

Allgemeine Beschreibung

OpenRefine ist ein Open-Source-Tool zur Datenbereinigung und -transformation, das ursprünglich als Google Refine bekannt war. Es bietet eine anwenderfreundliche grafische Benutzeroberfläche, mit der Daten in verschiedenen Formaten analysiert, bereinigt und strukturiert werden können.

OpenRefine eignet sich besonders gut für die Arbeit mit großen und unstrukturierten Datensätzen, wobei es das Filtern, Sortieren und Gruppieren von Daten sowie das Erkennen und Beheben von Fehlern und Unregelmäßigkeiten ermöglicht. Das Tool unterstützt außerdem auch die Zusammenführung von Datensätzen aus verschiedenen Quellen und das Aufteilen von Zellen, um Daten besser zu organisieren.

Der Vorteil von OpenRefine hinsichtlich digitaler Editionen ist, dass es nicht nur die Datenbereinigung, -transformation und -organisation großer unstrukturierter Datenmengen erleichtert, sondern vor allem auch Funktionen zur Normalisierung von Daten sowie zur Konsolidierung von Informationen bietet. Beim Export der Daten muss man auf die Möglichkeit, eine XML-Datei herunterzuladen, verzichten und auch etwas komplexere Datentransformationen beim Export - wie beispielsweise das Gruppieren von Daten - werden nicht unterstützt.

Anwendungsbereiche

- Bereinigung unstrukturierter und fehlerhafter Daten
- Zusammenführung und Konsolidierung von Daten aus verschiedenen Quellen
- Normalisierung von bestehenden Datenbeständen

Funktionsübersicht

- Datenbereinigung bei unstrukturierten und fehlerhaften Daten; erkennt Dubletten, Tippfehler, Inkonsistenzen und andere Unregelmäßigkeiten
- Datennormalisierung
- Datentransformation (z. B. Excel/CSV-Input zu JSON oder XML-Struktur)
- Datenzusammenführung, wenn verschiedene Quellen vorhanden sind
- Möglichkeit der Strukturierung von Metadaten
- Datenvisualisierung
- Automatisierung von wiederholten Datenbereinigungs- und Transformationsaufgaben durch die Erstellung von Skripten oder Aktionen für bestimmte Aufgaben

Voraussetzungen

Jedes Tool kann einerseits bestimmte Vorkenntnisse der Benutzer:innen voraussetzen und andererseits auch hinsichtlich der Software-Umgebung gewisse Anforderungen stellen.


Erforderliche Kenntnisse

- Ausdruckssprachen und Transformationstechniken von Vorteil

Benötigte Software

- Stabile Internetverbindung
- Webbrowser

Tool-Kompatibilität

	IIIF	Transkribus	FromThePage	ediarum	ba[sic?]	teiPublisher	ediarum.WEB
OpenRefine	✗	✗	✗		✗	✗	✗

Kostenübersicht

- kostenlos

Möglichkeiten & Grenzen

Da jedes Projekt unterschiedliche Anforderungen mit sich bringt, sollen nachfolgend mögliche Vor- und Nachteile des getesteten Tools dargestellt werden.

Stärken

- Benutzerfreundliche Bearbeitungsoberfläche und Wahrung der Datensicherheit durch die Bearbeitung am eigenen Rechner
- Datenbereinigung: OpenRefine kann bei der Bereinigung von unstrukturierten und fehlerhaften Daten helfen, indem es Dubletten, Tippfehler, Inkonsistenzen und andere Unregelmäßigkeiten erkennt und korrigiert. Außerdem bietet es die Möglichkeit, Arbeitsschritte wieder rückgängig zu machen, aber auch bereits getätigte Schritte wiederherzustellen oder den Änderungsverlauf zu exportieren und auf neue Daten anzuwenden.
- Datenerweiterung und -normalisierung: Über Reconciliation-Services können die Daten mit externen Datenbanken abgeglichen und mit Normdaten angereichert werden.
- Datentransformation: OpenRefine bietet Funktionen zur Transformation von Daten in andere Formate oder Strukturen. Dies kann nützlich sein, um Daten für bestimmte Anforderungen anzupassen oder in verschiedenen Datenbanken oder Plattformen zu verwenden.
- Datenzusammenführung: Wenn eine digitale Edition aus mehreren Quellen oder Versionen besteht, kann OpenRefine verwendet werden, um diese Daten zusammenzuführen, Dubletten zu entfernen und eine einheitliche Version zu erstellen.
- Strukturierung von Metadaten: OpenRefine ermöglicht die Organisation und Strukturierung von Metadaten, um eine bessere Durchsuchbarkeit und Navigation in der digitalen Edition zu gewährleisten. Dies kann die Indexierung und den Zugriff auf bestimmte Inhalte erleichtern.
- Datenvisualisierung: Diagramme, Grafiken und andere visuelle Darstellungen können Muster und Zusammenhänge in den Daten verdeutlichen.
- Qualitätssicherung: OpenRefine unterstützt die Überprüfung der Datenqualität, indem es Inkonsistenzen und Fehler identifiziert. Dies ermöglicht es den Herausgebern, Probleme zu beheben und sicherzustellen, dass die digitale Edition genau und zuverlässig ist.
- Automatisierung: OpenRefine kann auch bei der Automatisierung von wiederholten Datenbereinigungs- und Transformationsaufgaben helfen. Durch die Erstellung von Skripten oder Aktionen können bestimmte Aufgaben automatisiert werden, was Zeit und Mühe spart.

Herausforderungen & Probleme

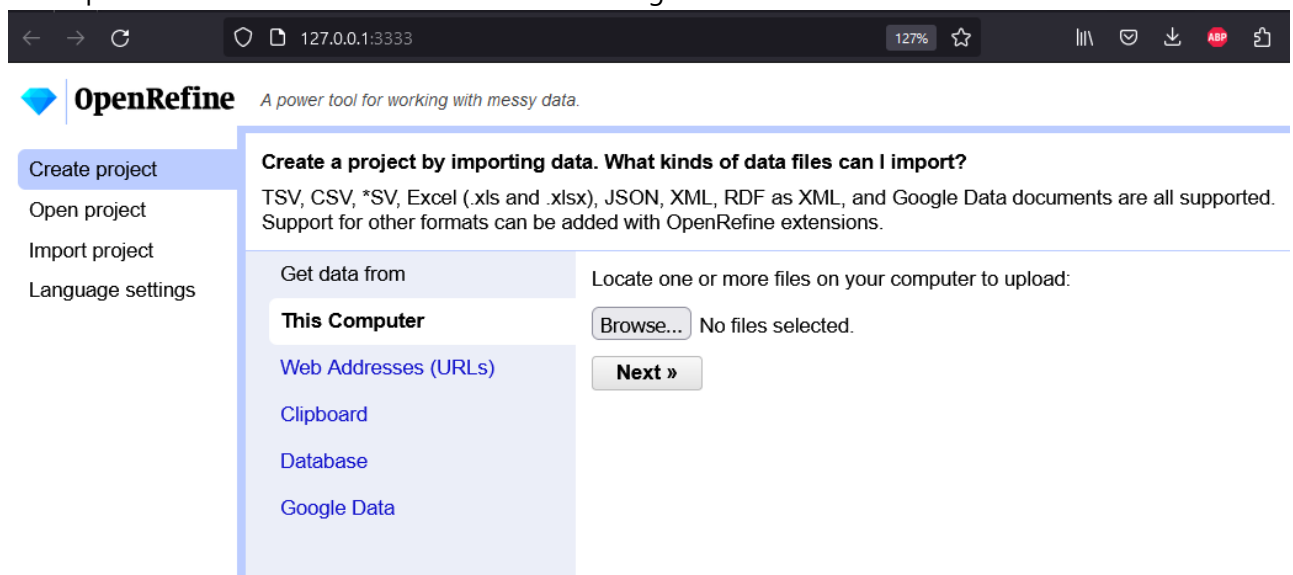
- Keine simultane Kollaborationsmöglichkeiten: Da OpenRefine für die lokale Verwendung konzipiert ist, können nicht mehrere Personen gleichzeitig an einem Projekt arbeiten. Die beste - aber bei vielen Mitarbeitenden relativ umständliche - Möglichkeit, mit einer anderen Person zusammenzuarbeiten, besteht daher darin, Projekte inklusive der gespeicherten Bearbeitungsschritte zu exportieren und daraufhin an einem anderen Rechner zu importieren, sodass man dort weitermachen kann, wo jemand anderes aufgehört hat.
- Teilweise mühsame Bedienung: Bei der manuellen Zuordnung von passenden Wikidata-Einträgen springt das Programm nach jeder einzelnen Übernahme zum Start der Tabelle, wodurch jedes Mal ein Scrollen zum zuletzt bearbeiteten Begriff notwendig ist.
- Keine direkte XML-Exportmöglichkeit: Der Export in ein XML-Dateiformat ist nicht vorgesehen. Über den Templating-Export können die Daten jedoch zumindest in einer XML-Struktur (als Plaintext-Datei) exportiert werden.
- Komplexere Datentransformationen - wie beispielsweise das Gruppieren von Datensätzen anhand des Inhalts einer Zelle - sind beim Export nicht möglich, wodurch Redundanzen in den Daten auftreten können und eine Nachbearbeitung erforderlich sein kann.

Einrichtung & Erste Schritte

Anhand unseres Beispielprojekts, das zum Ziel hat, Kochrezepte aus dem Mittelalter computergestützt zu analysieren und später über eine Forschungsplattform zur Verfügung zu stellen, soll nachfolgend ein möglicher Arbeitsablauf beschrieben werden. Die Manuskripte des Projektes wurden bereits mittels [FromThePage](#) transkribiert und mit [ediarum](#) erfolgten bereits erste Annotationen. In dieser Kurzanleitung erfolgt nun die Aufbereitung der Zutatenliste, die wir von einem Historiker im CSV-Format erhalten haben. Unser Ziel ist es, die Daten zu normalisieren und sie zusätzlich mit [Q-Nummern](#) - auch QID genannt - von Wikidata-Einträgen anzureichern.

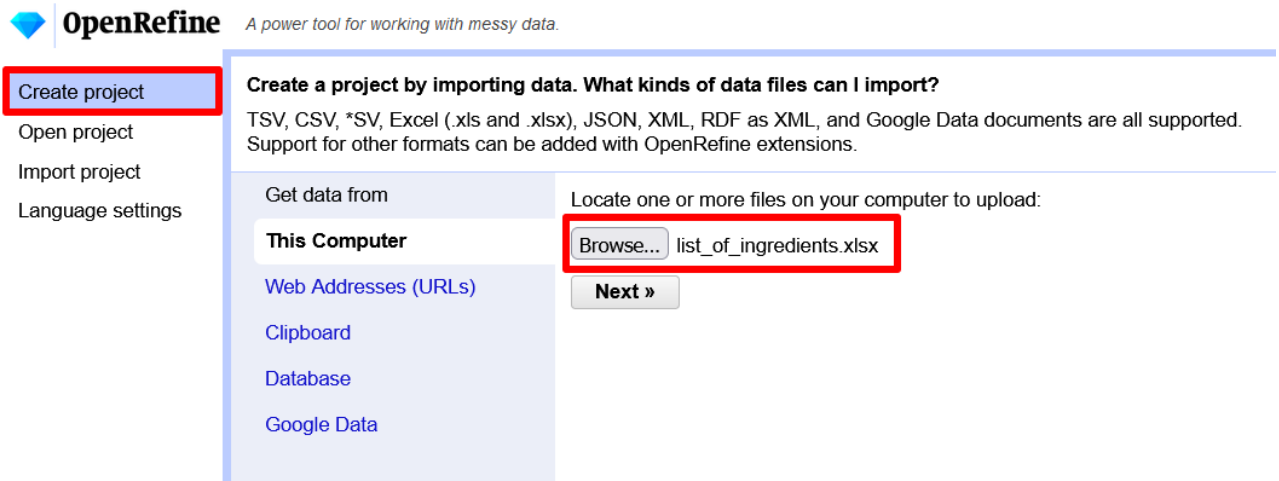
1. Installation

- Unser erster Schritt besteht darin, uns die entsprechende Version für unser Betriebssystem von [OpenRefine herunterzuladen](#). Nach dem Entpacken der ZIP-Datei haben wir openrefine.exe ausgeführt und OpenRefine hat sich direkt in unserem Browser geöffnet.



2. Einrichtung des Projekts

- Um ein Projekt erstellen zu können, werden wir aufgefordert, Daten zu importieren. Wir laden daher als erstes unsere **EXCEL-Datei mit der Zutatenliste** hoch.



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

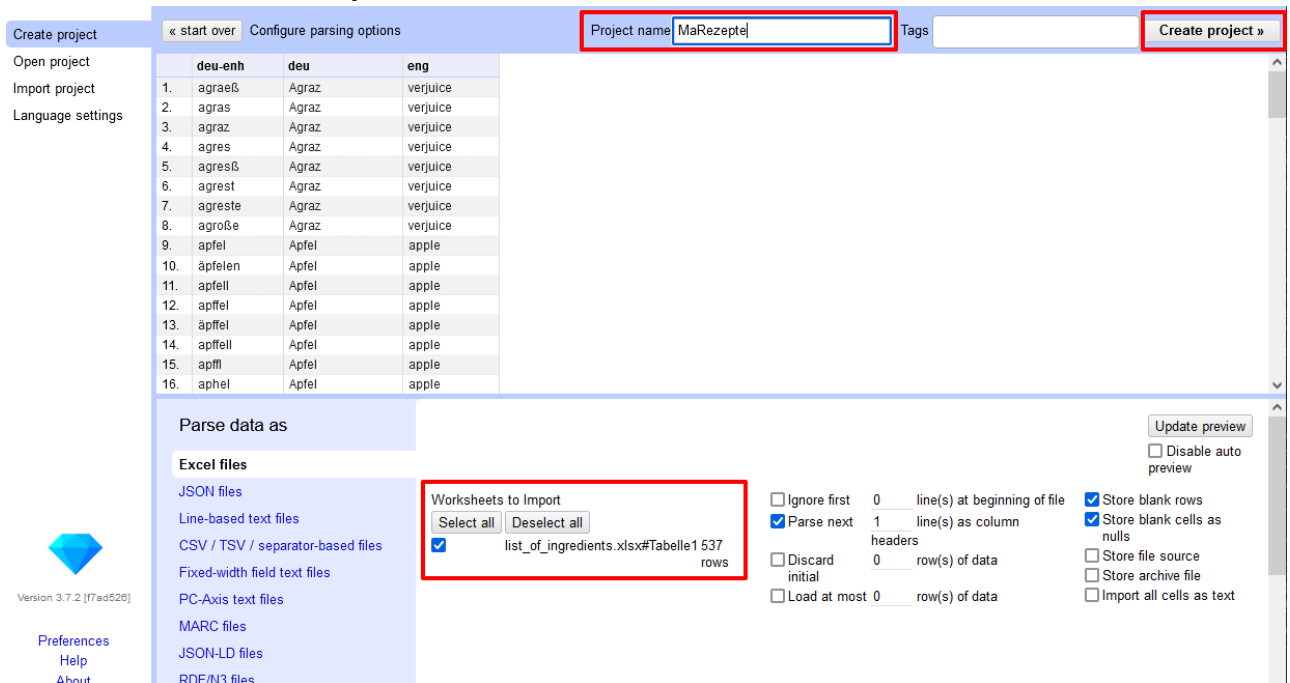
Web Addresses (URLs)
Clipboard
Database
Google Data

Locate one or more files on your computer to upload:

Browse... list_of_ingredients.xlsx

Next »

- Mit dem Button "Next" kommen wir in die darauffolgende Ansicht und können einige Einstellungen vornehmen, bevor unser Projekt erstellt wird.



« start over Configure parsing options

Project name **MaRezeptel** Tags **Create project »**

	deu-enh	deu	eng
1.	agraeß	Agraz	verjuice
2.	agras	Agraz	verjuice
3.	agraz	Agraz	verjuice
4.	agres	Agraz	verjuice
5.	agresß	Agraz	verjuice
6.	agrest	Agraz	verjuice
7.	agreste	Agraz	verjuice
8.	agroße	Agraz	verjuice
9.	apfel	Apfel	apple
10.	äpfelen	Apfel	apple
11.	apfell	Apfel	apple
12.	apffel	Apfel	apple
13.	äpfel	Apfel	apple
14.	apfell	Apfel	apple
15.	apffl	Apfel	apple
16.	aphel	Apfel	apple

Parse data as

Excel files

JSON files
Line-based text files
CSV / TSV / separator-based files
Fixed-width field text files
PC-Axis text files
MARC files
JSON-LD files
RDF/N3 files

Worksheets to Import

Select all Deselect all

☒ list_of_ingredients.xlsx#Tabelle1 537 rows

☐ 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

☒ Store blank rows

☒ Store blank cells as nulls

☐ Store file source

☐ Store archive file

☐ Import all cells as text

Update preview

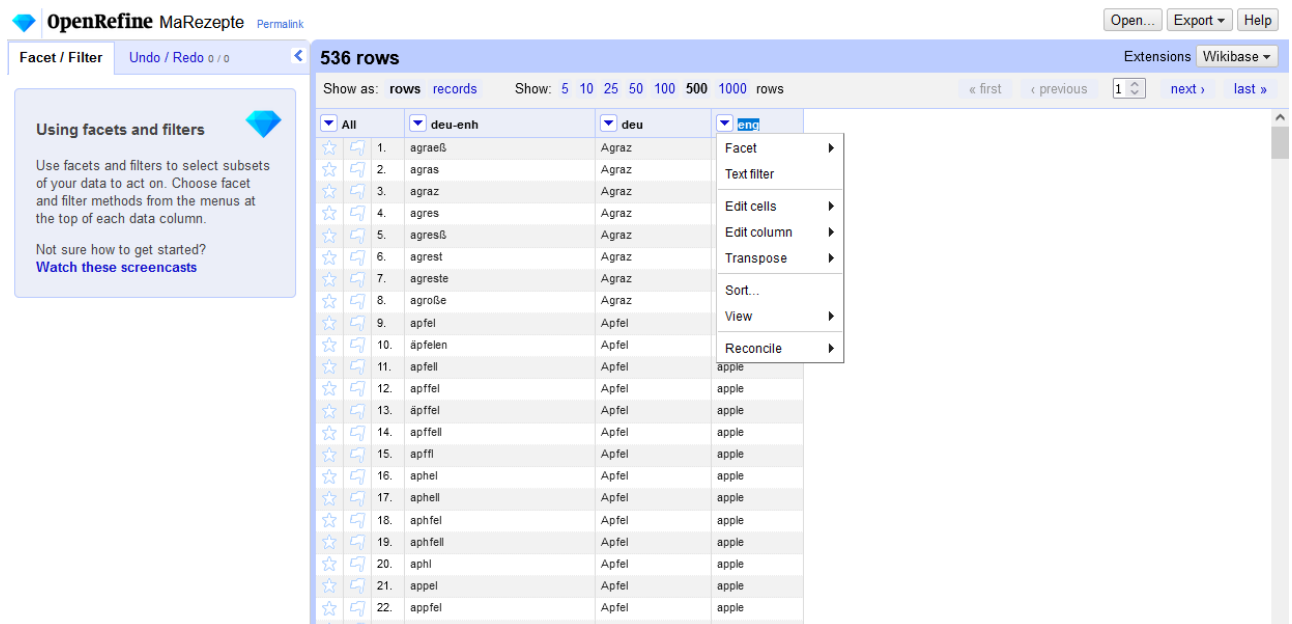
☐ Disable auto preview

Version 3.7.2 [f7ad526]

Preferences
Help
About

→ Für unser Projekt haben wir die vorausgewählten Einstellungen belassen und nur einen Projektnamen gewählt, bevor wir mit "Create project" fortgefahren sind.

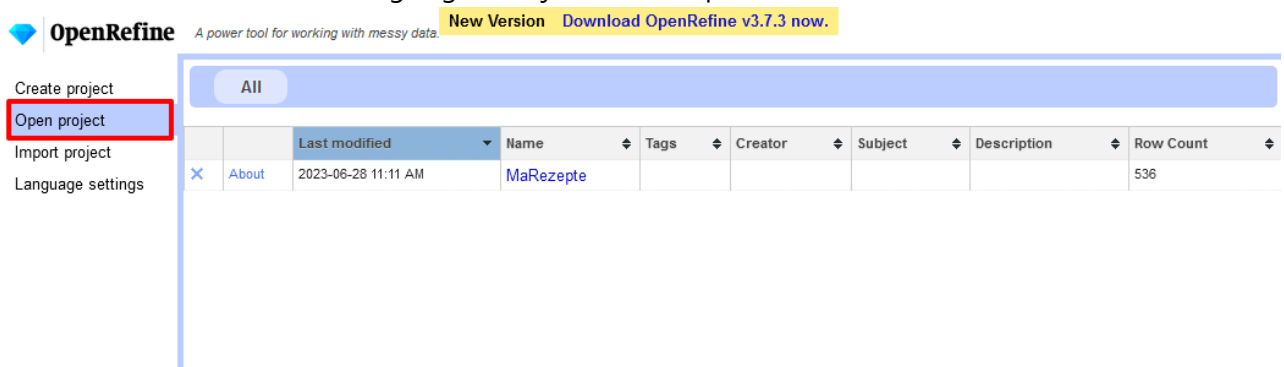
- Unsere Projektansicht sieht letztlich so aus:



→ Die Einträge aus der CSV-Datei werden tabellarisch dargestellt. In der ersten Spalte sind verschiedene frühneuhochdeutsche Schreibvarianten einzelner Zutaten, in der zweiten Spalte die heutige Schreibweise und in der dritten Spalte Übersetzungen in modernes Englisch. Jede Spalte verfügt über ein Drop-Down-Menü, das uns verschiedene Bearbeitungsmöglichkeiten bietet, wobei für uns vor allem die Funktion, die eine Anreicherung mit Normdaten (Reconciliation) ermöglicht, von Interesse ist.

3. Bearbeitung der Dokumente

- Sollten wir zwischenzeitlich unser Projekt geschlossen haben, müssen wir für die Arbeit in OpenRefine zuerst wieder unsere Datei openrefine.exe starten, über die erneut der Browser geöffnet wird. Unter **Open Project** in der Navigation auf der linken Seite können wir schließlich unsere Projekte einsehen. Wir öffnen hier unser bereits angelegtes Projekt "MaRezepte".



- Um unsere Zutatenliste mit Einträgen aus einer Normdatenbank anzureichern, überprüfen wir zuerst, welche Einträge auf Basis der Spalte mit den englischen Begriffen gefunden werden. Wir wählen hier das Englische, weil die englische Wikidata-Datenbank mit der größten Abdeckung an Begriffen zu einer höheren Trefferquote führt. Dafür wählen wir im Dropdown der Spalte mit der Überschrift "en" die

Option **Reconcile** und in der damit verbundenen Auswahl **Start Reconcile**.

OpenRefine MaRezepte [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

536 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 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	deu-enh	deu	eng
1.	agraeß	Agraz	Agraz	Facet
2.	agras	Agraz	Agraz	Text filter
3.	agraz	Agraz	Agraz	Edit cells
4.	agres	Agraz	Agraz	Edit column
5.	agresß	Agraz	Agraz	Transpose
6.	agrest	Agraz	Agraz	Sort...
7.	agreste	Agraz	Agraz	View
8.	agroße	Agraz	Agraz	
9.	apfel	Apfel	Apfel	
10.	äpfelen	Apfel	Apfel	Reconcile
11.	apfell	Apfel	Apfel	Start reconciling...
12.	apffel	Apfel	Apfel	Facets
13.	äpfel	Apfel	Apfel	Actions
14.	apffell	Apfel	Apfel	
15.	apffl	Apfel	Apfel	Copy reconciliation data...
16.	aphel	Apfel	Apfel	Use values as identifiers...
17.	aphell	Apfel	Apfel	Add entity identifiers column...
18.	aphfel	Apfel	Apfel	
19.	aphfell	Apfel	Apfel	
20.	aphl	Apfel	Apfel	
21.	appel	Apfel	Apfel	

- In dem neuen Fenster, das sich daraufhin öffnet, klicken wir in der linken Menüleiste auf "Wikidata (en)".

OpenRefine MaRezepte [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

536 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 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	deu-enh	deu	eng
1.	agraeß	Agraz	Agraz	Facet
2.	agras	Agraz	Agraz	Text filter
3.	agraz	Agraz	Agraz	Edit cells
4.	agres	Agraz	Agraz	Edit column
5.	agresß	Agraz	Agraz	Transpose
6.	agrest	Agraz	Agraz	Sort...
7.	agreste	Agraz	Agraz	View
8.	agroße	Agraz	Agraz	
9.	apfel	Apfel	Apfel	
10.	äpfelen	Apfel	Apfel	Reconcile
11.	apfell	Apfel	Apfel	Start reconciling...
12.	apffel	Apfel	Apfel	Facets
13.	äpfel	Apfel	Apfel	Actions
14.	apffell	Apfel	Apfel	
15.	apffl	Apfel	Apfel	Copy reconciliation data...
16.	aphel	Apfel	Apfel	Use values as identifiers...
17.	aphell	Apfel	Apfel	Add entity identifiers column...
18.	aphfel	Apfel	Apfel	
19.	aphfell	Apfel	Apfel	
20.	aphl	Apfel	Apfel	
21.	appel	Apfel	Apfel	

- In dem sich daraufhin öffnenden Fenster wählen wir folgende Einstellungen:
 - Bei der Kategorienzuordnung, mit der festgelegt werden kann, dass die Begriffe nur mit Entitäten einer bestimmten Kategorie abgeglichen werden, möchten wir uns nicht zu sehr einschränken. Wir könnten natürlich nur "food ingredients" auswählen, aber erstens sind nicht alle Entitäten einer Kategorie zugewiesen und zweitens ist die Kategorizuordnung nicht immer eindeutig, weshalb beispielsweise einer Zutat wie Petersilie anstelle der Kategorie "Zutat", auch einfach nur die Kategorie "Pflanze" zugeordnet sein könnte. Um zu verhindern, dass durch die Einschränkung auf eine bestimmte Kategorie möglicherweise unkategorisierte oder abweichend kategorisierte Entitäten nicht mit unseren Daten abgeglichen werden, nutzen wir die Option: "Reconcile against no particular type".

- Zusätzlich gibt es die Möglichkeit, über die Checkbox "Auto-match candidates with high confidence" einzustellen, dass bei jenen Begriffen, für die mit hoher Wahrscheinlichkeit eine passende Wikidata-Entität gefunden wurde, eine automatische Zuordnung vorgenommen wird.
- Mit diesen Einstellungen für unsere Daten wurde schließlich der Reconciliation-Prozess gestartet.

Reconcile column "eng"

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

- ☐ scholarly article (Q13442814)
- ☐ food ingredient (Q25403900)
- ☐ taxon (Q16521)
- ☐ edition of commercial catalogue (Q55089312)
- ☐ mountain (Q8502)
- ☐ enterprise (Q6881511)
- ☐ musical group (Q215380)
- ☐ dessert (Q182940)
- ☐ Japanese television drama

Also use relevant details from other columns:

Column Include? As property

deu-enh ☐

deu ☐

Reconcile against type:

☒ Reconcile against no particular type


☒ Auto-match candidates with high confidence

Maximum number of candidates to return

Add standard service... Discover services...

Start reconciling... Cancel

→ Dieser Prozess kann je nach Datenmenge ein paar Minuten dauern.

 **OpenRefine** MaRezepte [Permalink](#)

Reconcile cells in column eng to type null
33% complete [Cancel](#)

[Open...](#) [Export](#) [Help](#)

Facet / Filter

Undo / Redo 0 / 0

536 rows

Extensions Wikibase


Show as: rows records

Show: 5 10 25 50 100 500 1000 rows

« first < previous 1 next > last »

	All	deu-enh	deu	eng
☆	1.	agraeß	Agraz	verjuice
☆	2.	agras	Agraz	verjuice
☆	3.	agraz	Agraz	verjuice
☆	4.	agres	Agraz	verjuice
☆	5.	agresß	Agraz	verjuice
☆	6.	agrest	Agraz	verjuice
☆	7.	agreste	Agraz	verjuice
☆	8.	agroße	Agraz	verjuice
☆	9.	apfel	Apfel	apple
☆	10.	äpfelen	Apfel	apple
☆	11.	apfell	Apfel	apple

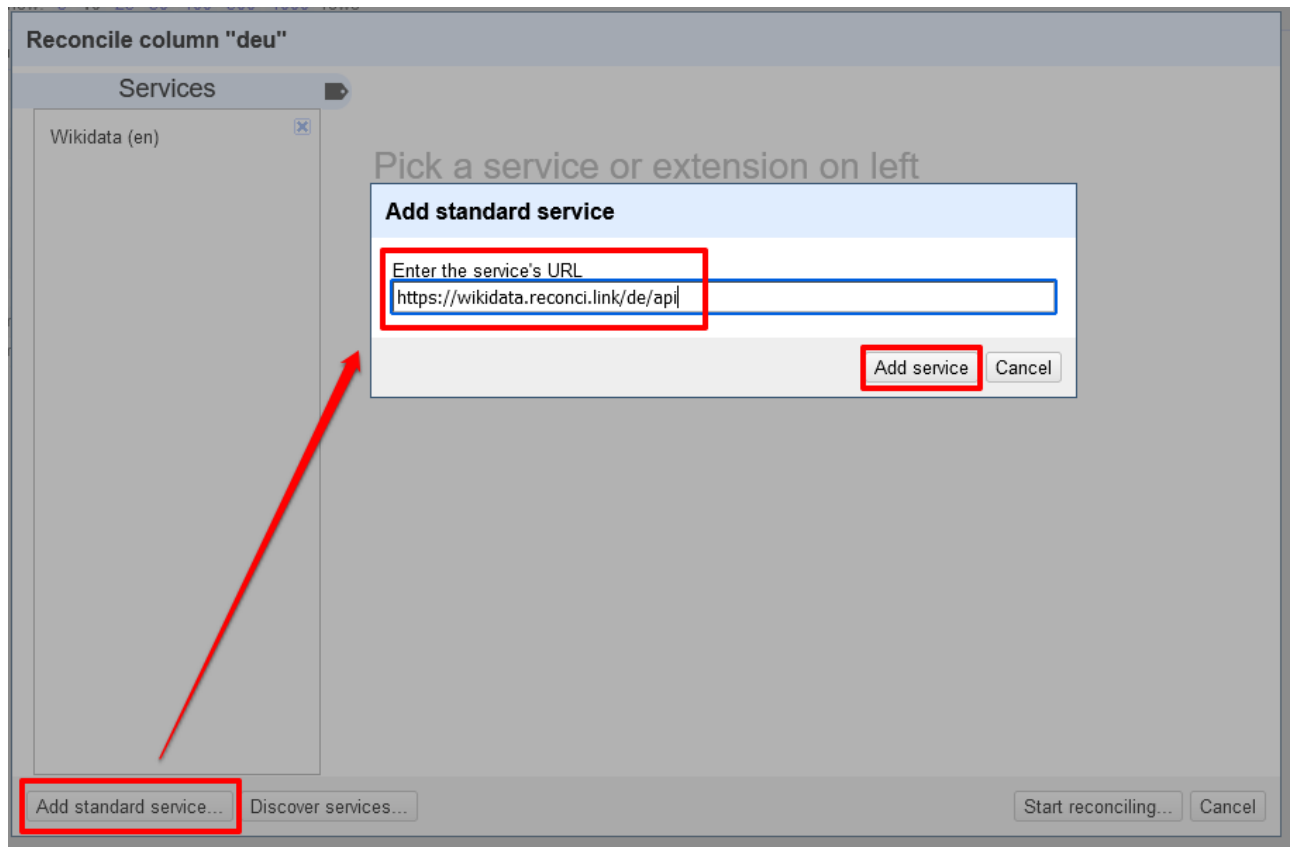
Using facets and filters



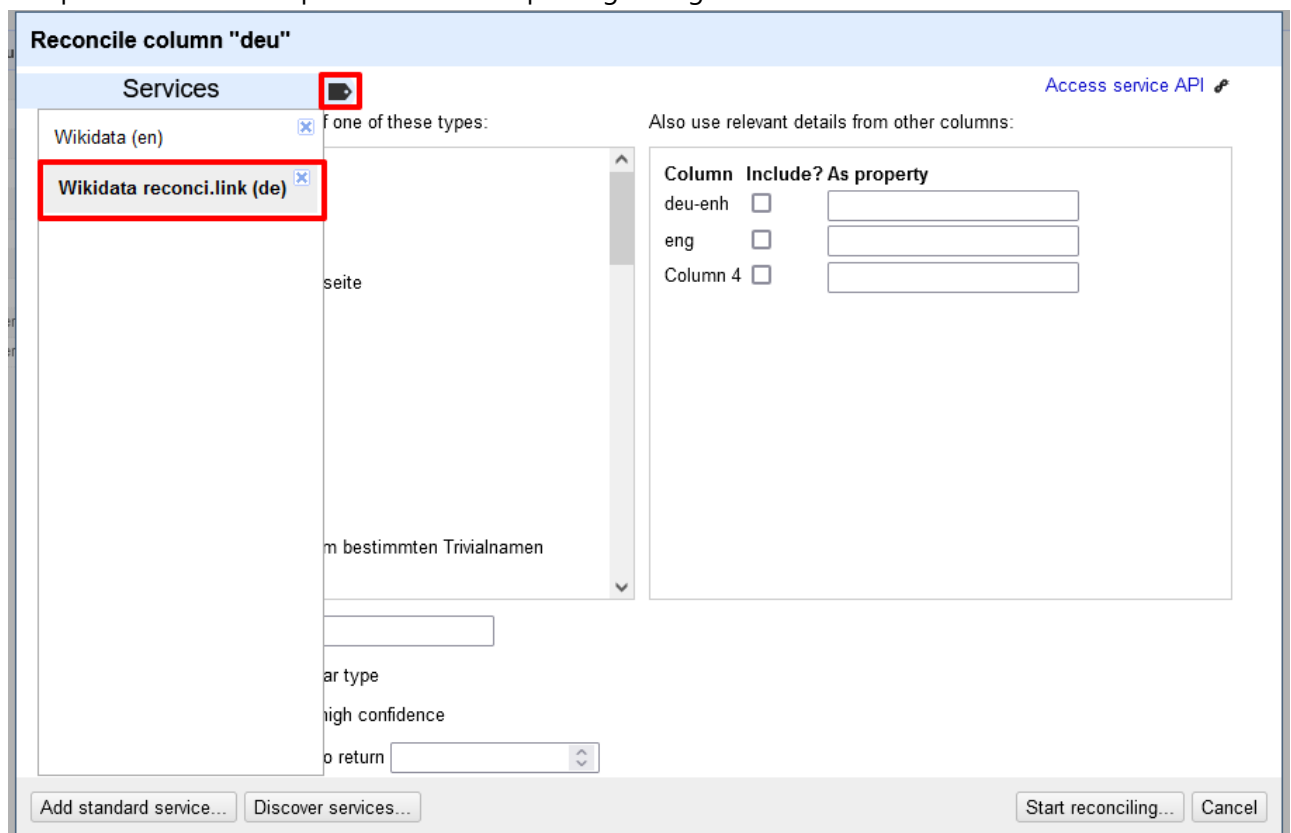
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

- **Kleiner Exkurs bei alternativen Daten:** Wenn wir die Begriffe nicht auch Englisch, sondern nur im Standarddeutsch hätten, müssten wir über den Button "Add standard service" ein weiteres Service für das deutsche Wikidata anlegen, indem wir die entsprechende URL zur API eingeben.



In unserer linken und über ein kleines Lesezeichen-Symbol ein- und ausklappbaren Liste erscheint nun ein Button für die Reconciliation von Begriffen mit deutschsprachigen Wikidata-Einträgen, die wir dann entsprechend für eine Spalte mit deutschsprachigen Begriffen auswählen könnten.



→ Hinter dem Button "Discover Services" verbergen sich außerdem [noch weitere Normdaten-Ressourcen](#).

- Sobald der Reconciliation-Prozess abgeschlossen ist, erhalten wir in der Header-Zeile der Spalte einen Überblick zu unserem Fortschritt in Form eines Balkens. Aus unserer Tabelle mit 536 Zeilen wurde knapp ein Fünftel automatisiert mit Normdaten angereichert und bei über 80% der Einträge ist noch eine

manuelle Überprüfung nötig, da es hier mehrere Entitäten gibt, die mit dem Begriff aus der jeweiligen Zeile übereinstimmen.

OpenRefine MaRezepte Permalink

Facet / Filter Undo / Redo 1 / 1

Refresh Reset all Remove all

536 rows

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

eng: judgment change

2 choices Sort by: name count

matched 100

none 436

Facet by choice counts

eng: best candidate's score change reset

72 — 101

	All	deu-enh	deu	eng
1.	agraeß	Agraz	verjuice	Choose new match
2.	agras	Agraz	verjuice	Choose new match
3.	agraz	Agraz	verjuice	Choose new match
4.	agres	Agraz	verjuice	Choose new match
5.	agresß	Agraz	verjuice	Choose new match
6.	agrest	Agraz	verjuice	Choose new match
7.	agreste	Agraz	verjuice	Choose new match
8.	agroße	Agraz	verjuice	Choose new match
9.	apfel	Apfel	apple	<input checked="" type="checkbox"/> apple (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> Muggsy Bogues (100) <input checked="" type="checkbox"/> Malus pumila (100) <input checked="" type="checkbox"/> Apple II series (100) <input checked="" type="checkbox"/> Apple Records (100) <input checked="" type="checkbox"/> Apple III (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> The Apple (100) <input checked="" type="checkbox"/> Apple River (100) <input checked="" type="checkbox"/> apple (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> Ariane Passenger Payload Experiment (100) <input checked="" type="checkbox"/> Apple II (100)

→ Zusätzlich bekommen wir in der linken Leiste Informationen zu den Matches und haben auch die Möglichkeit, den Prozess rückgängig zumachen.

- Bei allen Begriffen, für die nicht automatisch eine Entsprechung aus den Wikidata-Normaten übernommen wurde, müssen wir nun eine manuelle Zuordnung vornehmen. Durch die Übersetzung der verschiedenen Schreibweisen für einen konkreten Begriff haben wir im Englischen sehr viele gleiche Einträge. Damit wir nicht jeden Zeile einzeln durchgehen müssen, gibt es in OpenRefine die Möglichkeit, das Kästchen mit dem doppelten Häkchen zu verwenden, um den entsprechenden Wikidata-Eintrag für alle identischen Zellen zu übernehmen.

OpenRefine MaRezepte Permalink

Facet / Filter Undo / Redo 1 / 1

536 rows

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

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](#)

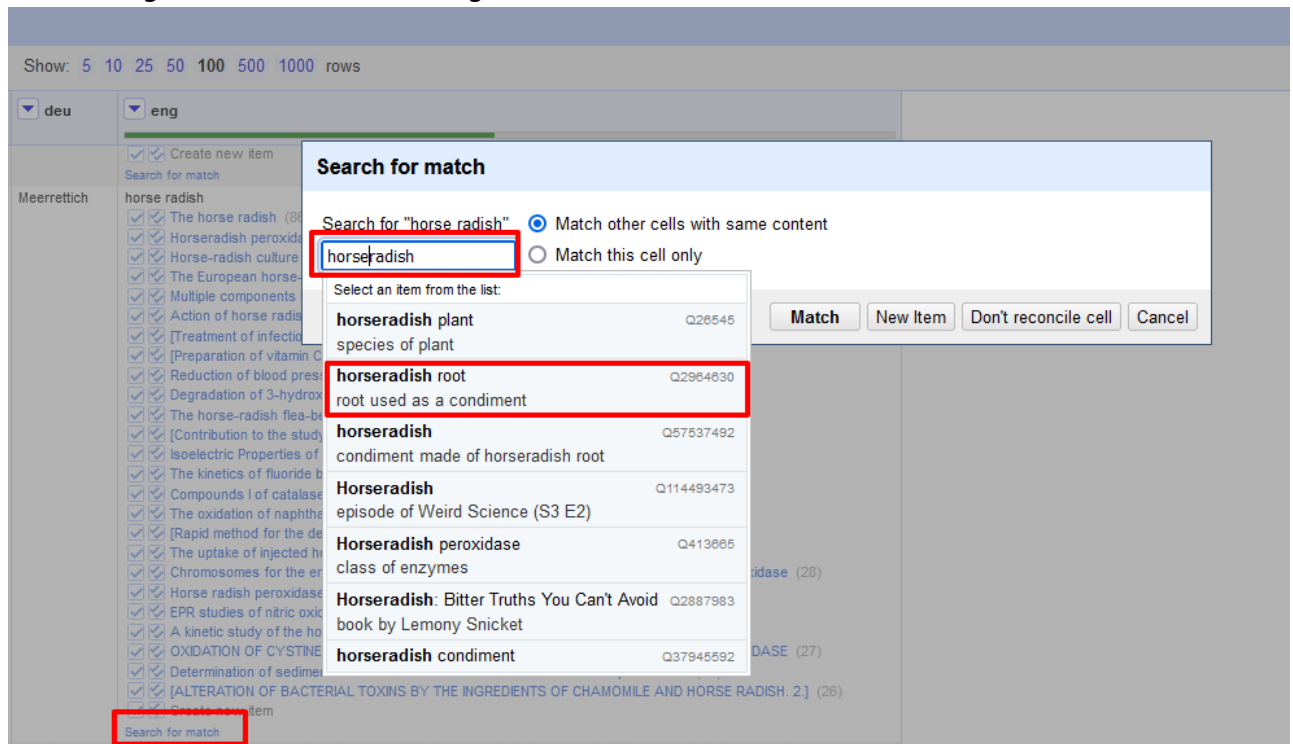
	All	deu-enh	deu	eng
1.	agraeß	Agraz	verjuice	Choose new match
2.	agras	Agraz	verjuice	Choose new match
3.	agraz	Agraz	verjuice	Choose new match
4.	agres	Agraz	verjuice	Choose new match
5.	agresß	Agraz	verjuice	Choose new match
6.	agrest	Agraz	verjuice	Choose new match
7.	agreste	Agraz	verjuice	Choose new match
8.	agroße	Agraz	verjuice	Choose new match
9.	apfel	Apfel	apple	<input checked="" type="checkbox"/> apple (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> Muggsy Bogues (100) <input checked="" type="checkbox"/> Malus pumila (100) <input checked="" type="checkbox"/> Apple II series (100) <input checked="" type="checkbox"/> Apple Records (100) <input checked="" type="checkbox"/> Apple III (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> The Apple (100) <input checked="" type="checkbox"/> Apple River (100) <input checked="" type="checkbox"/> apple (100) <input checked="" type="checkbox"/> Apple (100) <input checked="" type="checkbox"/> Ariane Passenger Payload Experiment (100) <input checked="" type="checkbox"/> Apple II (100)

Match this cell Match all identical cells Cancel

apple (Q89)
fruit of the apple tree

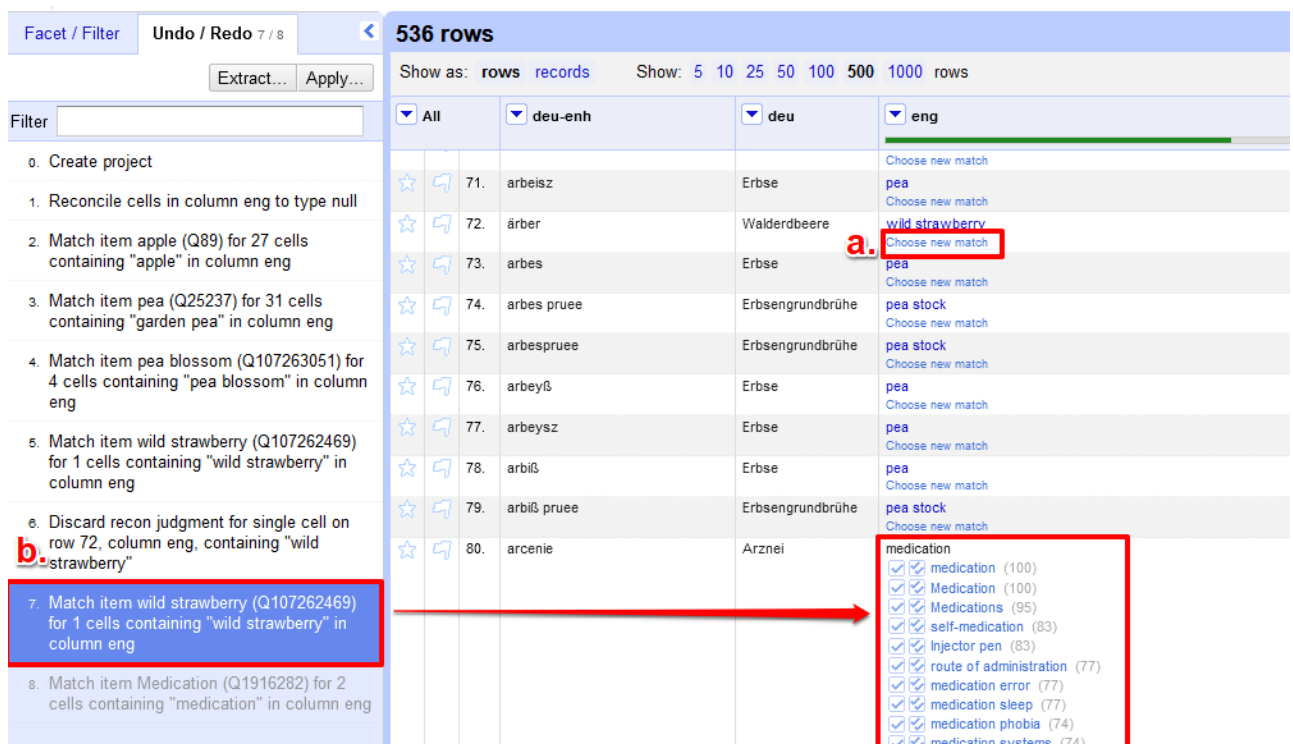
→ Etwas mühsam bei dieser manuellen Zuordnung ist, dass nach jeder Übernahme eines Wikidata-Eintrages das Programm anschließend zum Start der Tabelle hüpf, und man daher anschließend immer erneut zur nächsten, zur Bearbeitung ausstehenden Zeile scrollen muss.

- Sollte in den Vorschlägen eine passende Wikidata-Entsprechung fehlen, gibt es am Ende der Liste die Möglichkeit, nach weiteren Übereinstimmungen zu suchen und im neuen Suchfenster schließlich weitere Eingaben, unter denen ein Begriff auch zu finden sein könnte, vorzunehmen.



→ In unserem Datensatz wurde zum Beispiel das englische Wort "horse radish" mit einem Leerzeichen geschrieben, weshalb in der Liste mit Vorschlägen kein passender Eintrag zu finden war.

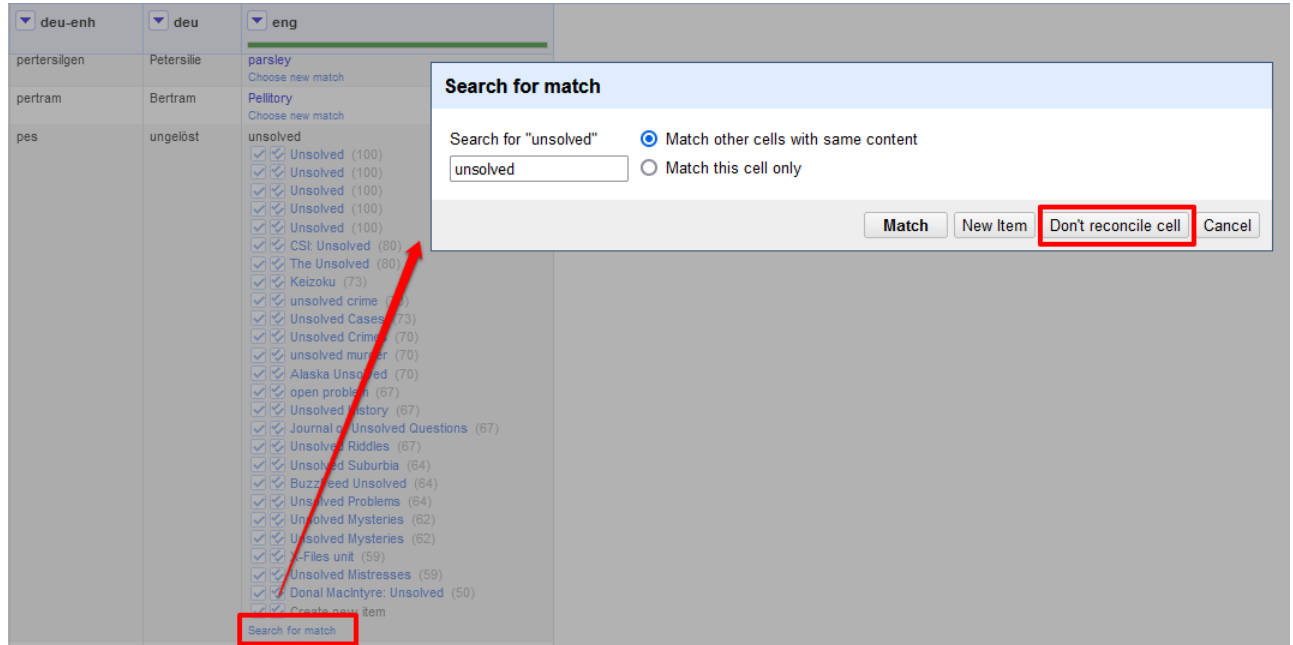
- Sollten wir mit einer unserer Zuordnungen nicht zufrieden sein, gibt es zwei Möglichkeiten, die Zuordnung wieder rückgängig zu machen. Entweder wir klicken einfach auf "Choose new match", direkt unter dem Begriff, der falsch zugeordnet wurde (a.), oder wir gehen in der linken Menüleiste in den Reiter **Undo/Redo** und wählen dort einen vorangegangenen Schritt aus, um dort wieder weiterzumachen (b.).



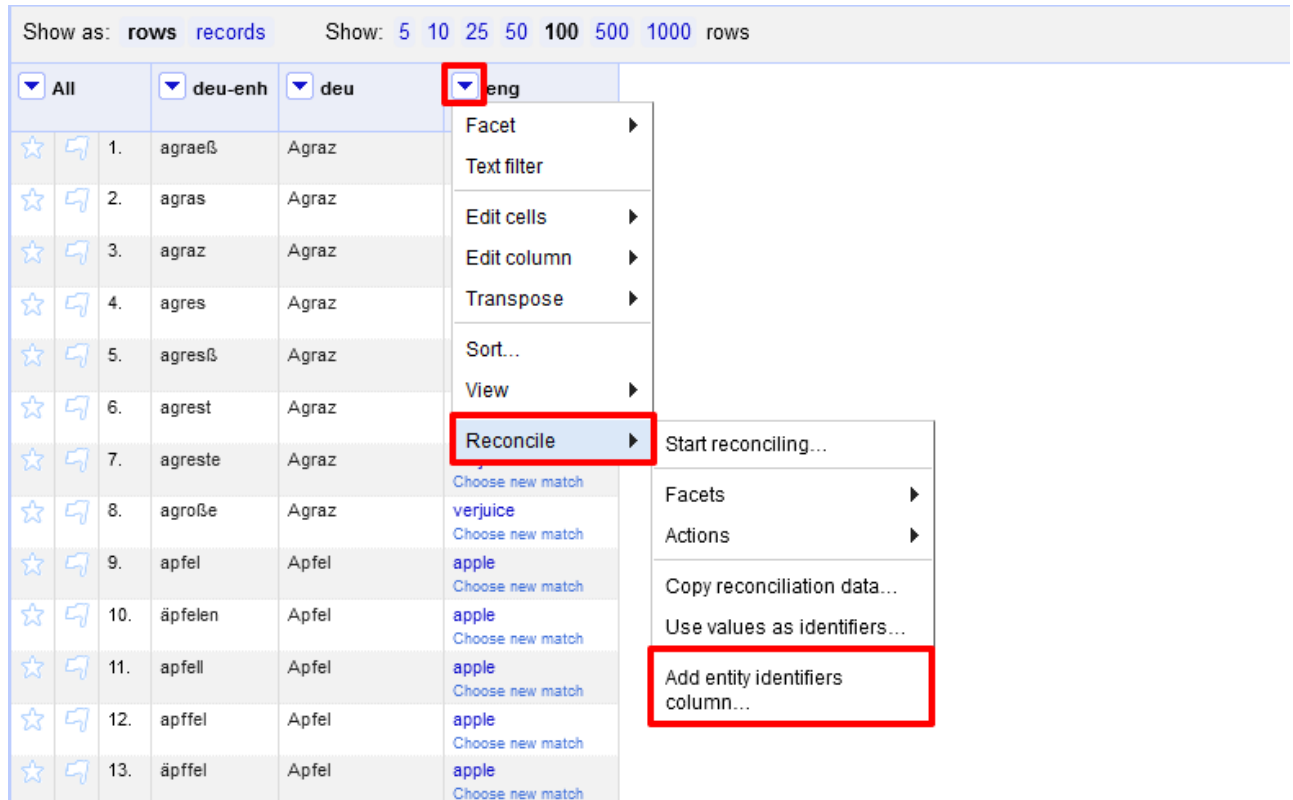
→ Mit dem "Extract"-Button in der linken Menüleiste ist es außerdem möglich, entweder alle oder einen Teil der bereits getätigten Schritte zu exportieren. Sollte sich die Liste beispielsweise erheblich

verändern, so könnte man ein neues Projekt erstellen, und den bisherigen Arbeitsfortschritt über den Import der Arbeitsschritte (mittels "Apply"-Button) wiederherstellen. Es müssten anschließend nur mehr die neu hinzugekommenen Einträge mit Wikidata-Normdaten angereichert werden.

- Für Einträge, die man nicht mit normalisieren möchte oder nicht kann, weil wie in unserem Beispielprojekt mitunter nicht jede Zutat entschlüsselt wurde, gibt es die Möglichkeit, über die Ansicht, die unter "Search for match" erscheint, auszuwählen, dass der Zelle kein Eintrag zugeordnet werden soll.



- Sobald wir all unsere Einträge mit Wikidata-Einträgen angereichert haben, können wir uns die Q-Nummern der Wikidata-Einträge in einer eigenen Spalte anzeigen lassen.



Wir müssen dieser Spalte nur mehr einen Namen geben und jede Zeile erhält eine weitere Zelle mit der entsprechenden Q-Nummer.

536 rows

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

			deu-enh	deu	eng	wikidata-qid
☆	🗨	1.	agraeß	Agraz	verjuice Choose new match	Q1060458
☆	🗨	2.	agras	Agraz	verjuice Choose new match	Q1060458
☆	🗨	3.	agraz	Agraz	verjuice Choose new match	Q1060458
☆	🗨	4.	agres	Agraz	verjuice Choose new match	Q1060458
☆	🗨	5.	agresß	Agraz	verjuice Choose new match	Q1060458
☆	🗨	6.	agrest	Agraz	verjuice Choose new match	Q1060458
☆	🗨	7.	agreste	Agraz	verjuice Choose new match	Q1060458
☆	🗨	8.	agroße	Agraz	verjuice Choose new match	Q1060458
☆	🗨	9.	apfel	Apfel	apple Choose new match	Q89
☆	🗨	10.	äpfeln	Apfel	apple Choose new match	Q89

→ Wir haben uns für commodity entschieden, da wir später beim Exportieren diesen Begriff direkt als Attributsbezeichnung übernehmen wollen und als Wert die entsprechende Q-Nummer eingefügt werden soll.

4. Export der Dokumente

- Um unsere angereicherte Tabelle bzw. normalisierten Daten zu exportieren, klicken wir auf den Button "Export" und wählen die Option "Templating...". Denn unser Ziel ist es, direkt eine XML-Struktur zu generieren, die wir in unser Register in ediarum übernehmen können.

536 rows

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

			deu-enh	deu	eng	commodity
☆	🗨	1.	agraeß	Agraz	verjuice Choose new match	Q1060458
☆	🗨	2.	agras	Agraz	verjuice Choose new match	Q1060458
☆	🗨	3.	agraz	Agraz	verjuice Choose new match	Q1060458
☆	🗨	4.	agres	Agraz	verjuice Choose new match	Q1060458
☆	🗨	5.	agresß	Agraz	verjuice Choose new match	Q1060458
☆	🗨	6.	agrest	Agraz	verjuice Choose new match	Q1060458
☆	🗨	7.	agreste	Agraz	verjuice Choose new match	Q1060458
☆	🗨	8.	agroße	Agraz	verjuice Choose new match	Q1060458
☆	🗨	9.	apfel	Apfel	apple Choose new match	Q89
☆	🗨	10.	äpfeln	Apfel	apple Choose new match	Q89
☆	🗨	11.	apfell	Apfel	apple Choose new match	Q89

Open... **Export** Help

- OpenRefine project archive to file
- Tab-separated value
- Comma-separated value
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Custom tabular...
- SQL...
- Templating...**
- OpenRefine project archive to Google Drive...
- Google Sheets...
- Wikibase edits...
- QuickStatements file
- Wikibase schema

- In der Ansicht für die Template-Erstellung haben wir nun die Möglichkeit, unsere Daten direkt in eine XML-Struktur zu überführen bzw. so zu gestalten, dass sie nur mehr in das ediarum-Sachregister kopiert werden müssen. Dafür tragen wir in das Prefix-Textfeld `<list>` und als Suffix `</list>` ein. Entsprechend des Schemas für Register in ediarum möchten wir für jede Zeile einen eigenen `<item>`-Eintrag erhalten. Als `@xml:id` soll die englische Übersetzung übernommen werden. Den Wikidata-Link übernehmen wir in Form eines `<idno>`-Elements innerhalb des `<item>`-Elements. Außerdem legen wir

auch 1-2 `<label>`-Elemente an, einmal mit dem Wert "reg" im @type-Attribut für die Übersetzung in Standarddeutsch, und ein weiteres mit dem Wert "alt", das die frühneuhochdeutschen Bezeichnung enthält.

In der Vorschau rechts sehen wir auch, wie unser Output schließlich aussehen wird.

Templating export

Prefix

Row template

```
<item xml:id="{ if(cells['eng'].value != 'unsolved', cells['eng'].value, cells['deu-enh'].value + '_unsolved') }" >
  {{ if(cells['wikidata-qid'].value != 'null', '<idno type="uri">https://www.wikidata.org/entity/' + cells['wikidata-qid'].value + '</idno>', '') }}
  {{ if(cells['deu'].value != 'ungelöst', '<label type="reg">' + cells['deu'].value + '</label>', '<label type="reg">' + cells['deu'].value + '(' + cells['deu-enh'].value + '</label>') }}
  <label type="alt">{{ cells['deu-enh'].value }}</label>
</item>
```

Row separator

Suffix

Reset template

Export Cancel

Preview:

```
<list>
<item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q1060458
  <label type="reg">Agraz</label>
  <label type="alt">agraes</label>
</item>
<item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q1060458
  <label type="reg">Agraz</label>
  <label type="alt">agraes</label>
</item>
<item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q1060458
  <label type="reg">Agraz</label>
  <label type="alt">agraes</label>
</item>
<item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q1060458
  <label type="reg">Agraz</label>
  <label type="alt">agraes</label>
</item>
```

→ Erläuterungen zum Code im Textfeld "Row Template": Unser Code, der über die einzelnen Zeilen unserer Tabelle iteriert, soll hier noch etwas genauer betrachtet werden. Mittels der [General Refined Expression Language \(GREL\)](#) haben wir unseren Code entsprechend unseren Anforderungen gestaltet.

```
<item xml:id="{ if(cells['eng'].value != 'unsolved', cells['eng'].value,
cells['deu-enh'].value + '_unsolved') }" >
  {{ if(cells['wikidata-qid'].value != 'null', '<idno
type="uri">https://www.wikidata.org/entity/' + cells['wikidata-qid'].value +
'</idno>', '') }}
  {{ if(cells['deu'].value != 'ungelöst', '<label type="reg">' +
cells['deu'].value + '</label>', '<label type="reg">' + cells['deu'].value +
'(' + cells['deu-enh'].value + '</label>') }}
  <label type="alt">{{ cells['deu-enh'].value }}</label>
</item>
```

Wir haben hier noch zusätzliche Bedingungen für folgende Spezialfälle eingeführt:

- **Fehlende Übersetzungen:** Sollten Zellen in unserem Datensatz in der englischen Spalte "unsolved" bzw. in der deutschen Spalte "ungelöst" beinhalten, weil man nicht weiß, welche Bedeutung der frühneuhochdeutsche Begriff hat, nutzen wir das frühneuhochdeutsche Wort als @xml:id.
- **Fehlende Q-Nummer:** Sollte eine Zeile keine Q-Nummer besitzen, wird auch kein `<idno>`-Element angelegt.

- Wenn unser Output so aussieht wie wir ihn gerne hätten, müssen wir nur mehr auf den "Export"-Button klicken und eine [TXT-Datei](#) wird heruntergeladen. Für unser Projekt müssen wir diesen Output aber noch ein wenig anpassen (siehe [Transition OpenRefine → ediarum](#)).

Kontakt

Weblink: <https://openrefine.org/>

Mail:

Allgemeiner Support

[Forum](#)

Christian Steiner (DH Craft)

christian.steiner@dhcraft.org

Ressourcen

Dokumentation

- [Offizielle OpenRefine Dokumentation](#)
- [Reconciliation mit Wikibase](#)
- [Github-Repository](#)

Tutorials

- [Using OpenRefine to Clean Your Data](#)
- [Get Started with OpenRefine: Explore, Clean, and Transform your Data!](#)
- [Reconciliation with Wikidata](#)

Projekte, die dieses Tool genutzt haben

- [Corema - Cooking Recipes of the Middle Ages](#)

Literatur

- Crossley, L. (2019, Oktober 29). *Text Mining Digital Humanities Blogs with APIs, OpenRefine, and R*. <https://mars.gmu.edu/handle/1920/11632>
- Delpuch, A. (2019). *A survey of OpenRefine reconciliation services* (arXiv:1906.08092). arXiv. <https://doi.org/10.48550/arXiv.1906.08092>
- Dreßen, A., & Sacher, E. (2023, März 6). *Querying Art History Data on the Web (5): Modelling Data with OpenRefine*. https://www.db-thueringen.de/receive/dbt_mods_00055804
- Engelhardt, F., Freitag, N., & Wildermuth, M. (2023). Die Migration der Bibliographia Cartographica: Daten aufräumen mit OpenRefine. *Bibliotheksdienst*, 57(2), 95–110. <https://doi.org/10.1515/bd-2023-0016>
- Gallant, K., Lorang, E., & Ramirez, A. (2014). *Tools for the digital humanities: a librarian's guide* (Emerging Technologies in Libraries). <https://mospace.umsystem.edu/xmlui/handle/10355/44544>
- Gutiérrez De la Torre, Silvia Eunice. (2021, Juni 15). *OpenRefine, Authority Control and Wikidata*. <https://zenodo.org/record/4950866>

- Handelman, M. (2015). Digital Humanities as Translation: Visualizing Franz Rosenzweig's Archive. *TRANSIT*, 10(1). <https://doi.org/10.5070/T7101029573>
- Ikonić Nešić, M., Stanković, R., Schöch, C., & Skoric, M. (2022). From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back). *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, 7–16. <https://aclanthology.org/2022.ldl-1.2>
- Krimmel, Erica, & Walker, Lindsay J. (2022, Mai 11). *Using OpenRefine for natural history collections data*. Society for the Preservation of Natural History Collections (SPNHC), Edinburgh, Scotland, UK, 5-10 June 2022, Edinburgh, Scotland, UK,. <https://zenodo.org/record/6574729>
- Mandal, S. (2022). Integration of Linked Open Data Authorities with OpenRefine: A Methodology for Libraries. *Library Philosophy and Practice (e-journal)*. <https://digitalcommons.unl.edu/libphilprac/7195>
- Myntti, J., & Neatrou, A. (2015). Use Existing Data First: Reconcile Metadata before Creating New Controlled Vocabularies. *Journal of Library Metadata*, 15(3–4), 191–207. <https://doi.org/10.1080/19386389.2015.1099989>
- Ransom, L., & Coladangelo, L. P. (2021, Dezember 3). Semantic Enrichment of the Schoenberg Database of Manuscripts Name Authority through Wikidata. *15th International Conference on Metadata and Semantics Research*. https://www.academia.edu/63137370/Semantic_Enrichment_of_the_Schoenberg_Database_of_Manuscripts_Name_Authority_through_Wikidata
- Sohmen, L., & Rossenova, L. (2022). Open refine to wikibase: a new data upload pipeline. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–2. <https://doi.org/10.1145/3529372.3530919>
- Steeg, F., & Pohl, A. (2021). Ein Protokoll für den Datenabgleich im Web am Beispiel von OpenRefine und der Gemeinsamen Normdatei (GND). In M. Franke-Maier, A. Kasprzik, A. Ledl, & H. Schürmann (Hrsg.), *Qualität in der Inhaltserschließung* (S. 259–278). De Gruyter. <https://doi.org/10.1515/9783110691597-013>
- Woitas, Kathi. (2021, Dezember 13). *OpenRefine*. <https://zenodo.org/record/5776098>

Factsheet

System	
Scope des Tools	Datenbereinigung & Normalisierung
Softwareumgebung/Softwaretyp (Remotesystem im Browser / Lokaler Client)	lokale Browser-Anwendung
Unterstützte Plattformen	Linux, Windows & Mac
Geräte	Desktop
Einbindung anderer Systeme (Interoperabilität)	<input checked="" type="checkbox"/> (Wikidata, Wikibase)
Accountsystem	<input checked="" type="checkbox"/> (keine Anmeldung erforderlich)
Kostenmodell (Kostenübersicht / Open Source)	kostenlos

Anforderungen & Methoden

Erforderte Code Literacy	sehr gering
Interface-Sprachen (ISO 639-1)	en
Unterstützte Zeichenkodierung	UTF-8, UTF-16, ASCII
Inkludierte Datenkonvertierung (Im Pre-Processing mögliche Anpassung der Daten an für die Software erforderliches Format)	☑
Abhängigkeit von anderer Software (Falls ja, wird diese Software automatisch mitinstalliert?)	✗
Erforderliche Plug-Ins (bei web-basierten Anwendungen)	✗

Dokumentation & Support

Wartung und ständige Erweiterung	☑
Einbindung der Community	☑ via Github & Forum
Dokumentation	☑
Dokumentationssprache	Englisch
Dokumentationsformat	HTML
Dokumentationsabschnitte	Introduction, Installing, Running, Starting a project, Exploring data, Transforming data, Reconciling, Wikibase, Wikidata, and Wikimedia Commons, Expressions, Exporting, Troubleshooting, GREL Reference, Technical Reference
Verfügbarkeit von Tutorials	☑ Blogbeitrag mit Sammlung an Tutorials
Aktiver Support/Community (Forum, Slack, Issue Tracker etc.)	☑ Forum, GitHub-Issues-Mechanismus

Nutzbarkeit & Nachhaltigkeit

Installationsablauf	sehr einfach
Test (Gibt es ein Test Suite, um zu überprüfen, ob die Installation erfolgreich war?)	☑
Lizenz, unter der das Tool veröffentlicht wurde	CC BY 4.0

Registrierung in einem Repository	<input checked="" type="checkbox"/> Github
Möglichkeit zur Software-Entwicklung beizutragen	<input checked="" type="checkbox"/>
Benutzerinteraktion & Benutzeroberfläche	
Benutzerprofil (erwartete Nutzer:innen)	Data Scientists, Datenbankbeauftragte
Benutzerinteraktion (erwartete Nutzung)	Hochladen von Dateien, Datenzusammenführung, -bereinigung, -strukturierung, -normalisierung, -transformation und -visualisierung, Export von Dateien
Benutzeroberfläche	browserbasiertes GUI
Visualisierungen (Analyse-, Input-, Outputkonfigurationen)	<input checked="" type="checkbox"/>
Benutzerverwaltung	
Personenverwaltung	✗
Interne Kommunikationsmöglichkeiten (z. B. Annotationsrichtlinien, Kommentarfunktionen, ...)	✗
Daten- und Toolverwaltung	
Zentrale/dezentrale Verwaltungsmöglichkeit	<input checked="" type="checkbox"/> mehrere Projekte möglich
Versionskontrolle	<input checked="" type="checkbox"/> jegliche Änderungen können nachverfolgt und rückgängig gemacht bzw. wiederhergestellt werden
Projektspezifische Einstellungen	<input checked="" type="checkbox"/>
API	<input checked="" type="checkbox"/> für Reconciliation
Möglichkeit auf simultanes Arbeiten	✗
Datenupload	
Unterstützte Dateiformate	CSV, TSV, TXT, JSON, XML, ODS, XLS, XLSX, PX, MARC, RDF(JSON-LD, N3, N-Triples, Turtle, RDF/XML), Wikitext Importmöglichkeiten auch über Weblinks, SQL-Datenbank oder Google Drive
Informationen zur Datensicherheit	[nicht anwendbar, da lokale Ausführung]
Zugänglichkeit von	✗

verschiedenen Standorten/Geräten	
Einschränkungen hinsichtlich der Datenmenge	max. 1 GB
Verlustfreier Upload von bereits bearbeiteten Dokumenten	<input checked="" type="checkbox"/>
Unterstützung von IIIF-Import	[nicht anwendbar]
Datenbearbeitung (Normalisierungstool)	
Komplexitätsgrad der Normalisierung (z. B. Verfügbarkeit von Buttons, Drag&Drop-Funktion, ...)	<input checked="" type="checkbox"/> Buttons für Übernahme von Vorschlägen, Liste für Standardservices verfügbar
Reconciliation-Möglichkeiten entsprechend bestimmten Standards für digitale Editionen	<input checked="" type="checkbox"/> Wikidata, GND, GeoNames, Pleiades, etc.
Anpassungsmöglichkeit und Validierung entsprechend projektspezifischen Konventionen/Schemata	<input checked="" type="checkbox"/>
Datenexport	
Unterstützte Dateiformate	TSV, CSV, XLS, XSLX, HTML, ODF, SQL, TXT (Templatingmöglichkeit für JSON, XML usw.)
Datenverlust (nicht vollständiger Erhalt von Annotationen, die bereits vor Verwendung des Tools gemacht wurden)	[nicht anwendbar]
Validierungsmöglichkeit für TEI-XML vor Export	[nicht anwendbar, da keine Möglichkeit auf XML-Export]
Datenaufbewahrung nach Export	[nicht anwendbar, da lokale Ausführung]