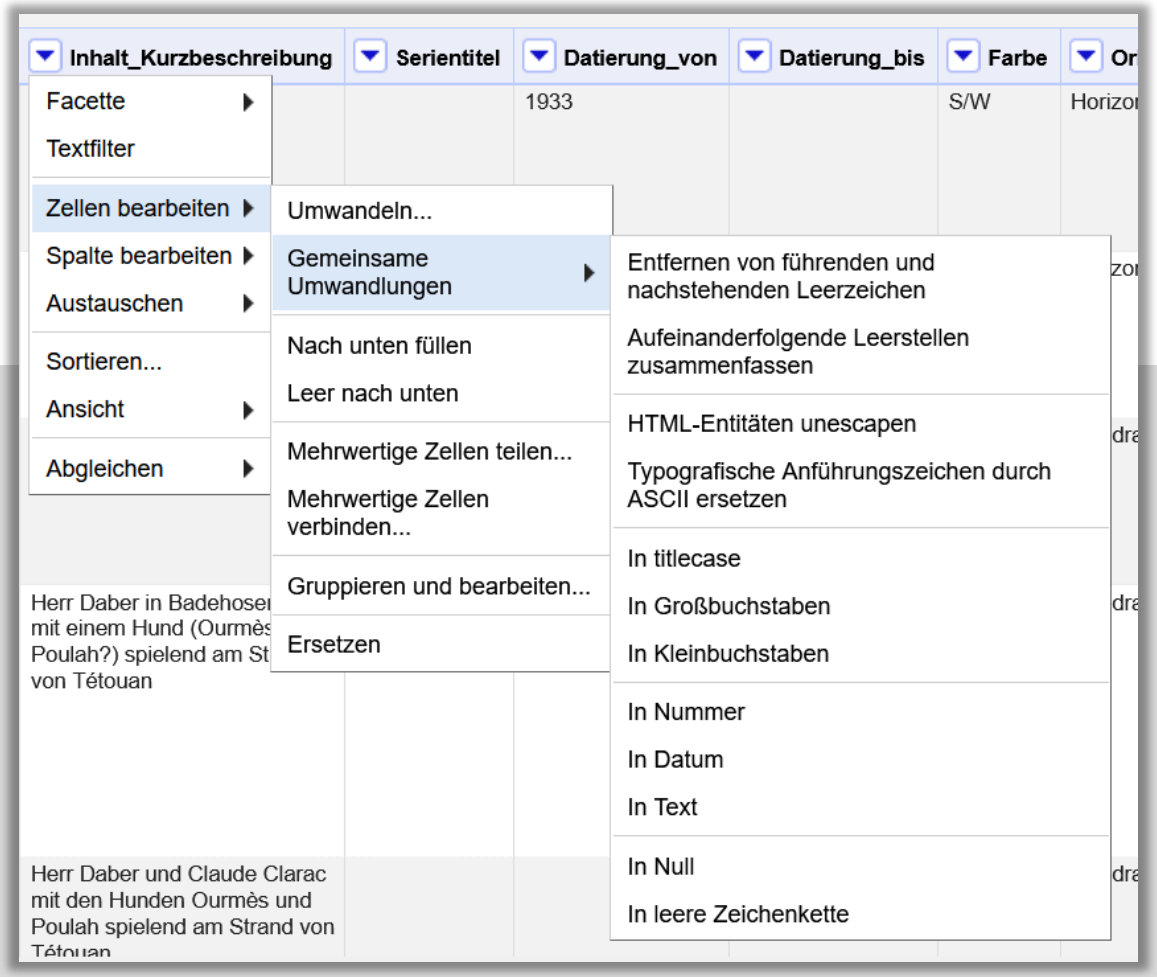
zutreffende („matching“) Zeilen werden automtisch ausgewählt = **Gesamtset eingeschränkt**!

Wofür?

- Existieren bestimmte Werte oder Muster in der Spalte?
- (Regex ist ein mächtiges Tool)

Was kann ich damit machen? Filtern





Was kann ich damit machen?

Standard-Bereinigungen bzw. Transformationen

→ für alle Werte in der Spalte

Was kann ich damit machen? Clustern

- Werte in einer Spalte **nach Ähnlichkeit gruppieren**
- quasi „Facettierung mit Unschärfe“
- Verschiedene Verfahren implementiert, z.T justierbar
- **Erkennen** von **ähnlichen Werten**…

- … aber vor allem auch von **Unregelmässigkeiten und Fehlern**
- bei komplexeren, und vielen verschiedenen Werten!
- Informationen zu den Verfahren: https://docs.openrefine.org/manual/celledit ing#cluster-and-edit (nur eng)

Cluster & Edit column "Inhalt_Kurzbeschreibung"

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 very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more...

Method: key collision | Keying Function: fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	3	• Margot Lind auf einem Geländer sitzend (2 rows) • Margot Lind sitzend auf einem Geländer	☐	Margot Lind auf einem Ge
2	3	• Claude Clarac (?) (2 rows) • Claude Clarac	☐	Claude Clarac (?)
2	2	• Zwei jünge Männer des Stammes Azande, einer mit Lanze der andere mit einem Velo auf der Schulter • Zwei jünge Männer des Stammes Azande, einer mit Lanze, der andere mit einem Velo auf der Schulter	☐	Zwei jünge Männer des St
2	2	• Junge sitzende Frau • Sitzende junge Frau	☐	Junge sitzende Frau
2	2	• Sitzende Männer • sitzende Männer	☐	Sitzende Männer
2	2	• Freitagsmoschee • Freitagsmoschee?	☐	Freitagsmoschee

Select All | Unselect All | Export Clusters | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Diese Funktion hilft Ihnen, Gruppen von verschiedenen Zellwerten zu finden, die alternative Darstellungen derselben Sache sein könnten. Beispielsweise beziehen sich die Konzept und haben nur Unterschiede in der Großschreibung, und „Gödel“ und „Godel“ beziehen sich wahrscheinlich auf dieselbe Person. Mehr herausfinden…

Methode nächster Nachbar | Levenshtein | Radius 1.0 | Block Zeichen 6

Clustergröße	Reihenanzahl	Werte im Cluster
2	2	• Blick auf Hafenanlage mit einigen (Wohn) Gebäuden im Vordergrund • Blick auf Hafenanlage mit einigen (Wohn)Gebäuden im Vordergrund
2	3	• Gebäude am Hafen (2 Zeilen) • Gebäude im Hafen
2	2	• Lagerhallen am Hafen • Lagerhallen im Hafen
2	2	• 2 Autos mit Nummernschild vor dem Senat • 3 Autos mit Nummernschild vor dem Senat
2	2	• Mönch auf der Klostertreppe • Mönche auf der Klostertreppe
2	2	• Zwei unterschiedlich uniformierte Soldaten zusammen mit einer Frau und mehreren Kindern im Stadtteil 'Belge' auf einem unasphaltierte • Zwei unterschiedlich uniformierte Soldaten zusammen mit einer Frau und mehreren Kindern im Stadtteil 'Belge'auf einem unasphaltierte
2	2	• Karteikarte: Zelt unter Bäumen

Alles markieren | Auswahl aufheben | Gruppen exportieren | **Auswahl zusammenführen & neu gruppieren** | Auswahl zusammenführen & schließen | Schließen

UNIVERSITÄT BERN

Was kann ich damit machen? URL abfragen

UNIVERSITÄT BERN

Spalte hinzufügen durch Abrufen von URLs basierend auf der Spalte Ort

Neuer Spaltenname: Wikipedia

Verzögerung drosseln: 5000 Millisekunden

Bei Fehler: auf leer gesetzt / Speicherfehler — Cache-Antworten

HTTP-Header, die beim Abrufen von URLs verwendet werden sollen: Anzeigen

Formulieren Sie die zu ladenden URLs:

Ausdruck — Sprache: General Refine Expression Language (GREL)

```
"https://de.wikipedia.org/wiki/" + value
```

Kein Syntaxfehler.

Vorschau | Verlauf | Mit Stern versehen | Hilfe

17.	Horgen	https://de.wikipedia.org/wiki/Horgen
18.	Horgen	https://de.wikipedia.org/wiki/Horgen
19.	Horgen	https://de.wikipedia.org/wiki/Horgen
20.	Horgen	https://de.wikipedia.org/wiki/Horgen
21.	Horgen	https://de.wikipedia.org/wiki/Horgen
22.	Luzern	https://de.wikipedia.org/wiki/Luzern
23.	Luzern	https://de.wikipedia.org/wiki/Luzern

Ort	Wikipedia
Horgen	<!DOCTYPE html> <html class="client-nojs" lang="de" dir="ltr"> <head> <meta charset="UTF-8"/> <title>Horgen – Wikipedia</title> <script>document.documentElement.className="client-js";R "wgRelevantPageIsProbablyEditable":true,"wgRestrictionEdit": "ready","ext.globalCssJs.user":"ready","user":"ready","user.op "ext.navigationTiming","ext.cx.eventlogging.campaigns","ext.c <script>(RLQ=window.RLQ||[]).push(function(){mw.loader.imple });});</script> <link rel="stylesheet" href="/w/load.php?lang=de&module <script async="" src="/w/load.php?lang=de&modules=st <meta name="ResourceLoaderDynamicStyles" content=""/> <link rel="stylesheet" href="/w/load.php?lang=de&module <meta name="generator" content="MediaWiki 1.38.0-wmf.9"/> <meta name="referrer" content="origin"/> <meta name="referrer" content="origin-when-crossorigin"/> <meta name="referrer" content="origin-when-cross-origin"/> <meta name="format-detection" content="telephone=no"/> <meta property="og:image" content="https://upload.wikimedia <meta property="og:image:width" content="1200"/>

Was kann ich damit machen? Daten parsen

UNIVERSITÄT BERN

Spalte basierend auf Spalte hinzufügen Wikipedia

Neuer Spaltenname: Wikipedia_Rohtext

Bei Fehler: (•) auf leer gesetzt ( ) Speicherfehler ( ) Wert aus Originalspalte kopieren

Ausdruck — Sprache: General Refine Expression Language (GREL)

```
value.parseHtml().select("body")[0].htmlText()
```

Kein Syntaxfehler.

Vorschau | Verlauf | Mit Stern versehen | Hilfe

row	value	value.parseHtml().select("body ...
14.	<!DOCTYPE html> <html class="client-nojs" lang="de" dir="ltr"> <head> <meta charset="UTF-8"/> <title>Horgen – Wikipedia</title> <script>document.documentElement.className="client- js";RLCONF= {"wgBreakFrames":false,"wgSeparatorTransformTable": ["",\t.","\t,"],"wgDigitTransformTable":	Horgen aus Wikipedia, der freien Enzyklopädie Zur Navigation springen Zur Suche springen Dieser Artikel beschreibt die Gemeinde in der Schweiz, weitere Bedeutungen siehe Horgen (Begriffsklärung). Horgen Staat: Schweiz Schweiz Kanton: Kanton Zürich Zürich (ZH) Bezirk: Horgen BFS-Nr.: 0295i1f3f4 Postleitzahl: 8810 Horgen 8815 Horgenberg 8816 Hirzel 8135 Sihlbrugg Station 8135 Sihlwald UN/LOCODE: CH HOE

Ort	Wikipedia_Rohtext
Horgen	Horgen aus Wikipedia, der freien Enzyklopädie Zur Navigation springen Zur Suche springen Dieser Artikel beschreibt die Gemeinde in der Schweiz, weitere Bedeutungen siehe Horgen (Begriffsklärung). Horgen Staat: Schweiz Schweiz Kanton: Kanton Zürich Zürich (ZH) Bezirk: Horgen BFS-Nr.: 0295i1f3f4 Postleitzahl: 8810 Horgen 8815 Horgenberg 8816 Hirzel 8135 Sihlbrugg Station 8135 Sihlwald UN/LOCODE: CH HOE Koordinaten: 687729 / 23509347.2609138.597777408Koordinaten: 47° 15′ 39″ N, 8° 35′ 52″ O; CH1903: 687729 / 235093 Höhe: 408 m ü. M. Höhenbereich: 405–914 m ü. M.[1] Fläche: 30,83 km²[2] Einwohner: i23'090 (31. Dezember 2020)[3] Einwohnerdichte: 749 Einw. pro km² Ausländeranteil: (Einwohner ohne Schweizer Bürgerrecht) 29,8 % (31. Dezember 2020)[4] Arbeitslosenquote: 2,2 % Gemeindepräsident: Theo Leuthold (SVP) Website: www.horgen.ch Lage der

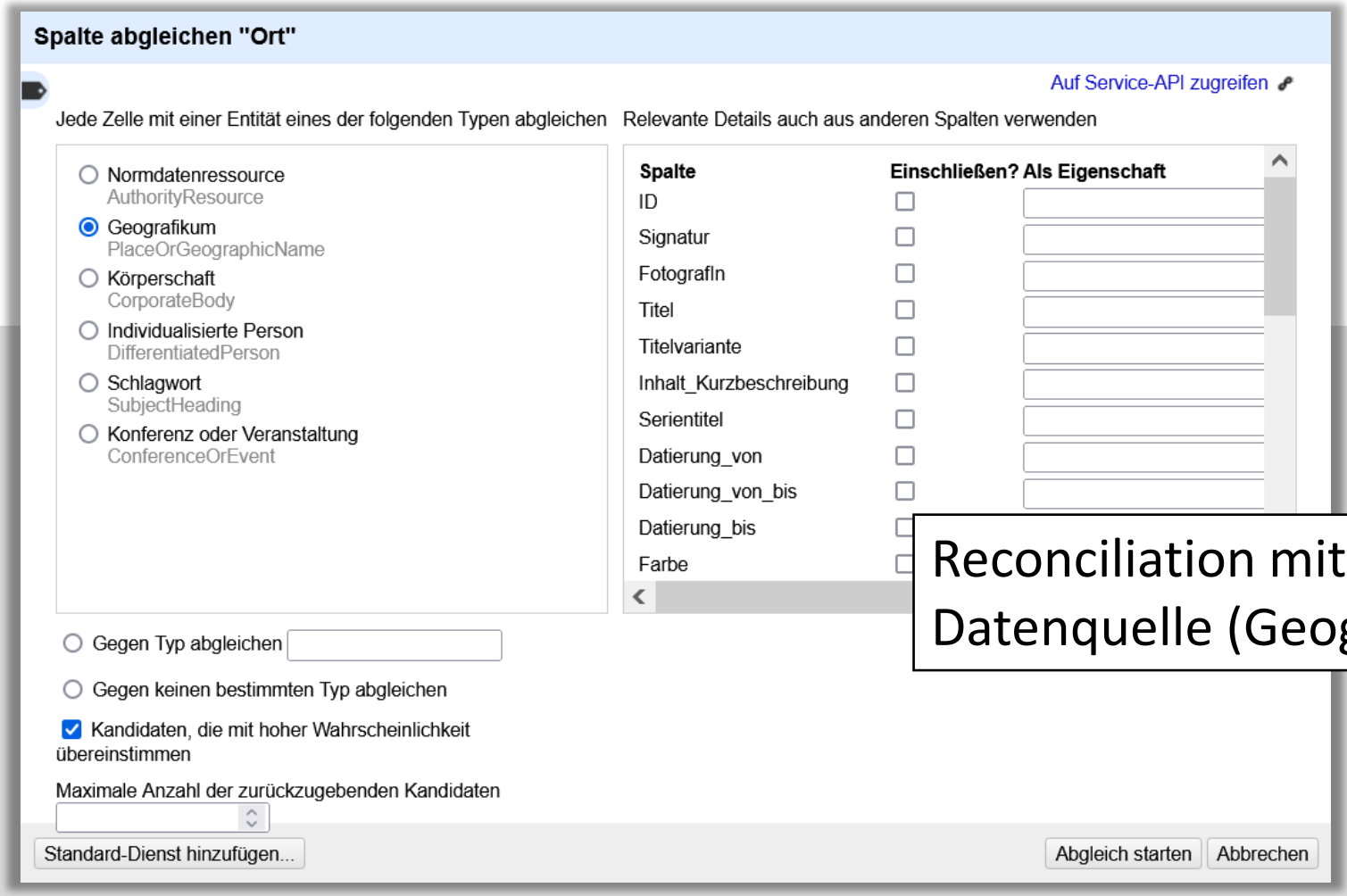
Was kann ich damit machen? Daten abgleichen/anreichern („Reconciliation“)

= halbautomatischer **Abgleich** mit externem Datenbestand

- bietet **interaktive Auswahl**, welche einzelnen Werte **geändert oder belassen** werden sollen
- Auswahlliste der zur Verfügung stehenden „Matches“

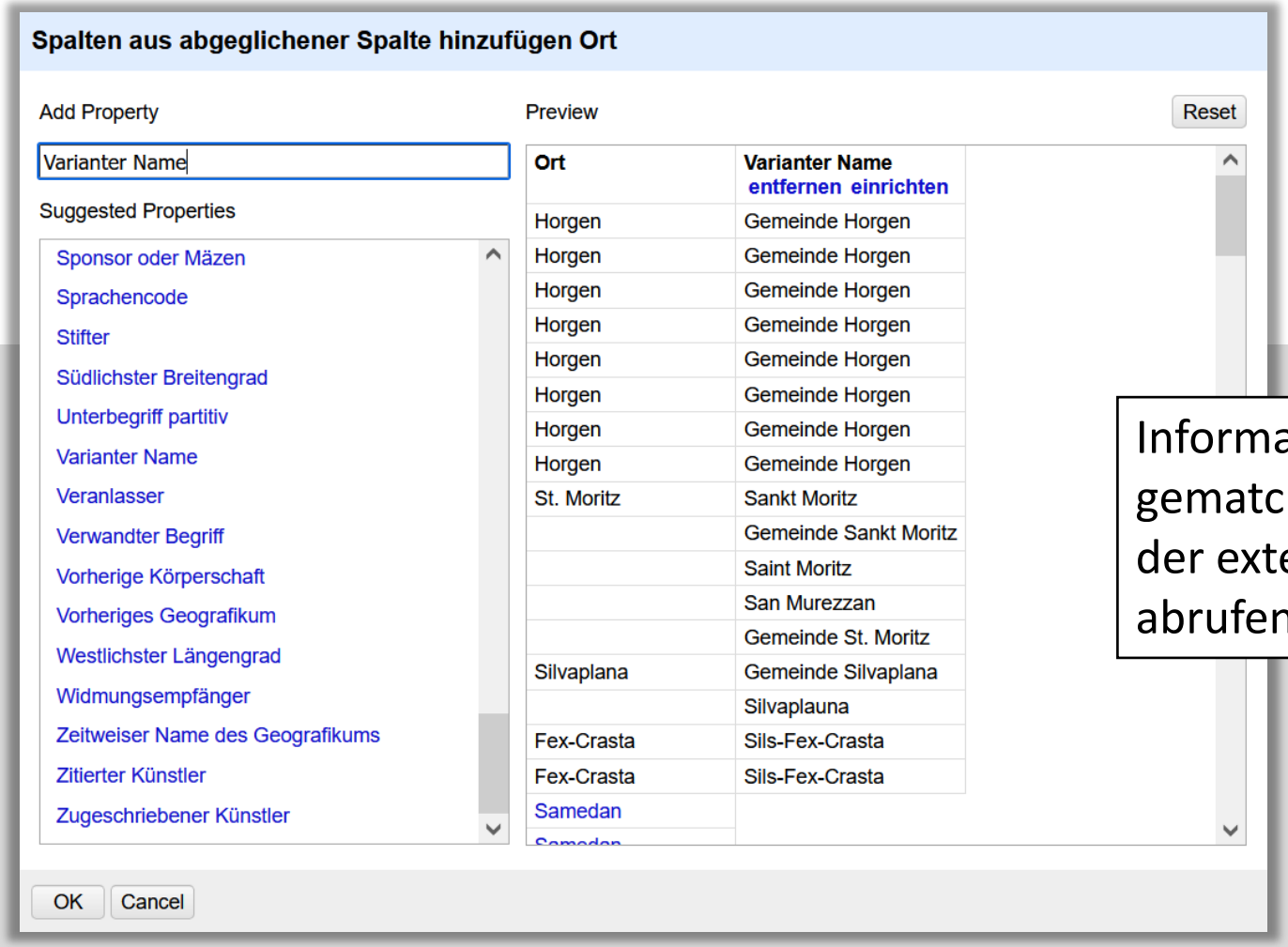
Sehr wertvoll um…

- vorhandene Werte gemäss externem **Standard zu vereinheitlichen**
- auf der Basis der externen Quelle die **eigenen Daten zu ergänzen (= anzureichern)**

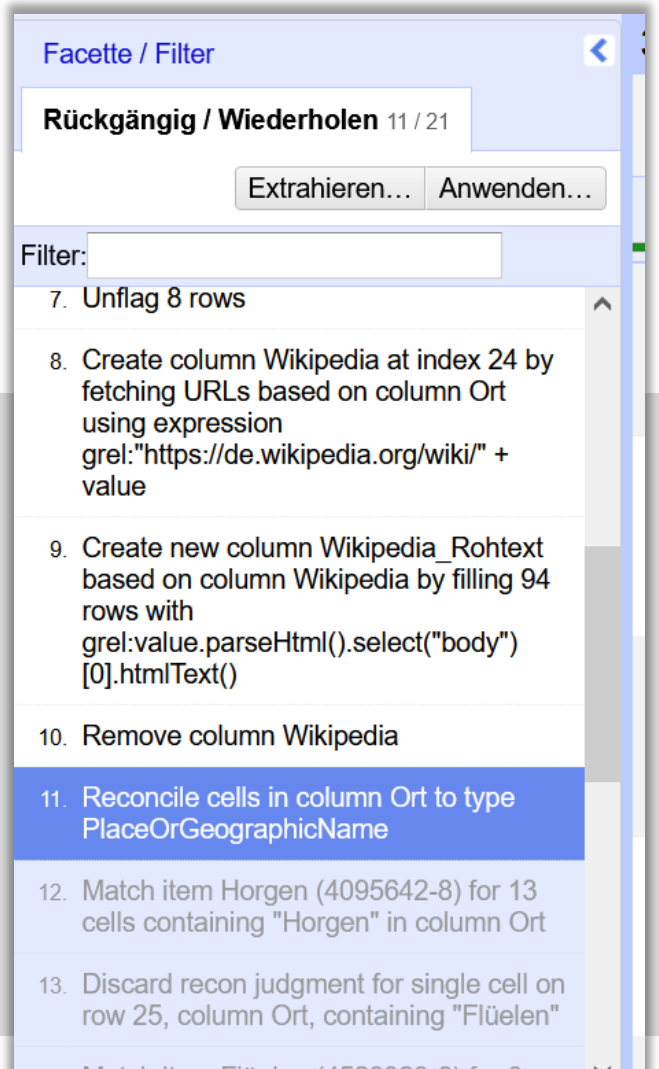


Reconciliation mit GND-Datenquelle (Geografika)…

- Auswahl der gematchten GND-Geografika
- Mouse-over liefert jeweils Details zur besseren Zuordnung
- Klick führt zur gematchten externen Entität



Informationen über die gematchten Entitäten von der externen Quelle abrufen (hier: GND)



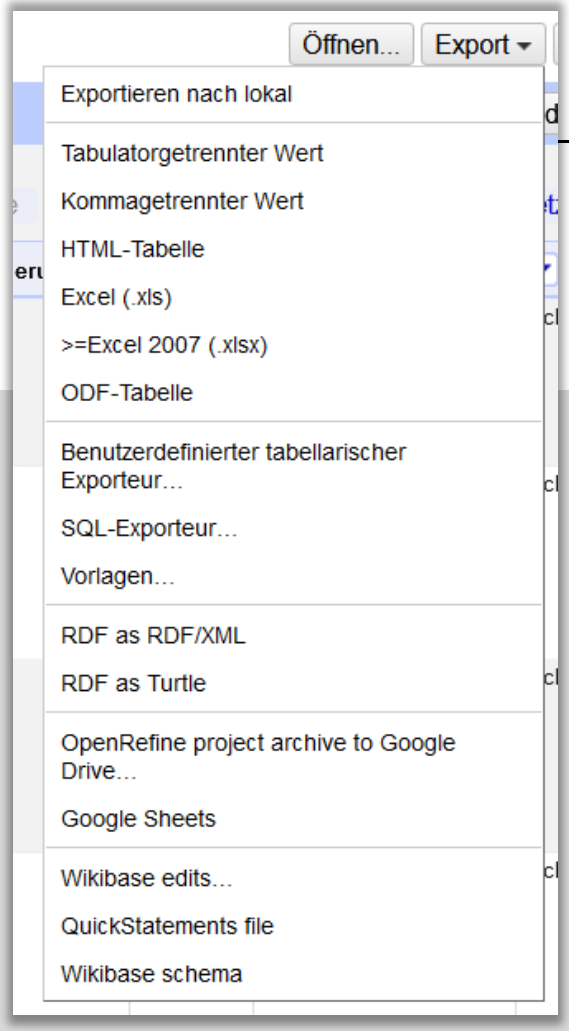
UNIVERSITÄT BERN

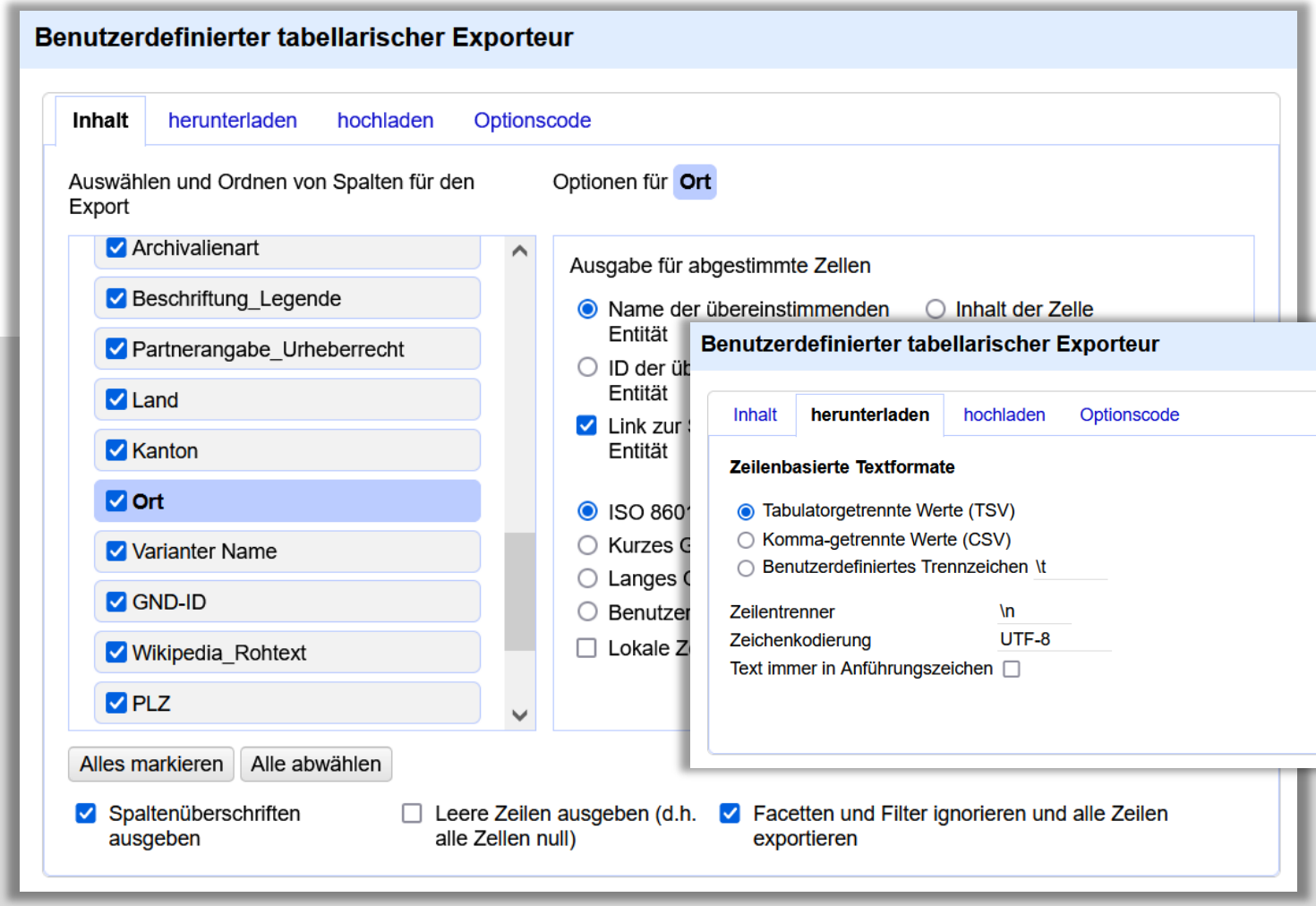
Was kann ich damit machen?

- Historie anschauen
- zurückspringen an Vorversion
- Historie exportieren (und auf neuen Daten nachnutzen)

Was kann ich damit machen?
Daten exportieren

…in viele Formate:
TSV, CSV, Excel, ODF, HTML, RDF, per Template (z.B. in JSON), als OpenRefine-Projekt (ZIP), SQL




Benutzerdefinierter tabellarischer Exporteur
Inhalt
herunterladen
hochladen
Optionscode
Auswählen und Ordnen von Spalten für den Export
Optionen für Ort
Archivalienart
Beschriftung_Legende
Partnerangabe_Urheberrecht
Land
Kanton
Ort
Varianter Name
GND-ID
Wikipedia_Rohtext
PLZ
Ausgabe für abgestimmte Zellen
Name der übereinstimmenden Entität
Inhalt der Zelle
Alles markieren
Alle abwählen
Spaltenüberschriften ausgeben
Leere Zeilen ausgeben (d.h. alle Zellen null)
Facetten und Filter ignorieren und alle Zeilen exportieren

Benutzerdefinierter tabellarischer Exporteur
Inhalt
herunterladen
hochladen
Optionscode
Zeilenbasierte Textformate
Tabulatorgetrennte Werte (TSV)
Komma-getrennte Werte (CSV)
Benutzerdefiniertes Trennzeichen \t
Zeilentrenner \n
Zeichenkodierung UTF-8
Text immer in Anführungszeichen
Andere Formate
Excel (.xls)
Excel in XML (.xlsx)
HTML-Tabelle
Vorschau

Zum Nachschauen & Weitermachen…

- OpenRefine User Manual
- Online-Tutorial Cleaning Data with OpenRefine (2018)
- Blogpost-Reihe zu OpenRefine (dt., 2017-19)

- General Refine Expression Language (GREL, GREL-Funktionen) → für elaborierte Datentransformationen mit etwas Code
- Reconciliation Services

UNIVERSITÄT
BERN

OpenRefine

Teil 2: Hands-On Workshop

Kathi Woitas, Digital Scholarship Services
Universitätsbibliothek Bern

Das haben wir vor…

- Daten laden
- Daten sortieren
- Daten facettieren
- Daten filtern

- Daten bereinigen (mutieren, transformieren)
- Daten clustern
- Daten anreichern
- Daten exportieren

Und nun: Hands-on!

1. Bitte OpenRefine starten (i.d.R. mit .exe-Datei)
2. Falls OpenRefine-Browser-Fenster *nicht* automatisch aufgeht, manuell localhost:3333 aufrufen
3. (Unter «Preferences» Sprache wechseln in «ger»)
4. In neuem Tab öffnen: https://opendata.swiss/de/dataset/bildersammlung-annemarie-schwarzenbach

Start: Daten laden

«Metadaten original» in OpenRefine laden:

1. Datei *schwarzenbach_metadata_original.csv* herunterladen
2. Datei in einfachem Texteditor öffnen
3. Alles markieren + kopieren (Win: `Ctrl + A`, `Crtl + C`)
3. Datenimport in OpenRefine mittels «Zwischenablage»
4. Daten in der Preview anschauen – alles richtig? Ansonsten Parameter anpassen
5. Projektname anpassen
6. «Projekt erstellen»

Daten sortieren

UNIVERSITÄT BERN

Aufgabe:

Suche die Spalte «Datierung_von».

Sortiere die Tabelle um, absteigend von den ältesten zu den jüngsten Fotografien.

(Sortierfunktion ist im Spaltenmenü zu finden.)

Wie seid ihr vorgegangen?

Ist euch etwas aufgefallen?

Neue Schaltfläche «Sortieren» im Kopf → Sortierung kann permanent gemacht werden

Daten facettieren und ändern

UNIVERSITÄT BERN

Aufgabe:

Erstelle die Textfacette aus der Spalte «Land».

In der Textfacette klicke «Switzerland» und danach «bearbeiten» an und ändere «Switzerland» in «Schweiz». Verfahre gleich mit «Turkey».

Sind die Werte in der Spalte «Land» entsprechend geändert?

Hinweis:

Facetten, die nicht mehr gebraucht werden, können mit Klick auf das «x-Kästchen» geschlossen werden.

Daten facettieren und ändern

UNIVERSITÄT BERN

Aufgabe:

Erstelle die Textfacette aus der Spalte «Orientierung_Form».

Erstelle die Wortfacette (unter «benutzerdefinierte Facetten») aus der Spalte «Orientierung_Form».

Was ist der Unterschied zwischen den beiden Facetten?

Frage:

Könnte man auch anders nach z.B. horizontalen Fotografien suchen?

Daten filtern

Aufgabe:

Filtere die Daten über die Spalte «Orientierung_Form» und suche nach horizontalen Bildern.

Filtere die Daten weiter nach Fotografien, die dem Titel nach im Iran aufgenommen wurden.

Wie viele Fotografien sind das?

Hinweis:

Filter entfernen, um zum Gesamtset zurückzukehren.

Spalten teilen

In der Spalte «Titel» sind die meisten Informationen über den Aufnahmeort in einer einheitlichen Syntax gehalten:

Land, *Ort*: *Motiv* oder

Land: Motiv

Doppelpunkt und z.T. **Komma** trennen die Bestandteile.

Für die Aufteilung der Bestandteile auf mehrere Spalten existiert die Funktion «In mehrere Spalten aufteilen».

Wir teilen zunächst nach dem Doppelpunkt – da dieser in den allermeisten Werten erscheint.

Spalten teilen

UNIVERSITÄT BERN

Aufgabe:

Die ehemalige Titel-Spalte wurde in «land_ort» und «motiv» aufgespalten. Trenne nun auch noch die Spalte «land_ort» in die neuen Spalten «land» und «ort» analog auf.

Überprüfe anhand der Textfacetten für «land» und «ort» die Ergebnisse.

Daten bereinigen

Die neu erstellten Spalten «land» und «ort» sehen insgesamt recht gut aus, benötigen aber etwas Feinschliff.

Die Werte, die ganz sicher nicht in ein Feld «land» gehören, sollten wir bereinigen.

Da dies für die Spalte «land» nur recht wenig, klare Einzelfälle sind, können wir diese direkt in den entsprechenden Feldern bearbeiten.

Clustern

Die Spalte «ort» gestaltet sich etwas unübersichtlicher.

Hier wollen wir daher Clustering-Verfahren ausprobieren, um ähnliche Werte vereinheitlichen zu können.

Die Clustering-Funktion lässt sich direkt auch aus der Facette aufrufen.

Aufgabe:

Probiere verschiedene Clustering-Methoden mit ihren Parametern iterativ aus und führe ähnliche Schreibweisen zusammen.

Welche Orte konntet ihr vereinheitlichen?

Es bitzeli Code: GREL

General Refine Expression Language

= einfache Sprache/Ausdrücke, um **Transformationen** auf den Daten auszuführen.

Diese funktionieren ähnliche wie mathematische **Funktionen**.

Funktion: `f(x, y) = 5x + y`^{2}

- Es wird eine spezifische Funktion/Transformation ausgewählt, und die Spalte, auf der diese angewendet werden soll.
- D.h. **alle Werte** in dieser **Spalte** werden **entsprechend der Funktion verarbeitet**.

Es bitzeli Code: GREL

Wie sieht das aus?

`f(x) = x + 5` (Addiere 5.)

Argument `x` ist der Wert in der Zelle → dieser heisst in GREL: `value`

Operation ist: `+ 5`

Formel gesamt: `value + 5`

«A» durch «B» ersetzen

nun 3 Argumente → `value, A, B`

Operation ist: `replace(x, y, z)`

x = wo?

y = Wert, der ersetzt werden soll

z = der stattdessen einzusetzende Wert

Formel gesamt:

`replace(value, "A", "B")` **oder:**

`value.replace("A", "B")`

Eine erste GREL-Aufgabe

Suche die Spalte «FotografIn» und wähle im Spaltenmenü «Zellen bearbeiten» und hier «Umwandeln».

Aufgabe:

Mache aus «Schwarzenbach, Annemarie» in allen Zellen «Schwarzenbach, Annegret»!

Für die ganz Schnellen…
Zusatzaufgabe:

Mache aus «Schwarzenbach, Annemarie» in allen Zellen «Weissenbach, Annegret»!

Was steht in der Historie als letzter Schritt?

Noch ein wenig GREL

Es sind nun zwei Spalten mit Ländern vorhanden: «land» und «Land».

Sinnvollerweise führen wir die beiden Angaben in einer Spalte zusammen.

Dabei übernehmen wir allerdings nur dort neue Angaben aus «land», wo nicht bereits ein Wert in «Land» vorhanden ist.

Der GREL-Code hierfür lautet:

```
if(value==null,
cells["land"].value, value)
```

Wie ist das zu «entschlüsseln»?

Letzte GREL-Aufgabe ☺

Führe nun die Zusammenführung der Spalten «land» und «Land» durch.

Hinweis Code:

```
if(value==null,
cells["land"].value, value)
```

Danach auch die Transformation mit den beiden Spalten «ort» und «Ort» durchführen.

Hierfür muss der Code leicht angepasst werden.

Danach können die Spalten «land» und «ort» gelöscht werden.

Daten abgleichen und anreichern

Die „Reconciliation“ ist eine OpenRefine-Spezialität!

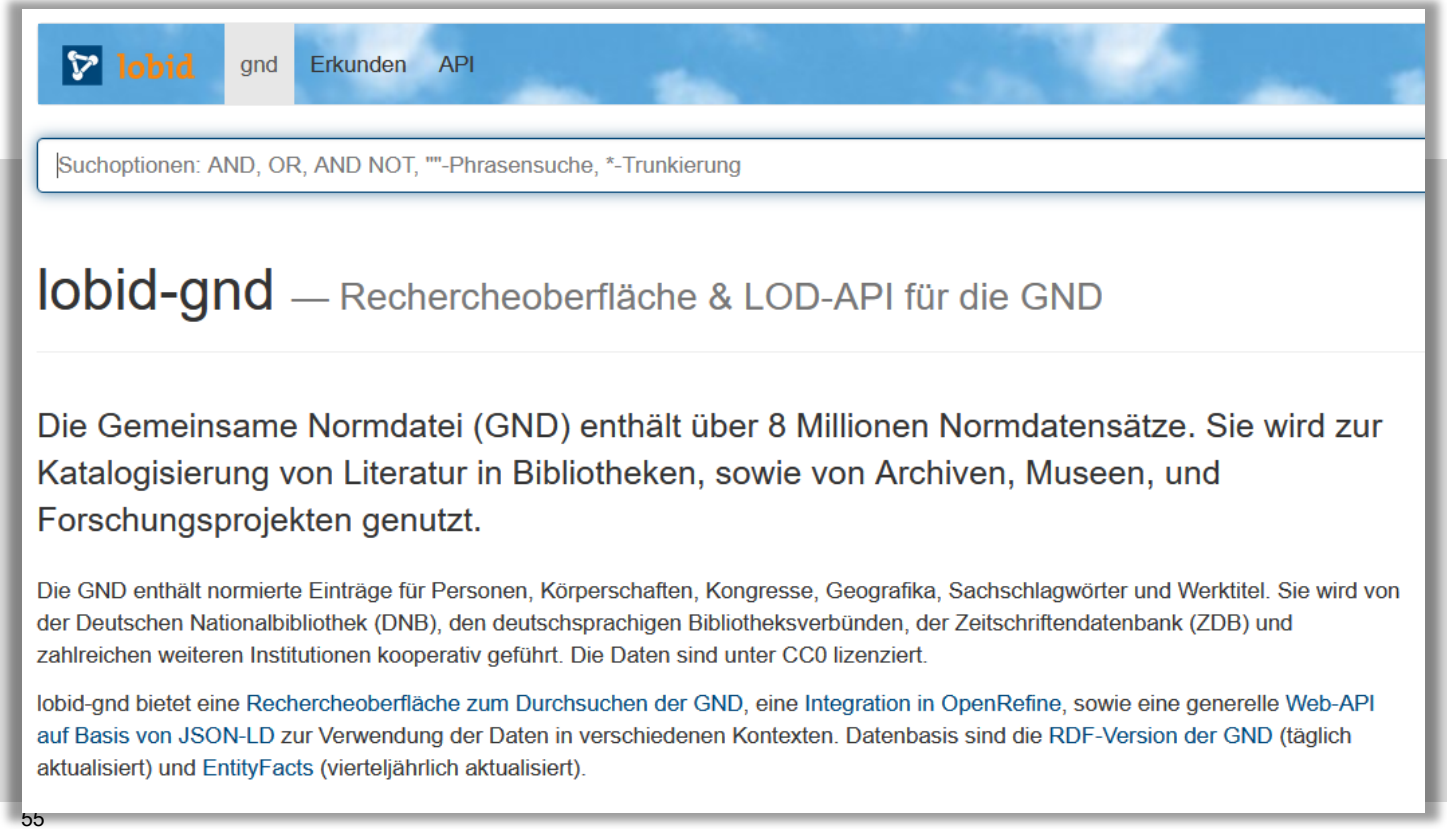
- eigene Daten mit einem externen Standard abgleichen und quasi verknüpfen
- auf der Grundlage externe Informationen in eigene Daten einfügen (anreichern)

Der Datenanbieter muss eine Reconciliation API anbieten. Bekannte Anbieter sind:

- Wikidata
- lobid.org (GND)
- VIAF
- ORCID

Weitere hier.

GND-Daten? lobid.org/gnd



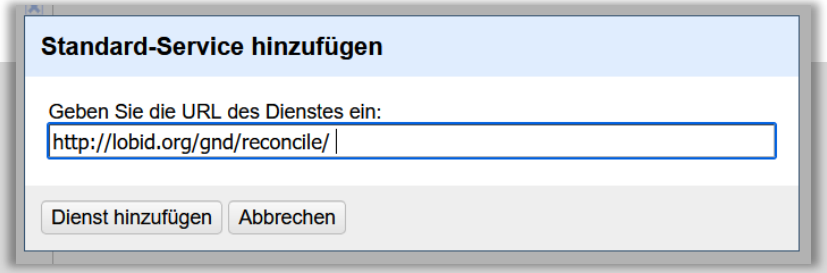
Daten abgleichen und anreichern

Wie funktioniert das?

Funktion «Abgleichen», «Starten Sie den Abgleich» ebenfalls im Spalten-Menü.

vorher nötig:

Einrichten eines neuen Standarddienstes (nur beim 1. Mal) per URL, z.B. http://lobid.org/gnd/reconcile/ (von Anbieter oder hier)

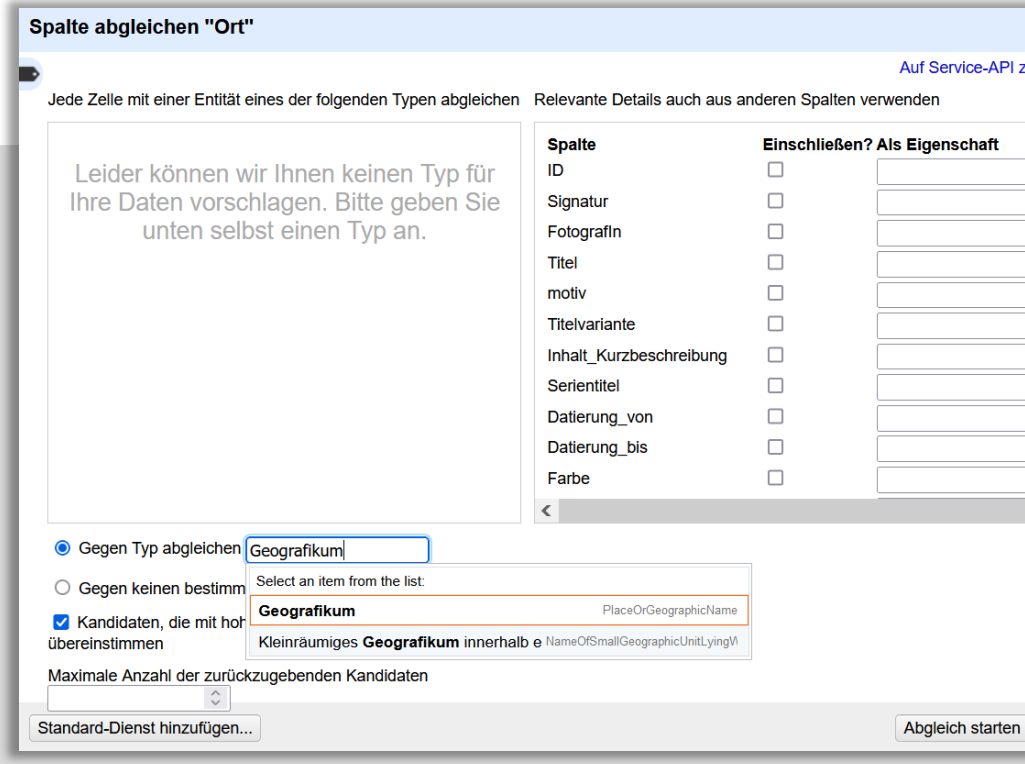


Wir richten den lobid.org-Reconciliation-Dienst ein, gehen dabei über die «Ort»-Spalte in das Menü.

Daten abgleichen und anreichern

Wir starten die Reconcilation über den Entitäten-Typ «Geografikum».

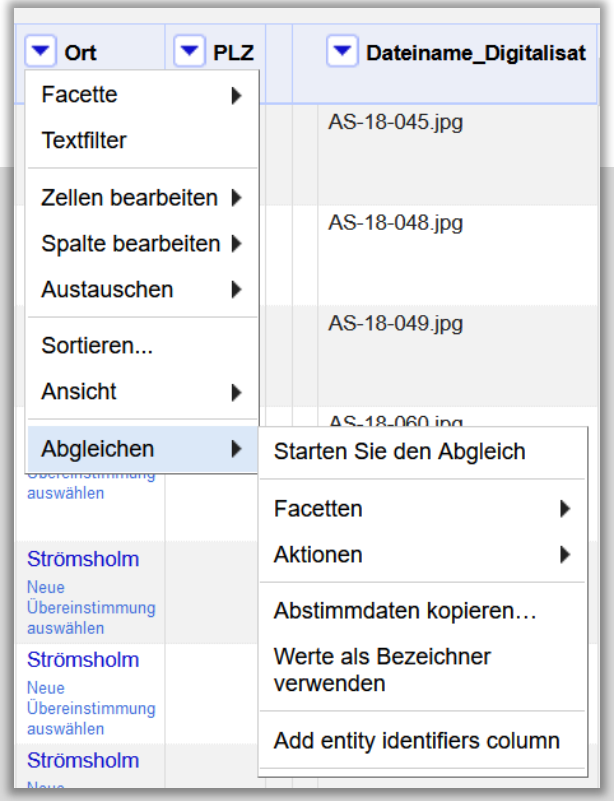
Nach erfolgten Abgleich erscheint eine Facette «judgement»/»Beurteilung, per dieser man die gematchten Werte auswählen kann.



Daten abgleichen und anreichern

Wir möchten auch noch gerne die GND-ID der gematchten Orte übernehmen.

Mit der untersten Auswahl «Add identifiers column» unter «Abgleichen» ist dies einfach möglich.



Daten abgleichen und anreichern

Schliesslich können wir aber auch alle anderen Informationen aus der GND für die gematchten Orte übernehmen.

Hierzu unter «Spalte bearbeiten» den Punkt «Spalte aus abgeglichen Werten hinzufügen» auswählen.

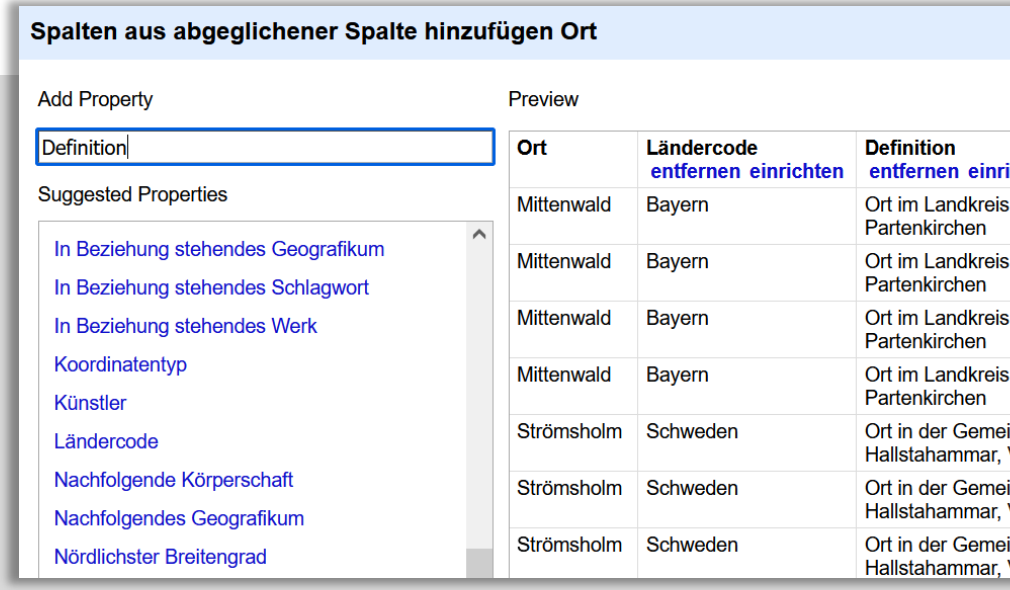
Daten abgleichen und anreichern

UNIVERSITÄT BERN

Aufgabe:

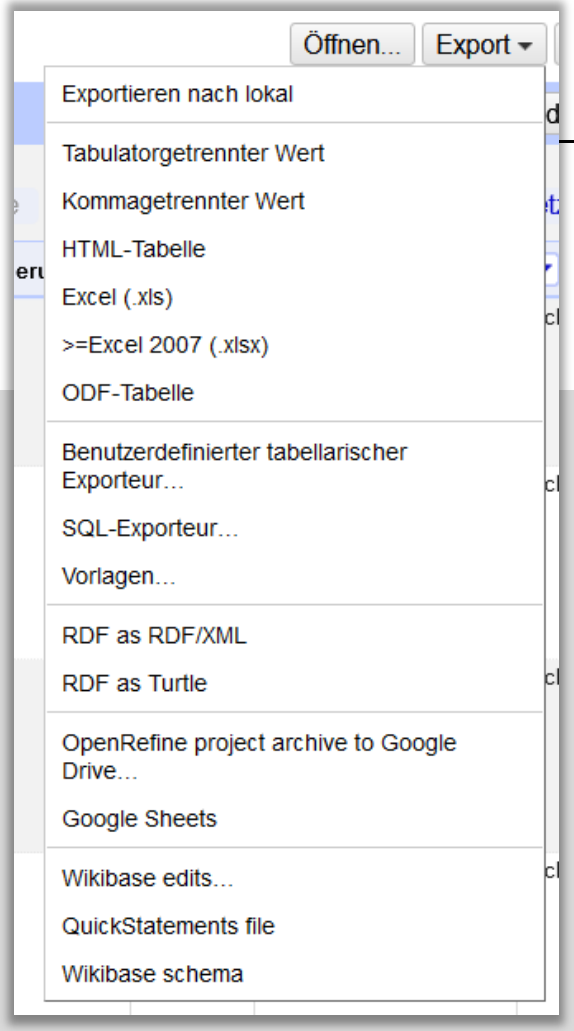
Teste verschiedene GND-Felder für eine Datenübernahme aus und importiere schliesslich die Informationen aus dem Feld «Ländercode».

Welche Unterschiede zum vorhandenen «Land» gibt es?



Daten exportieren

Aufgabe:
Führe einen Export in einem präferierten Format aus und öffne den Export dann in einer anderen passenden Anwendung.
Versuche danach den Export in einem neuen OpenRefine-Fenster (localhost:3333) wieder zu öffnen.



Zum Nachschauen & Weitermachen…

- OpenRefine User Manual
- Online-Tutorial Cleaning Data with OpenRefine (2018)
- Blogpost-Reihe zu OpenRefine (dt., 2017-19)

- General Refine Expression Language (GREL, GREL-Funktionen) → für elaborierte Datentransformationen mit etwas Code
- Reconciliation Services

Vielen Dank für eure Aufmerksamkeit!

Kathi Woitas, Digital Scholarship Services

Universitätsbibliothek Bern