

Introduction to OpenRefine

The Powerful Tool for Cleaning Messy Data



Sara Gonzales, MLIS

Data Librarian

Galter Health Sciences Library & Learning Center

Northwestern University Feinberg School of Medicine

Adapted from *Library Carpentry OpenRefine*, Copyright © 2016–
2019 [Library Carpentry](#), [CC BY 4.0](#).

What is OpenRefine?

- An open source tool first developed by Google (originally called GoogleRefine)
- A data-cleaning machine that requires no coding knowledge
 - (although it helps to learn a few ‘expressions’ – more on this later!
- A secure tool that runs in your local browser window, but does not share your data online

What can you do with OpenRefine?

- **Resolve inconsistencies in data**
 - Review large chunks of data at a glance
 - Correct errors in data entry
- **Split data into more granular elements**
 - Create new columns when two values have mistakenly been entered in one
- **Reformat data**
 - Change dates to different formats
 - Change spellings and capitalizations
- **Compare/match data against controlled sources**
 - Match author names against the Library of Congress Name Authority File
 - Match against other sources such as WikiData

Make all these changes and more without changing your original data file! Export a new version from OpenRefine



Download

Download OpenRefine from:
<http://openrefine.org/>

Download the dataset for this class:
<https://raw.githubusercontent.com/LibraryCarpentry/lc-open-refine/gh-pages/data/doaj-article-sample.csv>

OR

<https://github.com/LibraryCarpentry/lc-open-refine/raw/gh-pages/data/doaj-article-sample.csv>



A free, open source,
powerful tool for working
with messy data



Home
Community
Documentation
Download
Contact Us
Blog

Download

You will find on this page a list of OpenRefine distributions and extensions available for download. Are we missing something? Want to fix a typo? You can submit changes (pull request) [from here](#).

Official Distribution

Read the [installation instructions](#)

You can also download all official releases and source from our [GITHUB RELEASES PAGE](#)

OpenRefine 3.2 beta

The beta release of OpenRefine 3.2. Please BACKUP your workspace directory before installing and report any problems that you encounter.

The beta of 3.2 was released on March 1, 2019. A change log is provided on [the release page](#).

- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

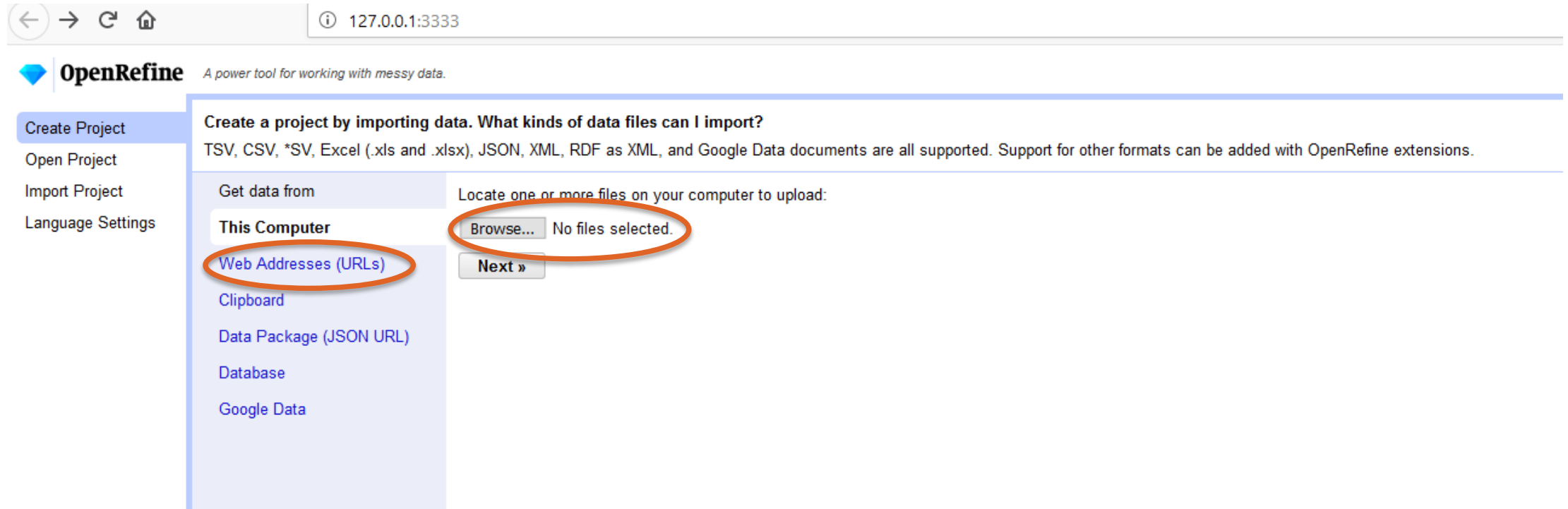
OpenRefine 3.1

The final release of OpenRefine 3.1. Please BACKUP your workspace directory before installing and report any problems that you encounter.

The final release of 3.1 was released on Nov 29, 2018. A change log is provided on [the release page](#).

- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

Open a Project



-Open CSV, Excel, JSON, and XML files stored on your computer

OR

-Fetch data from a Web Address

Rows, Records and Splitting cells

Split multi-valued cells

How to split multi-valued cells

☒ by separator

Separator

☐ regular expression

☐ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

OK Cancel

4009 rows

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


▼ All		▼ Title		▼ Authors	
☆	🗨	1.	The Fisher Thermodynamics of Quasi-Probabilities		Flavia Pennini
☆	🗨	2.			Angelo Plastino
☆	🗨	3.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective		Naveed Aslam
☆	🗨	4.			Peter C. Wynn
☆	🗨	5.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS		Rafael R. C. Cuadrat
☆	🗨	6.			Juliano C. Cury
☆	🗨	7.			Alberto M. R. Dávila
☆	🗨	8.	Synthesis and Reactivity of a Cerium(III) Scorpionate Complex Containing a Redox Non-Innocent 2,2'-Bipyridine Ligand		Fabrizio Ortu
☆	🗨	9.			Hao Zhu
☆	🗨	10.			Marie-Emmanuelle Boulon
☆	🗨	11.			David P. Mills

1001 records

Show as: rows records Show: 5 10

▼ All		▼ Title	
☆	🗨	1.	The Fisher Thermodynamics of Quasi-Probabilities
☆	🗨	2.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective
☆	🗨	3.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS

Joining Cells

 **OpenRefine** doaj article sample csv [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](#)

4009 rows

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

▼ All	▼ Title	▼ Authors	▼ DOI	▼ URL
☆	1.	The Fisher Thermodynamics of Quasi-Probabilities	0/e17127853	https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412
☆	2.			
☆	3.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective		https://doaj.org/article/572641c0bd45608812a34f9e
☆	4.			
☆	5.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS		https://doaj.org/article/5a0442382b84ba4f50007ee
☆	6.	Juliano C. Cury		
☆	7.	Alberto M. R. Dávila		
☆	8.	Synthesis and Reactivity of a Cerium(III) Scorpionate Complex Containing a Redox Non-Innocent 2,2'-Bipyridine Ligand	10.3390/inorganics3040534	https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3
☆	9.			

- Facet
 - Text filter
 - Edit cells
 - Edit column
 - Transpose
 - Sort...
 - View
 - Reconcile
- Transform...
 - Common transforms
 - Fill down
 - Blank down
 - Split multi-valued cells...
 - Join multi-valued cells...
 - Cluster and edit...
 - Replace

Edit cells → Join multi-valued cells to re-join our split cells

Facets and Filtering

OpenRefine doaj article sample.csv Permalink

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

Language change

4 choices Sort by: name count Cluster

EN 871
English 107
ES 7
FR 1
(blank) 15
Facet by choice counts

1001 rows

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

All	Title	Authors	DOI	URL	Date	Language
1.	The Fisher Thermodynamics of Quasi-Probabilities	Flavia Pennin Angelo Plastino	10.3390/e17127853	https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412	01/11/2015	English
2.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective	Naveed Aslam Peter C. Wynn	10.3390/agriculture5041172	https://doaj.org/article/0edc5af6672641c0bd45608812a34f9e	01/11/2015	English
3.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS	Rafael R. C. Cuadrat Juliano C. Cury Alberto M. R. Dávila	10.3390/ijms161226101	https://doaj.org/article/d9fe469f75a0442382b84ba4f50007ee	01/11/2015	English
4.	Synthesis and Reactivity of a Cerium(III) Scorpionate Complex Containing a Redox Non-Innocent 2,2'-Bipyridine Ligand	Fabrizio Ortù Hao Zhu Marie-Emmanuelle Boulon David P. Mills	10.3390/inorganics3040534	https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3	01/11/2015	EN
5.	Performance and Uncertainty Evaluation of	Magali Troin Richard Arenault Francis	10.3390/hydrology2040289	https://doaj.org/article/18b1470730d444573ab5c264b7f03a041	01/11/2015	EN

Facets: one of the most powerful features of OpenRefine. Faceting counts the values in every cell in a column and displays them in a box on the left side of the screen. This lets you see quickly where there might be inconsistencies in the data.

Examine values in the facets: for every listed value you have the choice to either include it or exclude it in any clustering operation you might do.

Clustering: If you have similar values in the facets, you can include them in a clustering operation, another powerful feature of OpenRefine. Clustering uses different algorithms to bring together values with similar words and spelling.

Clustering

Cluster & Edit column "Authors"

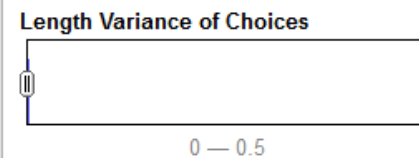
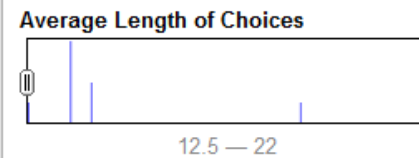
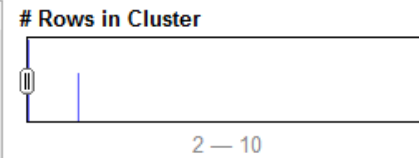
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision

Keying Function fingerprint

9 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none">A. Khan Vakeel (1 rows)Vakeel A. Khan (1 rows)	<input type="checkbox"/>	A. Khan Vakeel
2	3	<ul style="list-style-type: none">Chandra Naveen (2 rows)Naveen Chandra (1 rows)	<input type="checkbox"/>	Chandra Naveen
2	10	<ul style="list-style-type: none">B. K. Revathi (9 rows)B. K Revathi (1 rows)	<input type="checkbox"/>	B. K. Revathi
2	3	<ul style="list-style-type: none">Santiago Garcia-Granda (2 rows)Santiago García-Granda (1 rows)	<input type="checkbox"/>	Santiago Garcia-Granda
2	2	<ul style="list-style-type: none">Jian-Chao Yuan (1 rows)Jianchao Yuan (1 rows)	<input type="checkbox"/>	Jian-Chao Yuan
2	2	<ul style="list-style-type: none">Chang-Ge Zheng (1 rows)ChangGe Zheng (1 rows)	<input type="checkbox"/>	Chang-Ge Zheng
2	2	<ul style="list-style-type: none">Il'ya A. Gural'skiy (1 rows)Il'ya A. Gural'skiy (1 rows)	<input type="checkbox"/>	Il'ya A. Gural'skiy



Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

- From the top of your chosen column, choose "Edit Cells → Cluster and Edit."
- Examine the outputs of the various algorithms until you find what works best for your data
- Change the New Cell Value manually if the value suggested by OpenRefine is not correct.
- Check the check-box to merge all the values to the New Cell Value.
- Click "Merge Selected & Re-Cluster" at the bottom of the screen

Transformations

OpenRefine doaj article sample csv [Permalink](#)

Facet / Filter Undo / Redo 6 / 6

Refresh Reset All Remove All

4009 rows

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

Facet / Filter Publisher change

7 choices Sort by: name count Cluster

- Akshantala Enterprises 13
- Aurel Vlaicu University Editing House 17
- Consejo Superior de Investigaciones Cientificas 11
- International Union of Crystallography 858
- MDPI AG 93
- MDPI AG 3
- Society of Pharmaceutical Technocrats 6
- (blank) 3008

	All	Title	Authors	DOI	Publisher	URL	Date	Language	Subjects
★	1.	The Fisher Thermodynamics of Quasi-Probabilities	Flavia Pennini	10.3390/e17127853	Facet	doaj.org/article/5cca3f46cbbd63e91be5b32412	01/11/2015	English	Fisher information qua- probabilities compleme
★	2.		Angelo Plastino		Text filter				
★	3.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective	Naveed Aslam	10.3390/agriculture5041172	Edit cells	Transform...			
★	4.		Peter C. Wynn		Edit column	Common transforms			
★	5.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS	Rafael R. C. Cuadrat	10.3390/ijms161226101	Transpose	Fill down			
★	6.		Juliano C. Cury		Sort...	Blank down			
★	7.		Alberto M. R. Dávila		View	Split multi-valued cells...			
★	8.	Synthesis and Reactivity of a Cerium(III) Scorpionate Complex	Fabrizio Ortú	10.3390/inorganics3040534	Reconcile	Join multi-valued cells...			
						Cluster and edit...			
						Replace			

Trim leading and trailing whitespace

Collapse consecutive whitespace

Unescape HTML entities

To titlecase

To uppercase

To lowercase

To number

To date

To text

To null

To empty string

Transformations are another powerful aspect of OpenRefine. Some common transformations are pre-loaded into OpenRefine such as trimming leading and trailing whitespace, changing all text in a cell to titlecase, uppercase, or lowercase, or transforming values in cells to number, date, or text.

Writing Transformations

The screenshot shows the OpenRefine interface with a custom text transformation dialog open for the 'Publisher' column. The dialog is titled 'Custom text transform on column Publisher'. The 'Expression' field contains the GREL expression `value.toLowerCase()`, which is circled in red. The 'Language' dropdown is set to 'General Refine Expression Language (GREL)'. Below the expression field, there is a 'Preview' tab showing a table with two columns: 'value' and 'value.toLowerCase()'. The table lists several rows of publisher names, all of which are converted to lowercase in the transformed column. At the bottom of the dialog, there are options for 'On error' (keep original, set to blank, store error) and a checkbox for 'Re-transform up to 10 times until no change'. The 'OK' and 'Cancel' buttons are at the bottom left.

6 choices Sort by: name count Cluster

Custom text transform on column Publisher

Expression Language General Refine Expression Language (GREL)

`value.toLowerCase()` No syntax error.

Preview History Starred Help

row	value	value.toLowerCase()
537.	Akshantala Enterprises	akshantala enterprises
539.	Akshantala Enterprises	akshantala enterprises
544.	Akshantala Enterprises	akshantala enterprises
547.	Akshantala Enterprises	akshantala enterprises
550.	Akshantala Enterprises	akshantala enterprises
552.	Akshantala Enterprises	akshantala enterprises

On error ☒ keep original ☐ Re-transform up to 10 times until no change
☐ set to blank
☐ store error

OK Cancel

To make transformations that are not pre-programmed, use GREL (General Refine Expression Language) expressions. A list of GREL expression functions by type can be found here: <https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions>

Undo/Redo and Extract & Apply

The screenshot shows a data management interface with a table of 13 matching records (1001 total). The 'Undo / Redo' button is circled in red. The 'Extract Operation History' dialog box is open, showing a list of operations with checkboxes and a JSON representation of the history.

Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

- ☒ Split multi-valued cells in column Authors
- ☒ Join multi-valued cells in column Authors
- ☒ Split multi-valued cells in column Authors
- ☒ Mass edit cells in column Authors
- ☒ Reorder columns
- ☒ Reorder columns
- ☒ Text transform on cells in column Publisher using expression value.trim()

```
{
  "description": "Reorder columns",
  "columnNames": [
    "Title",
    "Authors",
    "DOI",
    "Publisher",
    "URL",
    "Date",
    "Language",
    "Subjects",
    "ISSNs",
    "Citation",
    "Licence"
  ],
  "op": "core/text-transform",
  "description": "Text transform on cells",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "columnName": "Publisher",
  "expression": "value.trim()",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10
}
```

Select All Unselect All

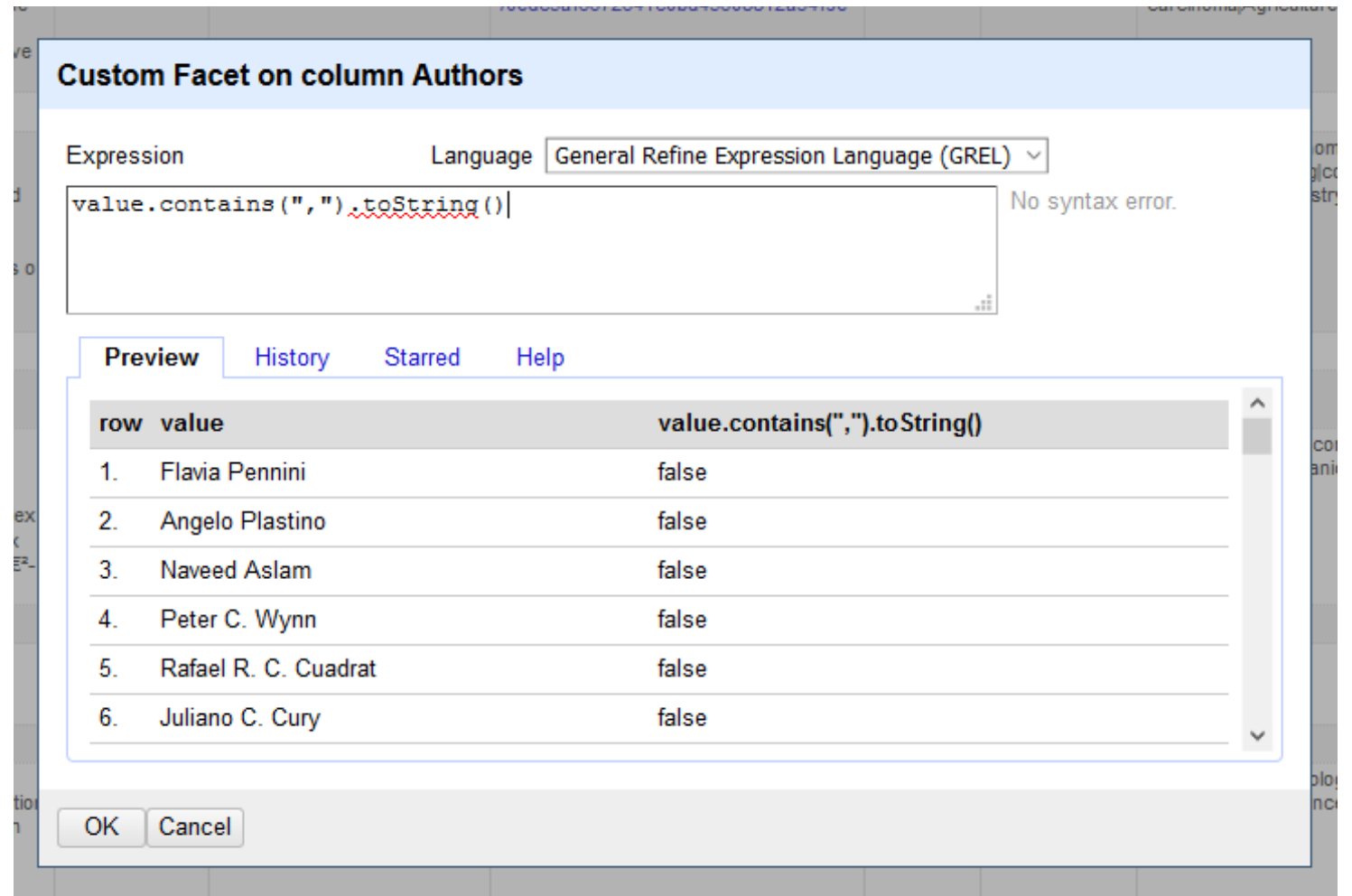
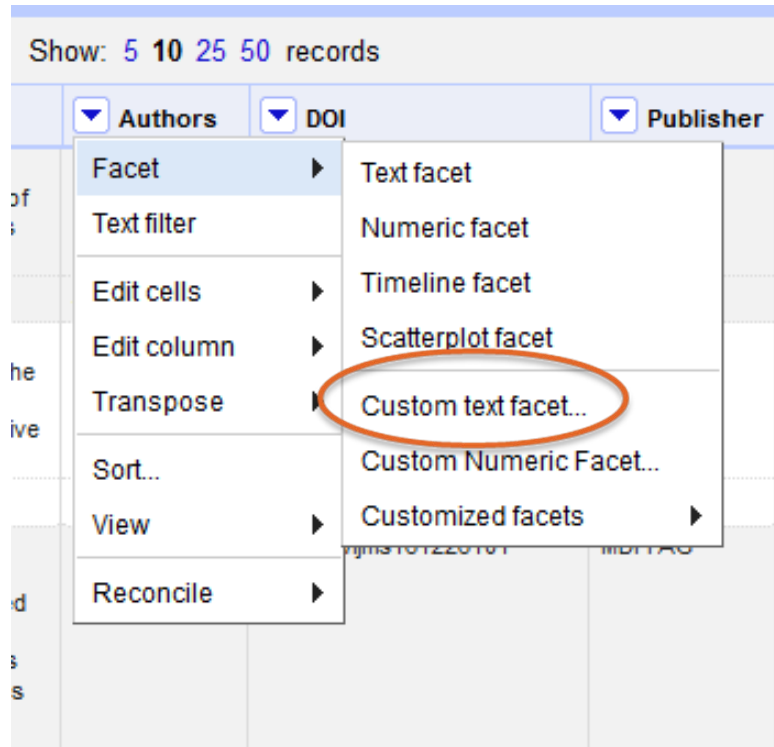
Close

- **Undo/Redo** in the upper left corner of the main screen will show every step you've taken in data transformation
- You can undo at any step, but if you undo before the last step, remember that all the changes beneath that step will also be lost
- **Extract:** Clicking extract will allow you to copy your entire change history in JSON format. You can apply these same changes again to the project, or use them in another project (by using "Apply")

2015-01-11T00:00:00Z

Arsehaute
Francois

Transforming Booleans



Transforming Arrays

Arrays: lists of values. OpenRefine represents these lists within square brackets, with the values surrounded by quotation marks and separated by commas. An example would be: ["cat", "dog", "fish", "turtle"]

Strings can be made into arrays by using a split function: `value.split()`. They can also be sorted and re-joined.

Custom text transform on column Authors

Expression Language General Refine Expression Language (GREL) ▾

`value.match(/(.*),(.*)/)`

Preview History Starred Help

row	value	value.match(/(.*),(.*)/)
493.	Martínez-García, B.	["Martínez-García", "B."]
494.	Suarez-Hernando, O.	["Suarez-Hernando", "O."]
495.	Rodríguez-Lázaro, J.	["Rodríguez-Lázaro", "J."]
496.	Pascual, A.	["Pascual", "A."]
497.	Ordiales, A.	["Ordiales", "A."]
498.	Murelaga, X.	["Murelaga", "X."]

On error

☒ keep original

☐ set to blank

☐ store error

☐ Re-transform up to times until no change

OK Cancel

Facet / Filter Undo / Redo 1 / 1

Refresh

Reset All

Remove All

Authors change invert reset

2 choices Sort by: name count

false 3965

true 44 exclude

Facet by choice counts

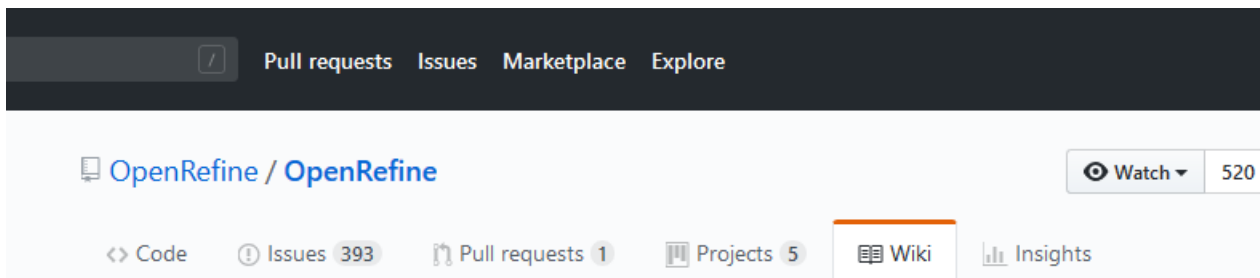
44 matching rows (4009 total)

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

All	Title	Authors
☆	493.	Las asociaciones de ostrácodos en secuencias aluviales como indicadores de cambios ambientales holocenos (Bardenas Reales de Navarra, Cuenca del Ebro, NE Península Ibérica)
☆	494.	Suarez-Hernando, O.
☆	495.	Rodríguez-Lázaro, J.
☆	496.	Pascual, A.
☆	497.	Ordiales, A.
☆	498.	Murelaga, X.
☆	499.	Sancho, C.
☆	500.	Muñoz, A.
☆	501.	Osácar, C.
☆	502.	La fracturation et les

Arrays

A Note about Arrays, and Exporting your Project



GREL Array Functions

Thad Guidry edited this page on Feb 15, 2018 · 8 revisions

Array functions supported by the [OpenRefine Expression Language \(GREL\)](#)

See also: [All GREL functions](#).

length(array a)

Returns the length of array `a`.

slice(array a, number from, optional number to)

Returns the sub-array of `a` from its index `from` up to but not including index `to`. If `to` is omitted, it is understood to be the end of the array `a`. For example, `slice([0, 1, 2, 3, 4], 1, 3)` returns `[1, 2]`, and `slice([0, 1, 2, 3, 4], 1)` returns `[1, 2, 3, 4]`.

slicing an array gives you another array.

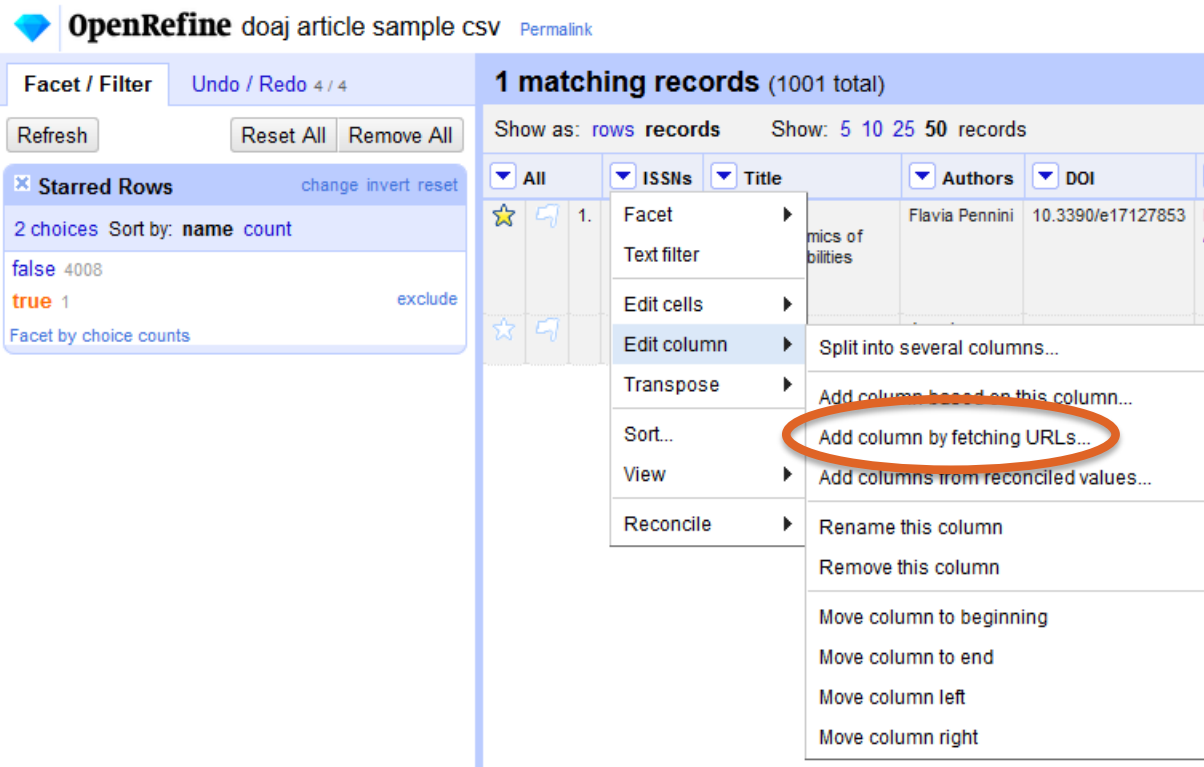
Arrays don't appear directly in OpenRefine cells. They are only seen when working in GREL Array Functions. Recipes for these functions are available from:

<https://github.com/OpenRefine/OpenRefine/wiki/>

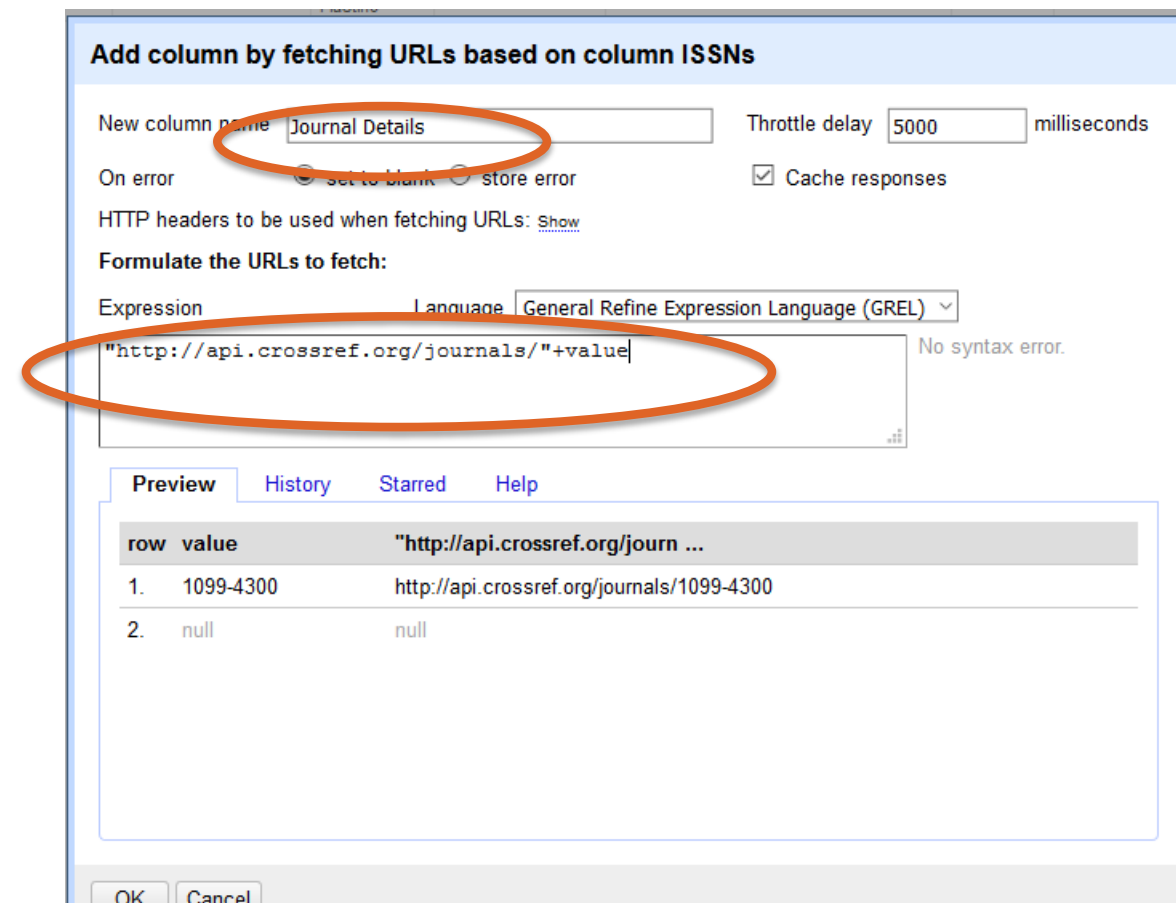
Exporting your file:

When you have finished making all data changes in your project, look for the export button in the upper right-hand corner of OpenRefine's main screen. There are multiple data options for export, including many comma-separated value options. Remember that OpenRefine will export the new, cleaned data file, which will be different from the original file you uploaded. The automatically-generated filename will be similar, so be sure to change it to reflect that this is a new, cleaned data file.

Advanced: Retrieving Data from URLs



The screenshot shows the OpenRefine interface with a dataset named 'doaj article sample csv'. The 'Facet / Filter' panel on the left shows a facet on the 'ISSNs' column, with 'true' selected. The main table shows 1001 records. A context menu is open for the 'ISSNs' column, and the option 'Add column by fetching URLs...' is highlighted with a red circle.



The screenshot shows the 'Add column by fetching URLs based on column ISSN's' dialog box. The 'New column name' is 'Journal Details'. The 'Expression' field contains the GREL expression: `"http://api.crossref.org/journals/"+value`. The 'Language' is set to 'General Refine Expression Language (GREL)'. The 'Throttle delay' is 5000 milliseconds. The 'Cache responses' checkbox is checked. The 'Formulate the URLs to fetch:' section shows a preview of the URLs generated by the expression.

row	value	"http://api.crossref.org/journ ..."
1.	1099-4300	http://api.crossref.org/journals/1099-4300
2.	null	null

Another powerful feature of OpenRefine is the ability to pull external data from URLs. Try this on the ISSN column. Star one record, then in the “All” column, facet by the star. Choose “true” in the Facet screen, then from the ISSN column drop-down choose “Edit column → Add column by fetching URLs.”

This brings you to a new Expression box. Title your new column “Journal Details.” In the expression box, type the GREL expression: `"http://api.crossref.org/journals/"+value` and click OK.

Parsing JSON in OpenRefine cells

1 matching records (1001 total)

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

▼ All ▼ ISSNs ▼ Journal Details

1.	1099-4300	<pre>{ "status": "ok", "message-type": "journal", "message-version": "1.0.0", "message": { "last-status-check-time": 1555573866743, "counts": { "total-dois": 4271, "current-dois": 2079, "backfile-dois": 2192, "breakdowns": { "dois-by-issued-year": [[2018, 998], [2017, 694], [2016, 459], [2015, 453], [2014, 328], [2013, 284], [2012, 137], [2010, 119], [2011, 101], [2009, 69], [2008, 56], [2007, 39], [2006, 18], [2005, 20], [2004, 33], [2003, 39], [2002, 10], [2001, 21], [2000, 12]] } }, "publisher": "MDPI AG", "coverage": { "affiliations-current": 0.0, "similarity-checking-current": 0.9995189905166626, "funders-backfile": 0.10492701083421707, "backfile": 0.9981752038002014, "funders-current": 0.3347763419151306, "affiliations-backfile": 0.0, "resource-links-backfile": 0.0, "orcid-backfile": 0.06021897867321968, "update-current": 0.0014430014416575432, "open-references-backfile": 1.0, "orcid-current": 0.46272245049476624, "similarity-checking-backfile": 0.9981752038002014, "references-backfile": 0.09990876168012619, "update-policies-backfile": 0.0, "licenses-current": 0.9995189905166626, "award-numbers-current": 0.32804232835769653, "abstracts-current": 0.0, "abstracts-backfile": 0.0, "astronomy": "ASJC": 3100, "time": 1555573861034, "affiliations-checking": 0.9985954761505, "time": 1555573859447, "affiliations-checking": 0.9981752038002, "time": 1555573857758, "affiliations-checking": 0.9995189905166626, "orcid-current": true, "deposit-links-current": false, "deposit-backfile": true, "deposits-award-backfile": false, "deposits-orcid": true, "issn-type": "p-ISSN" } } }</pre>

Add column based on column Journal Details

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	value.parseJson().message.titl ...
1.	<pre>{ "status": "ok", "message-type": "journal", "message-version": "1.0.0", "message": { "last-status-check-time": 1555573866743, "counts": { "total-dois": 4271, "current-dois": 2079, "backfile-dois": 2192, "breakdowns": { "dois-by-issued-year": [[2018, 998], [2017, 694], [2016, 459], [2015, 453], [2014, 328], [2013, 284], [2012, 137], [2010, 119], [2011, 101], [2009, 69], [2008, 56], [2007, 39], [2006, 18], [2005, 20], [2004, 33], [2003, 39], [2002, 10], [2001, 21], [2000, 12]] } }, "publisher": "MDPI AG", "coverage": { "affiliations-current": 0.0, "similarity-checking-current": 0.9995189905166626, "funders-backfile": 0.10492701083421707, "backfile": 0.9981752038002014, "funders-current": 0.3347763419151306, "affiliations-backfile": 0.0, "resource-links-backfile": 0.0, "orcid-backfile": 0.06021897867321968, "update-current": 0.0014430014416575432, "open-references-backfile": 1.0, "orcid-current": 0.46272245049476624, "similarity-checking-backfile": 0.9981752038002014, "references-backfile": 0.09990876168012619, "update-policies-backfile": 0.0, "licenses-current": 0.9995189905166626, "award-numbers-current": 0.32804232835769653, "abstracts-current": 0.0, "abstracts-backfile": 0.0, "astronomy": "ASJC": 3100, "time": 1555573861034, "affiliations-checking": 0.9985954761505, "time": 1555573859447, "affiliations-checking": 0.9981752038002, "time": 1555573857758, "affiliations-checking": 0.9995189905166626, "orcid-current": true, "deposit-links-current": false, "deposit-backfile": true, "deposits-award-backfile": false, "deposits-orcid": true, "issn-type": "p-ISSN" } } }</pre>	Entropy

OK Cancel

Journal Details column with long string of JSON code

GREL expression to parse the title from the rest of the JSON code.

For more on the parse JSON function, see: <https://github.com/OpenRefine/OpenRefine/wiki/GREL-Other-Functions>.

Reconciliation using URLs

Reconcile column "Publisher"

Services

- Wikidata Reconciliation for OpenRefine (en)
- Sindice
- Data.gov
- STW Thesaurus for Economics
- Real LCSH
- D2Refine
- SNOMED CT
- MeSH via SPARQL
- MeSH (Final)
- VIAF

Add Standard Service...

Reconcile column "Publisher"

» Access [Service API](#)

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

- ☒ Corporate Name
/organization/organization

Also use relevant details from other columns:

Column	Include?	As Property
ISSNs	<input type="checkbox"/>	
Journal Title	<input type="checkbox"/>	
Title	<input type="checkbox"/>	
Authors	<input type="checkbox"/>	
DOI	<input type="checkbox"/>	
URL	<input type="checkbox"/>	
Date	<input type="checkbox"/>	
Language	<input type="checkbox"/>	
Subjects	<input type="checkbox"/>	
Citation	<input type="checkbox"/>	
Licence	<input type="checkbox"/>	

☐ Reconcile against type:

☐ Reconcile against no particular type


☒ Auto-match candidates with high confidence

Maximum number of candidates to return

Add Standard Service...

Start Reconciling Cancel


Reconciliation Results

 **OpenRefine** doaj article sample csv [Permalink](#)

Facet / Filter [Undo / Redo](#) 16 / 16

[Refresh](#) [Reset All](#) [Remove All](#)

✕ Publisher: judgment [change](#)
2 choices Sort by: **name** count
(blank) 3008
none 1001
[Facet by choice counts](#)

✕ Publisher: best candidate's score [change](#) [reset](#)

0.57 — 0.97
☒ Numeric 954 ☐ Non-numeric 0 ☒ Blank 969 ☐ Error 0


1001 records
Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) records

★	🗨	1.	1099-4300	MDPI AG <input checked="" type="checkbox"/> MDPI (0.571) <input checked="" type="checkbox"/> Multidisciplinary Digital Publishing Institute (0.087) <input checked="" type="checkbox"/> Create new item	Enti
★	🗨	2.	2077-0472	MDPI AG <input checked="" type="checkbox"/> MDPI (0.571) <input checked="" type="checkbox"/> Multidisciplinary Digital Publishing Institute (0.087) <input checked="" type="checkbox"/> Create new item	
★	🗨	3.	1422-0067	MDPI AG <input checked="" type="checkbox"/> MDPI (0.571) <input checked="" type="checkbox"/> Multidisciplinary Digital Publishing Institute (0.087) <input checked="" type="checkbox"/> Create new item	

Single check-box: will accept the reconciled value only in this cell

Double check-box: will accept the reconciled value for every identical cell

Reconciliation Results, continued

 **OpenRefine** doaj article sample csv [Permalink](#)

Facet / Filter Undo / Redo 16 / 16

Refresh Reset All Remove All

858 matching records (1001 total)

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

Facet: Publisher: judgment change

2 choices Sort by: name count

(blank) 2554

none 858

Facet by choice counts

Facet: Publisher: best candidate's score change reset

0.57 — 0.97

☒ Numeric 858 ☐ Non-numeric 0 ☒ Blank 832 ☐ Error 0

Facet: Publisher change invert reset

7 choices Sort by: name count Cluster

Akshantala Enterprises 13

Aurel Vlaicu University Editing House 17

Consejo Superior de Investigaciones Científicas 11

International Union of Crystallography 858 exclude

MDPI AG 93

MDPI AG 3

Society of Pharmaceutical Technocrats 6

(blank) 963

All	ISSNs	Publisher
143.	2056-9890	International Union of Crystallography <input checked="" type="checkbox"/> International Union of Crystallography. (0.974) <input checked="" type="checkbox"/> International Union of Crystallography. Commission on Crystallographic Apparatus (0.475) <input checked="" type="checkbox"/> International union of crystallography. Commission on crystallographic computing (0.45) <input checked="" type="checkbox"/> Create new item
144.	2056-9890	International Union of Crystallography <input checked="" type="checkbox"/> International Union of Crystallography. (0.974) <input checked="" type="checkbox"/> International Union of Crystallography. Commission on Crystallographic Apparatus (0.475) <input checked="" type="checkbox"/> International union of crystallography. Commission on crystallographic computing (0.45) <input checked="" type="checkbox"/> Create new item

One new value is chosen. 858 values are matched.

Facet / Filter Undo / Redo 17 / 17

Refresh Reset All Remove All

858 matching records (1001 total)

Show as: rows records Show: 5 10 25

Facet: Publisher: judgment change

2 choices Sort by: name count

(blank) 2554

matched 858

Facet by choice counts

Facet: Publisher: best candidate's score change reset

0.57 — 0.97

☒ Numeric 858 ☐ Non-numeric 0 ☒ Blank 832 ☐ Error 0

Facet: Publisher change invert reset

7 choices Sort by: name count Cluster

Akshantala Enterprises 13

Aurel Vlaicu University Editing House 17

Consejo Superior de Investigaciones Científicas 11

International Union of Crystallography 858 exclude

MDPI AG 93

MDPI AG 3

Society of Pharmaceutical Technocrats 6

(blank) 963

All	ISSNs	Publisher	Journal
143.	2056-9890	International Union of Crystallography. Choose new match	
144.	2056-9890	International Union of Crystallography. Choose new match	
145.	2056-9890	International Union of Crystallography. Choose new match	
146.	2056-9890	International Union of Crystallography.	

Reconciliation, continued

To try reconciliation a different way, remove all filters and facets from the project so all the rows display again. From the Publisher column drop-down menu choose Reconcile → Actions → Match each cell to its best candidate

Cell values will be automatically matched, as you can see in the hyperlinked examples now populating the Publisher field.

1001 records

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

▼ All	▼ ISSN	▼ Publisher	▼ Journal Title	▼ Title	▼ Authors	▼ DOI
★	1.	1099-4300	Entropy	The Fisher Thermodynamics of Quasi-Probabilities	Flavia Pennini	10.3390/e17127853
★	2.	2077-0472		Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective	Angelo Plastino Naveed Aslam	10.3390/agriculture5
★	3.	1422-0067	MDPI AG			12261
★	4.	2304-6740	MDPI AG	Synthesis and Reactivity of a	Fabrizio Ortu	10.3390/inorganics3

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Start reconciling...

Facets

Actions

Copy reconciliation data...

Use values as identifiers

Match each cell to its best candidate

Create a new item for each cell

Create one new item for similar cells

Match all filtered cells to...

Discard reconciliation judgments

Clear reconciliation data

1001 records

Show as: rows records

Show: 5 10 25 50 records

▼ All		▼ ISSN	▼ Publisher	▼ Journal Title	▼ Title	
★	🗨	1.	1099-4300	MDPI. Choose new match	Entropy	The Fisher Thermodynamics of Quasi-Probabilities
★	🗨					
★	🗨	2.	2077-0472	MDPI. Choose new match		Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective
★	🗨					
★	🗨	3.	1422-0067	MDPI. Choose new match		Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS

Records

5 records Show: 5 10 25 50 records

674
.....
.....
.....
.....
533

Metagenomic
Upwelling-
Brazilian C
Seawater
Sequence
Type I PKS

Some final OpenRefine tips

- When working with larger datasets, you may want to allocate more memory to OpenRefine. Methods for increasing memory vary by operating system and also according to which version of Java you have.
 - For a guide on allocating more memory, see: the [OpenRefine FAQ: Allocate More Memory](#)
- Extensions can be added to OpenRefine to add increased functionality. Installing extensions might involve installing various additional software programs on your computer. Check with [NUIT](#) or [FSMIT](#) when in doubt.
 - A list of extensions is available on the OpenRefine [downloads](#) page
- There is a wealth of additional OpenRefine educational material online! Here is a small sample:
 - The Google Refine introduction series (pre-dates OpenRefine's name change, but still useful: https://www.youtube.com/watch?v=B70J_H_zAWM)
 - A library-focused OpenRefine [blog](#) by Owen Stephens
 - [Cleaning Data with OpenRefine](#) by Seth van Hooland, Ruben Verborgh, and Max DeWilde
 - [Getting Started with OpenRefine](#) by Thomas Padilla

Credits

The OpenRefine lesson on which this class is based can be found at Library Carpentry: OpenRefine, (<https://librarycarpentry.org/lc-open-refine/>), and is copyright © 2016–2019 [Library Carpentry](#). Class materials are available for adaptation under the [CC BY 4.0](#) license.

Cover slide image by kues1, accessed at: https://www.freepik.com/free-photo/old-files_1012333.htm

Thank you!