

# Allgemeine Beschreibung

---

*OpenRefine* ist ein Open-Source-Tool zur Datenbereinigung und Datentransformation, 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.

In Hinblick auf digitale Editionen ist ein Vorteil von *OpenRefine*, 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 jedoch auf die Möglichkeit, direkt eine XML-Datei herunterzuladen, verzichten und auch komplexere Datentransformationen beim Export - wie beispielweise 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 (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
- 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

- [EDV-Grundkenntnisse](#)
- Ausdruckssprachen und Transformationstechniken von Vorteil

### Benötigte Software

- Stabile Internetverbindung
- Webbrowser

## Tool-Kompatibilität

|            | IIIF | Transkribus | FromThePage | ediarum   | FairCopy | 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 Tools aufgelistet werden, die während der Durchführung des jeweiligen [Beispielprojekts](#) festgestellt wurden.

### Stärken

- Benutzerfreundliche Bearbeitungsoberfläche und Wahrung der Datensicherheit durch die lokale Bearbeitung und Speicherung
- Bereinigung von unstrukturierten und fehlerhaften Daten (Dubletten, Tippfehler, Inkonsistenzen) und damit Überprüfung der Datenqualität (Qualitätssicherung)
- Versionskontrolle durch die Möglichkeit, Arbeitsschritte wieder rückgängig zu machen oder bereits getätigte Schritte wiederherzustellen
- Datenerweiterung und Normalisierung über Reconciliation-Services, die den Datenabgleich mit externen Datenbanken ermöglichen
- Datentransformation in andere Formate oder Strukturen
- Datenzusammenführung bei mehreren Quellen oder Versionen
- Organisation und Strukturierung von Metadaten
- Automatisierung von wiederholten Datenbereinigungs- und Transformationsaufgaben durch die Möglichkeit, den Änderungsverlauf zu exportieren und auf neue Daten anzuwenden

### Herausforderungen & Probleme

- Keine simultane Kollaborationsmöglichkeit, da *OpenRefine* für die lokale Verwendung konzipiert ist und nicht mehrere Personen gleichzeitig an einem Projekt arbeiten können (eine - aber bei einer Vielzahl von Projektmitarbeitenden relativ umständliche - Möglichkeit, mit anderen Personen zusammenzuarbeiten, besteht aber darin, Projekte inklusive der gespeicherten Bearbeitungsschritte zu exportieren und daraufhin an einem anderen Rechner zu importieren)
- Teilweise mühsame Bedienung durch das Problem, dass *OpenRefine* bei der manuellen Zuordnung von passenden Wikidata-Einträgen nach jeder einzelnen Match-Bestätigung zum Start der Tabelle springt
- Keine direkte Exportmöglichkeit in eine XML-Datei, wobei über den Templating-Export die Daten jedoch zumindest in einer XML-Struktur (als Plaintext-Datei) exportiert werden können
- 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

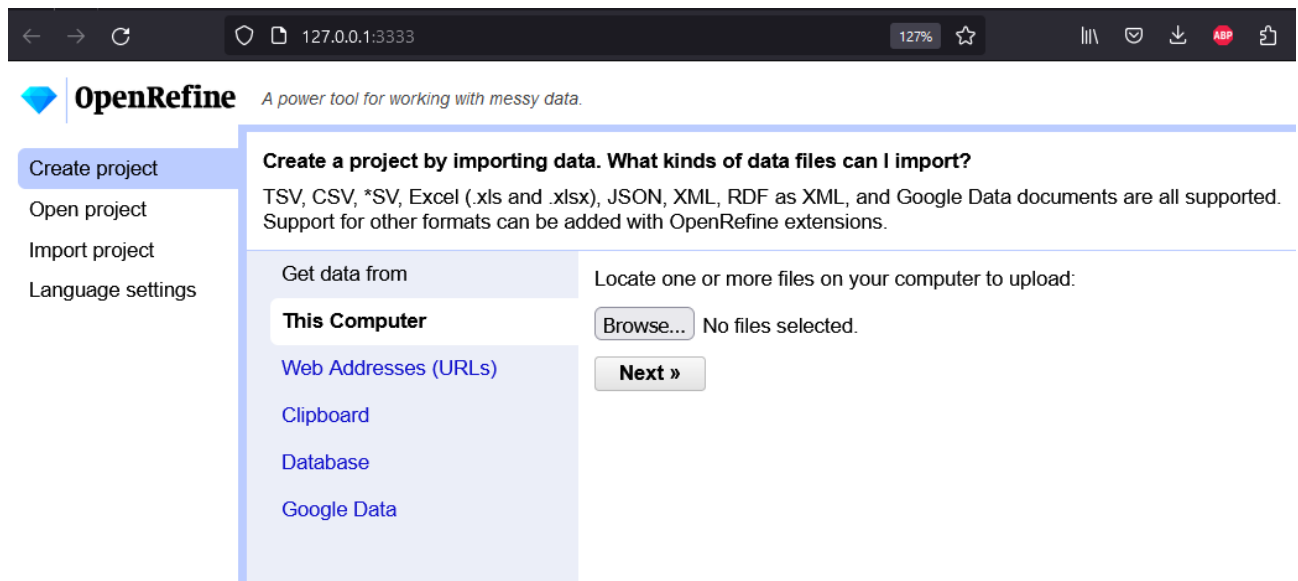
## Einrichtung & Erste Schritte

---

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](#) wurden bereits erste Annotationen vorgenommen. 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 den [Q-Nummern](#) - auch QID genannt - von Wikidata-Einträgen anzureichern.

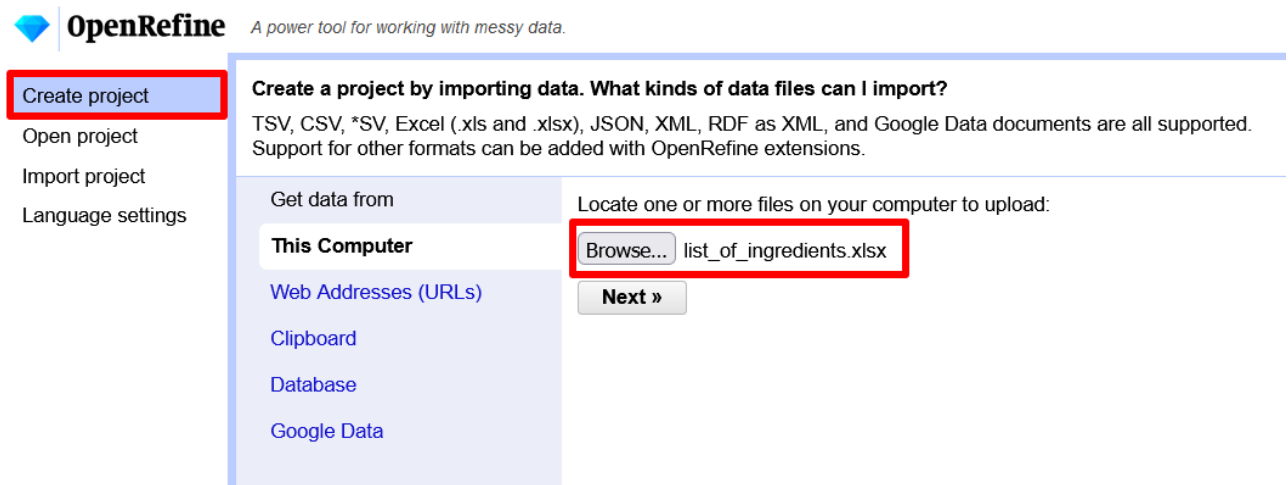
## 1. Installation

- Unser erster Schritt besteht darin, uns die entsprechende Version für unser Betriebssystem von [OpenRefine](#) [herunterzuladen](#). Nach dem Entpacken der ZIP-Datei führen wir openrefine.exe aus, wodurch sich *OpenRefine* direkt in unserem Browser öffnet.



## 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.



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

Create project

Open project

Import project

Language settings

« 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. | äpfell  | Apfel | apple    |
| 12. | äpfel   | Apfel | apple    |
| 13. | äpfel   | Apfel | apple    |
| 14. | äpfell  | Apfel | apple    |
| 15. | äpfll   | 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/JSON 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

→ Für unser Projekt haben wir an den vorausgewählten Einstellungen nichts geändert und nur einen Projektnamen gewählt, bevor wir mit "Create project" fortgefahren sind.

- Unsere Projektansicht sieht letztlich so aus:


**OpenRefine** MaRezepte [Permalink](#)
Open... Export Help


Facet / Filter    Undo / Redo 0 / 0    <

536 rows

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 | Facet       |
| ☆ ↗ 2.  | agras     | Agraz | Text filter |
| ☆ ↗ 3.  | agraz     | Agraz |             |
| ☆ ↗ 4.  | agres     | Agraz | Edit cells  |
| ☆ ↗ 5.  | agresß    | Agraz | Edit column |
| ☆ ↗ 6.  | agrest    | Agraz | Transpose   |
| ☆ ↗ 7.  | agreste   | Agraz |             |
| ☆ ↗ 8.  | agroße    | Agraz | Sort...     |
| ☆ ↗ 9.  | apfel     | Apfel | View        |
| ☆ ↗ 10. | äpfelen   | Apfel | Reconcile   |
| ☆ ↗ 11. | apfell    | Apfel | apple       |
| ☆ ↗ 12. | apffel    | Apfel | apple       |
| ☆ ↗ 13. | äpfel     | Apfel | apple       |
| ☆ ↗ 14. | apffel    | Apfel | apple       |
| ☆ ↗ 15. | apffi     | Apfel | apple       |
| ☆ ↗ 16. | aphel     | Apfel | apple       |
| ☆ ↗ 17. | aphell    | Apfel | apple       |
| ☆ ↗ 18. | aphfel    | Apfel | apple       |
| ☆ ↗ 19. | aphfell   | Apfel | apple       |
| ☆ ↗ 20. | aphl      | Apfel | apple       |
| ☆ ↗ 21. | appel     | Apfel | apple       |
| ☆ ↗ 22. | appfel    | 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.

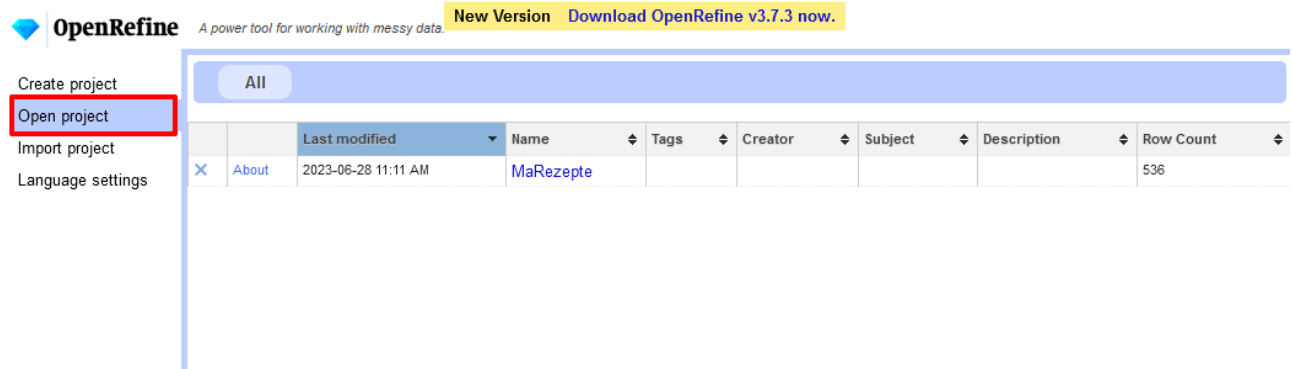
Not sure how to get started?  
[Watch these screencasts](#)

→ 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 befinden sich die Übersetzungen in modernes Englisch. Jede Spalte verfügt über ein Drop-Down-Menü, das uns verschiedene Bearbeitungsmöglichkeiten bietet, wobei für uns vor allem jene Funktion, die eine Anreicherung mit Normdaten (Reconciliation) ermöglicht, von Interesse ist.

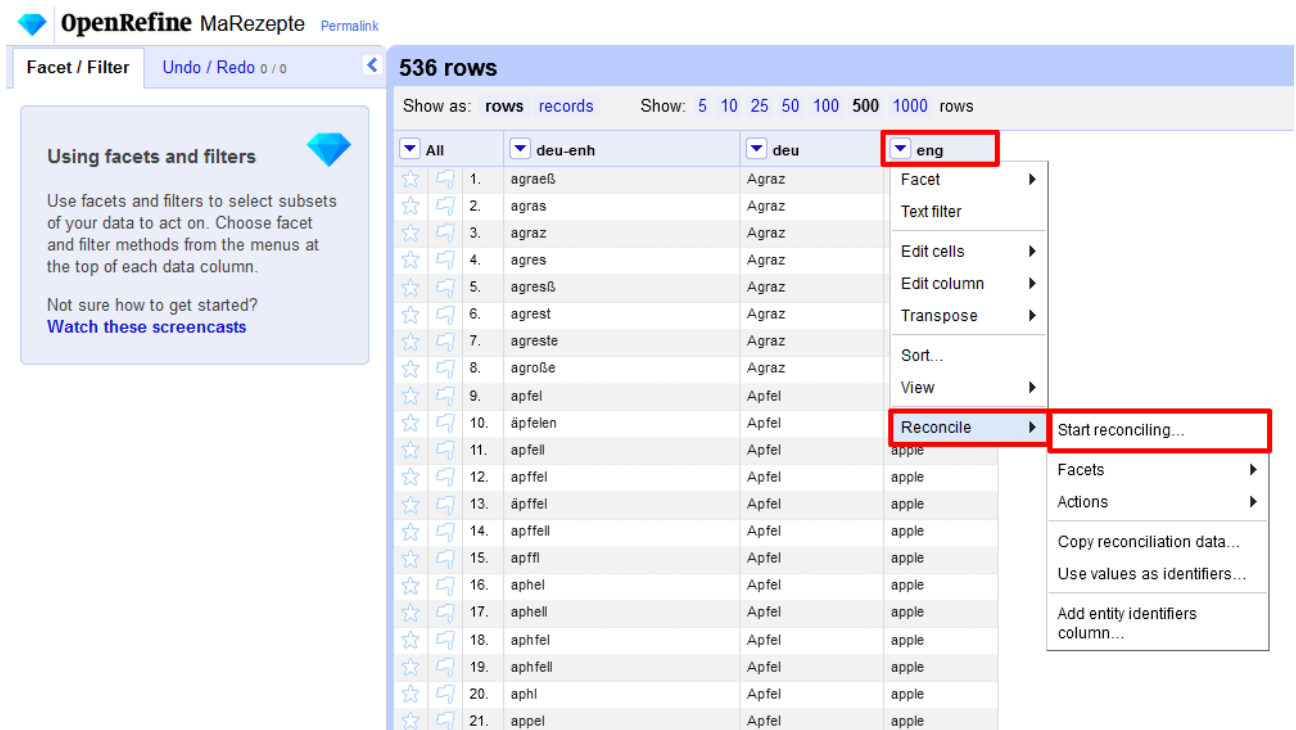
### 3. Bearbeitung der Dokumente

- 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 linken Navigationleiste können wir schließlich unsere Projekte einsehen. Wir öffnen hier unser bereits

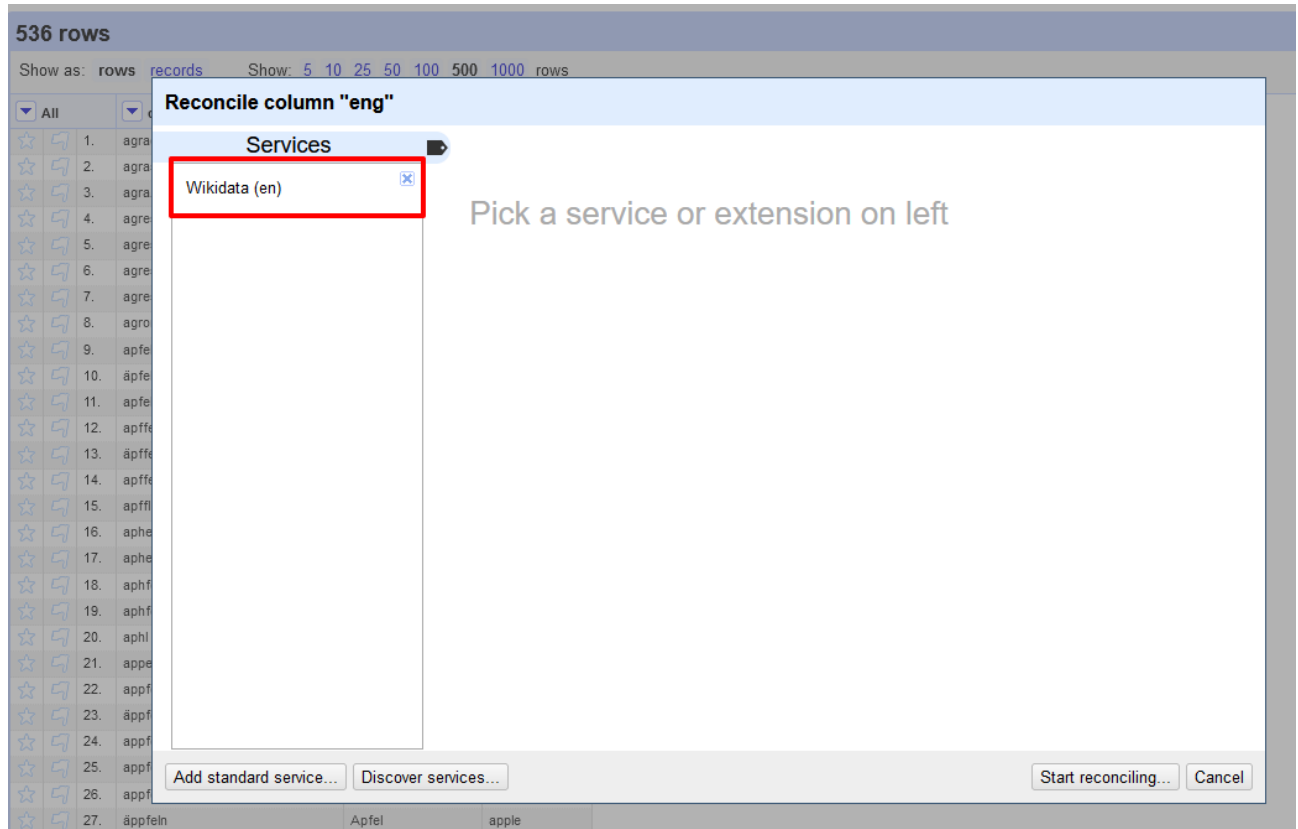
angelegtes Projekt "MaRezepte".



- 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 gehen wir auf die Spalte mit der Überschrift "eng", wählen im Dropdown die Option **Reconcile** und dann in der damit verbundenen Auswahl **Start Reconcile**.



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



- In dem nachfolgenden 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 Kategorienzuordnung 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 starten wir schließlich den Reconciliation-Prozess.

**Reconcile column "eng"**

Access service API

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

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 next > last »

Extensions Wikibase

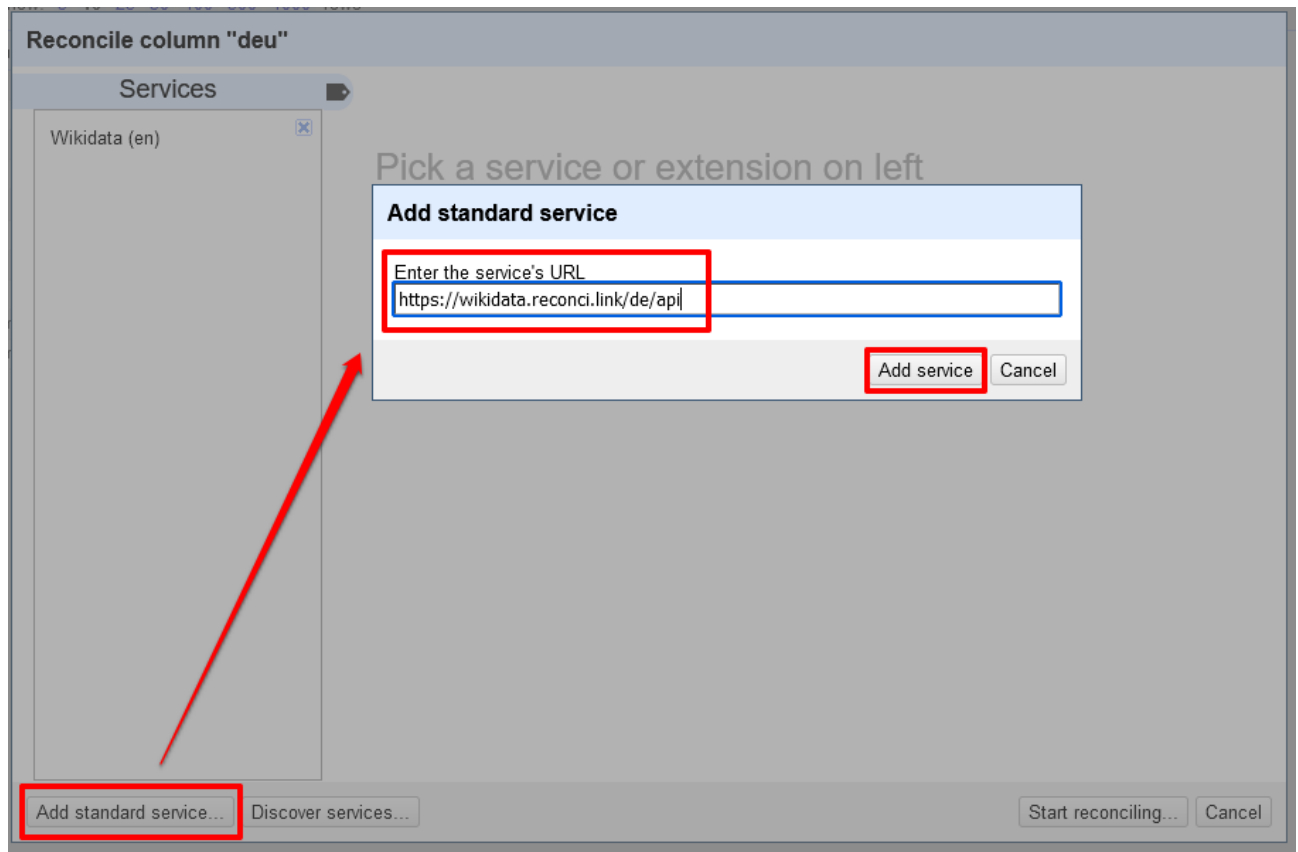
**Using facets and filters**

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.

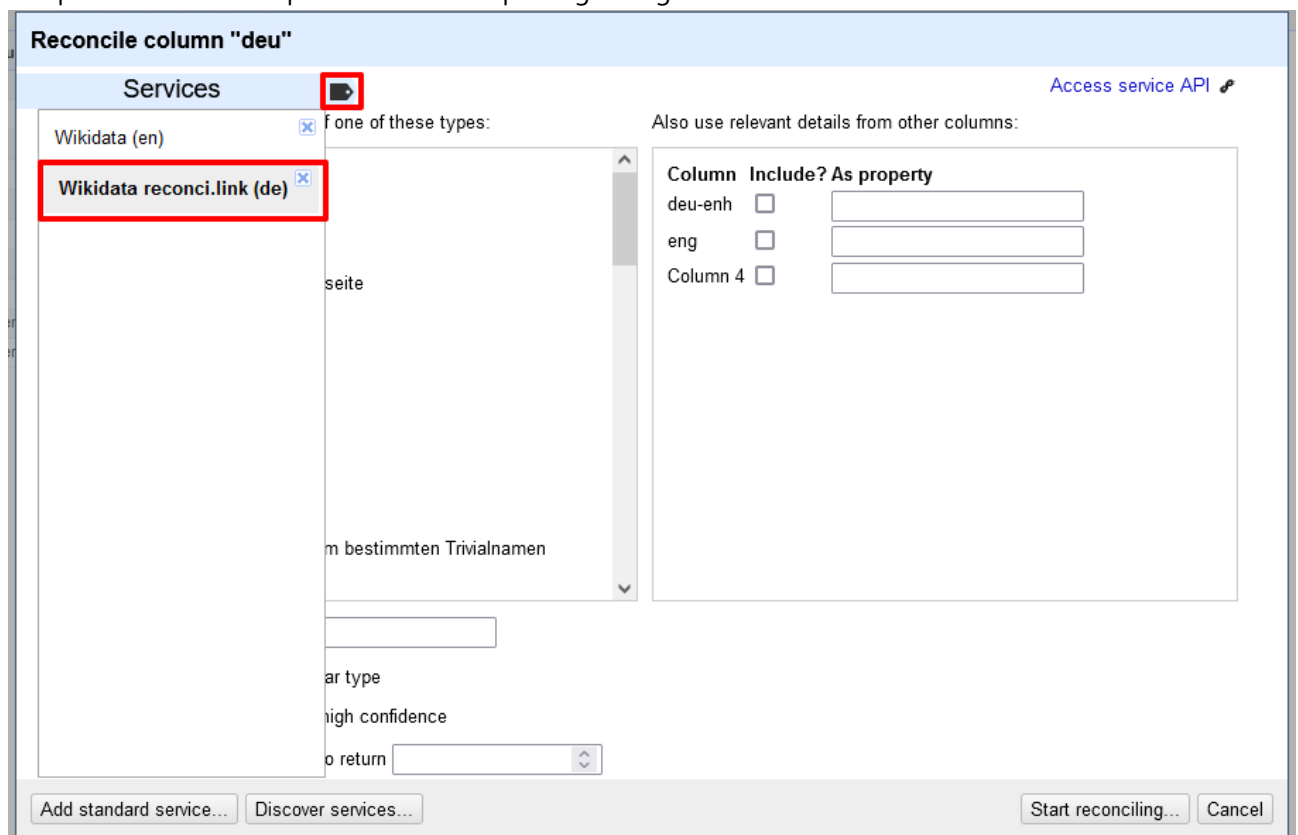
Not sure how to get started?  
[Watch these screencasts](#)

|     | All     | deu-enh | deu      | eng |
|-----|---------|---------|----------|-----|
| 1.  | agraeß  | Agraz   | verjuice |     |
| 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. | äpfeln  | Apfel   | apple    |     |
| 11. | apfell  | Apfel   | apple    |     |

- 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 für 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

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

536 rows

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

|    | 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)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> Muggsy Bogues (100)<br><input checked="" type="checkbox"/> Malus pumila (100)<br><input checked="" type="checkbox"/> Apple II series (100)<br><input checked="" type="checkbox"/> Apple Records (100)<br><input checked="" type="checkbox"/> Apple III (100)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> The Apple (100)<br><input checked="" type="checkbox"/> Apple River (100)<br><input checked="" type="checkbox"/> apple (100)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> Ariane Passenger Payload Experiment (100)<br><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 zu machen.

- 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

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

536 rows

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

|    | 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)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> Muggsy Bogues (100)<br><input checked="" type="checkbox"/> Malus pumila (100)<br><input checked="" type="checkbox"/> Apple II series (100)<br><input checked="" type="checkbox"/> Apple Records (100)<br><input checked="" type="checkbox"/> Apple III (100)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> The Apple (100)<br><input checked="" type="checkbox"/> Apple River (100)<br><input checked="" type="checkbox"/> apple (100)<br><input checked="" type="checkbox"/> Apple (100)<br><input checked="" type="checkbox"/> Ariane Passenger Payload Experiment (100)<br><input checked="" type="checkbox"/> Apple II (100) |

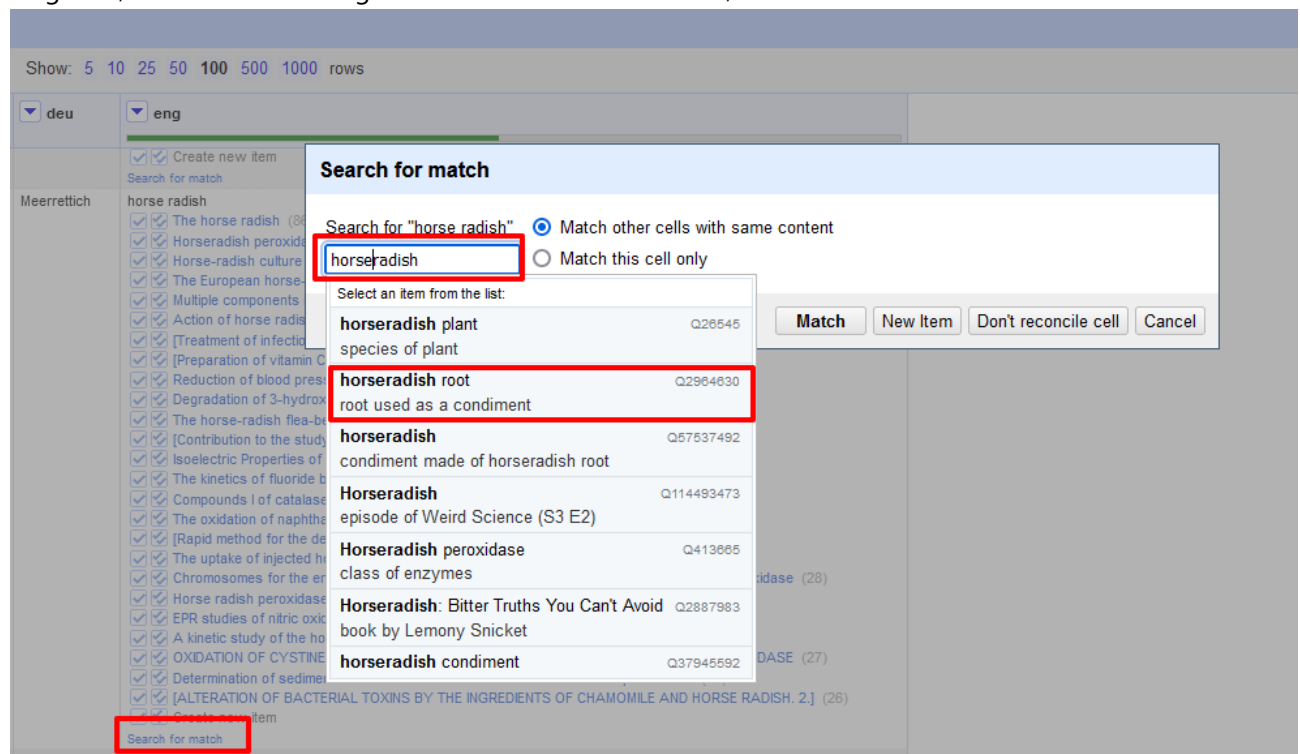
Match this cell Match all identical cells Cancel

apple (Q89)  
fruit of the apple tree

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

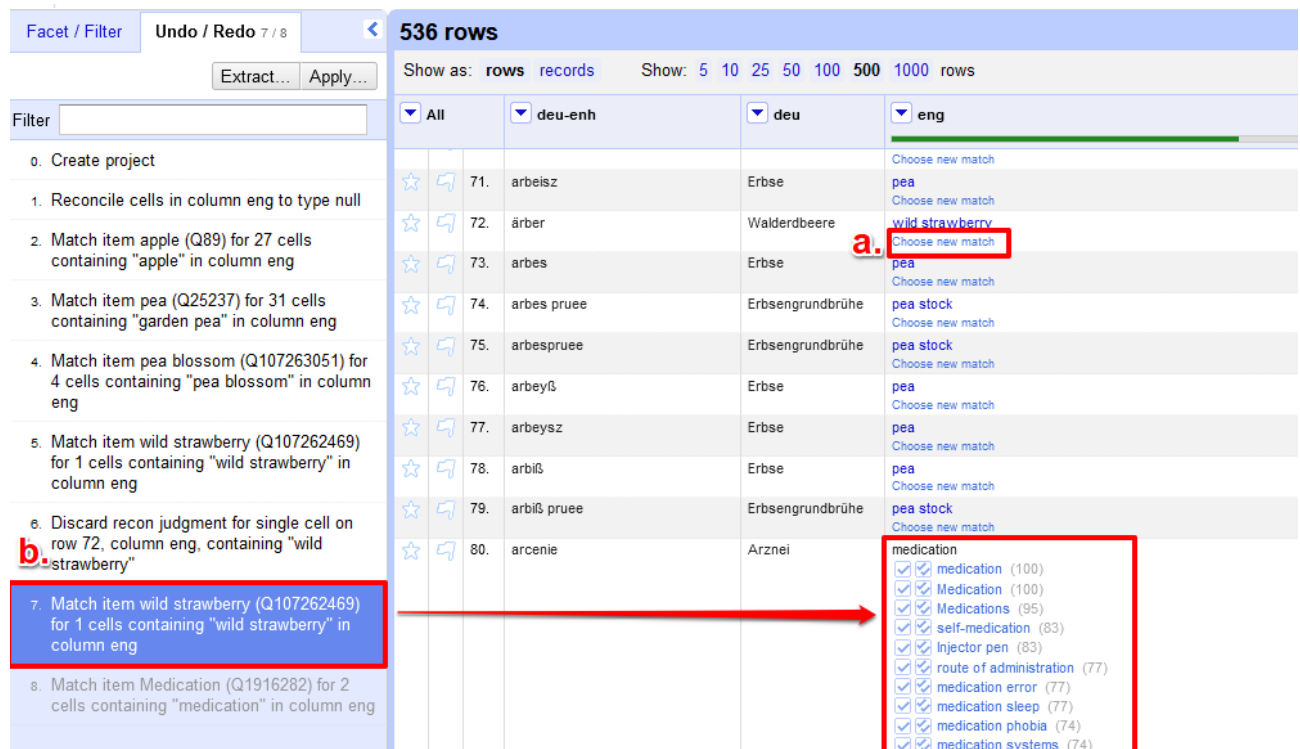
- Sollte in den Vorschlägen eine passende Wikidata-Entsprechung fehlen, gibt es am Ende der Liste die Möglichkeit, nach weiteren Übereinstimmungen zu suchen und im neuen Suchfenster schließlich weitere

Eingaben, unter denen ein Begriff auch zu finden sein könnte, vorzunehmen.



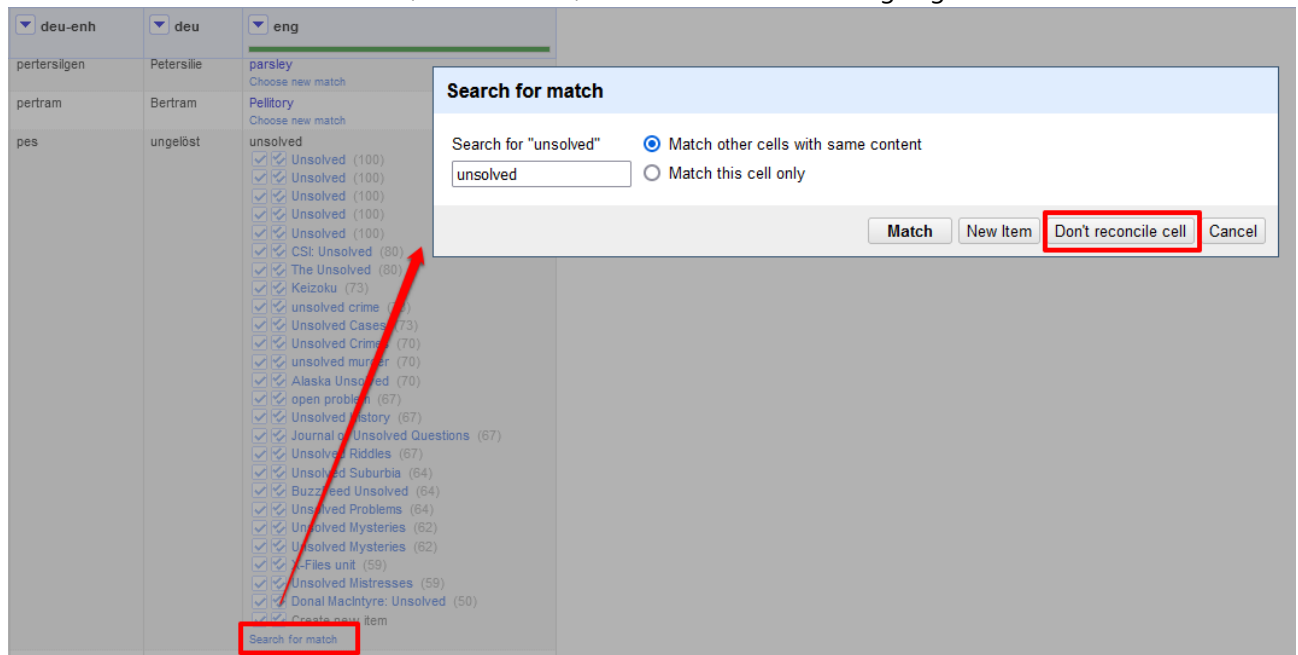
→ In unserem Datensatz wurde zum Beispiel das englische Wort "horse radish" mit einem Leerzeichen geschrieben, weshalb in der Liste mit Vorschlägen kein passender Eintrag zu finden war.

- 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 einen vorangegangenen Schritt aus, um dort wieder weiterzumachen (b.).

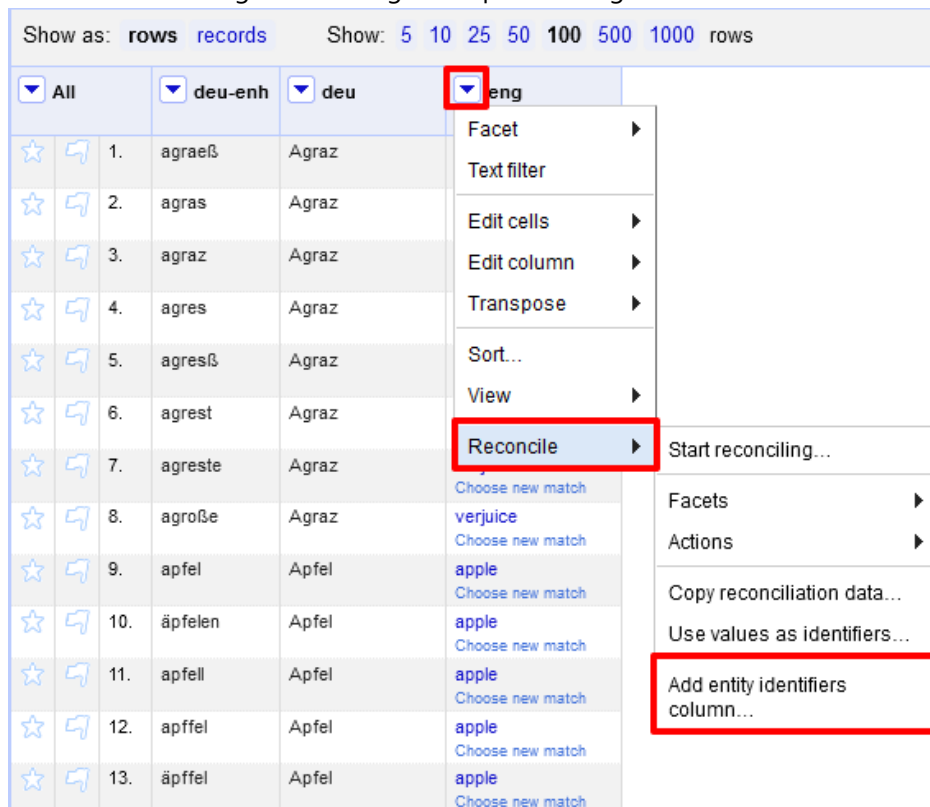


→ Mit dem "Extract"-Button in der linken Menüleiste ist es außerdem möglich, entweder alle oder einen Teil der bereits getätigten Schritte zu exportieren. Sollte sich die Liste beispielsweise erheblich verändern, so könnte man ein neues Projekt erstellen, und den bisherigen Arbeitsfortschritt über den Import der Arbeitsschritte (mittels "Apply"-Button) wiederherstellen. Anschließend müsste man anschließend nur mehr die neu hinzugekommenen Einträge mit Wikidata-Normdaten angereichert werden.

- Für Einträge, die man nicht mit Normdaten anreichern 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: <b>rows</b> records |         | Show: 5 10 25 50 100 500 1000 rows |       |                              |          |
| All                          | deu-enh | deu                                | eng   | idno                         |          |
| ☆                            | 1.      | agraeß                             | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 2.      | agras                              | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 3.      | agraz                              | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 4.      | agres                              | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 5.      | agresß                             | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 6.      | agrest                             | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 7.      | agreste                            | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 8.      | agroße                             | Agraz | verjuice<br>Choose new match | Q1060458 |
| ☆                            | 9.      | apfel                              | Apfel | apple<br>Choose new match    | Q89      |
| ☆                            | 10.     | äpfelen                            | Apfel | apple<br>Choose new match    | Q89      |

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

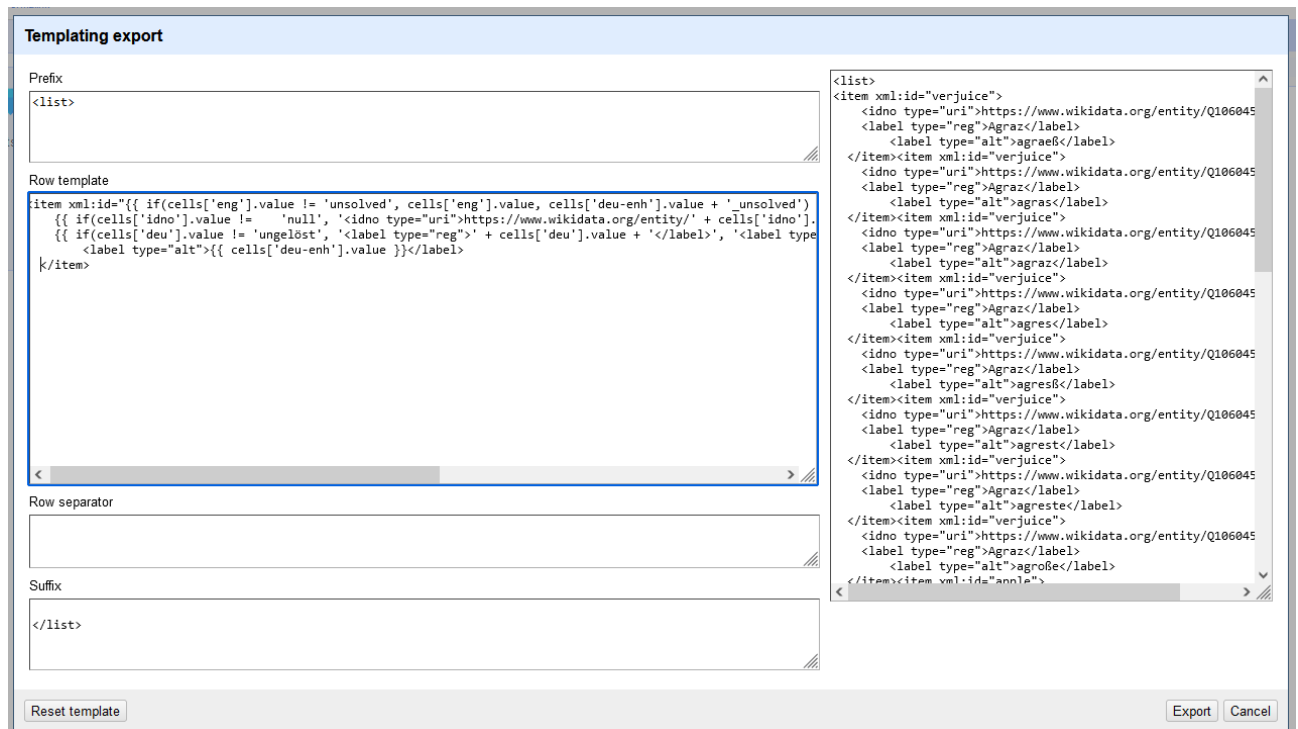
## 4. Export der Dokumente

- 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 anschließend in unser Register in *ediarum* übernehmen können.

The screenshot shows the OpenRefine interface with a table of 536 rows. The 'eng' column is highlighted. The 'Export' button is highlighted with a red box, and its dropdown menu is open, showing various export options. The 'Templating...' option is highlighted with a red box. The table data is as follows:

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

- In der Ansicht für die Template-Erstellung haben wir nun die Möglichkeit, unsere Daten 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 dem Schema 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>`-Elemente 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.



→ 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['idno'].value != 'null', '<idno
type="uri">https://www.wikidata.org/entity/' + cells['idno'].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 für unsere Daten 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 schließlich so aussieht wie wir ihn gerne hätten, müssen wir nur mehr auf den "Export"-Button klicken und eine **TXT-Datei** wird heruntergeladen. Für unser Beispielprojekt müssen wir diesen Output aber im Anschluss noch ein wenig anpassen (siehe [Transition OpenRefine → ediarum](#)).

## Kontakt

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

Allgemeiner Support

[Forum](#)

Christian Steiner (DH Craft) [christian.steiner@dhcraft.org](mailto: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](#): Im Projekt CoReMA wurden mittelalterliche Kochrezepte ediert. Mit den zugrundeliegenden Forschungsfragen sollten die kulinarischen Beziehungen zwischen Frankreich und den deutschsprachigen Ländern untersucht werden, auf welche die französische Kultur von jeher einen großen Einfluss hatte. Das Projekt stellt die transkribierten und annotierten Kochrezepte der deutschsprachigen Überlieferung vor 1500 zur Verfügung. In den historischen Texten sind Zutaten, Küchengeräte und Speisegattungen mit modernen Normdaten annotiert. Diese werden für unterschiedliche Analyse- und Auswertungsmethoden herangezogen.

## Literatur

- Crossley, L. (2019, Oktober 29). *Text Mining Digital Humanities Blogs with APIs, OpenRefine, and R*. <https://mars.gmu.edu/handle/1920/11632>
- Delpeuch, 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](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>
- Ikončić 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.lidl-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](https://www.academia.edu/63137370/Semantic_Enrichment_of_the_Schoenberg_Database_of_Manuscripts_Name_Authority_through_Wikidata)
- Sohmen, L., & Rossanova, 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 Inhaltsschließung* (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  |                                   |
|---|-----------------------------------|
| <b>Scope des Tools</b>  | Datenbereinigung & Normalisierung |
| <b>Softwareumgebung/Softwaretyp</b><br>(Remotesystem im Browser / Lokaler Client)   | lokale Browser-Anwendung          |
| <b>Unterstützte Plattformen</b>   | Linux, Windows & Mac              |
| <b>Geräte</b>   | Desktop                           |
| <b>Einbindung anderer Systeme (Interoperabilität)</b>   | ☑ (Wikidata, Wikibase)            |
| <b>Accountsystem</b>  | ✗ (keine Anmeldung erforderlich)  |
| <b>Kostenmodell</b><br>(Kostenübersicht / Open Source)  | kostenlos                         |
| Anforderungen & Methoden  |                                   |
| <b>Erforderte Code Literacy</b>   | sehr gering                       |
| <b>Interface-Sprachen (ISO 639-1)</b>   | en                                |
| <b>Unterstützte Zeichenkodierung</b>  | UTF-8, UTF-16, ASCII              |
| <b>Inkludierte Datenkonvertierung</b><br>(Im Pre-Processing mögliche Anpassung der Daten an für die Software erforderliches Format) | ☑                                 |
| <b>Abhängigkeit von anderer Software</b>  | ✗                                 |



(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**

✓ Github

**Möglichkeit zur Software-Entwicklung beizutragen**

✓

## Benutzerinteraktion & Benutzeroberfläche

**Benutzerprofil**

(erwartete Nutzer:innen)

Data Scientists, Datenbankbeauftragte

**Benutzerinteraktion**

(erwartete Nutzung)

Hochladen von Dateien, Datenzusammenführung, -bereinigung, -strukturierung, -normalisierung und -transformation, Export von Dateien

**Benutzeroberfläche**

browserbasiertes GUI

**Visualisierungen**

(Analyse-, Input-,

✓



Outputkonfigurationen)

| Benutzerverwaltung   |   |
|--|---|
| <b>Personenverwaltung</b>  | ✗   |
| <b>Interne Kommunikationsmöglichkeiten</b><br>(z. B. Annotationsrichtlinien, Kommentarfunktionen, ...)   | ✗   |
| Daten- und Toolverwaltung  |   |
| <b>Zentrale/dezentrale Verwaltungsmöglichkeit</b>  | <input checked="" type="checkbox"/> mehrere Projekte möglich  |
| <b>Versionskontrolle</b>   | <input checked="" type="checkbox"/> jegliche Änderungen können nachverfolgt und rückgängig gemacht bzw. wiederhergestellt werden  |
| <b>Projektspezifische Einstellungen</b>  | <input checked="" type="checkbox"/>   |
| <b>API</b>   | <input checked="" type="checkbox"/> für Reconciliation  |
| <b>Möglichkeit auf simultanes Arbeiten</b>   | ✗   |
| Datenupload  |   |
| <b>Unterstützte Dateiformate</b>   | CSV, TSV, TXT, JSON, XML, ODS, XLS, XLSX, PX, MARC, RDF(JSON-LD, N3, N-Triples, Turtle, RDF/XML), Wikitext<br>Importmöglichkeiten auch über Weblinks, SQL-Datenbank oder Google Drive |
| <b>Informationen zur Datensicherheit</b>   | [nicht anwendbar, da lokale Ausführung]   |
| <b>Zugänglichkeit von verschiedenen Standorten/Geräten</b>   | ✗   |
| <b>Einschränkungen hinsichtlich der Datenmenge</b>   | max. 1 GB   |
| <b>Verlustfreier Upload von bereits bearbeiteten Dokumenten</b>  | <input checked="" type="checkbox"/>   |
| <b>Unterstützung von IIIF-Import</b>   | [nicht anwendbar]   |
| Datenbearbeitung (Normalisierungstool)   |   |
| <b>Komplexitätsgrad der Normalisierung</b><br>(z. B. Verfügbarkeit von Buttons, Drag&Drop-Funktion, ...) | <input checked="" type="checkbox"/> Buttons für Übernahme von Vorschlägen, Liste für Standardservices verfügbar   |
| <b>Reconciliation-Möglichkeiten entsprechend bestimmten Standards für digitale Editionen</b>             | <input checked="" type="checkbox"/> Wikidata, GND, GeoNames, Pleiades, etc.   |

**Anpassungsmöglichkeit und  
Validierung entsprechend  
projektspezifischen  
Konventionen/Schemata**



---

## 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]