

DigEdTnT - Webinarreihe Tools & Transitions III

OpenRefine (vs. ba[sic?])

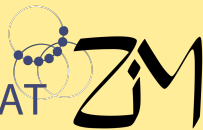
<https://digedtnt.github.io>

17.10.2023

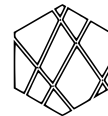
We work for
tomorrow



CLARIAH-AT



ZENTRUM FÜR
INFORMATIONSMODELLIERUNG
AUSTRIAN CENTRE FOR
DIGITAL HUMANITIES





Transkribus®



ediarum



ba[sic?]



Bild & Transkription

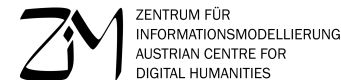
Annotation

Normalisierung

Publikation

<https://dightednt.github.io>

DigEdTnT Repository



Termine

- 19.09.2023 - FromThePage [Transkription] → ediarum.BASE [Annotation]
- 03.10.2023 - Transkribus [Transkription] → FairCopy [Annotation]
- 17.10.2023 - OpenRefine (vs. ba[sic?]) [Normalisierung]
- 31.10.2023 - ediarum.BASE [Annotation] → teiPublisher [Publikation]
- 14.11.2023 - FairCopy [Annotation] → ediarum.WEB [Publikation]

jeweils Dienstag, 17:00-18:00 Uhr

Programm

Tools & Transitions III: OpenRefine vs. ba[sic?]

- Einführung in OpenRefine
- Einführung in ba[sic?]
- Vergleich OpenRefine vs. ba[sic?]
- Fragerunde

OpenRefine

<https://openrefine.org/>

Tool: Datenbereinigung, Transformation, Organisation sowie Normalisierung und Reconciliation (Datenanreicherung) großer Datenmengen

Ziel: Normdaten-Anreicherung der Zutatenliste unseres [Beispielprojektes](#) (Kochrezepte aus dem Mittelalter)

- a) Normalisierung
- b) Reconciliation mit Wikidata Konzepten (Qid)

Kosten: Kostenlos und Open Source

OpenRefine

- Ursprünglich von Google entwickelt (Google Refine)
- Jetzt Open Source Data Wrangling Software
- “powerful tool for working with messy data
- **cleaning it**
- **transforming it** from one format into another
- and **extending it** with web services and external data.”



OpenRefine

- geschrieben in Java
- läuft **lokal** im Webbrowser

```
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF8
11:10:53.135 [      refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
11:10:53.198 [      refine_server] Initializing context: '/' from 'C:\Users\PC\Desktop\data\toc
efine-win-with-java-3.7.2\openrefine-3.7.2\webapp' (63ms)
11:10:56.692 [      refine] Starting OpenRefine 3.7.2 [f7ad526]... (3494ms)
11:10:56.693 [      refine] initializing FileProjectManager with dir (1ms)
11:10:56.693 [      refine] C:\Users\PC\AppData\Roaming\OpenRefine (0ms)
11:11:08.380 [      refine] POST /command/core/load-language (11687ms)
11:11:08.415 [      refine] GET /command/core/get-preference (35ms)
11:11:08.429 [      refine] POST /command/core/load-language (14ms)
11:11:08.447 [      refine] POST /command/core/load-language (18ms)
11:11:08.465 [      refine] POST /command/core/load-language (18ms)
11:11:08.540 [      refine] GET /command/core/get-importing-configuration (75ms)
11:11:08.562 [      refine] GET /command/core/get-all-project-tags (22ms)
11:11:08.582 [      refine] GET /command/core/get-all-project-metadata (20ms)
11:11:08.675 [      refine] GET /command/core/get-csrf-token (93ms)
11:11:08.697 [      refine] GET /command/core/get-languages (22ms)
11:11:09.005 [      refine] GET /command/core/get-version (308ms)
11:11:09.127 [      refine] GET /command/database/saved-connection (122ms)
```

OpenRefine - Kernfunktionalitäten

- **Datenbereinigung:** Ermöglicht die Korrektur von Inkonsistenzen, Duplikaten und Fehlern in Datensätzen
- **Datenfilterung und -sortierung:** Unterstützt das Filtern, Sortieren und Gruppieren von Daten zur effizienten Organisation und Analyse von Datensätzen
- **Datentransformation:** Unterstützt die Umwandlung von Daten von einem Format in ein anderes, z. B. von Excel/CSV zu JSON, XML, RDF ...
- **Reconciliation** (Datenanreicherung): Ermöglicht die semantische Erweiterung von Datensätzen über APIs oder anders zur Verfügung stehende externe Daten und dadurch der Zusammenführung von Datensätzen aus verschiedenen Quellen
- **OpenRefine API:** Ermöglicht die Kommunikation mit dem lokalen OpenRefine Server über Programmiersprachen (Python, Java, R, JS ...)

OpenRefine - Datenbereinigung

The screenshot shows the OpenRefine interface with a table of data. The 'Facet' menu is open, and the 'Duplicates facet' option is highlighted in yellow. The table has columns for 'president', 'prior', and 'party'. The 'president' column contains names like 'Franklin D. Roosevelt', 'Franklin Pierce', 'George H. W. Bush', etc. The 'prior' column contains titles like 'U.S. Representative for Illinois' 7th District (1847-1849)', 'U.S. Senator (Class 2) from Tennessee (1823-1825)', etc. The 'party' column contains 'Republican', 'Democratic', 'Republn', etc.

Duplikate entfernen

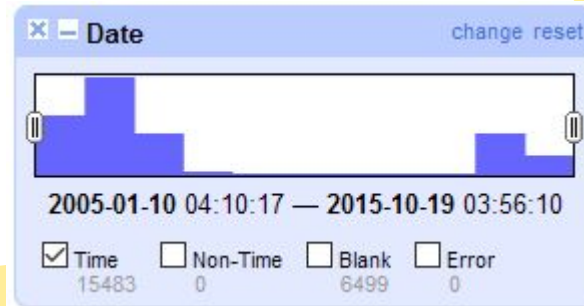
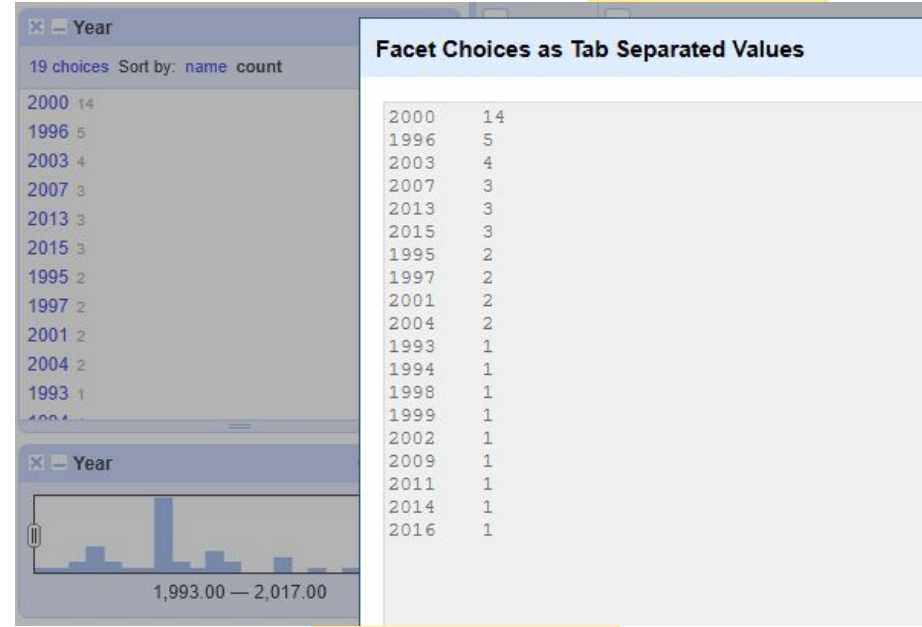
- Blank down oder
- Duplicates facet

The screenshot shows the OpenRefine interface with a table of 47 rows. The '47 rows' view is selected, and the 'Trim leading and trailing whitespace' option is highlighted in yellow. The table has columns for 'All', 'S.No.', 'president', 'prior', and 'party'. The 'president' column contains names like 'Franklin D. Roosevelt', 'Franklin Pierce', 'George H. W. Bush', etc. The 'prior' column contains titles like 'U.S. Representative for Illinois' 7th District (1847-1849)', 'U.S. Senator (Class 2) from Tennessee (1823-1825)', etc. The 'party' column contains 'Republican', 'Democratic', etc.

Zusammenführung von gleichen Einträgen

OpenRefine - Datenfilterung

- Text facet
- Numeric facet
- Timeline facet
- Scatterplot facet
- Custom facets



OpenRefine - Datentransformation

Templating export

Prefix

```
<list>
```

Row template

```
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'].
  {{ if(cells['deu'].value != 'ungelöst', '<label type="reg">' + cells['deu'].value + '</label>', '<label type
  <label type="alt">{{ cells['deu-enh'].value }}</label>
</item>
```

Row separator

Suffix

```
</list>
```

Reset template

Export Cancel

```
<list>
<item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agraeß</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agras</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agraz</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agres</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agresß</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agrest</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agreste</label>
</item><item xml:id="verjuice">
  <idno type="uri">https://www.wikidata.org/entity/Q106045
  <label type="reg">Agraz</label>
  <label type="alt">agreste</label>
</item><item xml:id="apple">
```

OpenRefine - Datentransformation

The screenshot shows the 'RDF Transform' dialog box in OpenRefine. The dialog is titled 'RDF Transform' and has tabs for 'Transform' and 'Preview'. Below the title bar, there are three buttons: 'Expand & Collapse' and 'Manage namespaces'. A text area contains the following information:

The RDF template below specifies how the RDF data is generated from your tabular data. The cells in each record of your data will get placed into nodes within the transform. Configure the transform by using column names and values, computed strings, or specified IRI as a subject, property, and object resources or literals. Compute strings using GREL. See the docs.

Base IRI: cocoa://info/ Edit

The main area of the dialog is divided into two sections: 'Transform' and 'Preview'. The 'Transform' section shows a list of available namespaces and a list of properties and classes. The 'Preview' section shows a list of nodes and their values. The dialog is annotated with several yellow boxes and red arrows:

- Base IRI control**: Points to the 'Base IRI' text.
- Expand & Collapse**: Points to the 'Expand & Collapse' button.
- Manage namespaces**: Points to the 'Manage namespaces' button.
- Classes**: Points to the 'R: [Index]' and 'choc:Entry' entries in the 'Available Namespaces' list.
- Properties**: Points to the 'choc:companyName', 'choc:beanLabel', 'choc:reference', 'choc:reviewDate', 'choc:percentCocoa', 'choc:companyLocation', 'choc:rating', 'choc:typeBean', and 'choc:beanOriginBroad' entries in the 'Available Namespaces' list.
- Delete any value**: Points to the 'x' icons next to the 'choc:companyName', 'choc:beanLabel', 'choc:reference', 'choc:reviewDate', 'choc:percentCocoa', 'choc:companyLocation', 'choc:rating', 'choc:typeBean', and 'choc:beanOriginBroad' entries.
- Data Values as Resources and Literals**: Points to the 'L: Company (Maker-if known)', 'L: Specific Bean Origin or Bar Name', 'R: REF', 'L: Review Date', 'L: Cocoa Percent', 'L: Company Location', 'L: Rating', 'R: Bean Type', and 'L: Broad Bean Origin' entries in the 'Preview' section.
- Import and export transform templates**: Points to the 'Import Template' and 'Export Template' buttons.
- Resize as needed**: Points to the bottom right corner of the dialog box.

The dialog also includes buttons for 'Add Root Node', 'Create multiple roots', 'Import Template', 'Export Template', and 'Save' at the bottom. The 'OK' and 'Cancel' buttons are at the bottom left.

OpenRefine - Reconciliation

- Reconciliation = Abgleich / Zusammenführung / Anreicherung
- Verbindungen zu Datenbanken herstellen
- Alle zusätzlichen Informationen in dieser DB werden verfügbar
- Wikidata (en) ist bereits integriert
- Services können hinzugefügt werden
 - Wikidata Deutsch: <https://wikidata.reconci.link/de/api>
 - Beschreibung: <https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources>
 - Liste: <https://reconciliation-api.github.io/testbench/>
 - z. B. GND <https://lobid.org/gnd/reconcile>
 - Dokumentation: <https://openrefine.org/docs/manual/reconciling>
 - Framework um selbst einen Service zu bauen: <https://github.com/codeforkjeff/conciliator>
 - API: <https://openrefine.org/docs/technical-reference/reconciliation-api> | <https://reconciliation-api.github.io>

OpenRefine - Reconciliation

5 10 25 50 100 500 1000 rows

| deu | eng |
|-------|-------|
| Agraz | |
| Agraz | |
| Agraz | |
| Agraz | |
| Agraz | |
| Agraz | |
| Agraz | |
| Agraz | |
| Apfel | |
| Apfel | |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |
| Apfel | apple |

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile
 - Start reconciling...

Reconcile column "deu"

Services

Wikidata (en)

Pick a service or extension on left

Add standard service

Enter the service's URL

Add service Cancel

Add standard service... Discover services... Start reconciling... Cancel

OpenRefine - Reconciliation

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. | agrosße | Agraz | verjuice Choose new match |
| ☆ | 9. | apfel | Apfel | apple edit |

apple (100)

Apple (100)

Muggsy Bogu

Malus pumila

Apple II serie

Apple Record

Apple III (100

Apple (100)

The Apple (1

Apple River (100)

apple (100)

Match this cell

Match all identical cells

Cancel



apple (Q89)

fruit of the apple tree

OpenRefine - Reconciliation

Kann ich auch selbst meine lokalen/externen Daten “reconcilen”? → Ja!

Reconciliation Service API v0.2

A protocol for data matching on the Web

[Final Community Group Report](#) 10 April 2023

This version:

<https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/>

Servers

Beyond publicly available hosted services, some software exposes reconciliation endpoints locally.

csv-reconcile

[csv-reconcile](#) is a Python based project can be used to run a reconciliation service off any CSV file.

reconcile-csv

[reconcile-csv](#) is a Clojure based project can be used to run a reconciliation service off any CSV file.

OpenRefine's RDF extension

The [RDF extension](#) can be used to wrap a SPARQL endpoint into a reconciliation service.

Open Reconcile

[Open Reconcile](#) is a project that can be used to run a reconciliation service on top of a SQL database.

OpenRefine - API

<https://raw.githubusercontent.com/opencultureconsulting/openrefine-client/master/openrefine-client-peek.gif>

<https://openrefine.org/docs/technical-reference/openrefine-api>

```
var project = openrefine
  .create('data_cleanup_project') // .create() auto-generates a project name
  .accept('csv')
  .accept({
    separator: ',',
    ignoreLines: 1
  })
  .expose('csv')
  .keep(true) // keep data after end() or pipe; default is not keeping
  .use([
    {
      "op": "core/column-split",
      "description": "Split column DATE by separator",
      "engineConfig": {
        "facets": [],
        "mode": "row-based"
      },
      "columnName": "DATE",
      "guessCellType": true,
      "removeOriginalColumn": true,
      "mode": "separator",
      "separator": "-",
      "regex": false,
      "maxColumns": 0
    }
  ])
  .use(customCleanupAddress()) // customCleanupAddress() returns an array of operations
```

Beispiel für Nutzung der OR API in NodeJS

OpenRefine - Extensions

- <https://openrefine.org/extensions>
- RDF
- GeoJSON
- Named-Entity Recognition
- ...

OpenRefine - Export

- Standardmethoden: TSV, CSV, Excel, ODF, HTML
- Custom tabular: Einstellungen ändern und Upload nach Google Spreadsheets ...
- RDF Extension: RDF/XML und RDF/Turtle
- Templating: customizable (JSON, XML ...)
- Database Export: SQL
- Projektexport
- Wikidata: Upload, QuickStatements, Schema



OpenRefine - Ressourcen

- van Hooland, S., Verborgh, R., & De Wilde, M. (2013). Cleaning Data with OpenRefine. In A. Crymble, P. Burns, & N. McGregor (Eds.), The Programming Historian. <https://doi.org/10.46430/phen0023>.
- YouTube: [Reconcillation in OpenRefine](#): Part 1 by Owen Stephens
- YouTube: [Reconcillation in OpenRefine](#): Part 2 by Owen Stephens
- [Open Refine for Librarians: http://liwong.blogspot.com/](http://liwong.blogspot.com/)
- Sample Datasets: <https://github.com/OpenRefine/OpenRefine/wiki/Sample-Datasets>
- Documentation: <https://openrefine.org/docs>
- FAQ: <https://github.com/OpenRefine/OpenRefine/wiki/FAQ>

ba[sic?]

<https://github.com/saw-leipzig/basic.app> bzw. <https://basicdemo.saw-leipzig.de/>

- Tool:** Verknüpfung von Named Entities (derzeit nur Personen, Organisationen, Orte) mit von Normdatenanbietern (derzeit nur GND und GeoNames) bezogenen Identifiern
- Ziel:** Normdaten-Anreicherung der Personen, Orte und Organisationen unseres [Beispielprojektes](#) (Briefe von/an Hugo Schuchardt)
a) Reconciliation mit GND und GeoNames Konzepten
- Kosten:** Kostenlos und Open Source

ba[sic?]

- Better Authorities [Search, Identify, Connect]
- Sächsische Akademie der Wissenschaften zu Leipzig
- Webapp - <https://github.com/saw-leipzig/basic.app> bzw. <https://basicdemo.saw-leipzig.de/>
- Verknüpfung der importierten Named Entities mit von Normdatenanbietern (derzeit nur GND und GeoNames) bezogenen Identifiern
- Optionale Anreicherung mit weiteren Informationen (manuell oder von GND/GeoNames)

ba[sic?]

ba[sic?]  Persons  Places  Organisations

Wolfgang Ambros

safe   

new Identifier URL

Search for



118839292

120255243

Name

Wolfgang Ambros

Alternate Names

W. Ambros

Gender

Mann

Date of Birth

1952

Place of Birth

Wien

Place of Activity

Waidring

Profession

Sänger, Musiker, Komponist, Schauspi...

Affiliation

Austria 3 (Musikgruppe)

Further Information

Österreichischer Liedermacher; DMA: F...

Identifier URL

Search for



Local Storage Configuration

New dataset name



Current Dataset: *lidal* 

Last modified: *Mon Oct 16 2023 13:33:46 GMT+0200 (Mitteleuropäische Sommerzeit)*

Objects: 2

Filter

safe **1**

unsafe **0**

unavailable **0**

unchecked **1**

duplicates **0**

incorrect **0**

Filter by title

Showing 2 of 2 objects

Start typing to filter results



Sort


Order: *ascending* 

Name


ID

Status

Actions


 Backup as JSON

 Restore from JSON

 Import from CSV **csv2cmi**

 Merge with CSV **csv2cmi**

 Import from XML

 Add person

OpenRefine vs. ba[sic?]

→ ?

- OpenRefine kann alles, was ba[sic?] kann
- ba[sic?] wurde in einem speziellen Forschungskontext entwickelt
- ba[sic?] für kleine Datensätze mit in der GND vorkommenden Entitäten
- ba[sic?] für bereits im CMIF ([Correspondence Metadata Interchange Format](#)) standardisiert vorliegende Briefdaten

```
<correspAction type="sent">
  <persName ref="http://d-nb.info/gnd/118629662">Weber, Carl Maria von</persName>
  <placeName ref="http://www.geonames.org/2935022">Dresden</placeName>
  <date when="1825-05-07"/>
</correspAction>
```



Noch mehr Normalisierung?

<https://reconciliation-api.github.io/census/>



Known clients

This page lists software packages which interact with reconciliation services using the reconciliation API.

- [OpenRefine](#)
- [Cocoda](#)
- [Alma Refine](#)
- [reconciler](#) (Python library)
- [reconciler](#) (R library)
- [SemTUI](#)
- [Testbench](#)
- [TEI Publisher](#)
- Any other?

Vielen Dank

<https://digedtnt.github.io/>

christian.steiner@dhcraft.org

We work for
tomorrow

