



OpenRefine

Kathi Woitas, Digital Scholarship Services

Universitätsbibliothek Bern



DOI: [10.5281/zenodo.5776098](https://doi.org/10.5281/zenodo.5776098)

This work is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)  

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.

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**

- 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/

OpenRefine *Ein leistungsstarkes Werkzeug für die Bearbeitung von ungeordneten Daten.*

Projekt erstellen
Projekt öffnen
Projekt importieren
Spracheinstellungen

Erstellen Sie ein Projekt, um Daten zu importieren. Welche Art von Daten können Sie importieren?
TSV, CSV, *SV, Excel (.xls und .xlsx), JSON, XML, RDF als XML, und Google Data Dokumente werden unterstützt. Die Unterstützung für andere Formate kann mit OpenRefine-Erweiterungen hinzugefügt werden.

Daten abrufen von

Dieser Computer

Webadressen (URLs)

Zwischenablage

Database

Google Data

Suchen Sie eine oder mehrere Dateien auf Ihrem Computer, die Sie hochladen möchten:

Durchsuchen... Keine Dateien ausgewählt.

Nächste »

Version 3.5.0 [d4209a2]

Einstellungen
Hilfe
Über

«Von vorne anfangen» Parsing-Optionen konfigurieren Projekt name 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 Zeichen kodierung [Aktualisieren Vorschau](#)

CSV-/ TSV- / trennzeichengetrennte Dateien

Textdateien (zeilenbasiert)
 Textdateien mit fester Feldbreite
 PC-Axis text files
 JSON-Dateien
 MARC-Dateien
 JSON-LD-Dateien
 RDF/N3-Dateien

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



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 »

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

> 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 »

<input type="checkbox"/> Alle	<input type="checkbox"/> ID	<input type="checkbox"/> Signatur	<input type="checkbox"/> FotografIn	<input type="checkbox"/> Titelvariante	<input type="checkbox"/> Inhalt_Kurzbeschreibung	<input type="checkbox"/> Serientitel
<input type="checkbox"/> <input type="checkbox"/>	1.	160952	SLA-Schwarzenbach-A-5-01/001	Schwarzenbach, Annemarie		Männer beim Pelota (Ball) spielen
<input type="checkbox"/> <input type="checkbox"/>	2.	96578	SLA-Schwarzenbach-A-5-01/011	Schwarzenbach, Annemarie		
<input type="checkbox"/> <input type="checkbox"/>	3.	96579	SLA-Schwarzenbach-A-5-01/012	Schwarzenbach, Annemarie	Barcelona: Stadtansicht	
<input type="checkbox"/> <input type="checkbox"/>	4.	96758	SLA-Schwarzenbach-A-5-01/013	Schwarzenbach, Annemarie	Spanien, Barcelona: Menschen	

- Facette ▶
- Textfilter
- Zellen bearbeiten ▶
- Spalte bearbeiten ▶**
 - In mehrere Spalten aufteilen...
 - Spalten verbinden...
 - Spalte basierend auf dieser Spalte hinzufügen...
 - Hinzufügen von Spalten durch Abrufen von URLs...
 - Spalten aus abgeglichenen Werten hinzufügen...
 - Diese Spalte umbenennen
 - Diese Spalte entfernen
 - Spalte an den Anfang verschieben
 - Spalte an das Ende verschieben
 - Spalte nach links verschieben
 - Spalte nach rechts verschieben
- Austauschen ▶
- Sortieren...
- Ansicht ▶
- Abgleichen ▶

Was kann ich damit machen?

Spalten aufteilen und verbinden

Spalten verbinden

Wähle und ordne Spalten zum verbinden

 Datierung_von **Datierung_bis** ID Signatur FotografIn Titel Titelvariante Inhalt_Kurzbeschreibung Serientitel

Alles markieren

Alle abwählen

Optionen auswählen

Trennzeichen zwischen dem Inhalt jeder Spalte:

Gib ein oder mehrere Zeichen an oder lass das Feld leer, um die Spalten ohne Trennzeichen zu verbinden.

Ersetze null-Werte mit...

Gib ein oder mehrere Zeichen an oder lass das Feld leer, um null-Werte durch leere Zeichenfolgen zu ersetzen.

Überspringe null-Werte.

Verwende im Ersatz für Separatoren und Null-Werte \n für neue Zeilen, \t für Tabulatoren, \\n für \n, \\t für \t.

Schreib Ergebnis in ausgewählte Spalte.

Schreib Ergebnis in neue Spalte mit Namen...

Verbundene Spalten löschen.

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

OpenRefine schwarzenbach [Permalink](#) Öffnen... Export ▾ H

Facette / Filter ← **3475 Zeilen** Erweiterungen: **RDF** Wikidata

Rückgängig / Wiederholen 2 / 3

Aktualisieren zurücksetzen Alles entfernen

Kanton ändern
5 Auswählen Sortieren nach: **Name** Anzahl Cluster

- Graubünden 162
- Luzern 13
- Tessin 3
- Uri 1
- Zürich 13
- (blank) 3283

Anzahl Facetten per Auswahl

Ort ändern
14 Auswählen Sortieren nach: **Name** Anzahl Cluster

- Airolo 3
- Avers 1
- Bever 1
- Chamues-ch 1
- Chapella 3

Anzeigen als: **Zeilen** Datensätze Anzeigen: 5 10 **25** 50 100 500 1000 Zeilen
« Erste < Vorherige 1 of 139 Seiten nächste > letzte

Partnerangabe_Urheberrecht	Land	Kanton	Ort	PLZ	Hochauflösendes_Bild	Ansichtsbild	Dateiname
					/wiki/File%3ACH-NB_-_Schweiz%2C_Horgen-_Gutshof_Bocken_-_Annemarie_Schwarzenbach_-_SLA-Schwarzenbach-A-5-01-077.jpg	/getImage.aspx?veid=97983&deid=10&sqznr=1&width=1000&klid=9	
	Switzerland	Zürich	Horgen	8810	https://commons.wikimedia.org/wiki/File%3ACH-NB_-_Schweiz%2C_Horgen-_Gutshof_Bocken_-_Annemarie_Schwarzenbach_-_SLA-Schwarzenbach-A-5-01-079.jpg	https://www.helveticaarchives.ch/getimage.aspx?veid=97985&deid=10&sqznr=1&width=1000&klid=9	AS-01-079.jpg
	Switzerland	Luzern	Luzern	6000	https://commons.wikimedia.org/wiki/File%3ACH-NB_-_Schweiz%2C_Luzern-_Reitveranstaltung_-_Annemarie_Schwarzenbach_-_SLA-Schwarzenbach-A-5-01-080.jpg	https://www.helveticaarchives.ch/getimage.aspx?veid=97986&deid=10&sqznr=1&width=1000&klid=9	AS-01-080.jpg
	Switzerland	Luzern	Luzern	6000	https://commons.wikimedia.org/wiki/File%3ACH-NB_-_Schweiz%2C_Luzern-_Reitveranstaltung_-_Annemarie_Schwarzenbach_-_SLA-Schwarzenbach-A-5-01-082.jpg	https://www.helveticaarchives.ch/getimage.aspx?veid=97989&deid=10&sqznr=1&width=1000&klid=9	AS-01-082.jpg

Land ändern umkehren Zurücksetzen

2 Auswahlen Sortieren nach: **Name** Anzahl Cluster

Switzerland	192	exclude
Turkey	1	
(blank)	3282	

Anzahl Facetten per Auswahl

Switzerland

Anwenden Abbrechen

Eingabe Esc

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 automatisch ausgewählt = **Gesamtset eingeschränkt!**

Wofür?

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

Was kann ich damit machen? Filtern

Facette / Filter <

Rückgängig / Wiederholen 2 / 3

Aktualisieren zurücksetzen Alles entfernen

Inhalt_Kurzbeschreibung umkehren Zurücksetzen

am strand

Groß-/Kleinschreibung beachten regulärer Ausdruck

Titel umkehren Zurücksetzen

(Marokko|Spanien)

Groß-/Kleinschreibung beachten regulärer Ausdruck

8 matching Zeilen (3475 total)

Anzeigen als: **Zeilen** Datensätze Anzeigen: 5 10 25 50 100 500 1000 Zeilen

« Erste < Vorherige

	▼ Signatur	▼ Fotografin	▼ Titel	▼ Titelvariante	▼ Inhalt_Kurzbeschreibung	▼ Serien
9	SLA-Schwarzenbach-A-5-01/014	Schwarzenbach, Annemarie	Spanien, Barcelona: Menschen		Paar am Strand von Barcelona	
80	SLA-Schwarzenbach-A-5-08/271	Schwarzenbach, Annemarie	Spanien, Mallorca: Menschen		Klaus und Erika Mann am Strand	
05	SLA-Schwarzenbach-A-5-26/129	Schwarzenbach, Annemarie	Spanisch-Marokko, Tétouan: Menschen		Herr Daber in Badehosen und mit einem Hund (Ourmès oder Poulah?) spielend am Strand von Tétouan	
06	SLA-Schwarzenbach-A-5-26/130	Schwarzenbach, Annemarie	Spanisch-Marokko, Tétouan: Menschen		Herr Daber und Claude Clarac mit den Hunden Ourmès und Poulah spielend am Strand von Tétouan	

Inhalt_Kurzbeschreibung	Serientitel	Datierung_von	Datierung_bis	Farbe	Or
Facette Textfilter		1933		S/W	Horizo
Zellen bearbeiten ▶	Umwandeln...				
Spalte bearbeiten ▶	Gemeinsame Umwandlungen ▶				
Austauschen ▶	Entfernen von führenden und nachstehenden Leerzeichen				
Sortieren...	Aufeinanderfolgende Leerstellen zusammenfassen				
Ansicht ▶	HTML-Entitäten unescapen				
Abgleichen ▶	Typografische Anführungszeichen durch ASCII ersetzen				
	In titlecase				
	In Großbuchstaben				
	In Kleinbuchstaben				
	In Nummer				
	In Datum				
	In Text				
	In Null				
	In leere Zeichenkette				
Herr Daber in Badehose mit einem Hund (Ourmès Poulah?) spielend am Strand von Tétouan					dra
Herr Daber und Claude Clarac mit den Hunden Ourmès und Poulah spielend am Strand von Tétouan					dra

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/cellediting#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

Keying Function

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	3	<ul style="list-style-type: none">Margot Lind auf einem Geländer sitzend (2 rows)Margot Lind sitzend auf einem Geländer	<input type="checkbox"/>	Margot Lind auf einem Ge
2	3	<ul style="list-style-type: none">Claude Clarac (?) (2 rows)Claude Clarac	<input type="checkbox"/>	Claude Clarac (?)
2	2	<ul style="list-style-type: none">Zwei junge Männer des Stammes Azande, einer mit Lanze der andere mit einem Velo auf der SchulterZwei junge Männer des Stammes Azande, einer mit Lanze, der andere mit einem Velo auf der Schulter	<input type="checkbox"/>	Zwei junge Männer des St
2	2	<ul style="list-style-type: none">Junge sitzende FrauSitzende junge Frau	<input type="checkbox"/>	Junge sitzende Frau
2	2	<ul style="list-style-type: none">Sitzende Männersitzende Männer	<input type="checkbox"/>	Sitzende Männer
2	2	<ul style="list-style-type: none">FreitagsmoscheeFreitagsmoschee?	<input type="checkbox"/>	Freitagsmoschee

Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

u^b

b
UNIVERSITÄT
BERN

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

 Levenshtein

 Radius

 Block Zeichen

Clustergröße	Reihenanzahl	Werte im Cluster
2	2	<ul style="list-style-type: none"> Blick auf Hafenanlage mit einigen (Wohn) Gebäuden im Vordergrund Blick auf Hafenanlage mit einigen (Wohn)Gebäuden im Vordergrund
2	3	<ul style="list-style-type: none"> Gebäude am Hafen (2 Zeilen) Gebäude im Hafen
2	2	<ul style="list-style-type: none"> Lagerhallen am Hafen Lagerhallen im Hafen
2	2	<ul style="list-style-type: none"> 2 Autos mit Nummernschild vor dem Senat 3 Autos mit Nummernschild vor dem Senat
2	2	<ul style="list-style-type: none"> Mönch auf der Klostertreppe Mönche auf der Klostertreppe
2	2	<ul style="list-style-type: none"> Zwei unterschiedlich uniformierte Soldaten zusammen mit einer Frau und mehreren Kindern im Stadtteil 'Belge' auf einem unasphaltierten Zwei unterschiedlich uniformierte Soldaten zusammen mit einer Frau und mehreren Kindern im Stadtteil 'Belge' auf einem unasphaltierten
2	2	<ul style="list-style-type: none"> Karteikarte: Zelt unter Bäumen

Was kann ich damit machen? URL abfragen

Spalte hinzufügen durch Abrufen von URLs basierend auf der Spalte Ort

Neuer Spaltenname Verzögerung drosseln 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 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	<pre><!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 "wgRelevantPagelsProbablyEditable".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"/></pre>

Was kann ich damit machen? **Daten parsen**

Spalte basierend auf Spalte hinzufügen Wikipedia

 Neuer Spaltenname

 Bei Fehler auf leer gesetzt Speicherfehler Wert aus Originalspalte kopieren

 Ausdruck Sprache

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-nojs";RLCONF={	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)**

Spalte abgleichen "Ort"

[Auf Service-API zugreifen](#)

Jede Zelle mit einer Entität eines der folgenden Typen abgleichen Relevante Details auch aus anderen Spalten verwenden

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

Spalte	Einschließen?	Als Eigenschaft
ID	<input type="checkbox"/>	<input type="text"/>
Signatur	<input type="checkbox"/>	<input type="text"/>
FotografIn	<input type="checkbox"/>	<input type="text"/>
Titel	<input type="checkbox"/>	<input type="text"/>
Titelvariante	<input type="checkbox"/>	<input type="text"/>
Inhalt_Kurzbeschreibung	<input type="checkbox"/>	<input type="text"/>
Serientitel	<input type="checkbox"/>	<input type="text"/>
Datierung_von	<input type="checkbox"/>	<input type="text"/>
Datierung_von_bis	<input type="checkbox"/>	<input type="text"/>
Datierung_bis	<input type="checkbox"/>	<input type="text"/>
Farbe	<input type="checkbox"/>	<input type="text"/>

Gegen Typ abgleichen

Gegen keinen bestimmten Typ abgleichen

Kandidaten, die mit hoher Wahrscheinlichkeit übereinstimmen

Maximale Anzahl der zurückzugebenden Kandidaten

Reconciliation mit GND-Datenquelle (Geografika)...

Kanton
 Ort
 Wikipedia_Rohtext
 PLZ

Zürich	Horgen	<input type="button" value="edit"/>	Horgen aus Wikipedia, der freien	8810
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen (61)	<div data-bbox="531 207 1342 595" style="border: 1px solid #ccc; padding: 5px;"> <p>Diese Zelle zuordnen</p> <p>Finde alle identischen Zellen <input type="button" value="Abbrechen"/></p> <div style="display: flex; align-items: center;">  <div> <p>Horgen (4095642-8)</p> <p>Gebietskörperschaft oder Verwaltungseinheit</p> </div> </div> </div>		
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Reformierte K...			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen (Horg...			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen / Scheller (50)			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen-Heub...			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen (Rottweil) (48			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Bezirk Horg...			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen / Dampfschiffst...			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Hirzel (Horgen) (47)			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen / Credit Suisse / Seminarzentrum			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Horgen / Seminargebäude (46)			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Villa Seerose (Horgen) (42)			
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	Neues Element anlegen			

Suche nach Übereinstimmung

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

Spalten aus abgeglicherer Spalte hinzufügen Ort

Add Property

Suggested Properties

Sponsor oder Mäzen
Sprachcode
Stifter
Südlichster Breitengrad
Unterbegriff partitiv
Varianter Name
Veranlasser
Verwandter Begriff
Vorherige Körperschaft
Vorheriges Geografikum
Westlichster Längengrad
Widmungsempfänger
Zeitweiser Name des Geografikums
Zitierter Künstler
Zugeschriebener Künstler

Preview

Reset

Ort	Varianter Name entfernen einrichten
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
Horgen	Gemeinde Horgen
St. Moritz	Sankt Moritz
	Gemeinde Sankt Moritz
	Saint Moritz
	San Murezzan
	Gemeinde St. Moritz
Silvaplana	Gemeinde Silvaplana
	Silvaplana
Fex-Crasta	Sils-Fex-Crasta
Fex-Crasta	Sils-Fex-Crasta
Samedan	
Samedan	

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

OK

Cancel

Facette / Filter

Rückgängig / Wiederholen 11 / 21

Extrahieren... Anwenden...

Filter:

- Unflag 8 rows
- Create column Wikipedia at index 24 by fetching URLs based on column Ort using expression
grel:"https://de.wikipedia.org/wiki/" + value
- Create new column Wikipedia_Rohtext based on column Wikipedia by filling 94 rows with
grel:value.parseHtml().select("body")[0].htmlText()
- Remove column Wikipedia
- Reconcile cells in column Ort to type PlaceOrGeographicName
- Match item Horgen (4095642-8) for 13 cells containing "Horgen" in column Ort
- Discard recon judgment for single cell on row 25, column Ort, containing "Flüelen"

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

Alles markieren Alle abwählen

- Spaltenüberschriften ausgeben
- Leere Zeilen ausgeben (d.h. alle Zellen null)
- Facetten und Filter ignorieren und alle Zeilen exportieren

Ausgabe für abgestimmte Zellen

- Name der übereinstimmenden Entität
- Inhalt der Zelle
- ID der übereinstimmenden Entität
- Link zur Entität
- ISO 8601
- Kurzes GND
- Langes GND
- Benutzerdefiniertes Trennzeichen
- Lokale Zeilen

Benutzerdefinierter tabellarischer Exporteur

Inhalt herunterladen hochladen Optionscode

Zeilenbasierte Textformate

- Tabulatorgetrennte Werte (TSV)
- Komma-getrennte Werte (CSV)
- Benutzerdefiniertes Trennzeichen

Zeilentrenner
Zeichenkodierung
Text immer in Anführungszeichen

Andere Formate

- Excel (.xls)
- Excel in XML (.xlsx)
- HTML-Tabelle

Vorschau Herunterladen

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

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`, `Ctrl + 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

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

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

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

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)`
 $x = \text{wo?}$

$y =$ 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!

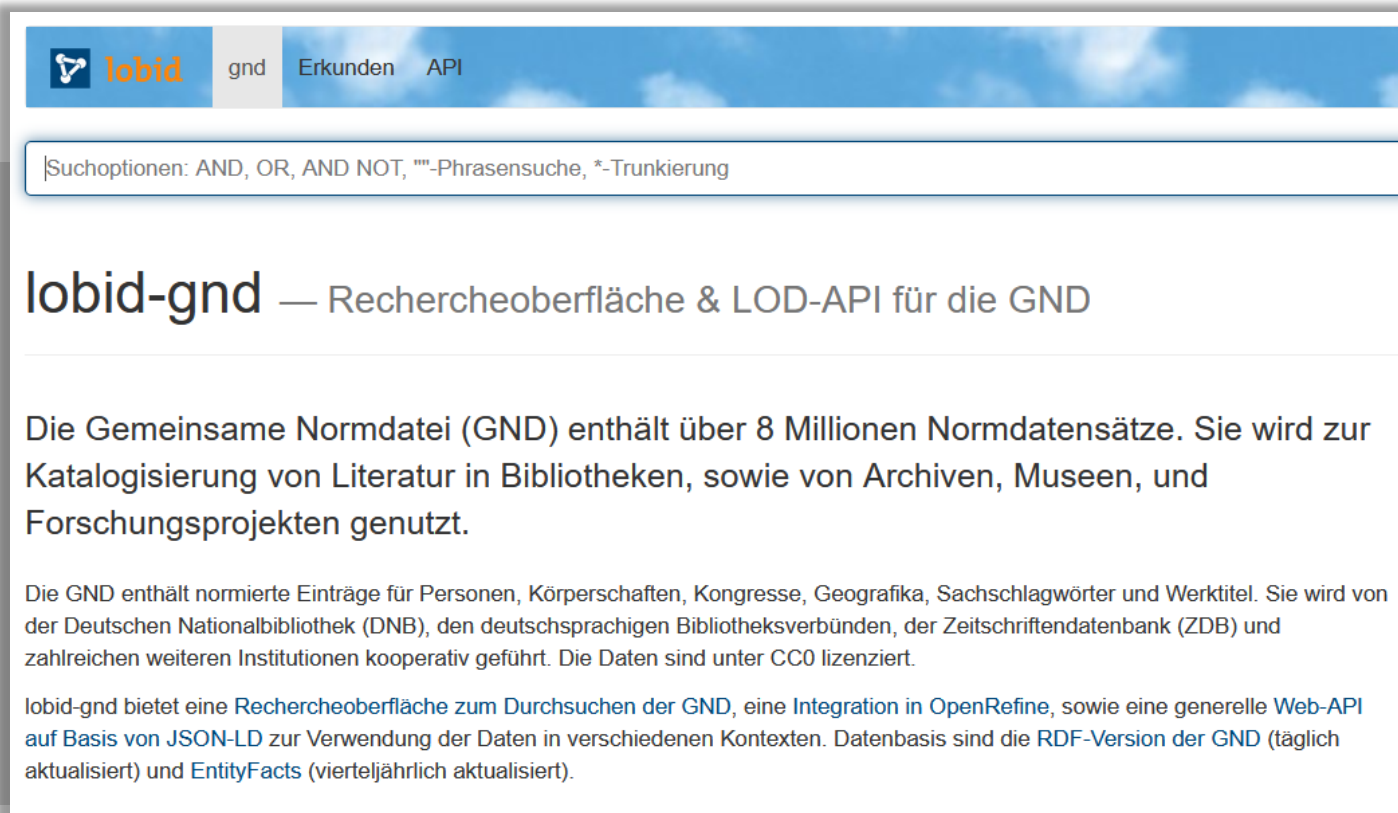
- eigene Daten mit einem externen Standard abgleichen und quasi verknüpfen
- auf der Grundlage externe Informationen in eigene Daten einfügen (anreichern)

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

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

Weitere [hier](#).

GND-Daten? lobid.org/gnd



lobid gnd Erkunden API

Suchoptionen: AND, OR, AND NOT, ""-Phrasensuche, *-Trunkierung

lobid-gnd — Rechercheoberfläche & LOD-API für die GND

Die Gemeinsame Normdatei (GND) enthält über 8 Millionen Normdatensätze. Sie wird zur Katalogisierung von Literatur in Bibliotheken, sowie von Archiven, Museen, und Forschungsprojekten genutzt.

Die GND enthält normierte Einträge für Personen, Körperschaften, Kongresse, Geografika, Sachschlagwörter und Werktitel. Sie wird von der Deutschen Nationalbibliothek (DNB), den deutschsprachigen Bibliotheksverbänden, der Zeitschriftendatenbank (ZDB) und zahlreichen weiteren Institutionen kooperativ geführt. Die Daten sind unter CC0 lizenziert.

lobid-gnd bietet eine [Rechercheoberfläche zum Durchsuchen der GND](#), eine [Integration in OpenRefine](#), sowie eine generelle [Web-API auf Basis von JSON-LD](#) zur Verwendung der Daten in verschiedenen Kontexten. Datenbasis sind die [RDF-Version der GND](#) (täglich aktualisiert) und [EntityFacts](#) (vierteljährlich aktualisiert).

Daten abgleichen und anreichern

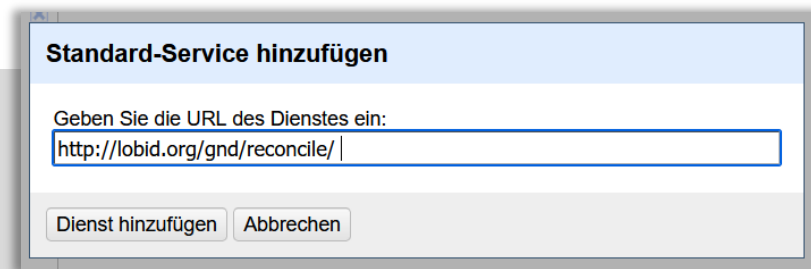
Wie funktioniert das?

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](#))



The screenshot shows a dialog box with a light blue header containing the title "Standard-Service hinzufügen". Below the header, there is a text prompt "Geben Sie die URL des Dienstes ein:" followed by a text input field containing the URL "http://lobid.org/gnd/reconcile/". At the bottom of the dialog, there are two buttons: "Dienst hinzufügen" and "Abbrechen".

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

Daten abgleichen und anreichern

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

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

Spalte abgleichen "Ort" Auf Service-API z...

Jede Zelle mit einer Entität eines der folgenden Typen abgleichen Relevante Details auch aus anderen Spalten verwenden

Leider können wir Ihnen keinen Typ für Ihre Daten vorschlagen. Bitte geben Sie unten selbst einen Typ an.

Spalte	Einschließen? Als Eigenschaft	
ID	<input type="checkbox"/>	<input type="text"/>
Signatur	<input type="checkbox"/>	<input type="text"/>
Fotografin	<input type="checkbox"/>	<input type="text"/>
Titel	<input type="checkbox"/>	<input type="text"/>
motiv	<input type="checkbox"/>	<input type="text"/>
Titelvariante	<input type="checkbox"/>	<input type="text"/>
Inhalt_Kurzbeschreibung	<input type="checkbox"/>	<input type="text"/>
Serientitel	<input type="checkbox"/>	<input type="text"/>
Datierung_von	<input type="checkbox"/>	<input type="text"/>
Datierung_bis	<input type="checkbox"/>	<input type="text"/>
Farbe	<input type="checkbox"/>	<input type="text"/>

Gegen Typ abgleichen

Gegen keinen bestimm...

Kandidaten, die mit hot übereinstimmen

Select an item from the list:

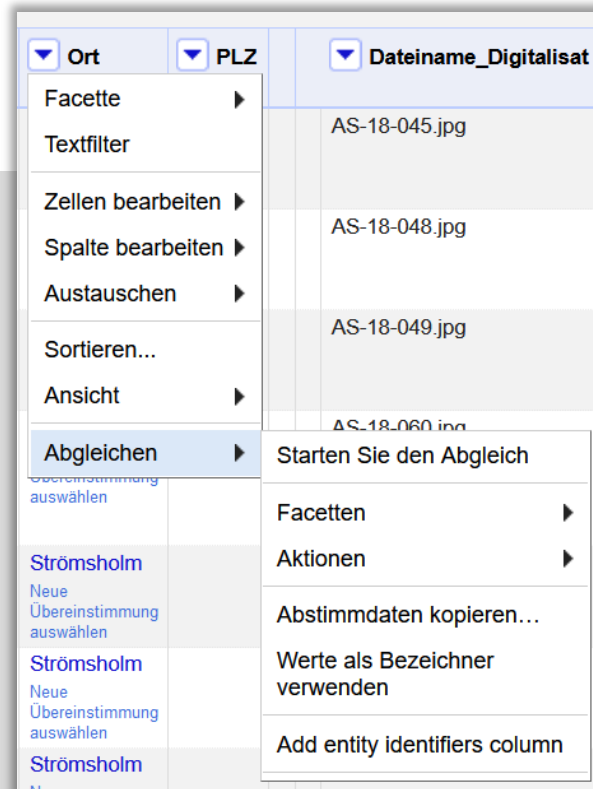
- Geografikum** PlaceOrGeographicName
- Kleinräumiges **Geografikum** innerhalb e NameOfSmallGeographicUnitLyingW

Maximale Anzahl der zurückzugebenden Kandidaten

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

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?

Spalten aus abgeglichener Spalte hinzufügen Ort

Add Property

Suggested Properties

- [In Beziehung stehendes Geografikum](#)
- [In Beziehung stehendes Schlagwort](#)
- [In Beziehung stehendes Werk](#)
- [Koordinatentyp](#)
- [Künstler](#)
- [Ländercode](#)
- [Nachfolgende Körperschaft](#)
- [Nachfolgendes Geografikum](#)
- [Nördlichster Breitengrad](#)

Preview

Ort	Ländercode entfernen einrichten	Definition entfernen einrichten
Mittenwald	Bayern	Ort im Landkreis Partenkirchen
Mittenwald	Bayern	Ort im Landkreis Partenkirchen
Mittenwald	Bayern	Ort im Landkreis Partenkirchen
Mittenwald	Bayern	Ort im Landkreis Partenkirchen
Strömsholm	Schweden	Ort in der Gemei Hallstahammar,
Strömsholm	Schweden	Ort in der Gemei Hallstahammar,
Strömsholm	Schweden	Ort in der Gemei Hallstahammar,

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

