

Datenaufbereitung für imdas pro mit OpenRefine

23. MusIS-Nutzertreffen

13.09.2023

Claus Werner

claus.werner@hdgbw.de



Haus der Geschichte
Baden
Württemberg

Agenda

- Was ist OpenRefine?
- Anwendungsbeispiele
- Empfohlene Tutorials



Was ist OpenRefine?

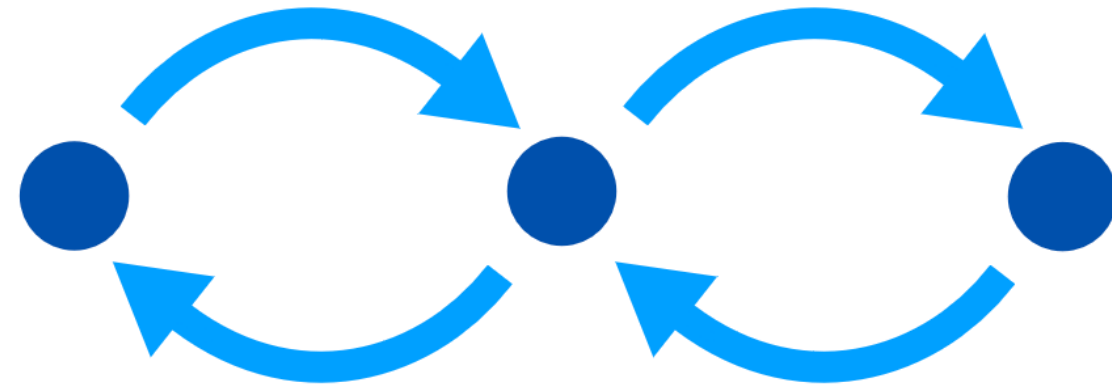


OpenRefine

- kostenlose Open-Source-Software zur Bereinigung und Aufbereitung von Daten
- ermöglicht es, Datensätze zu sichten, zu transformieren, mit externen Daten anzureichern und zu exportieren
- Entwickler: google, jetzt: Community
- Systemvoraussetzungen: Plattformunabhängig, Java, Browser
- **Wichtig:** Benutzeroberfläche öffnet sich im Browser, das Programm läuft aber lokal



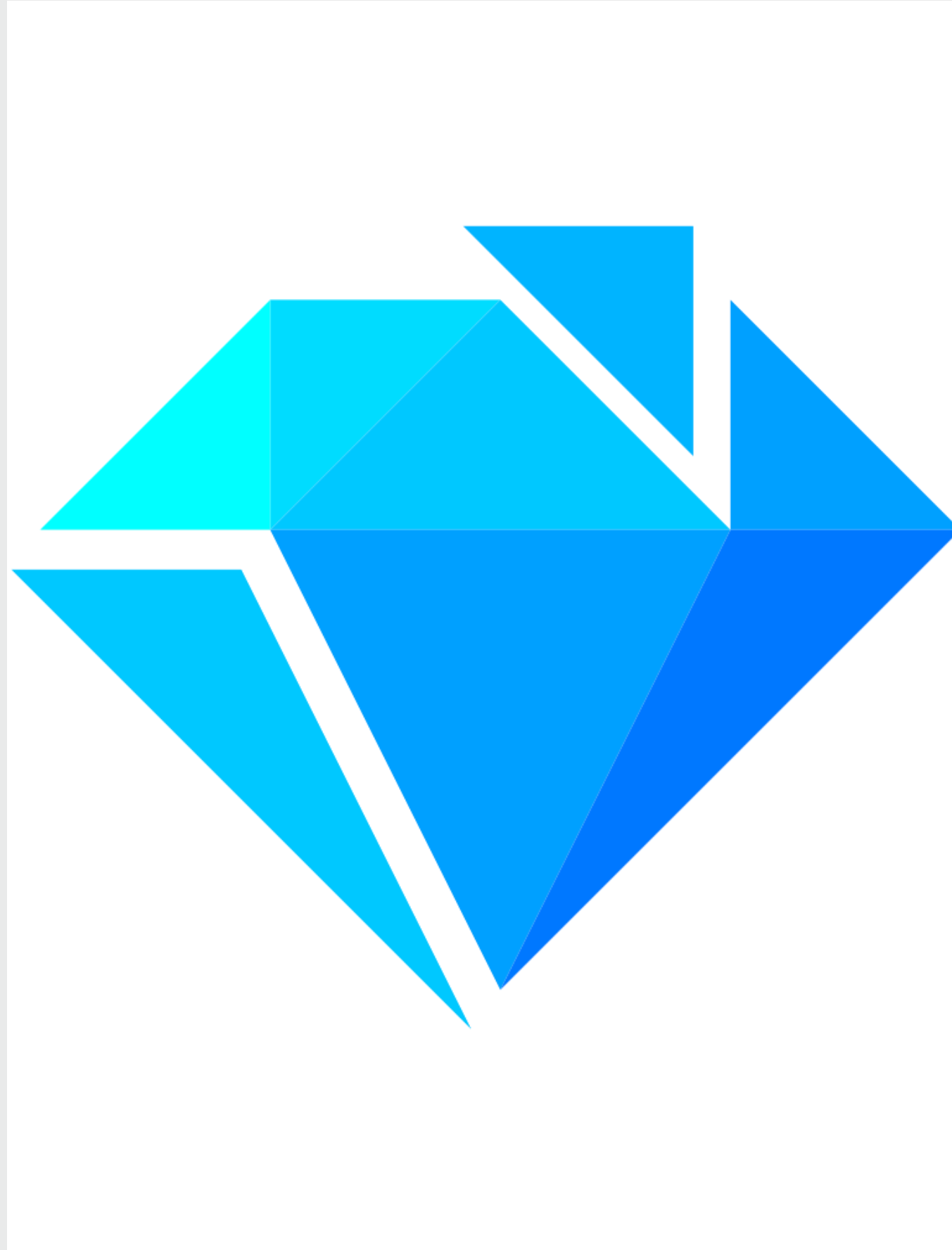
Gute Lernkurve



- Originaldatensatz wird nicht verändert
- Vorschau
- Arbeitsschritte automatisch geloggt, einfaches Undo
- Menüs
- Gute Dokumentation (intern und online), aktive Community (Tutorials, Youtube)



Vielfältigkeit



- In der Bedienung:
 - Voreingestellte Funktionen über Menüs
 - GREL
 - Jython / Clojure
- Import/Export:
 - Dateiformate: CSV, TSV, Text files, JSON, XML, ODS, XLS, XLSX , MARC, RDF, Wikitext, ...
 - Datenquellen: Lokale Datei(en), ZIP Archive, URL, Clipboard, SQL, Google Sheet, Wikidata, ...
- Extensions



Live Demo

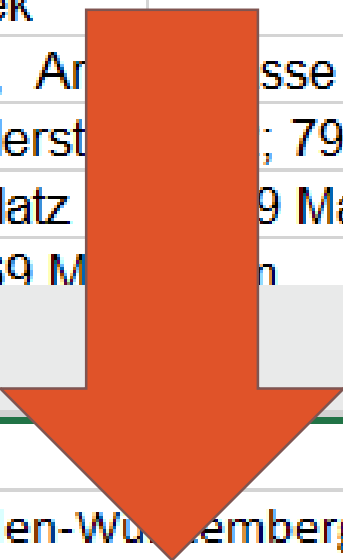
Adressangaben aus
Objektdatensätzen als
Körperschaften nach IMDAS
importieren

1. Daten vereinheitlichen
2. Informationen aufteilen
3. Normdatenabgleich
4. Thesauruspfad via
GeoNamesID ergänzen



<https://github.com/HdGBW>

Inventarnummer	Herkunft_Leihgeber
1989/1120/121	Landesmedienzentrum, Rotenbergstraße 111, 70190 Stuttgart
1989/1433	Institut für Stadtgeschichte, Markgrafenstr. 29, Karlsruhe
1991/0439	Institut für Stadtgeschichte, Markgrafenstr. 29, 76124 Karlsruhe
2001/1254/22824/02V	Landesmedienzentrum Baden-Württemberg, Rotenbergstraße 111, 70190 Stuttgart
2001/2635/01-03	Institut für Stadtgeschichte Karlsruhe, Markgrafenstr.29
2001/2643	Landesmedienzentrum Württemberg, Rotenbergstr.111, Stuttgart
2002/0530	Österreichische Nationalbibliothek, Josefsplatz 2, A-1015 Wien
2002/0860/01-02	Schweizerisches Bundesarchiv, Archivstr. 24, 3003 Bern
2003/0358	Österreichische Nationalbibliothek
2007/0445	Schweizerisches Bundesarchiv, Archivesstrasse 24, Ch-3003 Bern
2010/0373/0860	Archiv Soziale Bewegungen; Adlerstraße 12; 79098 Freiburg
2010/1859/04/02/10	Marchivum Mannheim; Archivplatz 1; 68169 Mannheim
2014/0041	MARCHIVUM Mannheim; Archivplatz 1; 68169 Mannheim



Nachname	Strasse	PLZ	Ort	Normdaten
Archiv Soziale Bewegungen	Adlerstraße 12	79098	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezirk Freiburg <RB>§Südlicher Oberrhein <Reg>§Stadtkreis Freiburg im Breisgau <Kr>§Freiburg im	O-GND~68855-1~https://d- nb.info/gnd/68855-1
Bundesarchiv Berlin	Finckensteinalle e 63	12205	Deutschland <DE>§Berlin <BL>§Berlin <Kr>§Berlin <Gm>§Berlin <O>	O-GND~10187558-7~https://d- nb.info/gnd/10187558-7
Maison de Victor Hugo	6 Place des Vosges	F-75004	Erde§Europa§Frankreich§Île-de-France§Paris§Paris	O-GND~1019177-X~https://d- nb.info/gnd/1019177-X
MARCHIVUM	Archivplatz 1	68169	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezirk Karlsruhe <RB>§Rhein-Neckar <Reg>§Stadtkreis Mannheim <Kr>§Mannheim	O-GND~1175461512~https://d- nb.info/gnd/1175461512
Österreichische				O-GND~2020893-5~https://d-



ermalink

25 rows

Show a

Cluster and edit column "Herkunft_Leihgeber"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method **Key collision** Keying function **Fingerprint** **3 clusters found**

Cluster size	Row count	Values in cluster	Merge?	New cell v	# Choices in cluster
3	4	<ul style="list-style-type: none"> Bundesarchiv Berlin; Finckensteinallee 63; 12205 Berlin (2 rows) BUnDesarchiv; Finckensteinallee 63; 12205 Berlin Bundesarchiv ; Finckensteinallee 63; 12205 Berlin 	<input type="checkbox"/>	Bundesarc	
2	3	<ul style="list-style-type: none"> Schweizerisches Bundesarchiv; Archivstrasse 24; Ch-3003 Bern (2 rows) Schweizerisches Bundesarchiv Bern; Archivstrasse 24; CH-3003 Bern 	<input type="checkbox"/>	Schweizer	
2	2	<ul style="list-style-type: none"> Marchivum Mannheim; Archivplatz 1; 68169 Mannheim Marchivum; Archivplatz 1; 68169 Mannheim Browse this cluster	<input type="checkbox"/>	Marchivum	

Select all Deselect all Export clusters **Merge selected & re-cluster** Merge selected & Close Close

1. Daten vereinheitlichen

- Einfache Datenbereinigungen (whitespace, replace)
- Daten sichten und angleichen (Text Facet, cluster)
- Dubletten entfernen (sort, blank down)



11 rows

Show as: **rows** records Show: 5 10 **25** 50 100 500 1000 rows

▼ All ▼ Herkunft_Leihgeber

☆	🗨	1.	Archiv Soziale Bewegungen; Adlerstraße 12; 79098 Freiburg
☆	🗨	2.	Bibliothèque du film française; 51 Rue du Bercy; F-75012 Paris
☆	🗨	3.	Bundesarchiv; Finckensteinallee 63; 12205 Berlin
☆	🗨	4.	Institut für Stadtgeschichte; Markgrafenstr. 29; 76124 Karlsruhe
☆	🗨	5.	Landesmedienzentrum Baden-Württemberg; Bergstraße 111; 70190 Stuttgart
☆	🗨	6.	Maison de Victor Hugo; 6 Place des Vosges; F-75004 Paris
☆	🗨	7.	MARCHIVUM; Archivplatz 1; D-68169 Mannheim

7 rows

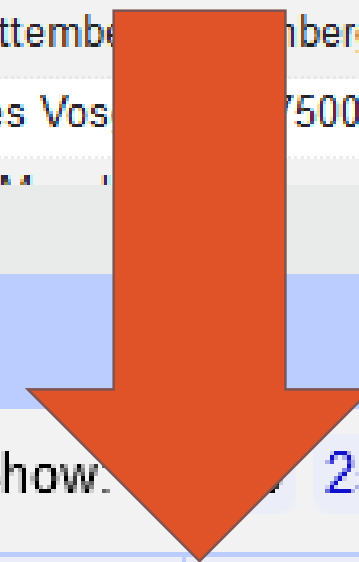
Show as: **rows** records Show: 25 50 100 500 1000 rows

▼ All ▼ Nachname ▼ Adresse ▼ PLZ ▼ Ort

☆	🗨	1.	Archiv Soziale Bewegungen	Adlerstraße 12	79098	Freiburg
☆	🗨	2.	Bundesarchiv Berlin	Finckensteinallee 63	12205	Berlin
☆	🗨	3.	Maison de Victor Hugo	6 Place des Vosges	F-75004	Paris
☆	🗨	4.	MARCHIVUM	Archivplatz 1	68169	Mannheim
☆	🗨	5.	Österreichische Nationalbibliothek	Josefsplatz 2	A-1015	Wien
☆	🗨	6.	Schweizerisches Bundesarchiv	Archivstrasse 24	Ch-3003	Bern
☆	🗨	7.	Stadtarchiv Dornbirn	Marktplatz 11	A-6850	Dornbirn

2. Informationen aufteilen

- Auf mehrere Spalten verteilen (split, rename)



3. Normdatenabgleich

11 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

« first < previous 1

All	Nachname	Strasse	P
★	1. Archiv Soziale Bewegungen <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Archiv für Soziale Bewegungen (106) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Archiv Soziale Bewegungen in Baden (Freiburg im Breisgau) (62) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item Search for match	Adlerstraße 12	79
★	2. Bibliothèque du film française <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item Search for match		F-750
★	3. Bundesarchiv <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv (Bern) (54) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Schweizerisches Bundesarchiv (54) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv (Koblenz) (54) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv Potsdam (54) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv-Militärarchiv (54) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Schweizerisches Bundesarchiv. Dienst GEVER (49) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv (Koblenz). Abteilung Potsdam (49) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Bundesarchiv (Koblenz). Außenstelle Ludwigsburg (47)		1:

Match this cell Match all identical cells Cancel

Archiv Soziale Bewegungen in Baden (Freiburg im Breisgau) (68855-1)

Freiburg im Breisgau

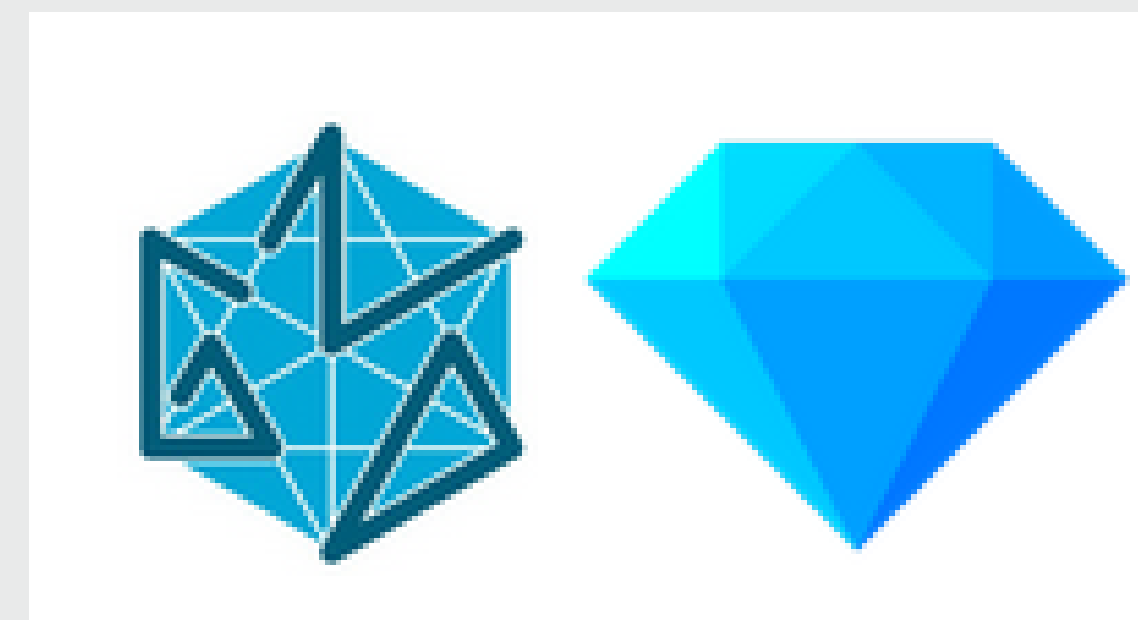
Körperschaft

- GND-ID der Institutionen via [Webservice](#) ermitteln (reconcile)

- Für die GND:
<https://lobid.org/gnd/reconcile/>

- Syntax für Imdas-Import:

O-GND~{GND-ID}~https://d-nb.info/gnd/{GND-ID}



	A	B	C	D	E
1	THESAURUS	KATEGORIE	NAME	PFAD	GeoNamesID
2	Orte	BSZ-Ortsthesaurus_DE	Berlin <O>	Deutschland <DE>§Schleswig-Holstein <BL>§Kreis Segeberg <	2950158
3	Orte	BSZ-Ortsthesaurus_DE	Berlin <O>	Deutschland <DE>§Berlin <BL>§Berlin <Kr>§Berlin <Gm>§Berli	2950159
4	Orte	BSZ-Ortsthesaurus_DE	Freiburg <O>	Deutschland <DE>§Niedersachsen <BL>§Landkreis Stade <Kr>	2925181
5	Orte	BSZ-Ortsthesaurus_DE	Freiburg <O>	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezi	2925177
6	Orte	BSZ-Ortsthesaurus_DE	Karlsruhe <O>	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezi	2892794
7	Orte	BSZ-Ortsthesaurus_DE	Karlsruhe <O>	Deutschland <DE>§Brandenburg <BL>§Landkreis Prignitz <Kr>	2892793
8	Orte	BSZ-Ortsthesaurus_DE	Karlsruhe <O>	Deutschland <DE>§Mecklenburg-Vorpommern <BL>§Mecklen	2892792
9	Orte	BSZ-Ortsthesaurus_DE	Mannheim <O>	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezi	2873891
10	Orte	BSZ-Ortsthesaurus_DE	Stuttgart <O>	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezi	2825297
11	Orte	Geonames	Paris	Erde§Europa§Frankreich§Île-de-France§Paris§Paris	2988507

4. Thesauruspfad

- Import von Thesaurusbegriffen in IMDAS über den Thesauruspfad
- Bsp. Berlin:
Deutschland <DE>§Berlin
<BL>§Berlin <Kr>§Berlin
<Gm>§Berlin <O>
- Benötigt: IMDAS -Export des Ortsthesaurus mit GeoNamesID
- Reconciliation mit GeoNames:
<https://fornpunkt.se/apis/reconciliation/geonames>
- Abgleich durch cross-Funktion in OpenRefine

Custom text transform on column GeoNamesID

Expression Language No syntax error.

```
cell.cross("Thesauruspfade", "GeoNamesID")
[0].cells["PFAD"].value
```

Preview History Starred Help

row	value	cell.cross("Thesauruspfade", " ...
1.	2925177	Deutschland <DE>§Baden-Württemberg <BL>§Regierungsbezirk Freiburg <RB>§Südlicher Oberrhein <Reg>§Stadtkreis Freiburg im Breisgau <Kr>§Freiburg im Breisgau <Gm>§Freiburg <O>
2.	2988507	Erde§Europa§Frankreich§Île-de-France§Paris§Paris



Tutorials und Dokumentation

- Offizielles Manual: <https://openrefine.org/docs>
- Liste mit Tutorials: <https://github.com/OpenRefine/OpenRefine/wiki/External-Resources>
- Library Carpentry: <https://librarycarpentry.org/lc-open-refine/>
- Programming Historian: <https://programminghistorian.org/en/lessons/>
- FDMLab@LABW:
 - <https://fdmlab.landesarchiv-bw.de/workshop/openrefine-einsteiger/warum-openrefine/>
 - <https://fdmlab.landesarchiv-bw.de/workshop/openrefine-fortgeschrittene/warum-openrefine/>
 - <https://fdmlab.landesarchiv-bw.de/post/>

