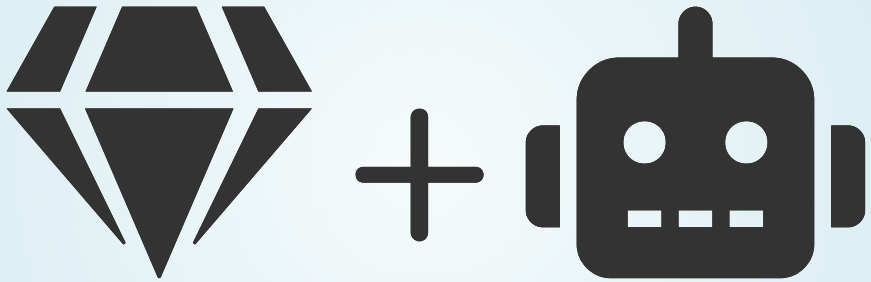
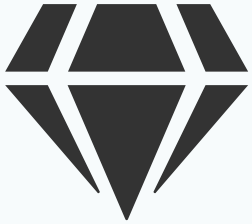
AUTOMATISIERUNGSMÖGLICHKEITEN DER SOFTWARE OPENREFINE



FELIX LOHMEIER

#BIBTAG21, 16.06.2021

OPENREFINE



- grafische Oberfläche, die einer klassischen Tabellenverarbeitungssoftware ähnelt
- dient der Analyse, Bereinigung, Konvertierung und Anreicherung von Daten
- wird in der Regel lokal auf einem Computer installiert und über den Browser bedient
- Open-Source-Software mit aktiver Community

LIVE-VORFÜHRUNG TEIL 1

- Szenario: Mit OpenRefine Liste von Schriftsteller*innen (Name, Geburtsdatum) mit Wikidata abgleichen und darüber GND-ID und Geburtsort ermitteln.
- Vorgehen:
  1. Import CSV-Datei
  2. Ableitung Spalte Geburtsjahr aus Geburtsdatum
  3. Abgleich mit Wikidata und Anreicherung
  4. Export in TSV-Datei

ver tool for working with messy data.

art Over | Configure Parsing Options | Project name Clipboard | Tags

lame	Geburtsdatum
rnst Jünger	29.03.1895
ilde Domin	27. Juli 1909
ermann Hesse	2.7.1877
ertrud von LeFort	11.10.1876
ohann Wolfgang von Goethe	1749
arah Kirsch	1935
riedrich von Schiller	10. November 1759
icarda Huch	18.7.1864
arl Wolfskehl	1869
uise Rinser	30. April 1911

IMPORT

arse data as

SV / TSV / separator-based files

- e-based text files
- ed-width field text files
- -Axis text files
- ON files
- ARC files
- ON-LD files
- F/N3 files
- F/N-Triples files
- F/Turtle files

Character encoding

Columns are separated by
- commas (CSV)
- tabs (TSV)
- custom: ,

Trim leading & trailing whitespace from strings

Escape special characters with \

Column names (comma separated):

Ignore first 0 line(s) at beginning of f

Parse next 1 line(s) as column head

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Use character " to enclose cells contain

Attempt to parse cell text into numbers

Store bla

Store bla

Store file

Store arc

oard Permalink

Show: 5 10 25 50 rows

« first ‹ previous 1

	Geburtsdatum
	29.03.1895
	27. Juli 1909
	2.7.1877
	11.10.1876
n Goethe	1749
	1935
	10. November 1759
	18.7.1864
	1869
	30. April 1911

Add column based on column Geburtsdatum

New column name: Geburtsjahr

On error: set to blank | store error | copy value from original column

Expression — Language: General Refine Expression Language (GREL)

```
value.slice(-4)
```

No syntax error.

TRANSFORMATION

Previe

row	value	value.slice(-4)
1.	29.03.1895	1895
2.	27. Juli 1909	1909
3.	2.7.1877	1877
4.	11.10.1876	1876
5.	1749	1749
6.	1935	1935

OK Cancel

oard Permalink

3 / 3

All Remove All

change

change reset

10 rows

Show as: rows records

All	Name
1.	Ernst Jünger Choose new match
2.	Hilde Domin Choose new match
3.	Hermann Hesse Choose new match
4.	Gertrud von Le F Choose new match
5.	Johann Wolfgang Choose new match
6.	Sarah Kirsch Choose new match
7.	Friedrich Schiller Choose new match
8.	Ricarda Huch Choose new match
9.	Karl Wolfskehl Choose new match
10.	Luise Rinser Choose new match

Add columns from reconciled column Name

Add Property

GND ID

Suggested Properties

child
country of citizenship
employer
ethnic group

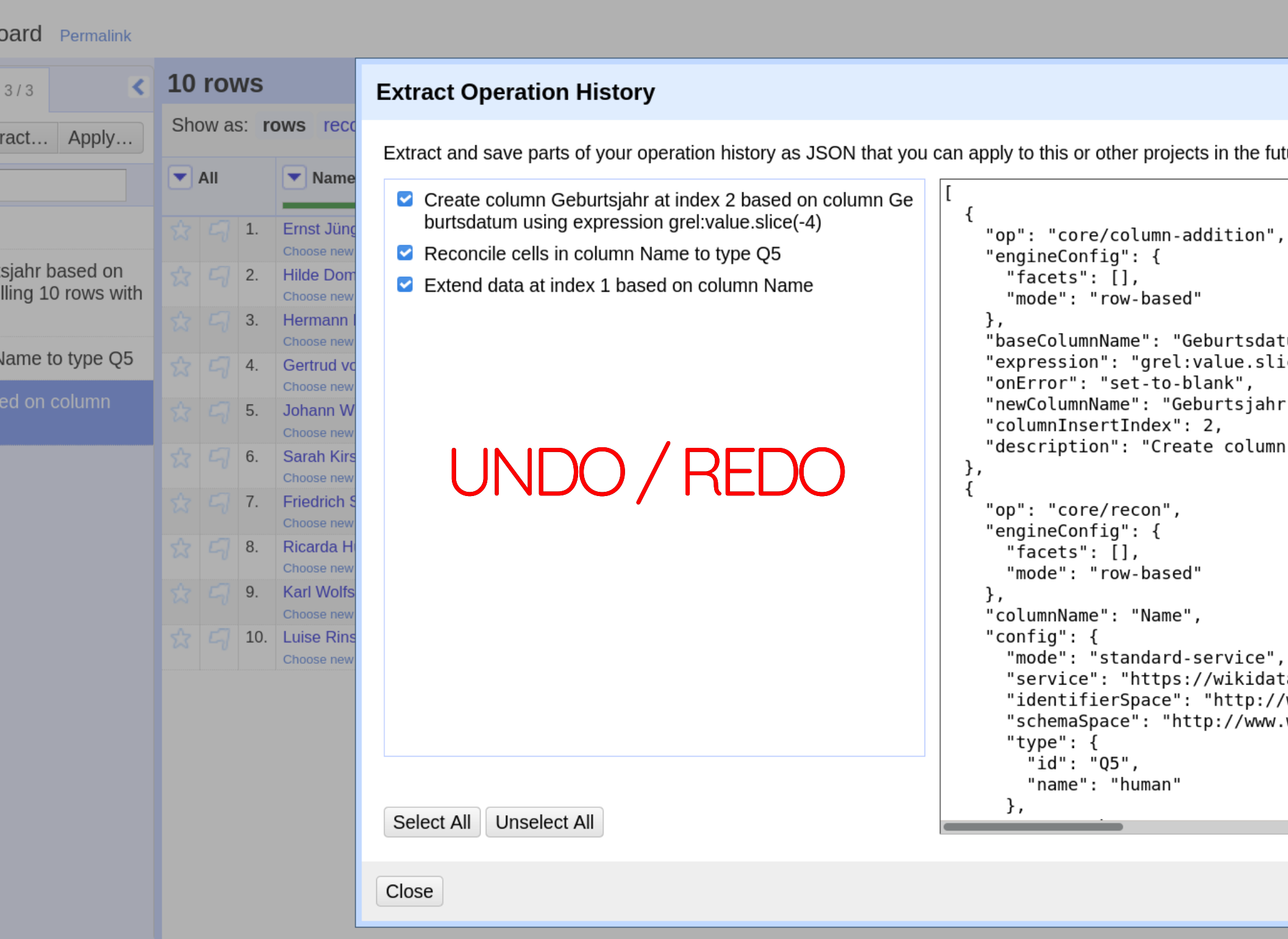
ANREICHERUNG

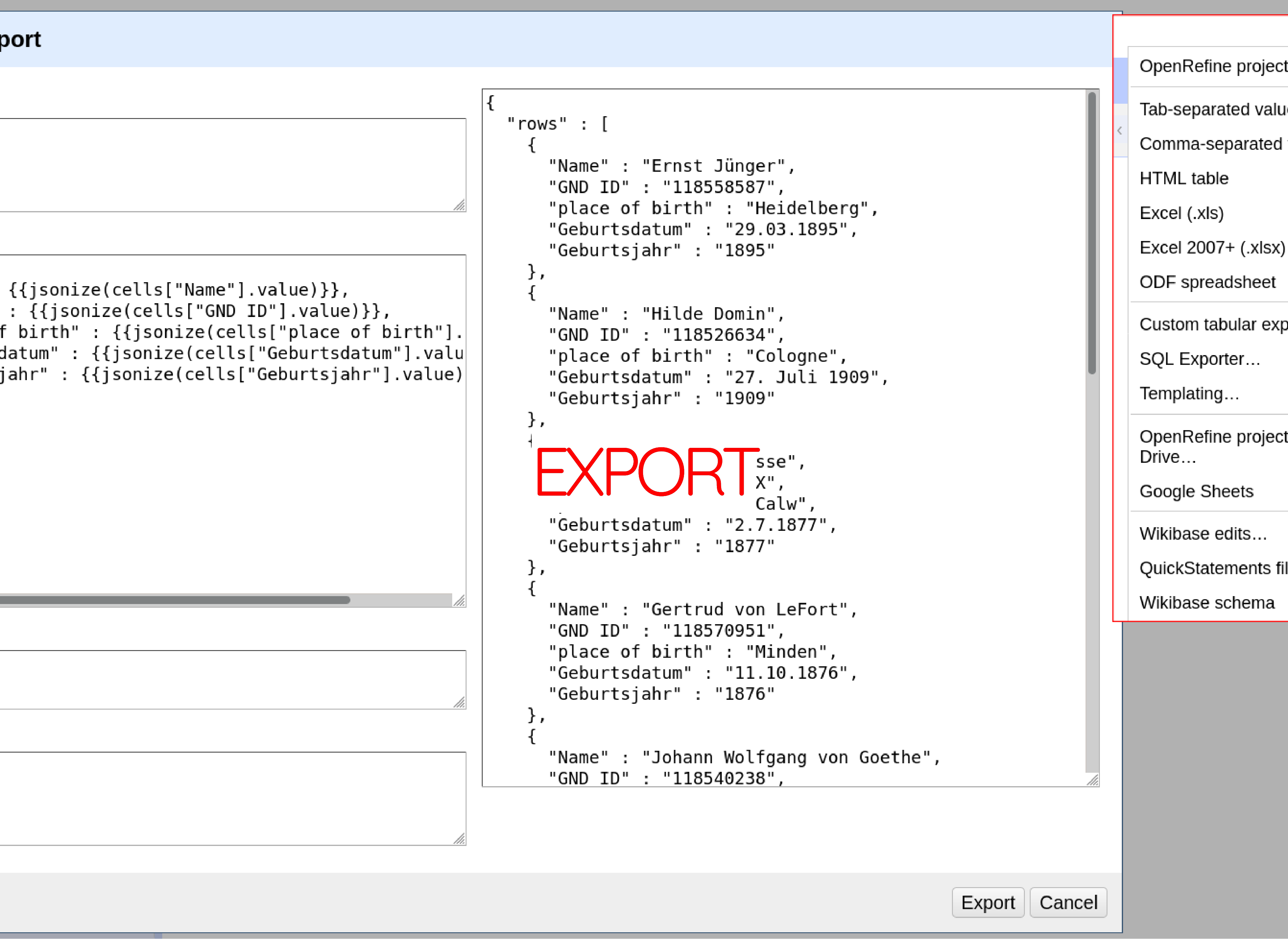
image
member of
member of political party
mother
native language
occupation
place of birth
place of burial
place of death

Preview

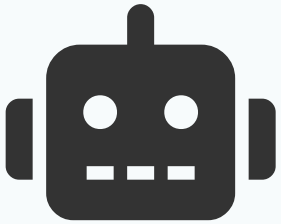
Name	GND ID remove configu
Ernst Jünger	118558587
Hilde Domin	118526634
Hermann Hesse	11855042X
Gertrud von Le Fort	118570951
Johann Wolfgang von Goethe	118540238
Sarah Kirsch	118562487
Friedrich Schiller	118607626
Ricarda Huch	118554190
Karl Wolfskehl	118634976
Luise Rinser	118601172

OK Cancel


Extract Operation History
Extract and save parts of your operation history as JSON that you can apply to this or other projects in the fut
Create column Geburtsjahr at index 2 based on column Ge burtsdatum using expression grel:value.slice(-4)
Reconcile cells in column Name to type Q5
Extend data at index 1 based on column Name
UNDO / REDO
Select All
Unselect All
[
{
"op": "core/column-addition",
"engineConfig": {
"facets": [],
"mode": "row-based"
},
"baseColumnName": "Geburtsdat
"expression": "grel:value.sli
"onError": "set-to-blank",
"newColumnName": "Geburtsjahr
"columnInsertIndex": 2,
"description": "Create column
},
{
"op": "core/recon",
"engineConfig": {
"facets": [],
"mode": "row-based"
},
"columnName": "Name",
"config": {
"mode": "standard-service",
"service": "https://wikidat
"identifierSpace": "http://
"schemaSpace": "http://www.
"type": {
"id": "Q5",
"name": "human"
},
Close


port
{{jsonize(cells["Name"].value)}},
: {{jsonize(cells["GND ID"].value)}},
f birth" : {{jsonize(cells["place of birth"].
datum" : {{jsonize(cells["Geburtsdatum"].valu
jahr" : {{jsonize(cells["Geburtsjahr"].value)
{
"rows" : [
{
"Name" : "Ernst Jünger",
"GND ID" : "118558587",
"place of birth" : "Heidelberg",
"Geburtsdatum" : "29.03.1895",
"Geburtsjahr" : "1895"
},
{
"Name" : "Hilde Domin",
"GND ID" : "118526634",
"place of birth" : "Cologne",
"Geburtsdatum" : "27. Juli 1909",
"Geburtsjahr" : "1909"
},
EXPORT
sse",
X",
Calw",
"Geburtsdatum" : "2.7.1877",
"Geburtsjahr" : "1877"
},
{
"Name" : "Gertrud von LeFort",
"GND ID" : "118570951",
"place of birth" : "Minden",
"Geburtsdatum" : "11.10.1876",
"Geburtsjahr" : "1876"
},
{
"Name" : "Johann Wolfgang von Goethe",
"GND ID" : "118540238",
Export
Cancel
OpenRefine project
Tab-separated valu
Comma-separated
HTML table
Excel (.xls)
Excel 2007+ (.xlsx)
ODF spreadsheet
Custom tabular exp
SQL Exporter…
Templating…
OpenRefine project
Drive…
Google Sheets
Wikibase edits…
QuickStatements fil
Wikibase schema

WARUM AUTOMATISIEREN?



- Zeitersparnis bei sich wiederholenden Aufgaben
- Reproduzierbarkeit durch eindeutige Dokumentation im Code
- Arbeitsteilung zwischen Metadaten-Expert*innen und IT-Personal

CLIENT-SERVER-ARCHITEKTUR

- Web-Applikation mit klarer Trennung zwischen Datenhaltung und Benutzerschnittstelle
  - Server: Java servlet ausgeführt in Jetty Webserver; Datenhaltung In-Memory bzw. dateibasiert
  - Client: HTML, CSS und Javascript
- Server und Client kommunizieren über HTTP GET und POST (vgl. OpenRefine API)

VERFÜGBARE CLIENTS

- Python *
- R *
- Java *
- Bash *
- .NET *
- PHP
- Ruby
- NodeJS
- Rust

* unterstützt OpenRefine 3.3 und neuer

OPENREFINE-CLIENT

- Fork des Python-Clients mit erweitertem Kommandozeilen-Interface
- Als kleine ausführbare Datei (5 MB) erhältlich für Windows, Mac und Linux
- Kann somit unabhängig von der Programmiersprache in Workflows genutzt werden

LIVE-VORFÜHRUNG TEIL 2

- Automatisierung der manuellen Schritte aus Teil 1
- Wer es direkt selbst mit nachvollziehen möchte, kann jetzt eine Arbeitsumgebung via mybinder.org starten: bibtag21-automatisierung-openrefine
- Vorgehen:
  1. OpenRefine starten
  2. openrefine-client ausführen
     - Import CSV-Datei
     - Anwendung der Undo/Redo-Historie
     - Export in TSV

Schritt 1: OpenRefine Server starten

Starten Sie hier als erstes den OpenRefine-Server, in dem Sie die vorbereiteten Befehle mit dem Play-Button "abspielen".

[1]:
```
mkdir data
```

[*]:
```
openrefine/refine -d data
```

```
Using refine.ini for configuration
You have 52331M of free memory.
Your current configuration is set to use 1400M of memory.
OpenRefine can run better when given more memory. Read our FAQ on how to allocate more memory here:
https://github.com/OpenRefine/OpenRefine/wiki/FAQ:-Allocate-More-Memory
/usr/bin/java -cp server/classes:server/target/lib/* -Xms1400M -Xmx1400M -Drefine.memory=1400M -Drefine.max_form_content_size
osity=info -Dpython.pa                                                                    ocal/share/google/refine/cachedir -D
Drefine.webapp=main/we                                                                    440 com.google.refine.Refine
Starting OpenRefine at

19:41:04.844 [           refine_server] Starting Server bound to /127.0.0.1:3333 (0ms)
19:41:04.845 [           refine_server] refine.memory size: 1400M JVM Max heap: 1419116544 (1ms)
19:41:04.854 [           refine_server] Initializing context: '/' from '/home/jovyan/openrefine/webapp' (9ms)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/jovyan/openrefine/server/target/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j/impl/StaticLo
SLF4J: Found binding in [jar:file:/home/jovyan/openrefine/webapp/WEB-INF/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j/impl/StaticL
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
19:41:06.826 [                  refine] Starting OpenRefine 3.4.1 [437dc4d]... (1972ms)
19:41:06.826 [                  refine] initializing FileProjectManager with dir (0ms)
19:41:06.826 [                  refine] data (0ms)
19:41:06.833 [       FileProjectManager] Failed to load workspace from any attempted alternatives. (7ms)
19:41:16.417 [                  refine] Sorry, some error prevented us from launching the browser for you.

 Point your browser to http://127.0.0.1:3333/ to start using Refine. (9584ms)
19:41:20.715 [                  refine] GET /command/core/get-csrf-token (4298ms)
19:41:20.803 [                  refine] POST /command/core/create-project-from-upload (88ms)
19:41:21.405 [                  refine] GET /command/core/get-models (602ms)
19:41:21.542 [                  refine] POST /command/core/get-rows (137ms)
19:41:23.427 [                  refine] GET /command/core/get-csrf-token (1885ms)
```

OPENREFINE STARTEN

Schritt 2: openrefine-client ausführen

Wenn der Server gestartet ist (dauert 10-15 Sekunden), dann können Sie hier mit den vorbereiteten Befehlen OpenRefine über die Kommandozeile bedienen.

[1]:
```
openrefine-client --create schriftstellerinnen.csv
```

```
id: 2615731062966
rows: 10
```

[2]:
```
openrefine-client --apply history.json schriftstellerinnen
```

```
File history.json has been successfully applied to project 2615731062966
```

[3]:
```
openrefine-client --export schriftstellerinnen
```

```
Name    GND ID  place of birth  Geburtsdatum    Geburtsjahr
Ernst Jünger    118558587       Heidelberg      29.03.1895      1895
Hilde Domin     118526634       Cologne 27. Juli 1909   1909
Herma
Gertr
Johar
Sarah
Friedrich Schiller      118607626       Marbach am Neckar       10. November 1759       1759
Ricarda Huch    118554190       Brunswick       18.7.1864       1864
Karl Wolfskehl  118634976       Darmstadt       1869    1869
Luise Rinser    118601172       Pitzling        30. April 1911  1911
```

OPENREFINE-CLIENT AUSFÜHREN

WORKFLOW

Für eine vollständige Automatisierung müsste der gezeigte Ablauf noch als Workflow eingerichtet werden.

- Logging
- Scheduling
- Ergebnisvalidierung
- Fehlerbehandlung
- ggf. Parallelisierung
- ggf. Caching

VERFÜGBARE WORKFLOW-TOOLS FÜR OPENREFINE

- openrefine-batch: Shell-Script für einfache Workflows
- openrefine-task-runner: Vorlagen für komplexere Workflows, nutzt go-task als task runner.

EINSCHRÄNKUNGEN

- Für größere Datenmengen wird viel Arbeitsspeicher benötigt (Limitierung entfällt für OpenRefine 4.0, das auf Spark basieren wird)
- Undo/Redo-Historie im JSON-Dateiformat ist schwierig anzupassen, was die Nachnutzbarkeit einschränkt
- Das Entwicklerteam rät eher davon ab, komplexe Workflows mit OpenRefine umzusetzen (vgl. Einschätzung auf der Mailingliste)

PRAXISBEISPIELE

- Konvertierung von Bibliotheca und Alephino nach PICA+ für die Berufsakademien Sachsen: ba-sachsen-pica bei GitHub
- Harvesting von OAI-PMH-Schnittstellen und Transformation in METS/MODS für das Portal noah.nrw: noah bei GitHub
- Datenintegration für den neuen Katalog des DLA Marbach: Präsentation auf der ELAG 2019

FRAGEN / DISKUSSION

- Für die Vorführung der Workflow-Tools fehlte die Zeit, daher separate Videokonferenz mit Live-Vorführung von openrefine-task-runner und Gelegenheit zu vertiefter Diskussion: Terminumfrage (28.6. - 9.7.)
- Spontane Fragen und Kommentare gerne jetzt ☻

CREDITS / LIZENZ

- Icons: Font Awesome (MIT Licence)
- Präsentations-Framework: Reveal.js (MIT Licence)
- Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.