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

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

Northwestern Medicine
Feinberg School of Medicine

OpenRefine

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.

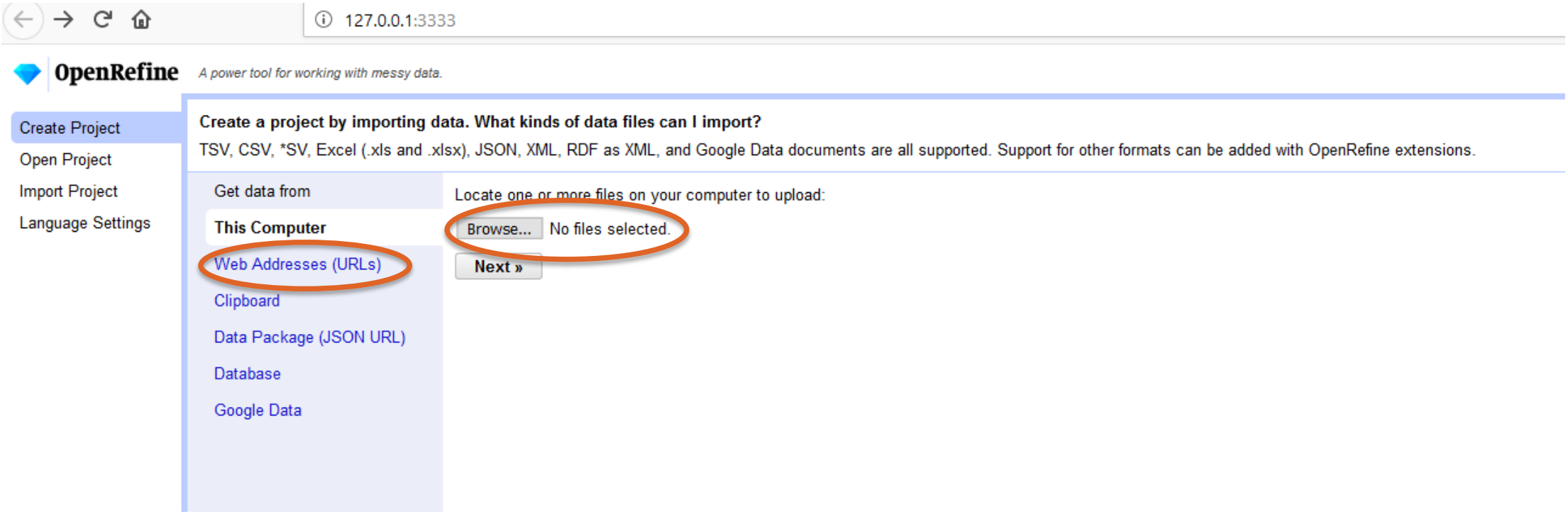
OpenRefine 3.1

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

OpenRefine *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- Language Settings

« Start Over | Configure Parsing Options | Project name: doaj article sample csv | Tags | Create Project »

	Title	Authors	DOI	URL	Date	Language	Subjects	ISSNs	Publisher	Citation
1.	The Fisher Thermodynamics of Quasi-Probabilities	Flavia Penninl|Angelo Plastino	10.3390/e17127853	https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412	01/11/2015	English	Fisher information|quasi-probabilities|complementarity|Physics|QC1-999|Science|Q	1099-4300	MDPI AG	Entropy, Vol 1 Iss 12, Pp 7848-7858 (2015)
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	aflatoxins|AFM1|AFB1|milk marketing chains|hepatocellular carcinoma|Agriculture (General)|S1-972|Agriculture|S	2077-0472	MDPI AG	Agriculture (Basel), Vol 5 Iss 4, Pp 1172-1182 (2015)
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	PKS|NRPS|metagenomics|environmental genomics|upwelling|coastal environment|Chemistry|QD1-999|Science|Q	1422-0067	MDPI AG	International Journal of Molecular Sciences, Vo 16, Iss 12, P 28285-28299 (2015)
4.	Synthesis and Reactivity of a Cerium(III) Scorpionate Complex Containing a Redox Non-Innocent 2,2'-Bipyridine Ligand	Fabrizio Ortu|Hao Zhu|Marie-Emmanuelle Boulon|David P. Mills	10.3390/inorganics3040534	https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3	01/11/2015	EN	lanthanide|cerium|scorpionate|tris(pyrazolyl)borate|radical|redox non-innocent|Inorganic chemistry|QD146-197	2304-6740	MDPI AG	Inorganics (Basel), Vol 3 Iss 4, Pp 534-553 (201
5.	Performance and Uncertainty Evaluation of Snow Models on Snowmelt Flow Simulations over a Nordic Catchment (Mistassibi, Canada)	Magali Troin|Richard Arsenault|François Brissette	10.3390/hydrology2040289	https://doaj.org/article/18b1d70730d44573ab5c264b7f93e041	01/11/2015	EN	snow models|hydrological models|snowmelt|uncertainty|Canada|Science|Q	2306-5338	MDPI AG	Hydrology, Vo 2, Iss 4, Pp 289-317 (201
6.	Dihydrochalcone Compounds Isolated from Crabapple Leaves Showed Anticancer Effects on Human Cancer Cell Lines	Xiaoxiao Qin|Yun Feng Xing|Zhiqin Zhou|Yuncong Yao	10.3390/molecules201219754	https://doaj.org/article/5765b418183c4b70bb0b774f4627beb5	01/11/2015	English	Malus crabapples|leaves|dihydrochalcone compounds|anticancer activity|human cancer cell lines|Organic chemistry|QD241-441|Chemistry|QD1-999|Science|Q	1420-3049	MDPI AG	Molecules, V 20, Iss 12, P 21193-21203 (2015)
7.	Ionic Liquids as Carbene Catalyst Precursors in the One-Pot Four-Component Assembly of Oxo Triphenylhexanoates (OTHOs)	Anton Axelsson|Linda Ta|Henrik Sundén	10.3390/catal5042052	https://doaj.org/article/d1d39464834447c8bd9c269710e1e17d	01/11/2015	English	ionic liquid|NHC|OTHO|multicomponent reaction|carbene|organocatalysis|organogel|MCR|1,6-ketoester|Physical and theoretical chemistry|QD450-801|Chemistry|QD1-999|Science|Q	2073-4344	MDPI AG	Catalysts, Vo Iss 4, Pp 2052-2067 (2015)
8.	Characterization of	Weihong Min|Huiying	10.3390/ijms161226098	https://doaj.org/article	01/11/2015	English	Corynebacterium pekinense|aspartate kinase|characterization|molecular	1422-0067	MDPI AG	International

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

Character encoding: UTF-8

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

Escape special characters with \

[ ] Column names (comma separated):

Update Preview

[ ] Ignore first 0 line(s) at beginning of file
[x] Parse next 1 line(s) as column headers
[ ] Discard initial 0 row(s) of data
[ ] Load at most 0 row(s) of data
[x] Use character " to enclose cells containing column separators

[ ] Parse cell text into numbers, dates, ...

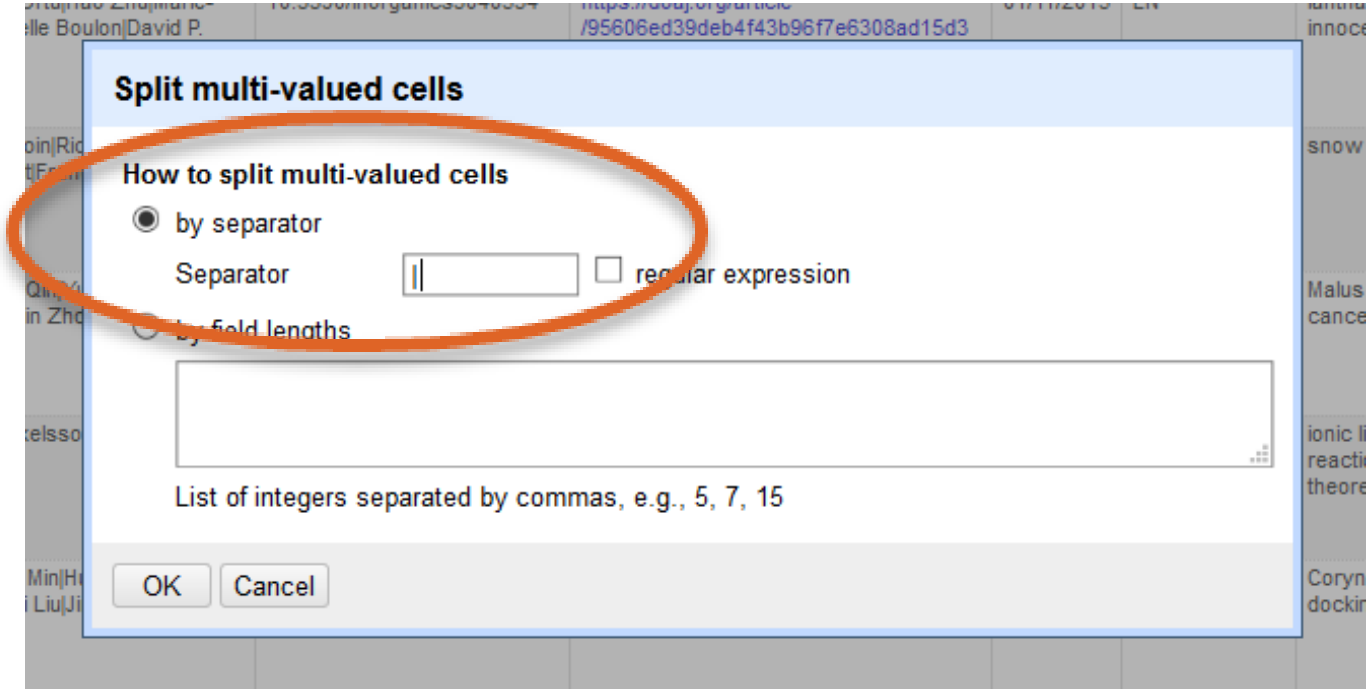
[x] Store blank rows
[x] Store blank cells as nulls
[ ] Store file source (file names, URLs) in each row

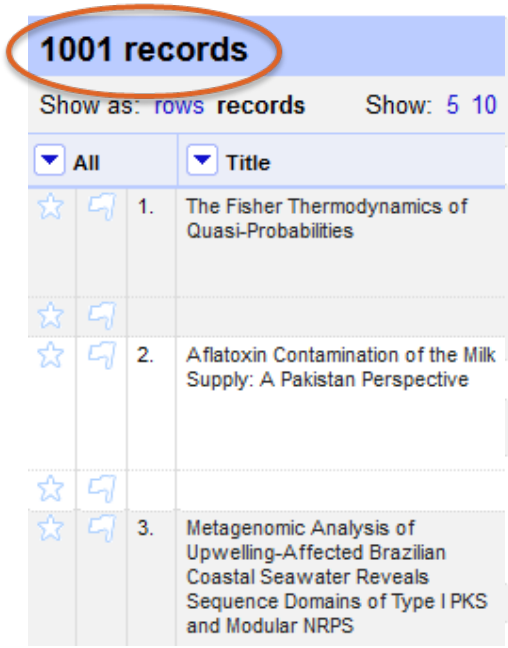
Unccheck this box →

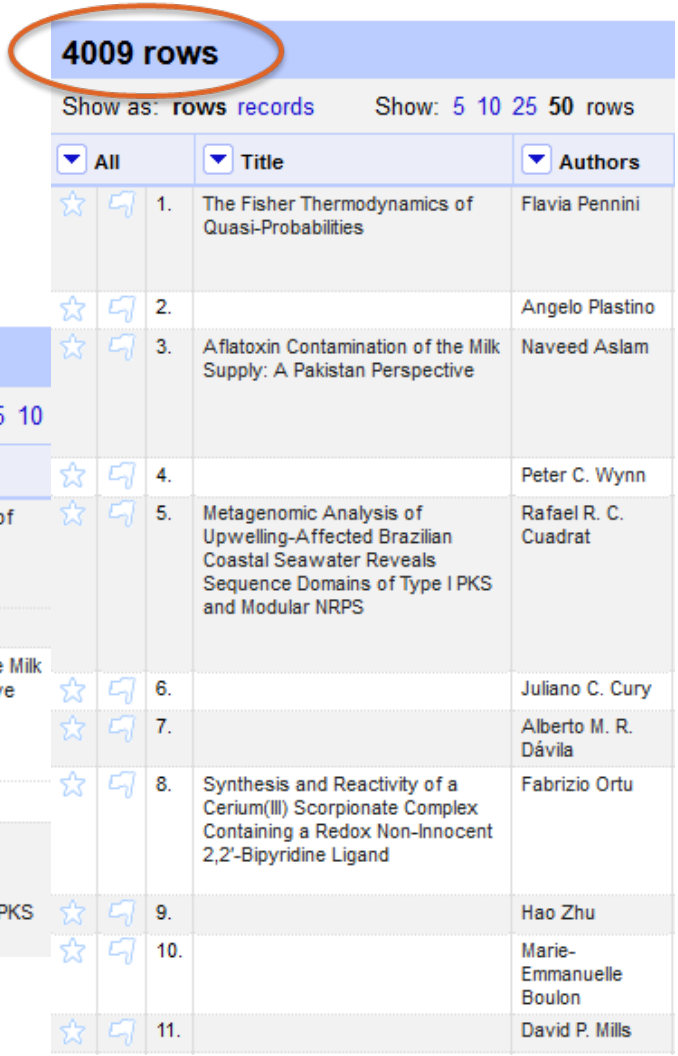
Version 3.1 [b90e413]

Preferences
Help

Rows, Records and Splitting cells







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

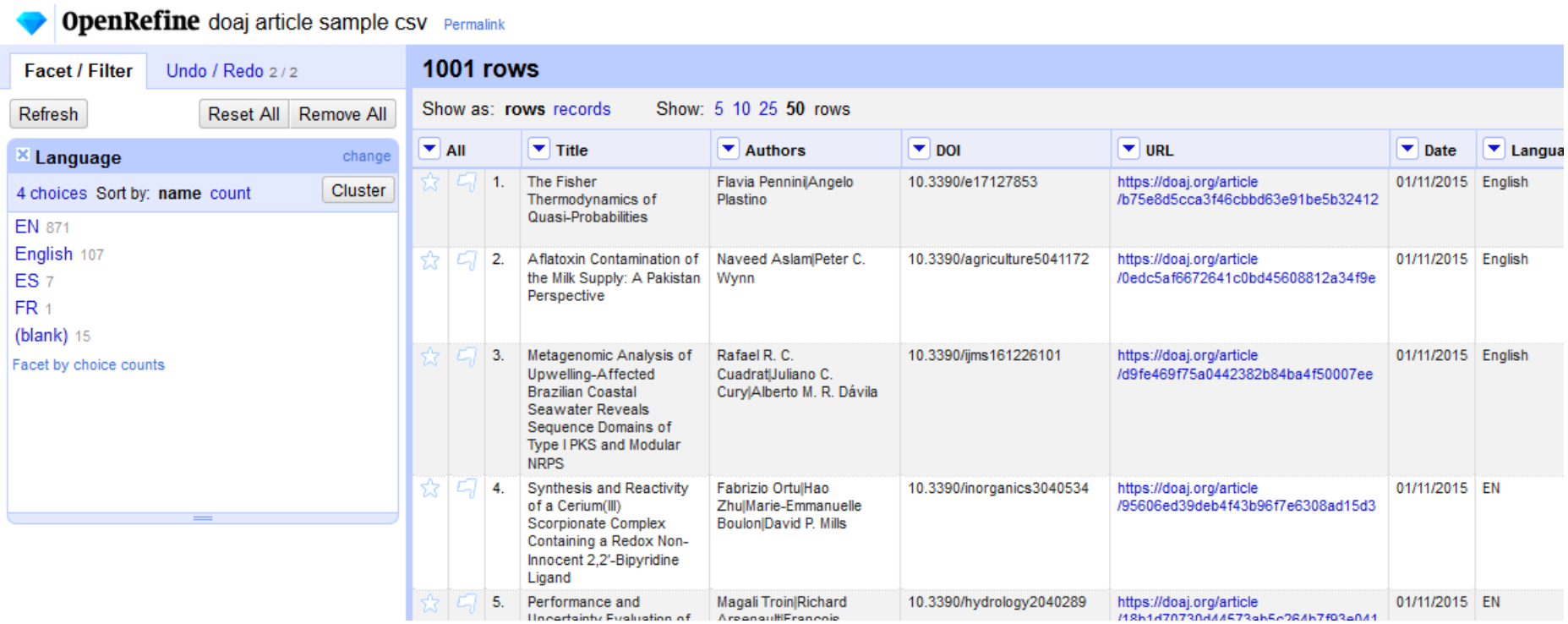
All	Title	Authors	DOI	URL
1.	The Fisher Thermodynamics of Quasi-Probabilities		0/e17127853	https://doaj.org/article/b75e8d5cca3f46cbbd63e91be5b32412
2.				
3.	Aflatoxin Contamination of the Milk Supply: A Pakistan Perspective			j.org/article 572641c0bd45608812a34f9e
4.				
5.	Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS			j.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	Fabrizio Ortu	10.3390/inorganics3040534	https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3

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

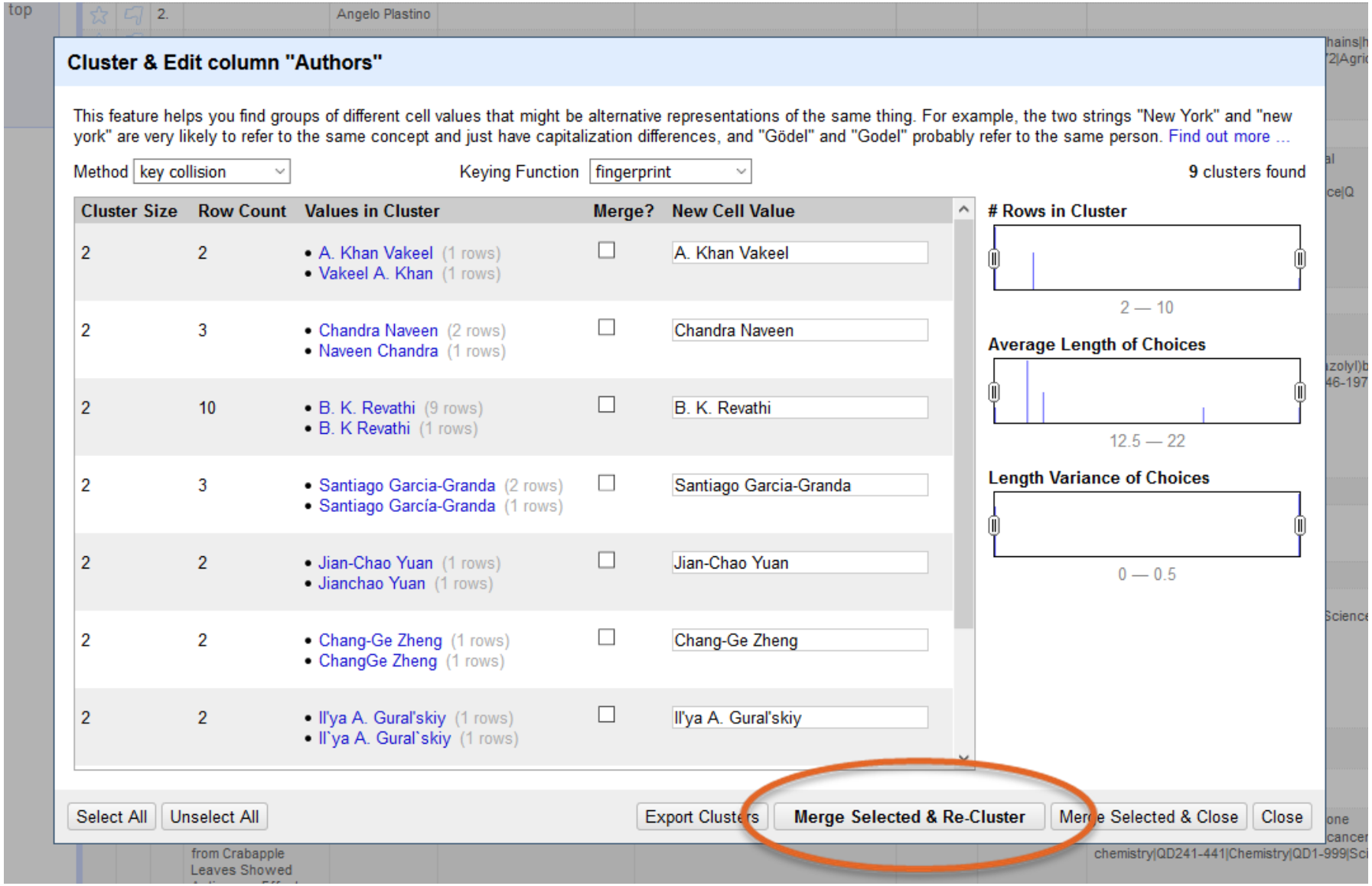
Edit cells → Join multi-valued cells to re-join our split cells

Facets and Filtering



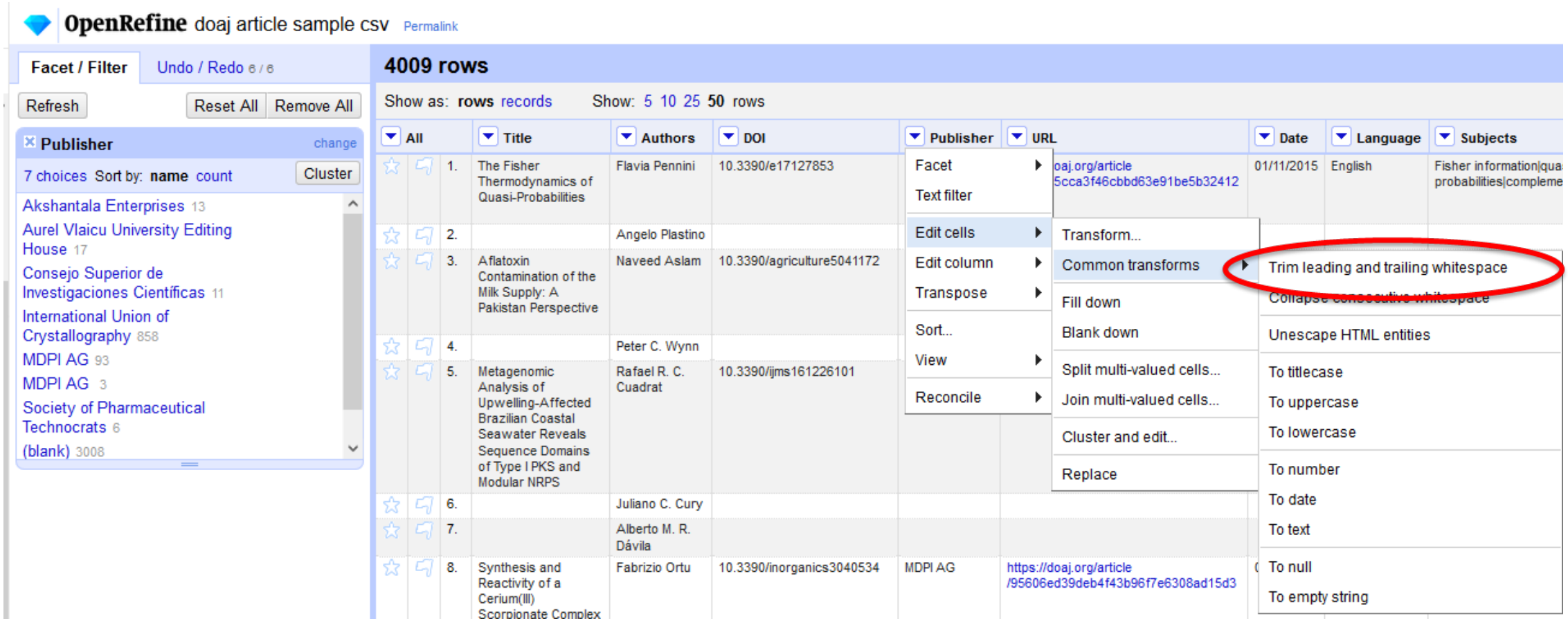
Facets: one of the most powerful features of OpenRefine. Faceting counts the values in every cell in a column and displays them in a box on the left side of the screen. This lets you see quickly where there might be inconsistencies in the data.
Examine values in the facets: for every listed value you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



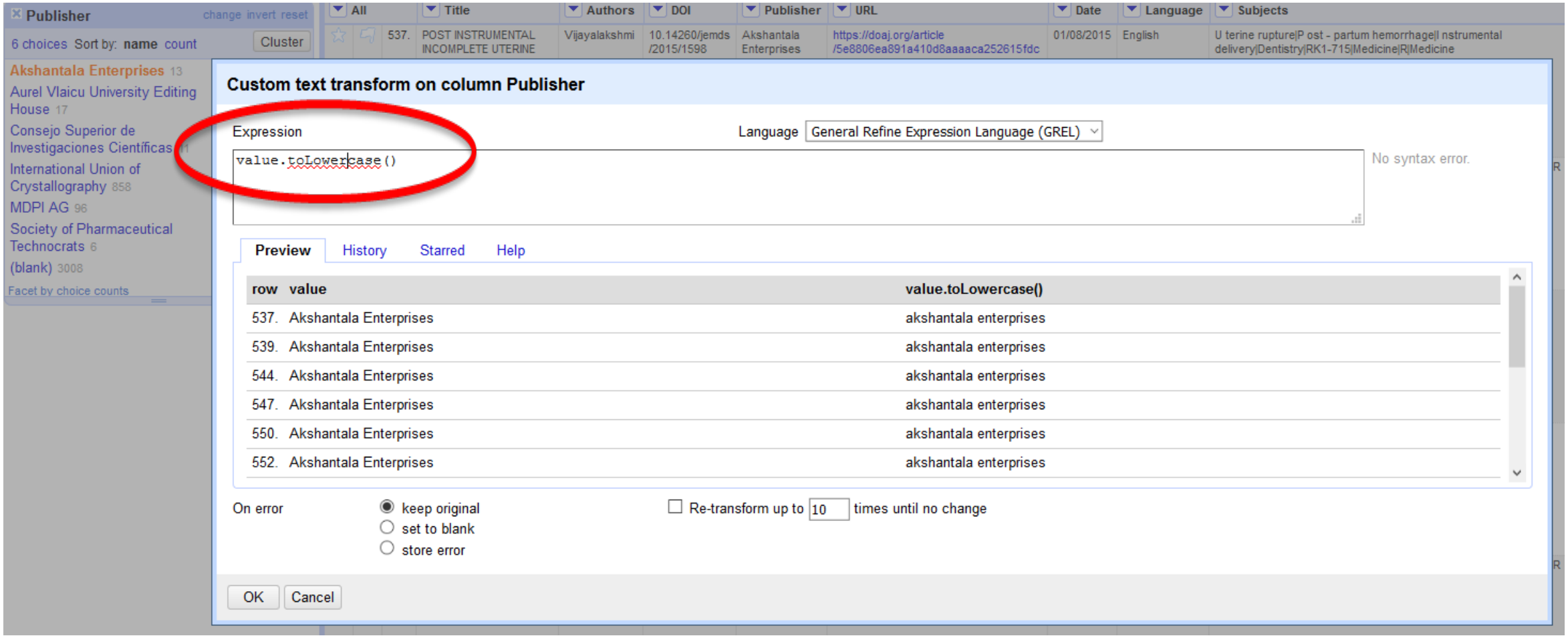
- From the top of your chosen column, choose "Edit Cells → Cluster and Edit."
- Examine the outputs of the various algorithms until you find what works best for your data
- Change the New Cell Value manually if the value suggested by OpenRefine is not correct.
- Check the check-box to merge all the values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

Transformations



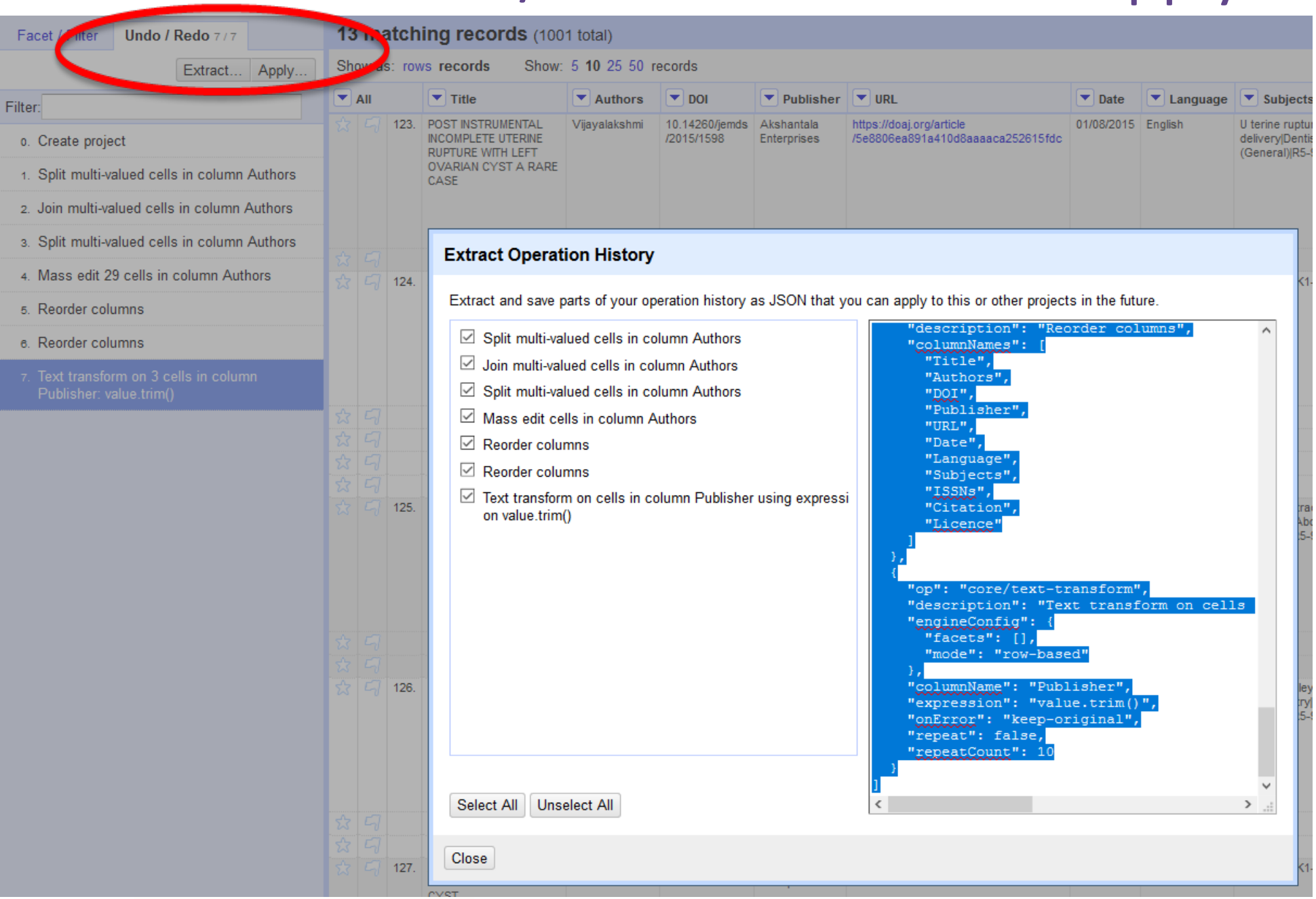
Transformations are another powerful aspect of OpenRefine. Some common transformations are pre-loaded into OpenRefine such as trimming leading and trailing whitespace, changing all text in a cell to titlecase, uppercase, or lowercase, or transforming values in cells to number, date, or text.

Writing Transformations



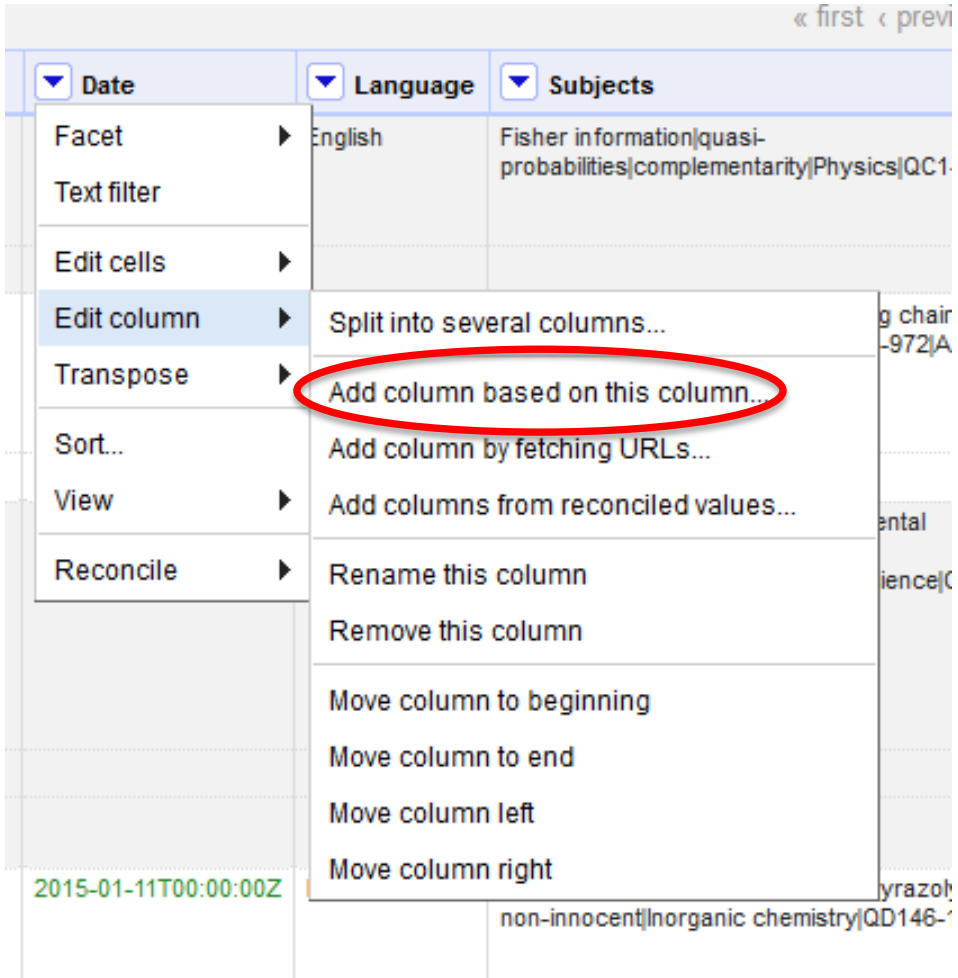
To make transformations that are not pre-programmed, use GREL (General Refine Expression Language) expressions. A list of GREL expression functions by type can be found here: https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

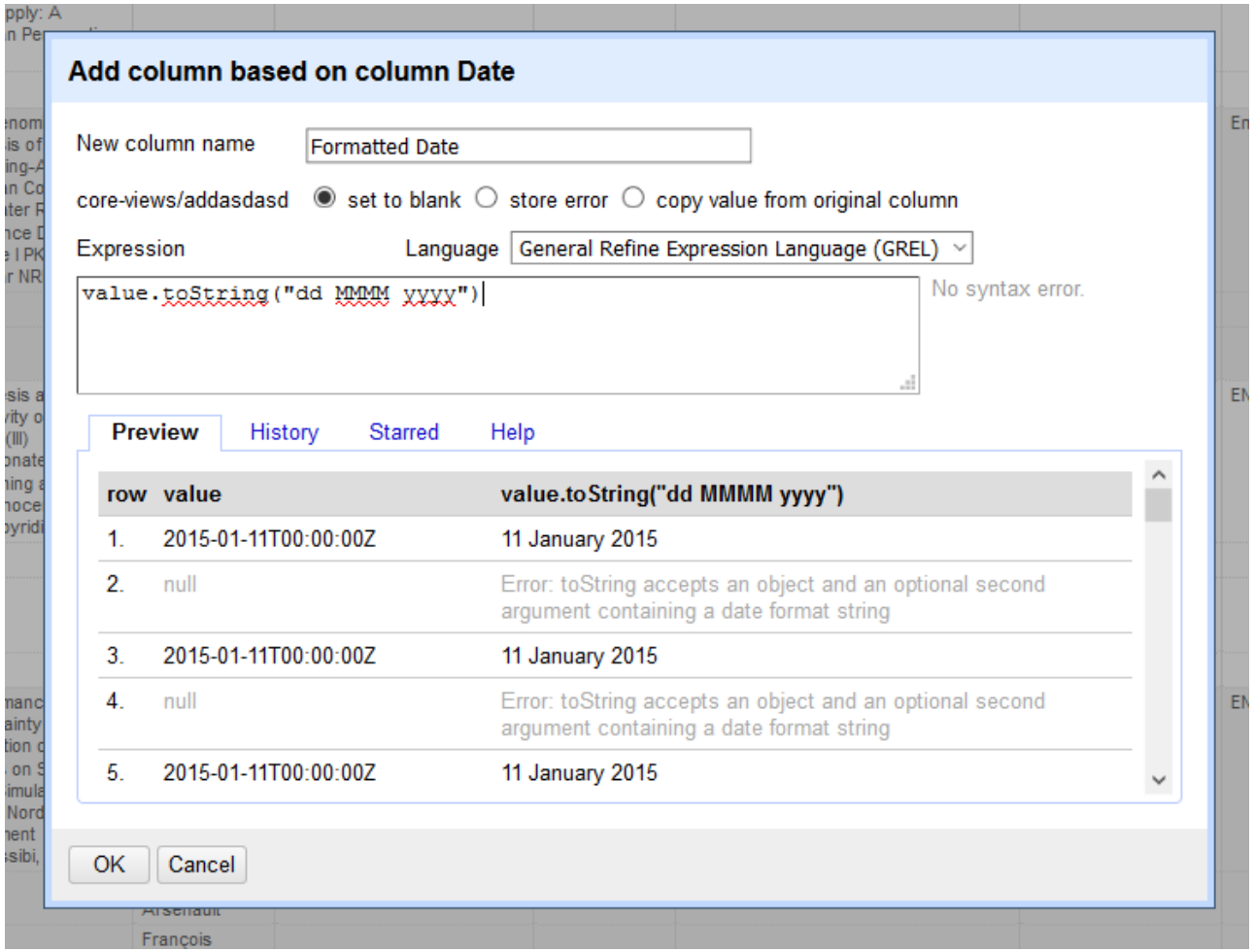
Undo/Redo and Extract & Apply



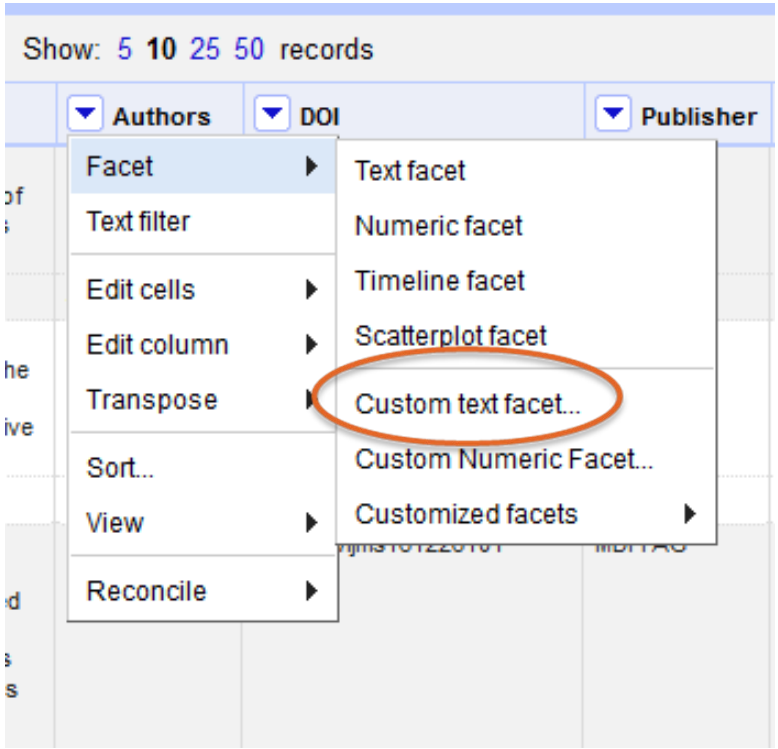
- **Undo/Redo** in the upper left corner of the main screen will show every step you've taken in data transformation
- You can undo at any step, but if you undo before the last step, remember that all the changes beneath that step will also be lost
- **Extract:** Clicking extract will allow you to copy 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)

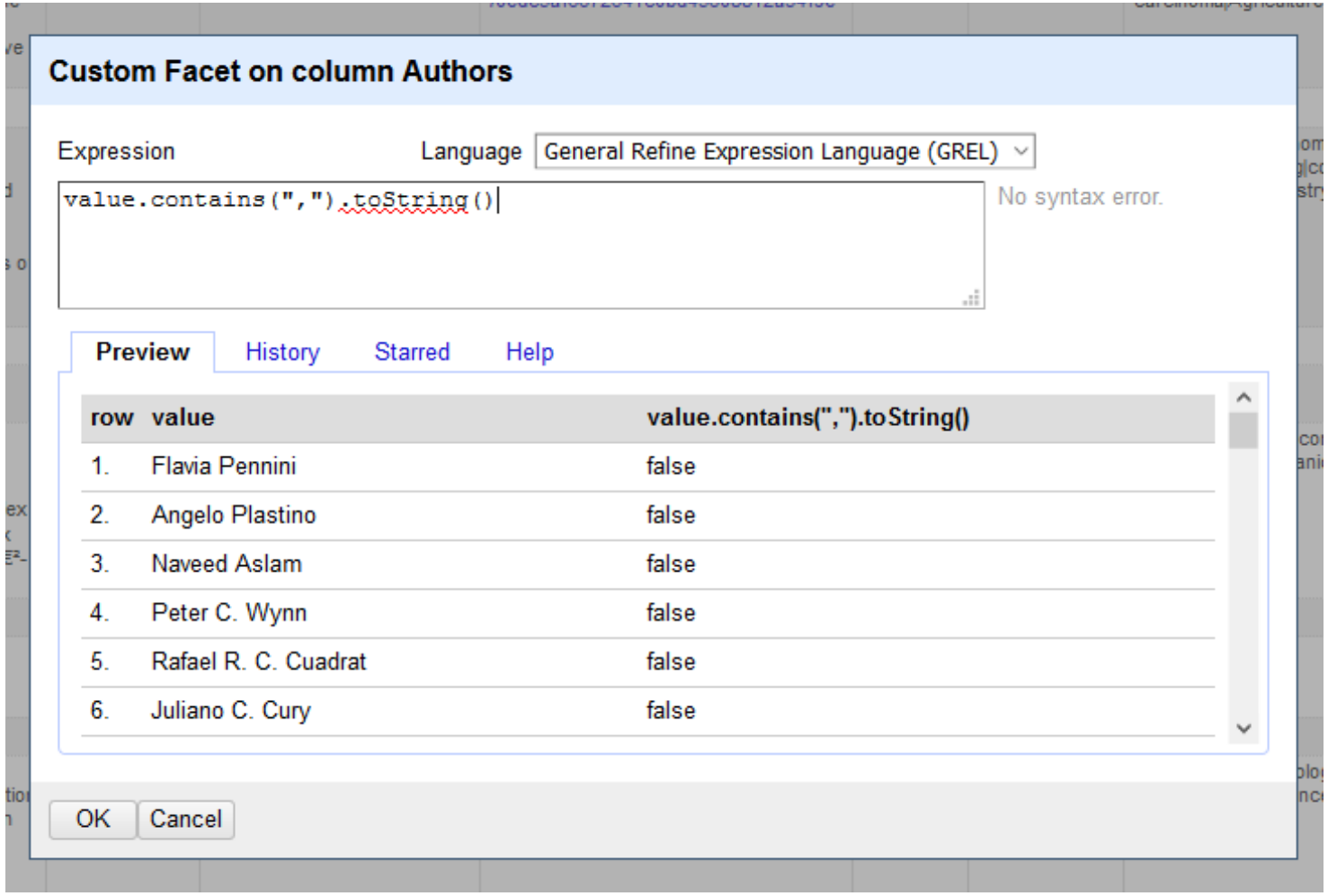
Transforming Dates and Numbers





Transforming Booleans



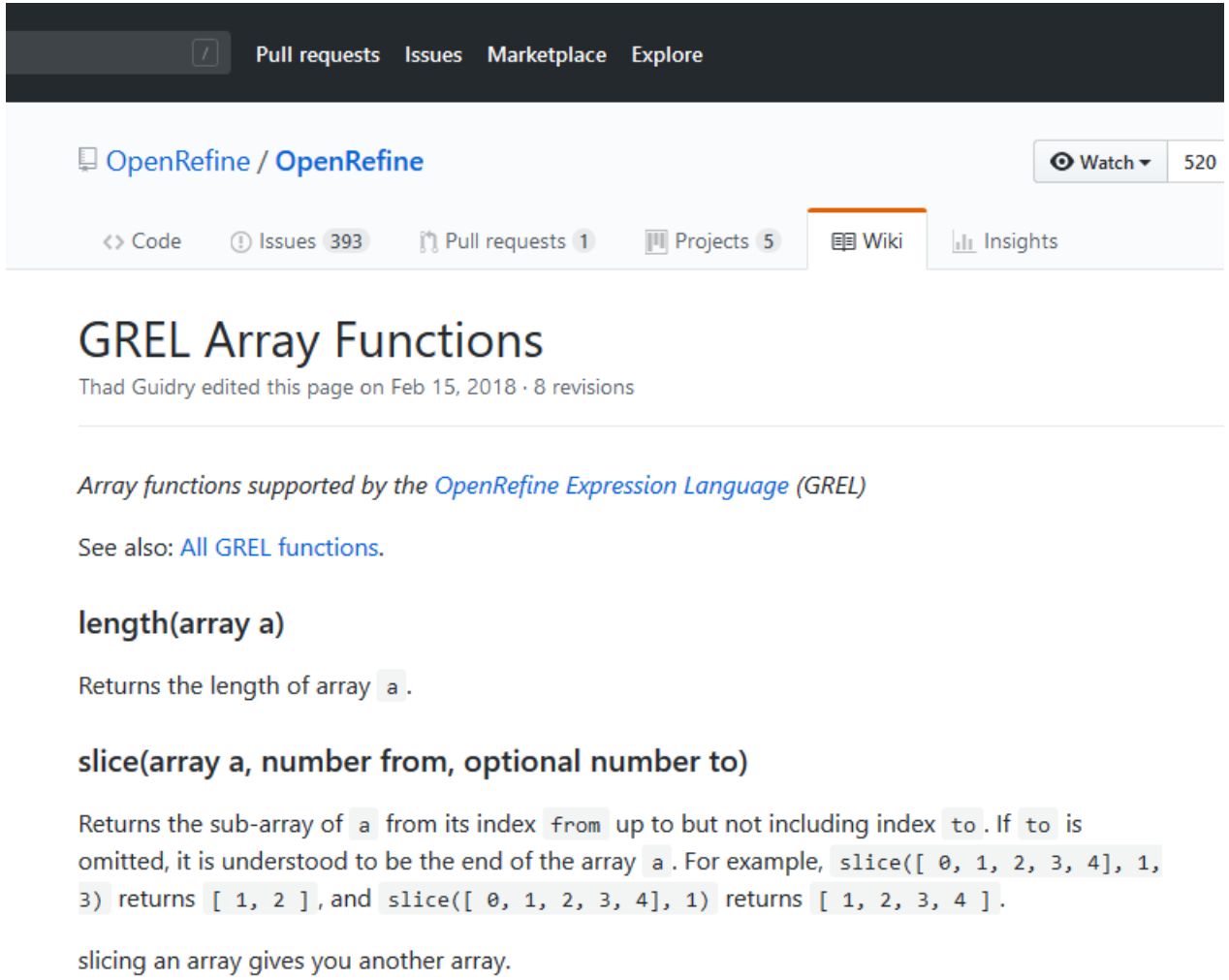


Transforming Arrays

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.

A Note about Arrays, and Exporting your Project

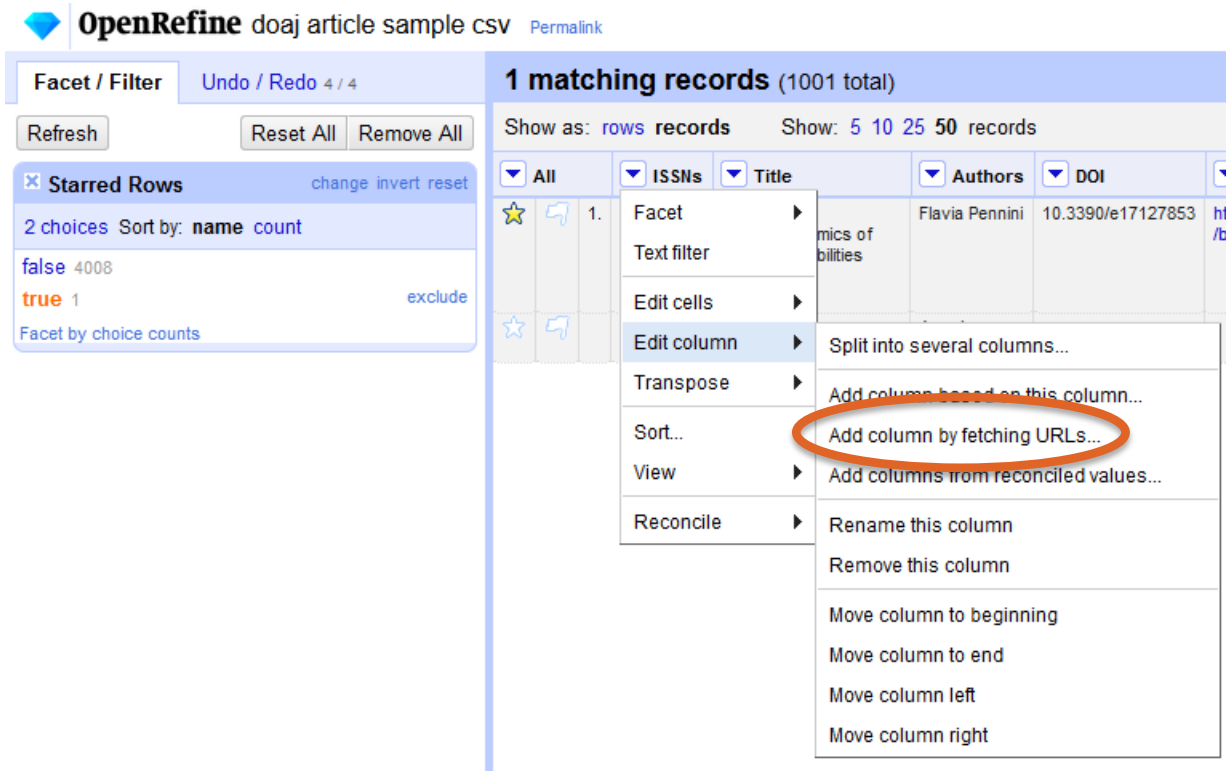


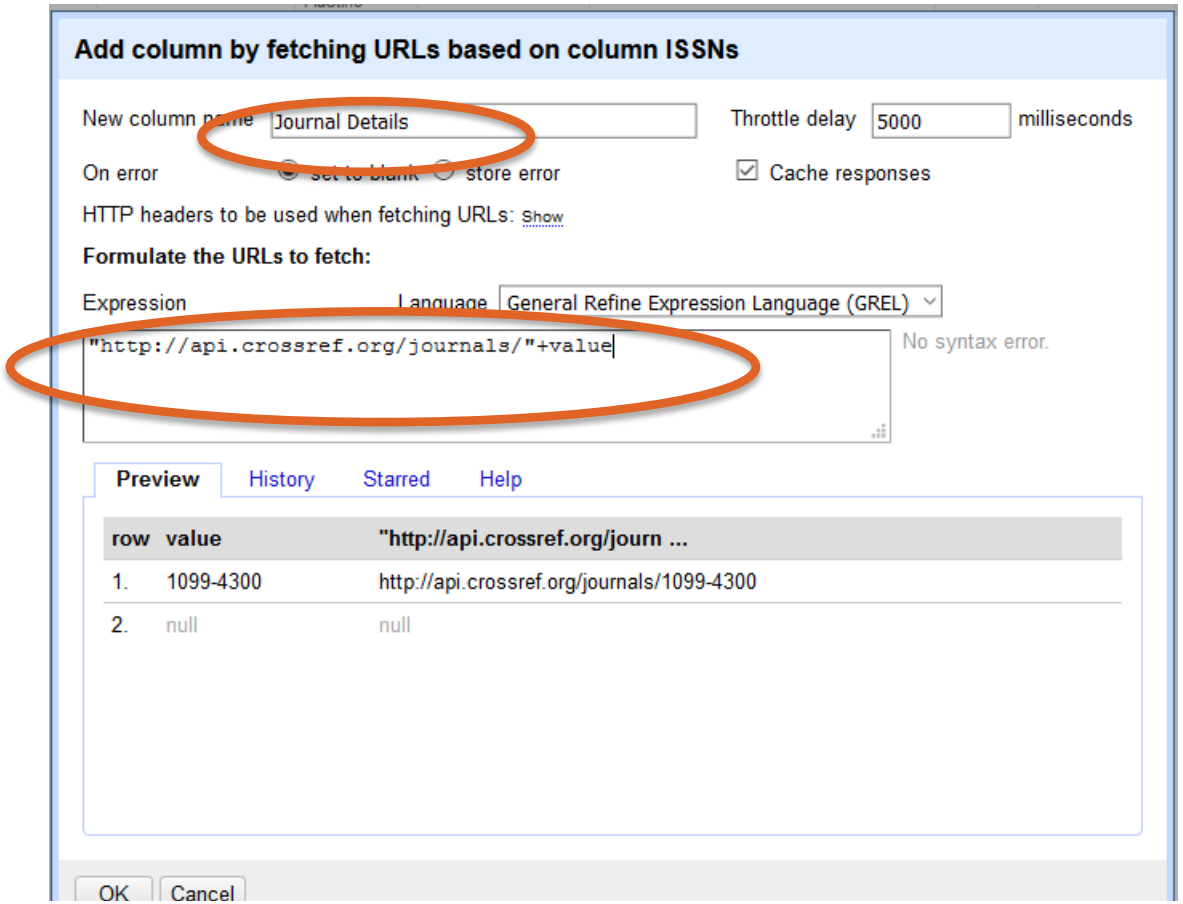
Arrays don't appear directly in OpenRefine cells. They are only seen when working in GREL Array Functions. Recipes for these functions are available from: https://github.com/OpenRefine/OpenRefine/wiki/

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

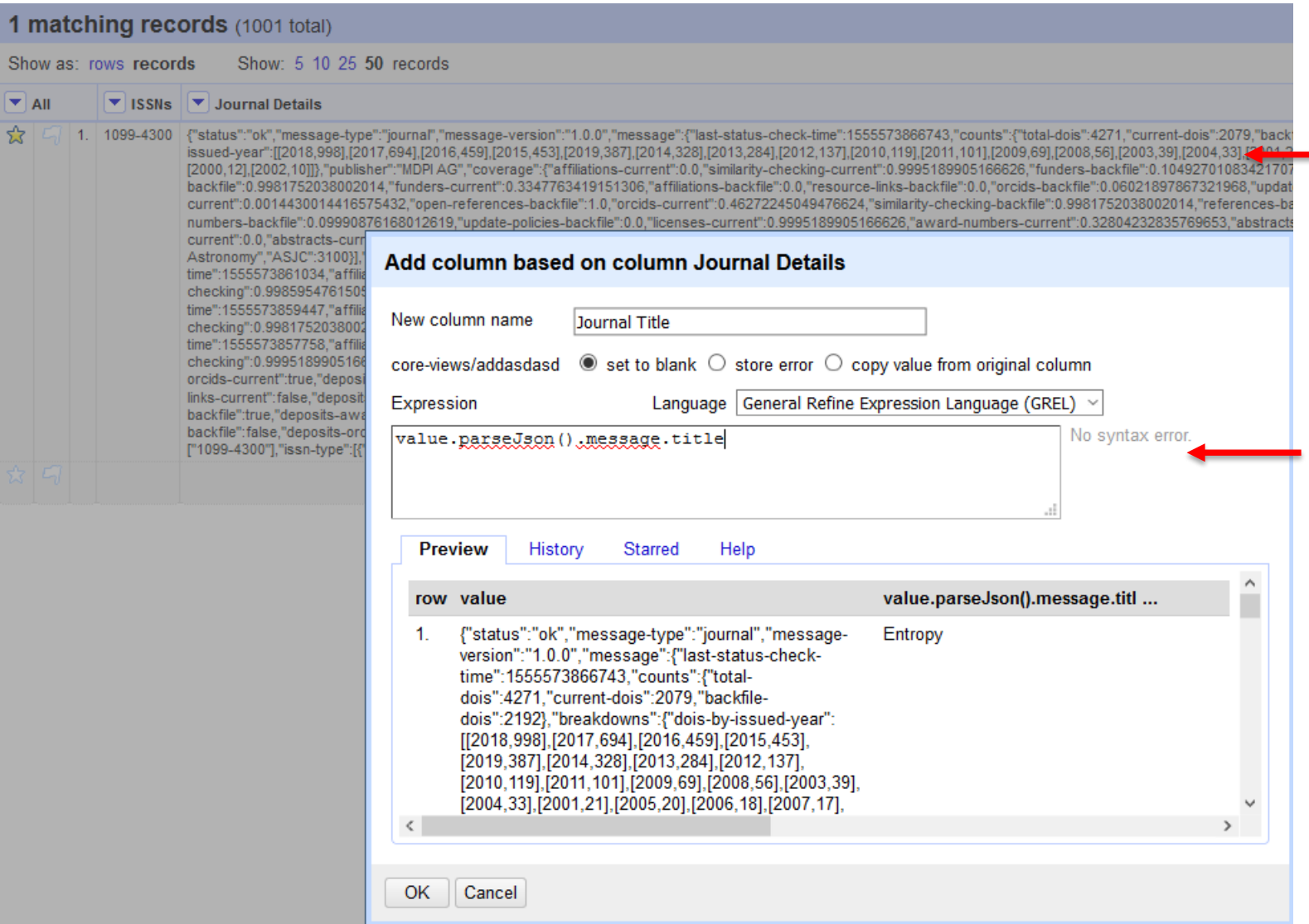




Another powerful feature of OpenRefine is the ability to pull external data from URLs. Try this on the ISSN column. Star one record, then in the “All” column, facet by the star. Choose “true” in the Facet screen, then from the ISSN column drop-down choose “Edit column → Add column by fetching URLs.”

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

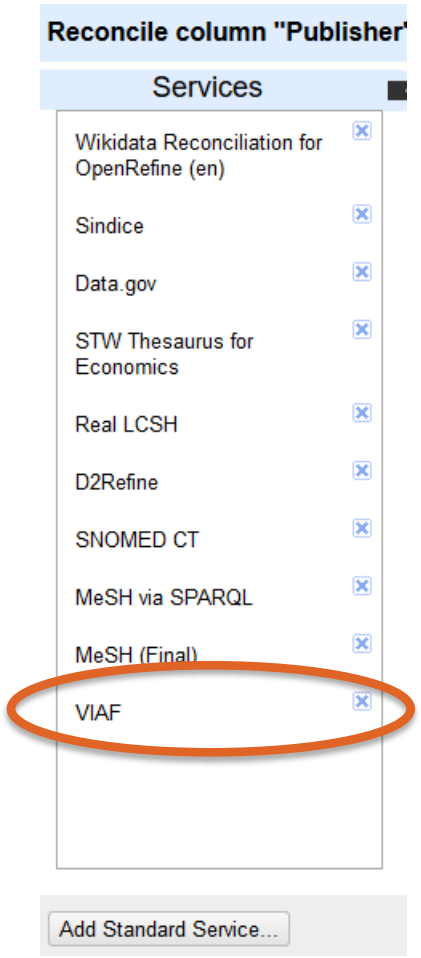


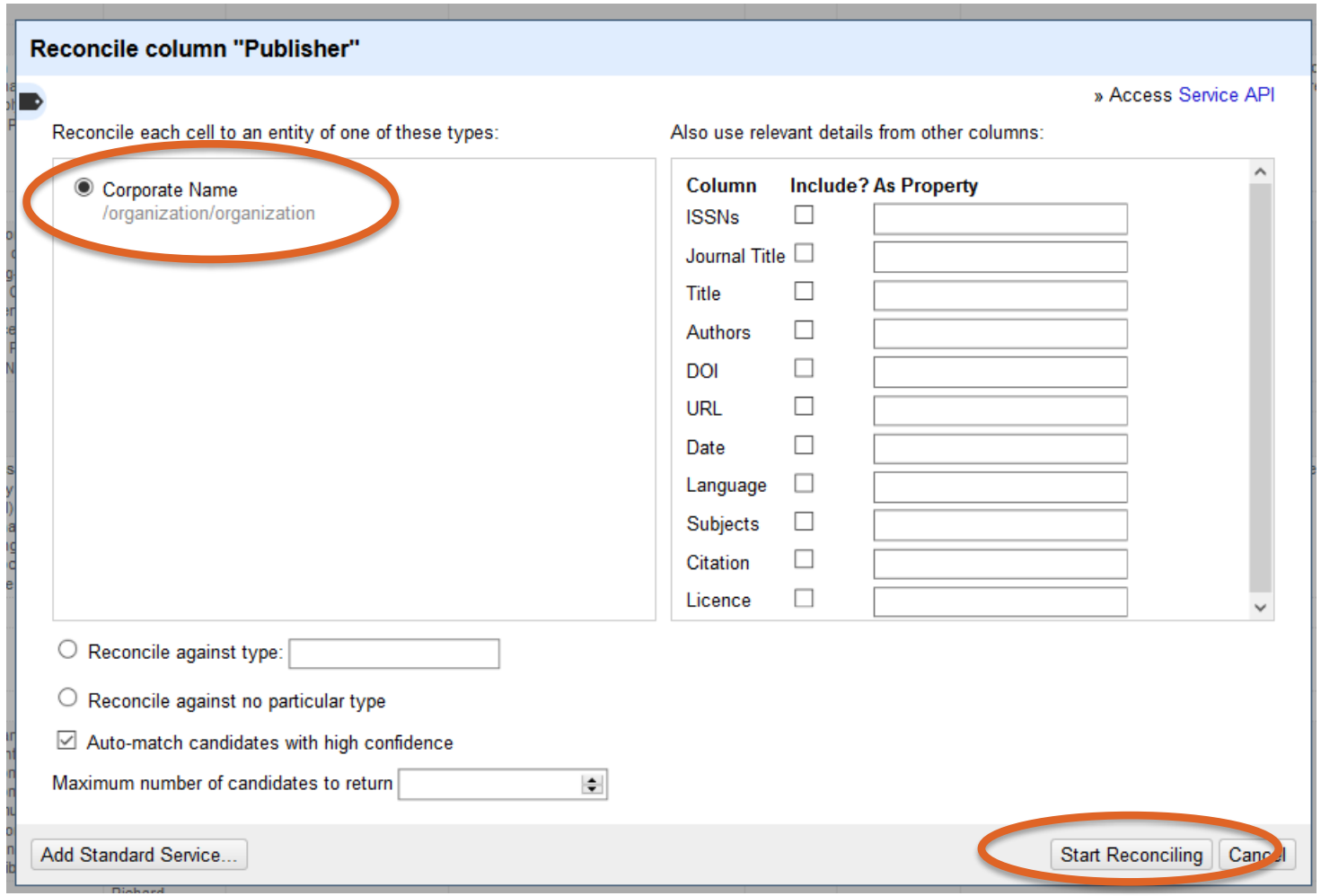
Journal Details column with long string of JSON code

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

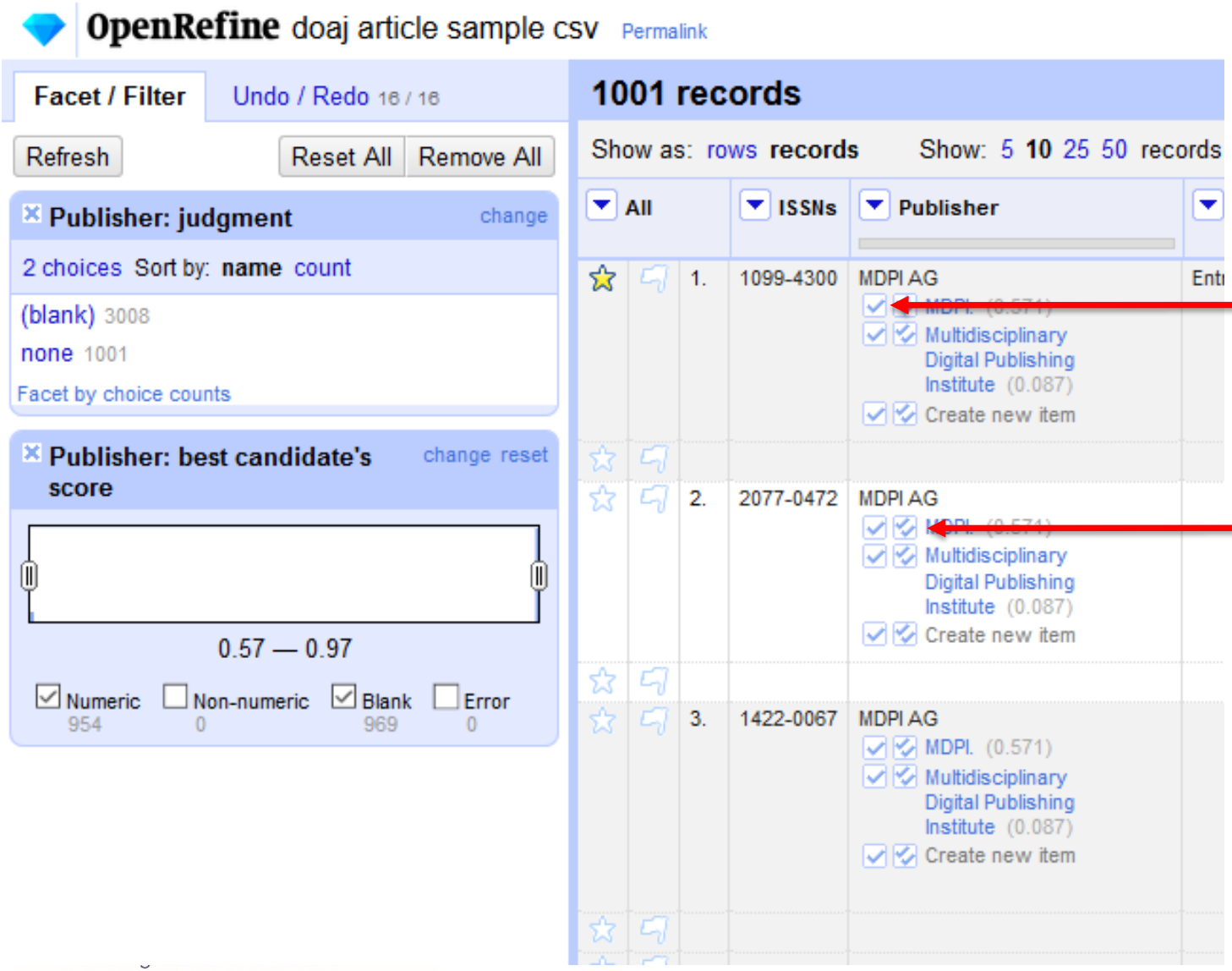
For more on the parse JSON function, see: https://github.com/OpenRefine/OpenRefine/wiki/GREL-Other-Functions.

Reconciliation using URLs





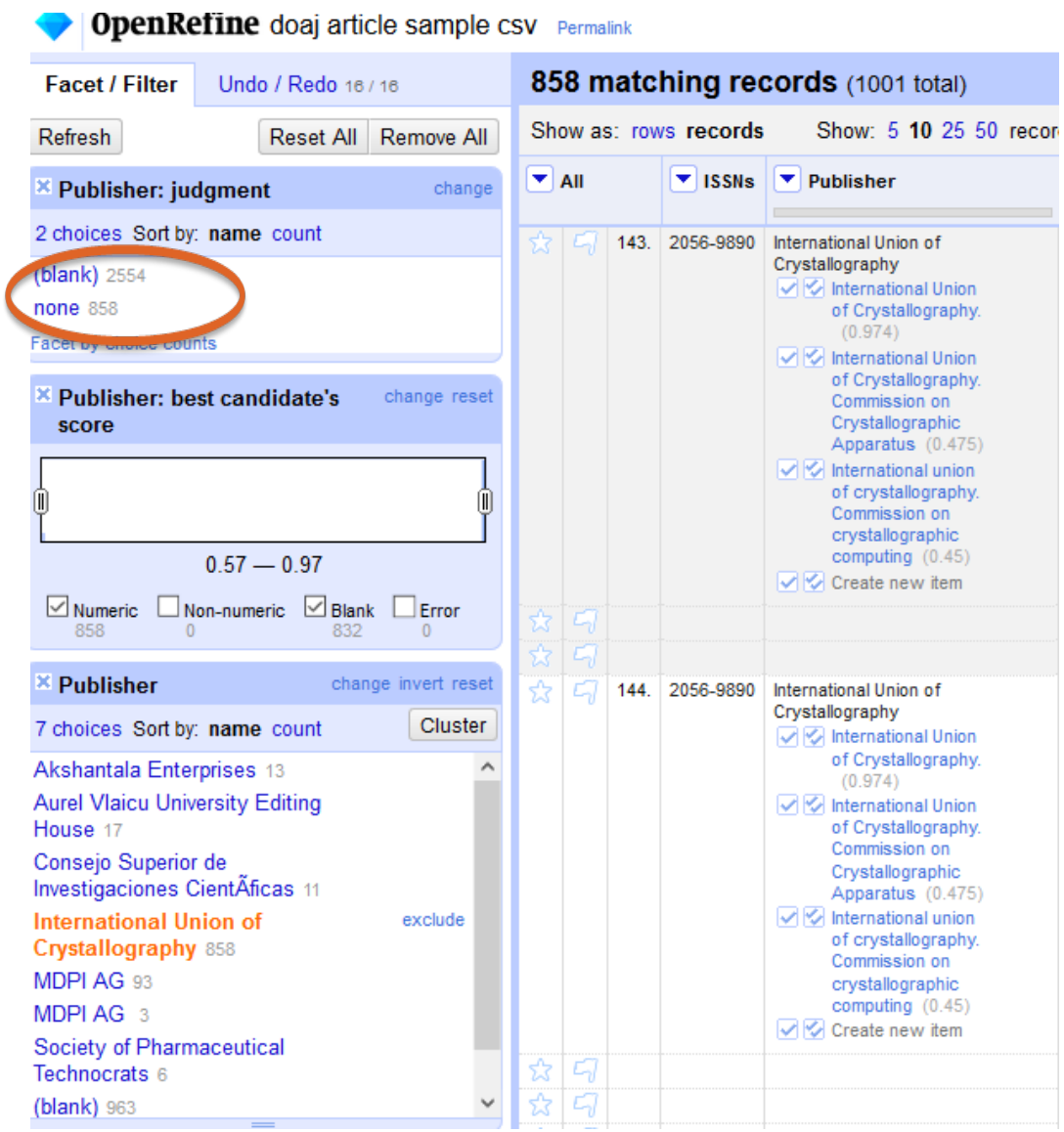
Reconciliation Results



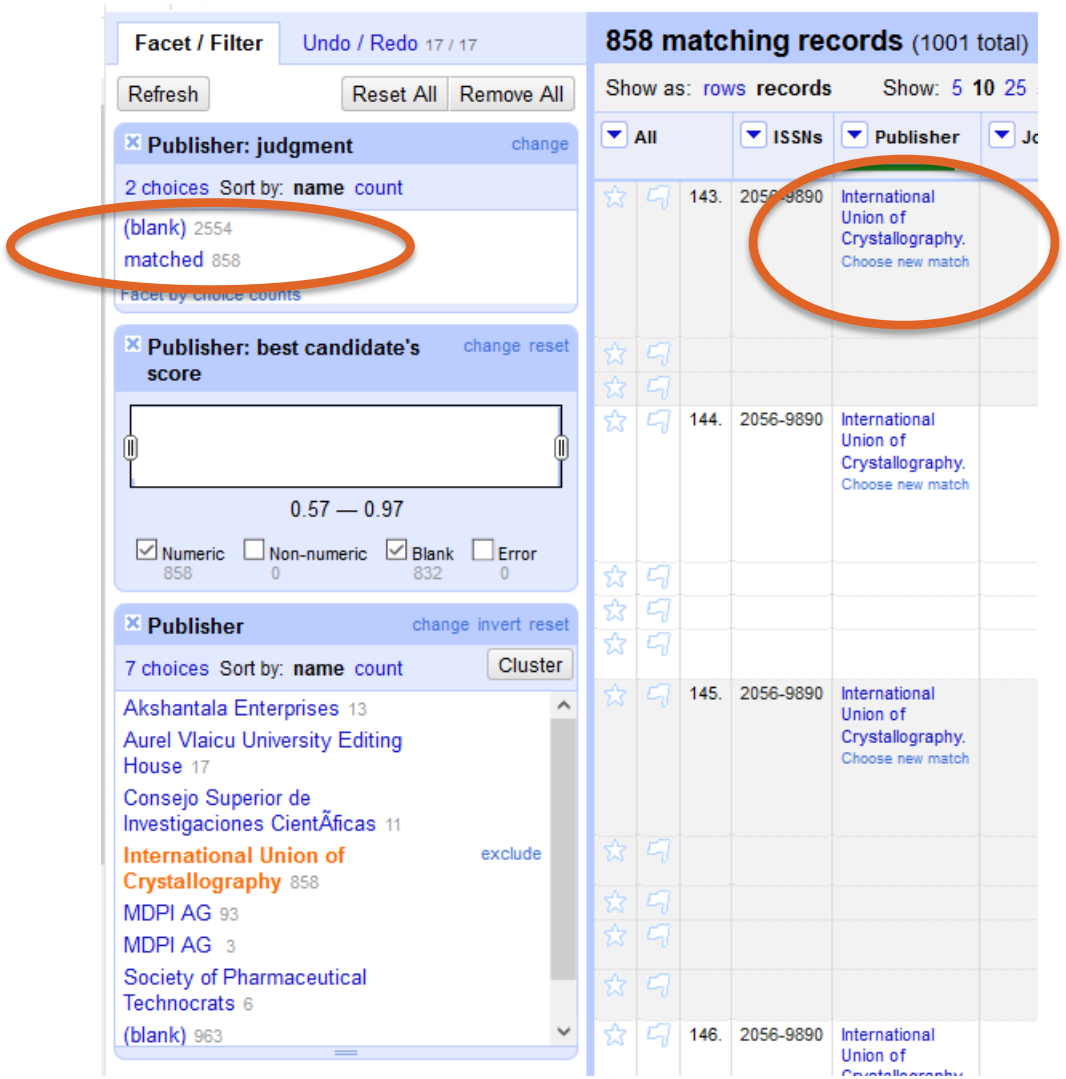
Single check-box: will accept the reconciled value only in this cell

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

Reconciliation Results, continued



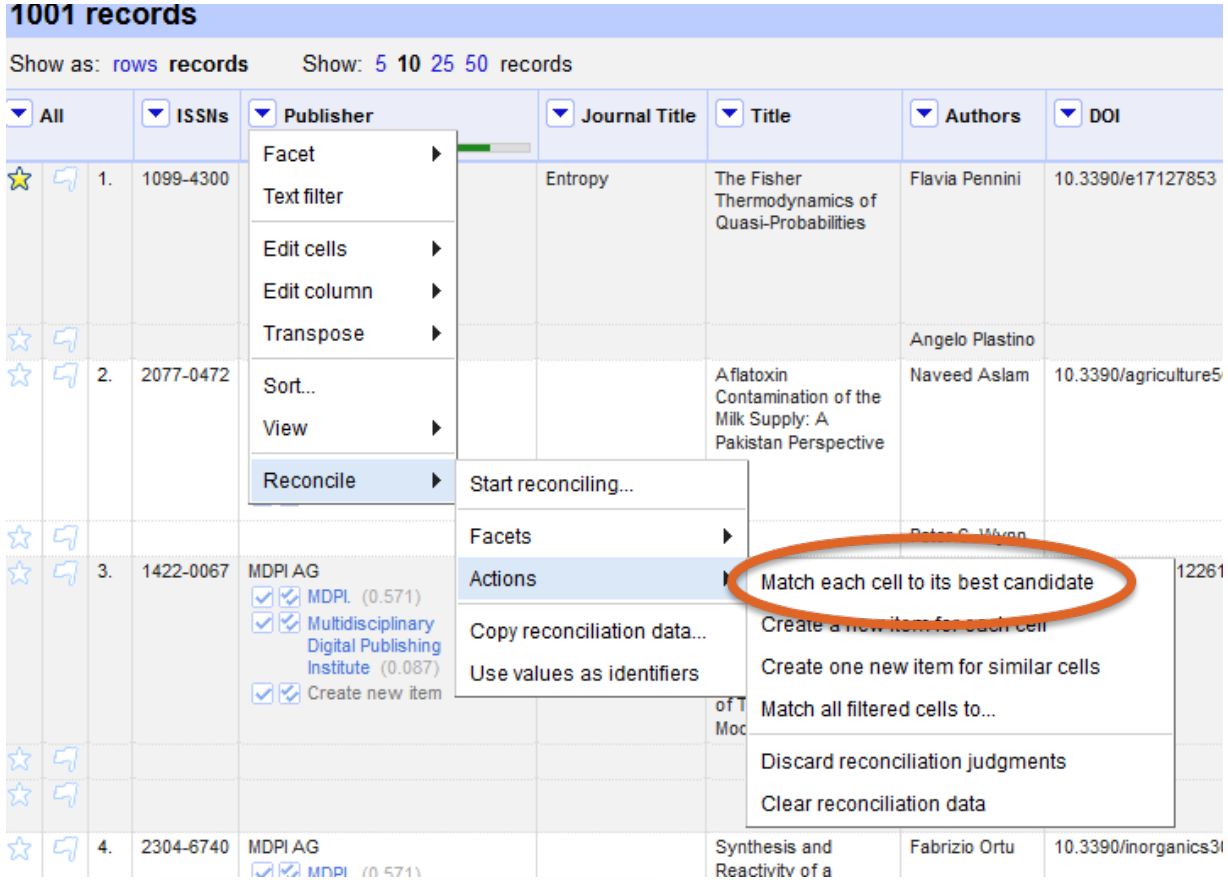
One new value is chosen. 858 values are matched.

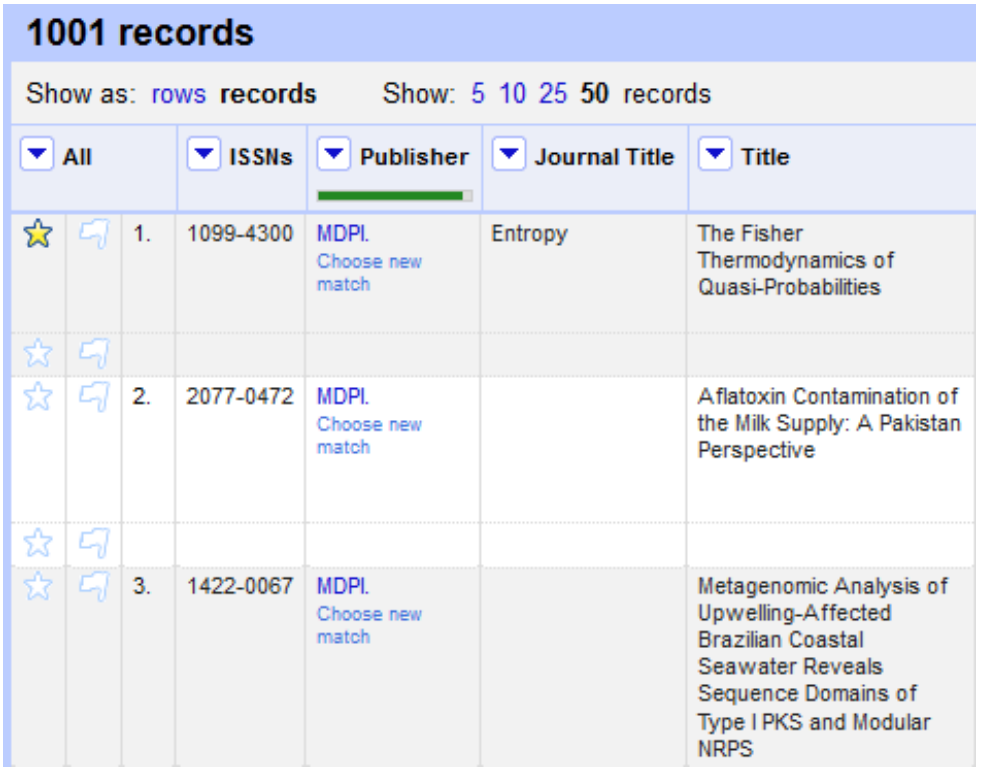


Reconciliation, continued

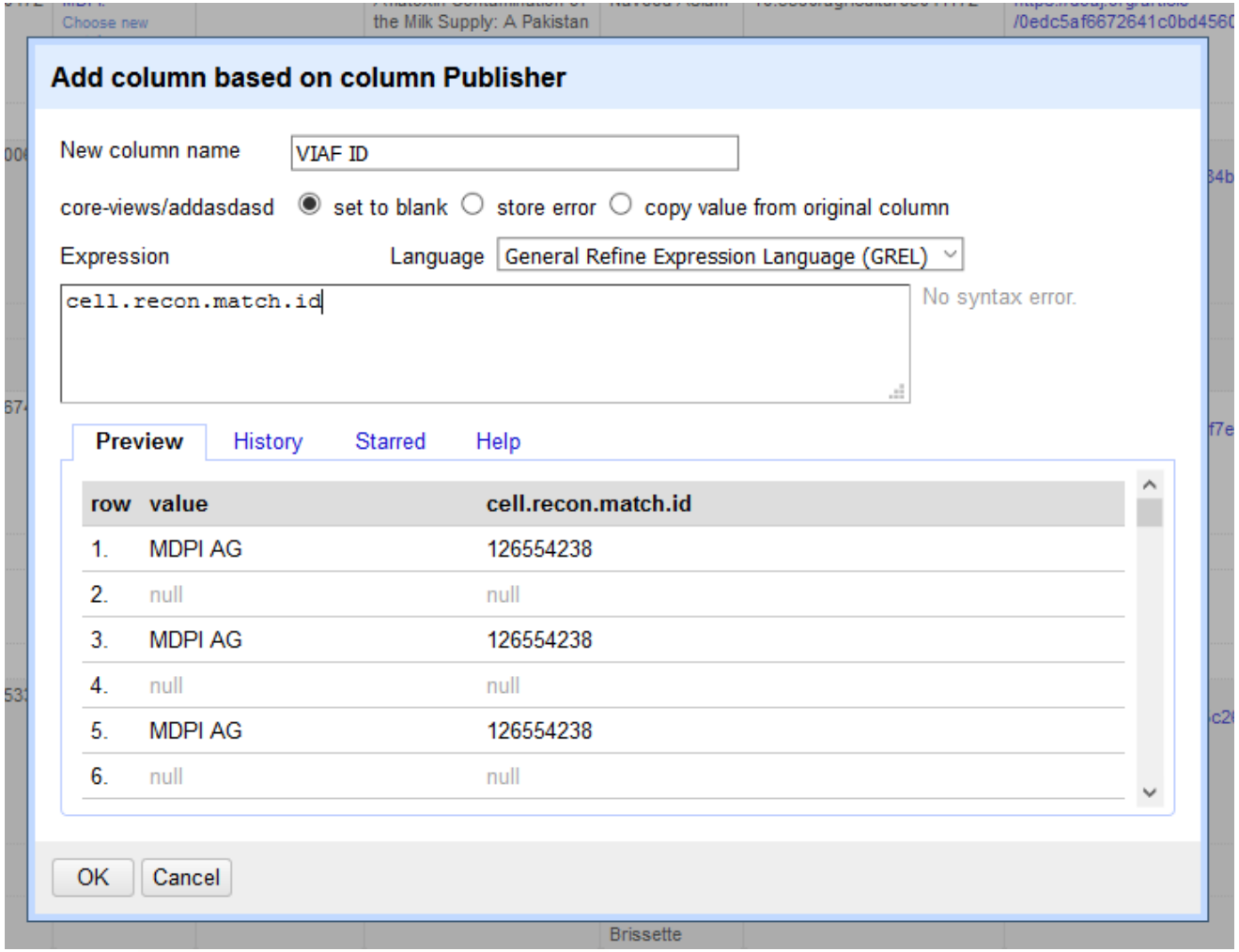
To try reconciliation a different way, remove all filters and facets from the project so all the rows display again. From the Publisher column drop-down menu choose Reconcile → Actions → Match each cell to its best candidate

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





Reconciliation, continued



VIAF ID number column added

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

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

Thank you!