

OpenRefine Tool for Data Reconciliation

Jonathan Ward
Editor, Getty Vocabulary Program

OpenRefine

→ **Open Source: <http://openrefine.org>**

Available in multiple languages; runs in a browser.

→ **New GVP service for data reconciliation & cleaning**

◆ **Tutorial here:**

<http://www.getty.edu/research/tools/vocabularies/obtain/openrefine.html>


◆ **Uses GREL language (General Refine Expression Language)**

<https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions>

OpenRefine – “Create Project”

The screenshot shows the OpenRefine web interface in a browser window. The browser's address bar shows the URL `http://127.0.0.1:3333`. The page title is "OpenRefine" with the subtitle "A free tool for working with messy data." The main content area is titled "Create a project by importing data. What kinds of data files can I import?" and lists supported formats: "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." Below this, there are four sections for "Get data from": "This Computer" (with a "Browse..." button and "No files selected." text), "Web Addresses (URLs)" (with a "Next »" button), "Clipboard", and "Google Data". On the left side, a navigation menu includes "Create Project" (highlighted with an orange arrow), "Open Project", "Import Project", and "Language Settings". At the bottom left, there is a blue diamond logo and the text "Version 3.2:beta [8d89a2a]". At the bottom right of the menu, there are links for "Preferences", "Help", and "About".

OpenRefine A power tool for working with messy data.

Project name: belvedere.csv Tags  Create Project »

	ID	family name	given name	variant names	artist group	date of birth	date of death	place of birth	place of death	sex (m=male/w=female)	in Austria since	lived in	roles	teacher (ID)
1.	4965	Abandio				0000-00-00	0000-00-00			m	0000-00-00	Steiermark	Maurermeister	
2.	4964	Abart	Franz			1769-12-22	1863-09-10	Schlinig (Tirol)	Sankt Niklausen (Schweiz)	m	0000-00-00			
3.	4963	Abb	Johann			0000-00-00	0000-00-00			m	0000-00-00	1697 - 1704 Mariazell	Bildhauerei	
4.	4966	Abbati	Vincenz			0000-00-00	0000-00-00	Neapel (Italien)		m	0000-00-00	um 1843 Graz, später Florenz, venedig, Neapel	Malerei	
5.	4967	Abbiati	Julius			0000-00-00	0000-00-00	wohl aus der Mailände Familie Abbiati stammend		m	0000-00-00	1844 Wien	Malerei	
6.	4968	Abdank	Alois			0000-00-00	0000-00-00			m	0000-00-00	7.8. 1752: Aufnahme in wiener Akademie	Stukkateur	
7.	4969	Abdank	Gottfried			0000-00-00	vor 1770-00-00			m	0000-00-00	2.5. 1749: Aufnahme in die Wiener Akademie	Stukkateur	
8.	4970	Abdank	Paul Franz			1765-02-00	0000-00-00	Erlau (Ungarn)	Erlau (Ungarn)	m	0000-00-00	ab 1. 10. 1735 an der Wiener Akademie		
9.	4971	Abdank	Thomas Christian			1677-00-00	1743-08-07			m	0000-00-00			
10.	4972	Abdullah- Hammerschmidt	Anni Marie			1873-10-29	2/13/1916	Wien	Wien	w	0000-00-00			
11	19857	Ahal	Arnold			0000-00-00	1564-02-14			m	0000-00-00			

Parse data as

Character encoding

Update Preview

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

RDF/N-Triples files

Columns are separated by

commas (CSV)

tabs (TSV)

custom: ,

Escape special characters with \

Column names (comma separated):

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

Use character " " to enclose cells containing column separators

Parse cell text into numbers, dates, ...

Store blank rows

Store blank cells as nulls

Store file source (file names, URLs) in each row

Facet / Filter Undo / Redo 0 / 0

20799 rows

Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows

« first « previous 1 - 50 next » last »

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	ID	family name	given name	variant names	artist group	date of birth	date of death	place of birth	place of death	sex (m=male/w)	in Austria since	lived in
1.	4965	Abandio					0000-00-00	0000-00-00			m	0000-00-00	Steiermark
2.	4964	Abart	Franz				1769-12-22	1863-09-10	Schlinig (Tirol)	Sankt Niklausen (Schweiz)	m	0000-00-00	
3.	4963	Abb	Johann				0000-00-00	0000-00-00			m	0000-00-00	1697 - 1704 Mariazell
4.	4966	Abbati	Vincenz				0000-00-00	0000-00-00	Neapel (Italien)		m	0000-00-00	um 1843 Graz, später Florenz, venedig, Neapel
5.	4967	Abbiati	Julius				0000-00-00	0000-00-00	wohl aus der Mailände Familie Abbiati stammend		m	0000-00-00	1844 Wien
6.	4968	Abdank	Alois				0000-00-00	0000-00-00			m	0000-00-00	7.8. 1752: Aufnahme in wiener Akademi
7.	4969	Abdank	Gottfried				0000-00-00	vor 1770-00-00			m	0000-00-00	2.5. 1749: Aufnahme in die Wiener Akademi
8.	4970	Abdank	Paul Franz				1765-02-00	0000-00-00	Erlau (Ungarn)	Erlau (Ungarn)	m	0000-00-00	ab 1. 10. 1735 a der Wiener Akademie
9.	4971	Abdank	Thomas Christian				1677-00-00	1743-08-07			m	0000-00-00	
10.	4972	Abdullah-Hammerschmidt	Anni Marie				1873-10-29	2/13/1916	Wien	Wien	w	0000-00-00	
11.	19857	Abel	Arnold				0000-00-00	1564-02-14			m	0000-00-00	
12.	19858	Abel	Bernhard				0000-00-00	1563-10-13			m	0000-00-00	
13.	4974	Abel	Franz				1860-06-07	0000-00-00	Laa an der Thaya (Niederösterreich)		m	0000-00-00	
14.	4975	Abel	Friedrich				0000-00-00	0000-00-00			m	0000-00-00	
15.	4976	Abel	Gustav				1/25/1902	6/2/1963	Wien	Wien	m	0000-00-00	
16.	19493	Abel	Josef (Joseph)				1764-08-22	1818-10-07	Aschach an der Donau (Oberösterreich)	Wien	m	0000-00-00	
17.	4977	Abel	Leona				1872-08-11	0000-00-00	Budapest (Ungarn)		w	0000-00-00	
18.	4978	Abel	Lothar				1841-02-15	1896-06-24	Wien	Wien	m	0000-00-00	
19.	20506	Abel	Maria				0000-00-00	0000-00-00			w	0000-00-00	
20.	20505	Abel	Marie				0000-00-00	0000-00-00			w	0000-00-00	
21.	4979	Abels D'Albert	Erika				1896-11-03	3/7/1975	Berlin	Paris	w	1898-00-00	Wien, Paris
22.	4980	Abenthung	Josef				1719-00-00	1802-00-00	Götzens		m	0000-00-00	
23.	20867	Aberer	Ilse				1954-00-00	0000-00-00	Dornbirn		w	0000-00-00	Götzis (Vorarlberg)
24.	2	Aberle	Peter				8/21/1945	0000-00-00	Wien		m	0000-00-00	

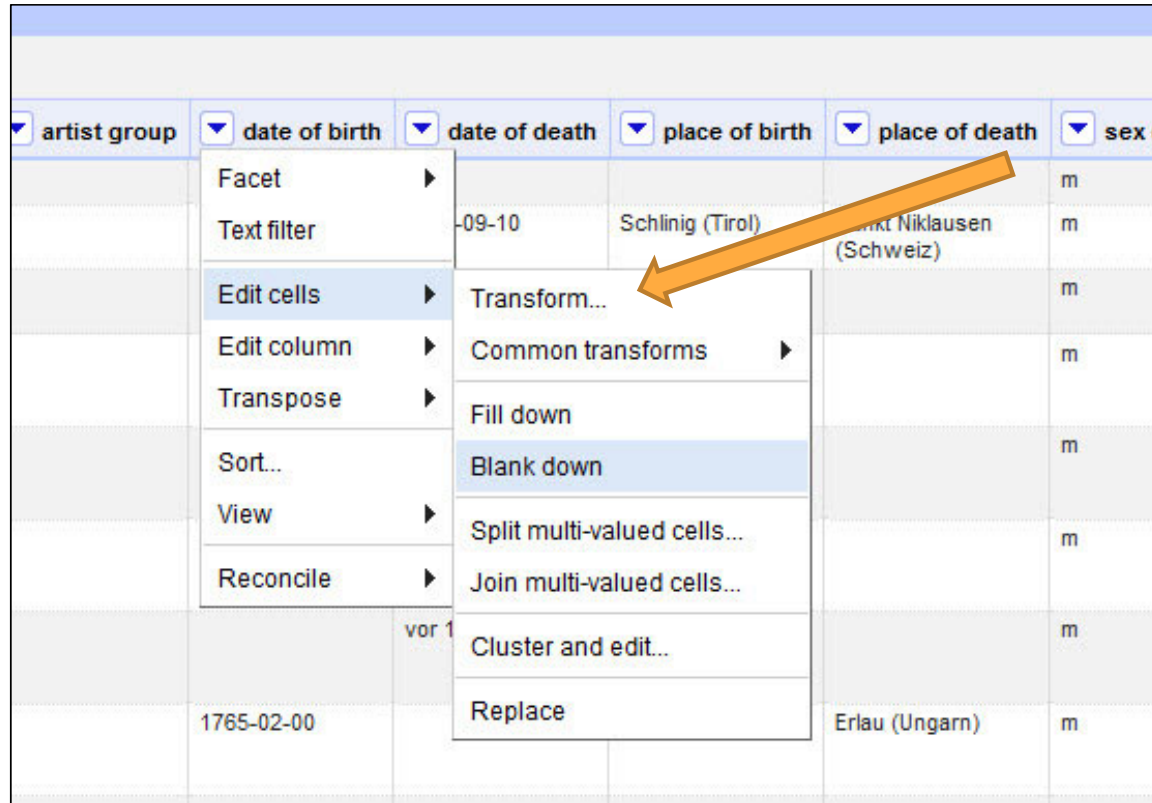
OpenRefine: Applying a global fix Simple cleanup of “date” column

The screenshot shows the OpenRefine interface with a global fix dialog open over a table. The dialog has a 'Data type' dropdown set to 'text' and a text input field containing '0000-00-00'. Below the input field are three buttons: 'Apply' (with 'Enter' below it), 'Apply to All Identical Cells' (with 'Ctrl-Enter' below it), and 'Cancel' (with 'Esc' below it). The background table has columns for 'date of birth', 'date of death', 'place of birth', 'place of death', 'sex (m=male/w)', and 'in Au'. The 'date of death' column contains several '0000-00-00' entries. One row is partially visible with the text 'wohl aus der Mailände Familie' and 'Abbi...'. Navigation controls '« first < pre' are visible in the top right corner.

Simple cleanup of “date” column

group	date of birth	date of death	p
	1769-12-22	1863-09-10	Schlin
			Neape
			wohl Mailän Abbia
		vor 1770-00-00	
	1765-02-00		Erlau
	1677-00-00	1743-08-07	
	1873-10-29	2/13/1916	Wien
		1564-02-14	
		1563-10-13	
	1860-06-07		Laa a (Niede
	1/25/1902	6/2/1963	Wien
	1764-08-22	1818-10-07	Asche Donau (Ober
	1872-08-11		Budap
	1841-02-15	1896-06-24	Wien

Simple cleanup of “date” column Transform using GREL expression



The screenshot shows a data table with columns: artist group, date of birth, date of death, place of birth, place of death, and sex. A context menu is open over the 'date of birth' column, with the 'Transform...' option highlighted. An orange arrow points to this option. The menu also includes options like Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, Reconcile, Common transforms, Fill down, Blank down, Split multi-valued cells..., Join multi-valued cells..., Cluster and edit..., and Replace.

artist group	date of birth	date of death	place of birth	place of death	sex
			Schlinig (Tirol)	Sankt Niklausen (Schweiz)	m
					m
					m
					m
					m
					m
	1765-02-00			Erlau (Ungarn)	m

Simple cleanup of “date” column Transform using GREL expression

Custom text transform on column date of birth

Expression Language General Refine Expression Language (GREL) ▾

```
value.toDate('yyyy-MM-DD').toString('yyyy')
```

No syntax error.

Preview History Starred Help

row	value	value.toDate("yyyy-MM-DD").toS
1.		Error: Unable to parse as date
2.	1769-12-22	1769
3.		Error: Unable to parse as date
4.		Error: Unable to parse as date
5.		Error: Unable to parse as date
6.		Error: Unable to parse as date

On error keep original set to blank store error Re-transform up to times until no change

OK Cancel

Simple cleanup of “date” column

p	date of birth	date of death	
	1769	1863	Schlinig (
			Neapel (H
			wohl aus Mailände Abbiati st
		vor 1770-00-00	
	1764		Erlau (Un
	1676	1743	
	1873	2/13/1916	Wien
		1564	
		1563	
	1860		Laa an d (Niederö:
	1/25/1902	6/2/1963	Wien
	1764	1818	Aschach Donau (Oberöst
	1872		Budapes
	1841	1896	Wien
	1896	3/7/1975	Berlin
	1718	1801	Götzens
	1953		Dornbirn



Getty Vocabularies Reconciliation Service Using OpenRefine

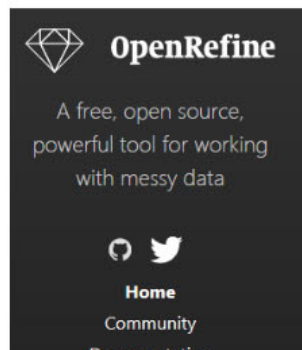
Step-by-step process from program installation to implementation

May 2019

Gregg Garcia, Getty Digital
Documentation by Lindsey Gant

Step 1. Download OpenRefine

OpenRefine is an open source tool that can be downloaded at openrefine.org.



Welcome!

OpenRefine (formerly Google Refine) is a 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 is available in English, Chinese, Spanish, French, Russian, Portuguese (Brazil), German, Japanese, Italian, Hungarian, Hebrew, Filipino, Cebuano, Tagalog

OpenRefine is supported by:

Reconciliation: ULAN artist names

Show as: **rows** records Show: 5 10 25 **50** rows

▼ All	▼ artistName	▼ artistNonPrefName	▼ artistBirth	▼ artistDeath	▼ artistNationality	▼ artistBirthPlace
☆	1. Facet	אהרונסון, יעקב	1924	2100	Israeli	
☆	2. Text filter	עבדי, עבד	1942	2100	Israeli	
☆	3. Edit cells	עבאדי, שי	1965	2100	Israeli	Jerusalem
☆	4. Edit column	אבאקנוביץ' מגדאלנה	1930	2017	Polish	
☆	5. Transpose	אבא, עירית	1953	2100	Israeli	
☆	6. Sort...	אבאסי, ריזה, (מיוחס ל)	1200	2100	Iranian	
☆	7. View	אבאסי, ריזה (אסכולה)	1200	2100	Iranian	
☆	8. Reconcile	אבאסי, ריזה	1565	1635	Iranian	
☆	9. Start reconciling...		1991		American	
☆	10. Emil Aboud, Jumana			2100	Israeli	Shefaram
☆	11. Abdullah Freres		1899		Turkish	
☆	12. Abecassis, Raphael			2100	Moroccan	Marrakesh
☆	13. Abel, Myer		1946		American	
☆	14. Abelardo, Morrell			2100	Cuban	
☆	15. Abeles, Gabi	אבלס, גבי	1200	2100		

Reconciliation: ULAN artist names

Reconcile column "artistName"

Services

Wikidata (en)

Pick a Service or Extension on Left

core-recon/add-std-srv


Enter the service's URL:


Add Service Cancel



The image shows a software interface for reconciling data. At the top, a header reads "Reconcile column 'artistName'". Below this is a "Services" panel with a list containing "Wikidata (en)". The main area of the interface is dimmed and contains the text "Pick a Service or Extension on Left". A modal dialog box is open in the center, titled "core-recon/add-std-srv". It prompts the user to "Enter the service's URL:" and has a text input field containing "http://services.getty.edu/vocab/reconcile/". At the bottom of the dialog are two buttons: "Add Service" and "Cancel".

Reconciliation: ULAN artist names

Reconcile column "artistName"

Services  » [Access Service API](#)

Wikidata (en)  one of these types:

Getty Vocabularies Reconciliation Service  

Also use relevant details from other columns:

Column	Include?	As Property
artistNonPrefName	<input type="checkbox"/>	<input type="text"/>
artistBirth	<input type="checkbox"/>	<input type="text"/>
artistDeath	<input type="checkbox"/>	<input type="text"/>
artistNationality	<input type="checkbox"/>	<input type="text"/>
artistBirthPlace 1	<input type="checkbox"/>	<input type="text"/>
artistBirthPlace 2	<input type="checkbox"/>	<input type="text"/>
artistBirthPlace 3	<input type="checkbox"/>	<input type="text"/>
artistBirthPlace 4	<input type="checkbox"/>	<input type="text"/>
artistBirthPlaceNation	<input type="checkbox"/>	<input type="text"/>
artistRole	<input type="checkbox"/>	<input type="text"/>
artistNonPrefRole	<input type="checkbox"/>	<input type="text"/>
artistNonPrefRole2	<input type="checkbox"/>	<input type="text"/>

Reconciliation: ULAN artist names

Reconcile column "artistName"

» Access [Service API](#)

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

- ULAN search
/ulan
- TGN search
/tgn
- AAT search
/aat

Also use relevant details from other columns:

Column	Include?	As Property
artistNonPrefName	<input type="checkbox"/>	
artistBirth	<input type="checkbox"/>	birthDate
artistDeath	<input type="checkbox"/>	deathDate
artistNationality	<input type="checkbox"/>	nationality
artistBirthPlace 1	<input type="checkbox"/>	
artistBirthPlace 2	<input type="checkbox"/>	
artistBirthPlace 3	<input type="checkbox"/>	
artistBirthPlace 4	<input type="checkbox"/>	
artistBirthPlaceNation	<input type="checkbox"/>	
artistRole	<input type="checkbox"/>	agentType
artistNonPrefRole	<input type="checkbox"/>	
artistNonPrefRole2	<input type="checkbox"/>	

Reconcile against type:

Reconcile against no particular type

Auto-match candidates with high confidence

Reconciliation: ULAN artist names

11 matching rows (19 total)					
Show as: rows records Show: 5 10 25 50 rows					
All	artistName	artistNonPrefName	artistBirth	artistDeath	artistNationality
☆	3. Abadi, Shay <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Abadi, Shay (46) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Alkalay, Shay (22) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Abadi, Abed (20) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Peri, Shay Frisch (20) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Cleary, Shay (17) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Docking, Shay (17) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Shay, Art (16) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Abadi, Fritzie (15) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Shay, James (15) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Shay, Patricia (15) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item	עבאדי, שי	1965	2100	Israeli
☆	4. Abakanowicz, Magdalena <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Abakanowicz, Magdalena (43) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Hambrouch, Magdalena (20) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Wiecek, Magdalena (18) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Jetelová, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Pažourková, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Franciskovic, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Correa, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Pozas, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Rădulescu, Magdalena (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Jitrik, Magdalena (14)	אבאקנוביץ' מגדאלנה	1930	2017	Polish

Reconciliation: ULAN artist names

Add column based on column artistName

New column name

core-views/addasdasd set to blank store error copy value from original column

Expression Language ▾

No syntax error.

Preview [History](#) [Starred](#) [Help](#)

row	value	cell.recon.best.id
3.	Abadi, Shay	ulan/500486311
4.	Abakanowicz, Magdalena	ulan/500084577
6.	Abbasi, Riza (attributed to)	ulan/500309512
7.	Abbasi, Riza (school of)	ulan/500309512

Reconciliation: ULAN artist names

11 matching rows (19 total)							
Show as: rows records		Show: 5 10 25 50 rows					
<input type="checkbox"/> All	<input type="checkbox"/> artistName	<input type="checkbox"/> ULAN ID	<input type="checkbox"/> artistNonPrefNa	<input type="checkbox"/> artistBirth	<input type="checkbox"/> artistDeath	<input type="checkbox"/> artistNationality	
<input type="checkbox"/>	3. Abadi, Shay <small>Choose new match</small>	ulan/500486311	עבאדי, שי	1965	2100	Israeli	Jerusalem
<input type="checkbox"/>	4. Abakanowicz, Magdalena <small>Choose new match</small>	ulan/500084577	אבאקנוביץ' מגדאלנה	1930	2017	Polish	
<input type="checkbox"/>	6. Riza 'Abbasi Museum <small>Choose new match</small>	ulan/500309512	אבאסי, ריזה, (מיזחס ל)	1200	2100	Iranian	
<input type="checkbox"/>	7. Abbasi, Riza (school of) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Riza 'Abbasi Museum (37) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Riza (31) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Imam Riza Shrine Museum (23) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Aqa Riza (23) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Shaykh 'Abbasi (22) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> University of Wisconsin, School of Nursing (15) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Museum of Contemporary Art (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Tahmāsp I, Shah of Iran (14) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Dublin Institute of Technology School	ulan/500309512	אבאסי, ריזה (אסכולה)	1200	2100	Iranian	

Thank you.

Jonathan Ward
Editor, Getty Vocabulary Program

Getty