MESSY DATA? CLEAN IT UP WITH OpenRefine

Tricia Lampron
Metadata Services Specialist
University of Illinois Library

GUIDES.LIBRARY.ILLINOIS.EDU/OPENREFINE

Created by Brinna Michael

OpenRefine

A free, open source, power tool
for working with messy data.

- About OpenRefine
- Hands-on:
  - *Importing your data*
  - *Manipulating and cleaning your data*
  - *Exporting your data*
- Resources
- Questions?

OpenRefine
as a Data Exploration Tool

- OpenRefine can *normalize* and *visualize* your data!
- OpenRefine looks like a *spreadsheet*, but it acts like a *database*.
- Use OpenRefine to *explore*, *clean*, and *link* your data.
  - *Pull out greater granular details when there's a lot of information*
  - *Pinpoint very specific criteria using a combination of facets*

Strengths:

- Free and Open Source
- Much more powerful than Excel
- Platform independent
- Great history tracking
- Can export commonly used functions for reuse

Weaknesses:

- Can be a little unstable (some queries will run slowly)
- Some methods require light programming*
- Some tasks that are easy in Excel are more difficult/impossible in OpenRefine
  - *Adding new rows / New Data*
  - *Editing individual cells*

Hands-On OpenRefine

- Creating a New Project
- Records vs. Rows
- Reordering Columns
- Basic Normalization
- Tracking Operation Histories
- Faceting and Clustering
- Exporting your work

To get started:

- Download and install OpenRefine
  - http://openrefine.org/download.html

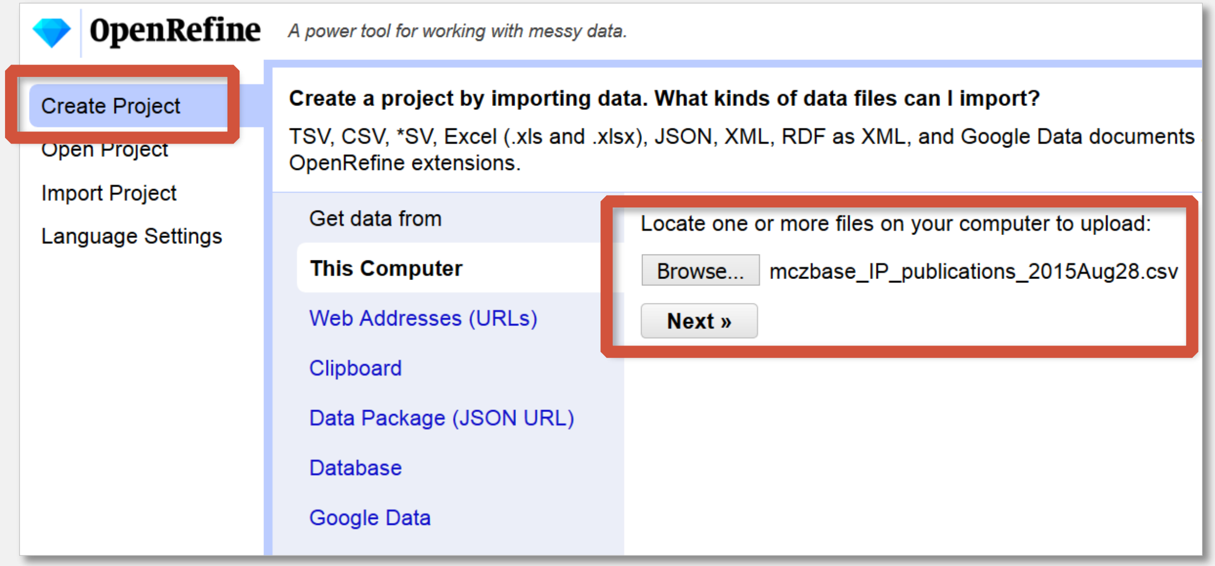
- PLEASE NOTE: OpenRefine runs in a browser, like a web app (though it does not require an internet connection). It works best with Chrome, Firefox, or Safari, and will automatically run on your default browser. OpenRefine is unstable when run on Internet Explorer and does not run at all on Edge, so it is best to choose another default browser.

SAMPLE DATA

goo.gl/CBqj2n

Creating a project

- TSV, CSV, other separated values
- Line-based; Fixed-width text files
- Spreadsheets
  - *Excel files (.xls, .xlsx)*
  - *Open document format spreadsheets (.ods)*
  - *Google Sheets*
- XML
- JSON



Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents
OpenRefine extensions.

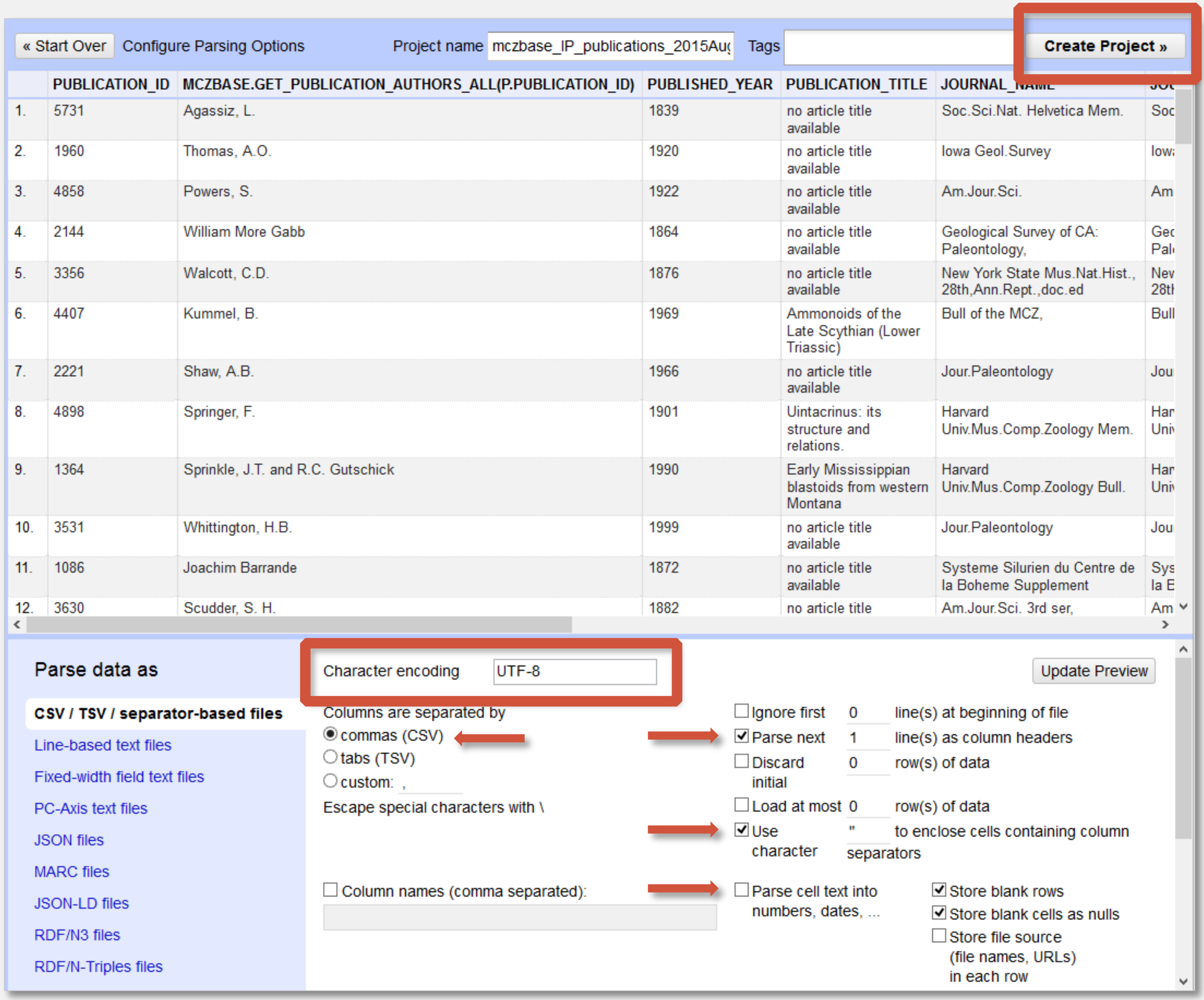
Locate one or more files on your computer to upload:

mczbase_IP_publications_2015Aug28.csv

Creating a project

Project Settings

- Check your encoding!
- Choose the correct separators!
- Parse your column headers!
- Don’t parse text into numbers, dates, etc.! (Unless you are VERY confident about your data)



Project name mczbase_IP_publications_2015Au(

Tags

	PUBLICATION_ID	MCZBASE.GET_PUBLICATION_AUTHORS_ALL(P.PUBLICATION_ID)	PUBLISHED_YEAR	PUBLICATION_TITLE	JOURNAL_NAME	JOU
1.	5731	Agassiz, L.	1839	no article title available	Soc.Sci.Nat. Helvetica Mem.	Soc
2.	1960	Thomas, A.O.	1920	no article title available	Iowa Geol.Survey	Iowa
3.	4858	Powers, S.	1922	no article title available	Am.Jour.Sci.	Am
4.	2144	William More Gabb	1864	no article title available	Geological Survey of CA: Paleontology,	Geo Pal
5.	3356	Walcott, C.D.	1876	no article title available	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.ed	New 28th
6.	4407	Kummel, B.	1969	Ammonoids of the Late Scythian (Lower Triassic)	Bull of the MCZ,	Bull
7.	2221	Shaw, A.B.	1966	no article title available	Jour.Paleontology	Jou
8.	4898	Springer, F.	1901	Uintacrinus: its structure and relations.	Harvard Univ.Mus.Comp.Zoology Mem.	Har Univ
9.	1364	Sprinkle, J.T. and R.C. Gutschick	1990	Early Mississippian blastoids from western Montana	Harvard Univ.Mus.Comp.Zoology Bull.	Har Univ
10.	3531	Whittington, H.B.	1999	no article title available	Jour.Paleontology	Jou
11.	1086	Joachim Barrande	1872	no article title available	Systeme Silurien du Centre de la Boheme Supplement	Sys la E
12.	3630	Scudder, S. H.	1882	no article title	Am.Jour.Sci. 3rd ser,	Am

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

Character encoding UTF-8

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

OpenRefine Layout

- OpenRefine displays data in a tabular format (like a spreadsheet):
  - *Each line will represent a **‘record’** or **‘row’** in the data*
  - *Each **column** represents a type of information*
- OpenRefine only displays a limited number of lines of data at a time.
  - *You can adjust the number to **5, 10** (default), **25,** or **50**.*
  - *When you manipulate your data, you are manipulating the whole set, not just the first 50!*

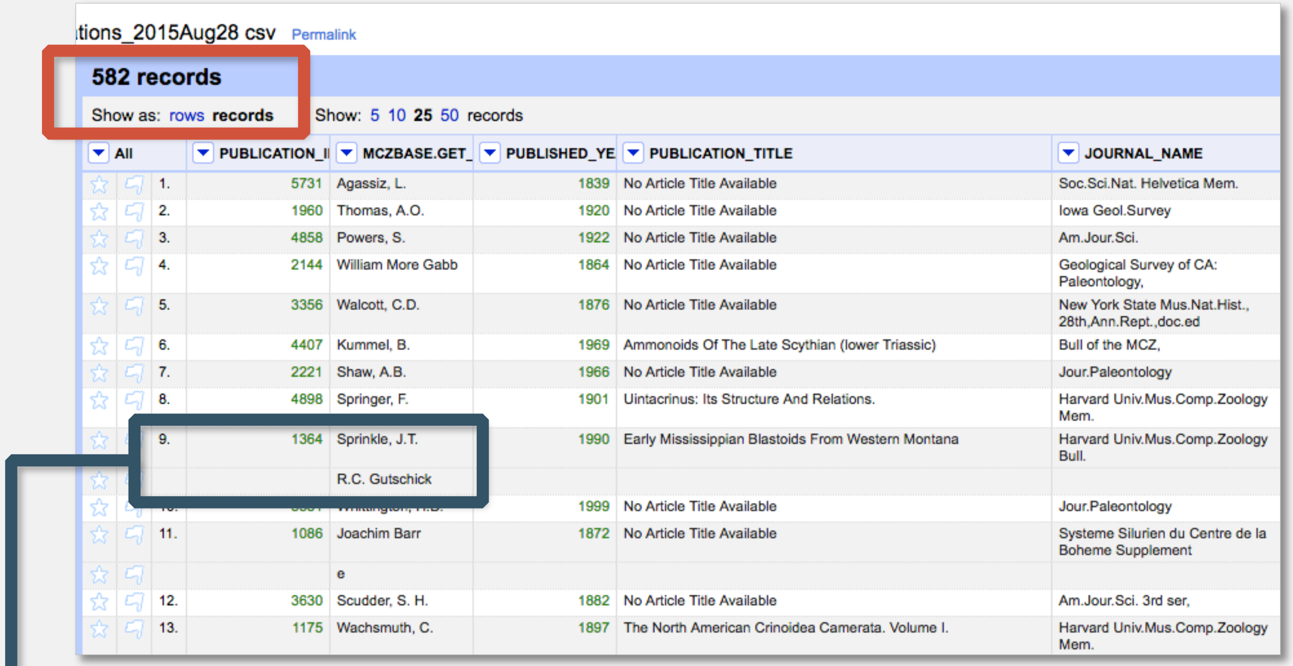
Records vs. Rows

ROWS

- Individual lines of data, not linked by relationships between the data in neighboring lines

RECORDS

- Multi-line groupings of data, linked by relationships between data in those lines

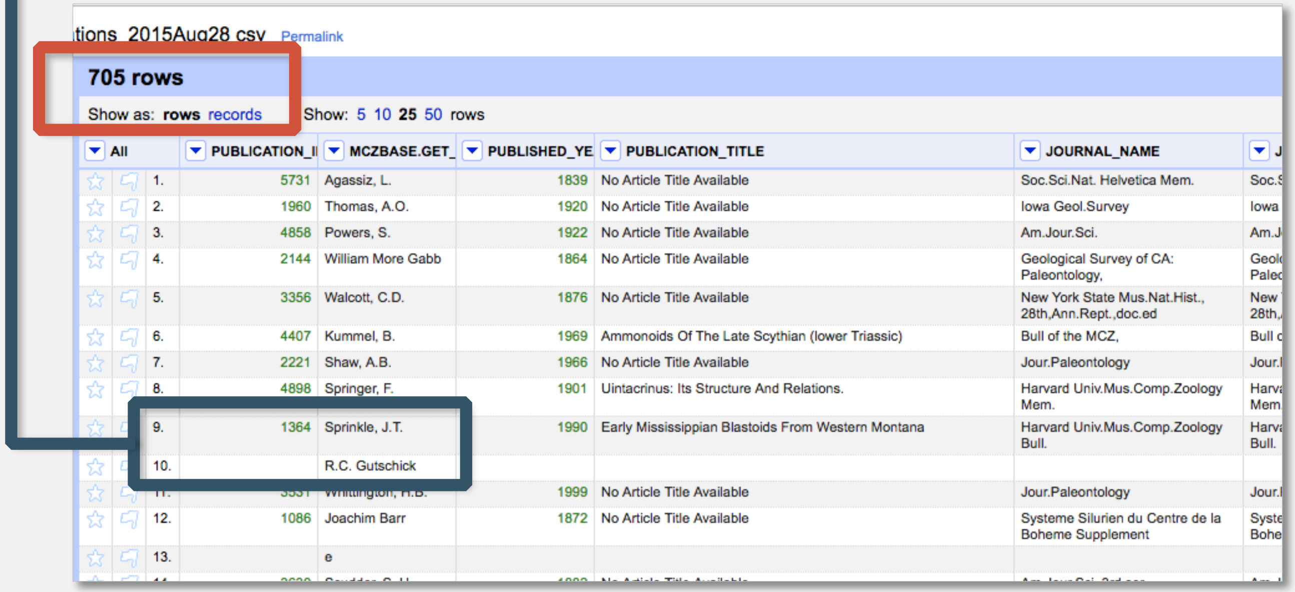


tions_2015Aug28 csv Permalink

582 records

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

All			PUBLICATION_I	MCZBASE.GET_	PUBLISHED_YE	PUBLICATION_TITLE	JOURNAL_NAME
☆		1.	5731	Agassiz, L.	1839	No Article Title Available	Soc.Sci.Nat. Helvetica Mem.
☆		2.	1960	Thomas, A.O.	1920	No Article Title Available	Iowa Geol.Survey
☆		3.	4858	Powers, S.	1922	No Article Title Available	Am.Jour.Sci.
☆		4.	2144	William More Gabb	1864	No Article Title Available	Geological Survey of CA: Paleontology,
☆		5.	3356	Walcott, C.D.	1876	No Article Title Available	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.ed
☆		6.	4407	Kummel, B.	1969	Ammonoids Of The Late Scythian (lower Triassic)	Bull of the MCZ,
☆		7.	2221	Shaw, A.B.	1966	No Article Title Available	Jour.Paleontology
☆		8.	4898	Springer, F.	1901	Uintacrinus: Its Structure And Relations.	Harvard Univ.Mus.Comp.Zoology Mem.
☆		9.	1364	Sprinkle, J.T.	1990	Early Mississippian Blastoids From Western Montana	Harvard Univ.Mus.Comp.Zoology Bull.
☆				R.C. Gutschick			
☆		10.	[illegible]	Whittington, H.B.	1999	No Article Title Available	Jour.Paleontology
☆		11.	1086	Joachim Barr	1872	No Article Title Available	Systeme Silurien du Centre de la Boheme Supplement
☆				e			
☆		12.	3630	Scudder, S. H.	1882	No Article Title Available	Am.Jour.Sci. 3rd ser,
☆		13.	1175	Wachsmuth, C.	1897	The North American Crinoidea Camerata. Volume I.	Harvard Univ.Mus.Comp.Zoology Mem.



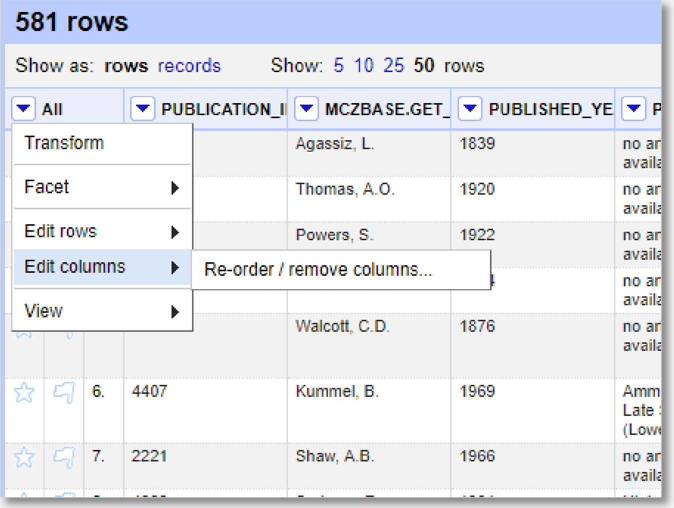
tions_2015Aug28 csv Permalink

705 rows

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

All			PUBLICATION_I	MCZBASE.GET_	PUBLISHED_YE	PUBLICATION_TITLE	JOURNAL_NAME	J
☆		1.	5731	Agassiz, L.	1839	No Article Title Available	Soc.Sci.Nat. Helvetica Mem.	Soc.S
☆		2.	1960	Thomas, A.O.	1920	No Article Title Available	Iowa Geol.Survey	Iowa
☆		3.	4858	Powers, S.	1922	No Article Title Available	Am.Jour.Sci.	Am.J
☆		4.	2144	William More Gabb	1864	No Article Title Available	Geological Survey of CA: Paleontology,	Geol Paleo
☆		5.	3356	Walcott, C.D.	1876	No Article Title Available	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.ed	New 28th,
☆		6.	4407	Kummel, B.	1969	Ammonoids Of The Late Scythian (lower Triassic)	Bull of the MCZ,	Bull o
☆		7.	2221	Shaw, A.B.	1966	No Article Title Available	Jour.Paleontology	Jour.I
☆		8.	4898	Springer, F.	1901	Uintacrinus: Its Structure And Relations.	Harvard Univ.Mus.Comp.Zoology Mem.	Harva Mem.
☆		9.	1364	Sprinkle, J.T.	1990	Early Mississippian Blastoids From Western Montana	Harvard Univ.Mus.Comp.Zoology Bull.	Harva Bull.
☆		10.		R.C. Gutschick				
☆		11.	[illegible]	Whittington, H.B.	1999	No Article Title Available	Jour.Paleontology	Jour.I
☆		12.	1086	Joachim Barr	1872	No Article Title Available	Systeme Silurien du Centre de la Boheme Supplement	Syste Bohe
☆		13.		e				
☆		14.	3630	Scudder, S. H.	1882	No Article Title Available	Am.Jour.Sci. 3rd ser,	Am.J

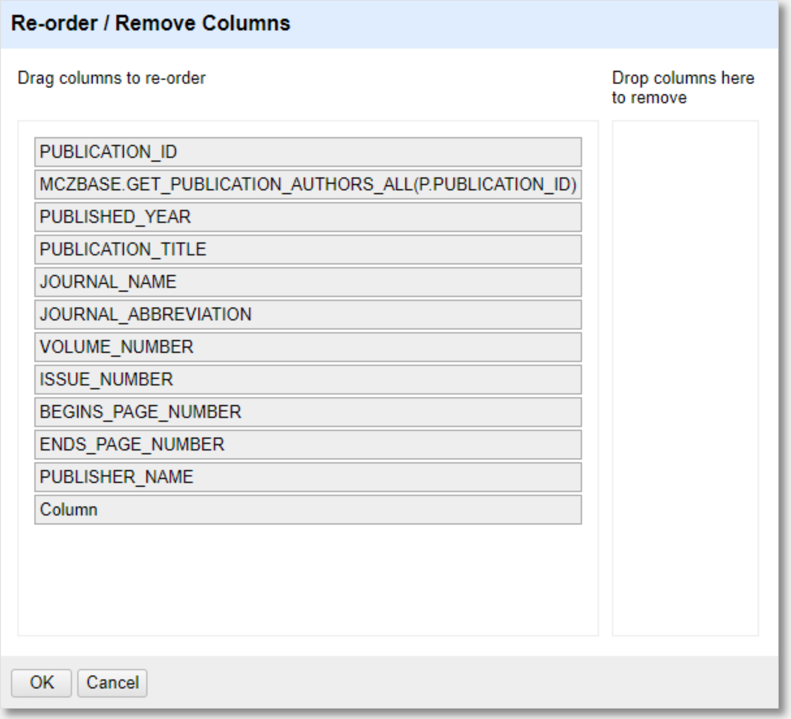
Reordering columns



581 rows

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

	PUBLICATION_I	MCZBASE.GET_	PUBLISHED_YE	P
		Agassiz, L.	1839	no ar availa
		Thomas, A.O.	1920	no ar availa
		Powers, S.	1922	no ar availa
				no ar availa
		Walcott, C.D.	1876	no ar availa
6.	4407	Kummel, B.	1969	Amm Late (Lowe
7.	2221	Shaw, A.B.	1966	no ar availa



Re-order / Remove Columns

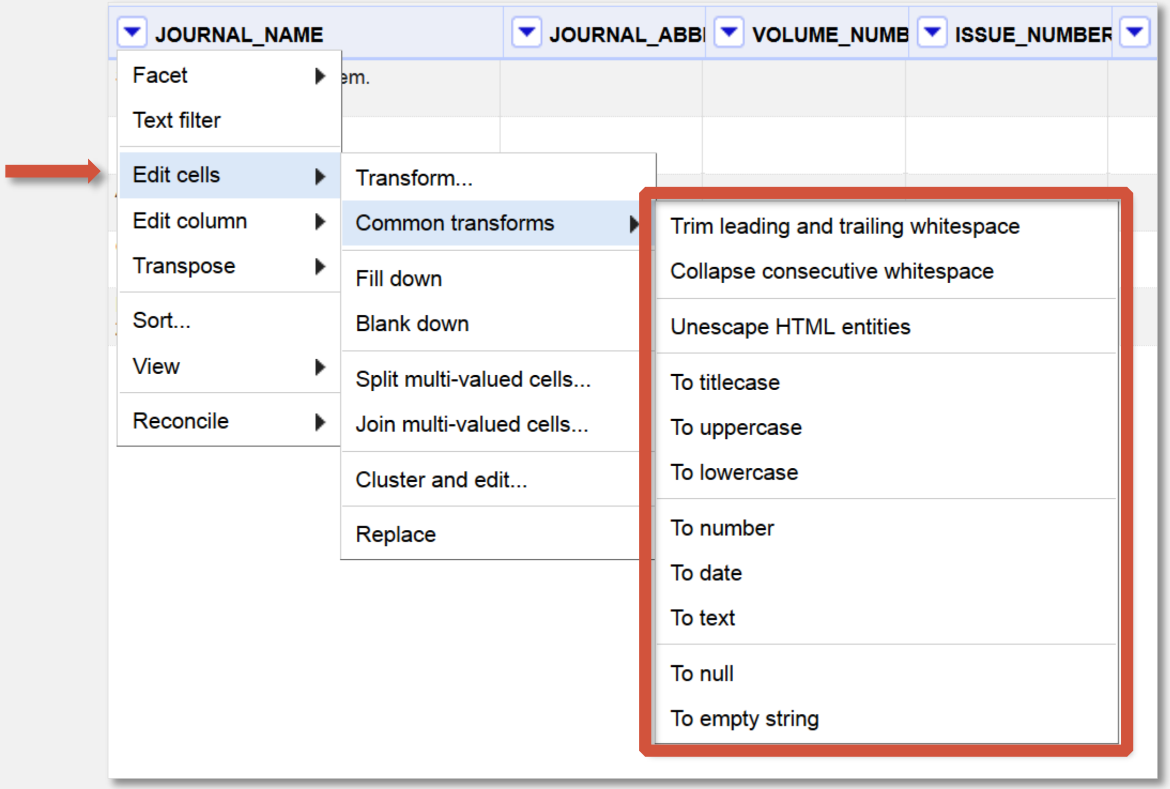
Drag columns to re-order

Drop columns here to remove

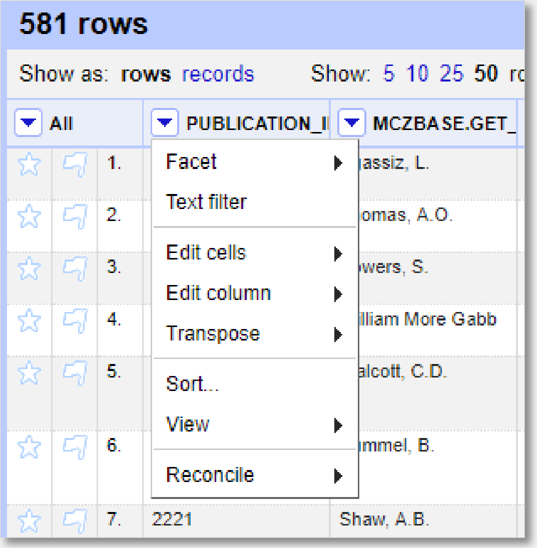
- PUBLICATION_ID
- MCZBASE.GET_PUBLICATION_AUTHORS_ALL(P.PUBLICATION_ID)
- PUBLISHED_YEAR
- PUBLICATION_TITLE
- JOURNAL_NAME
- JOURNAL_ABBREVIATION
- VOLUME_NUMBER
- ISSUE_NUMBER
- BEGINS_PAGE_NUMBER
- ENDS_PAGE_NUMBER
- PUBLISHER_NAME
- Column

Basic normalization

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
  - &nbsp;
  - &amp;
- titlecase, uppercase, lowercase
- number, date, text
- null, empty string



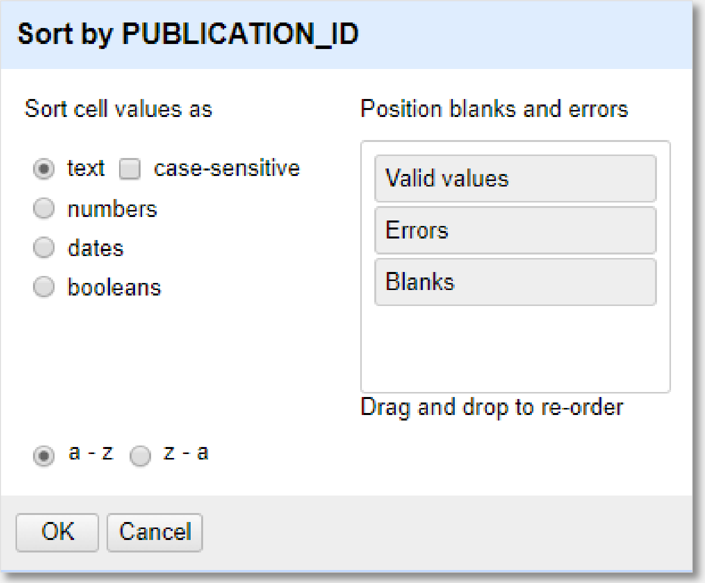
Sorting Data



581 rows

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

			All	PUBLICATION_I	MCZBASE.GET_
☆		1.			assiz, L.
☆		2.			omas, A.O.
☆		3.			wers, S.
☆		4.			lliam More Gabb
☆		5.			alcott, C.D.
☆		6.			mmel, B.
☆		7.		2221	Shaw, A.B.



Sort by PUBLICATION_ID

Sort cell values as

● text □ case-sensitive

○ numbers

○ dates

○ booleans

Position blanks and errors

Valid values

Errors

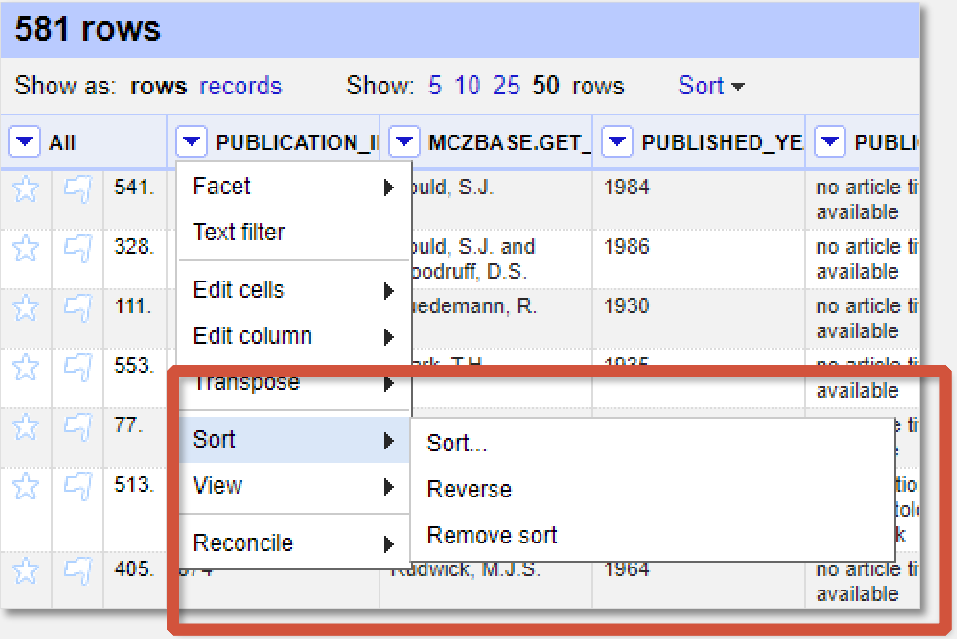
Blanks

Drag and drop to re-order

● a - z ○ z - a

Sorting Data

- Reverse sort
- Remove sort



581 rows

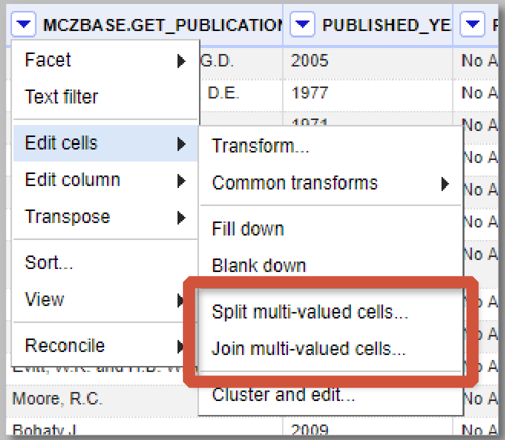
Show as: rows records   Show: 5 10 25 50 rows   Sort ▾

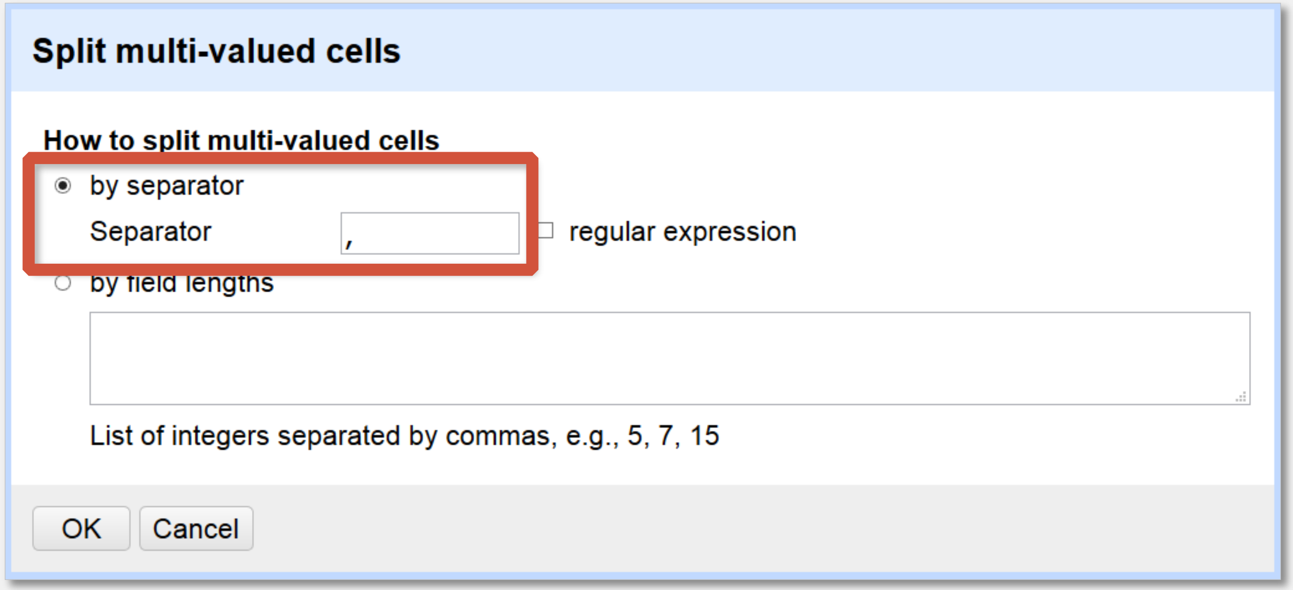
	All	PUBLICATION_I	MCZBASE.GET_	PUBLISHED_YE	PUBLI
☆	541.		uld, S.J.	1984	no article ti available
☆	328.		uld, S.J. and oodruff, D.S.	1986	no article ti available
☆	111.		edemann, R.	1930	no article ti available
☆	553.		ark, T.H.	1935	no article ti available
☆	77.				e ti
☆	513.				tio tol k
☆	405.	574	Rudwick, M.J.S.	1964	no article ti available

Splitting multiple values

Identifier	Author
123456	Smith, Joe; Green, Jane
123457	Collins, Mark; Prestley, Susan

Identifier	Author
123456	Smith, Joe
	Green, Jane
123457	Collins, Mark
	Prestley, Susan




Split multi-valued cells

How to split multi-valued cells

● by separator

Separator , □ regular expression

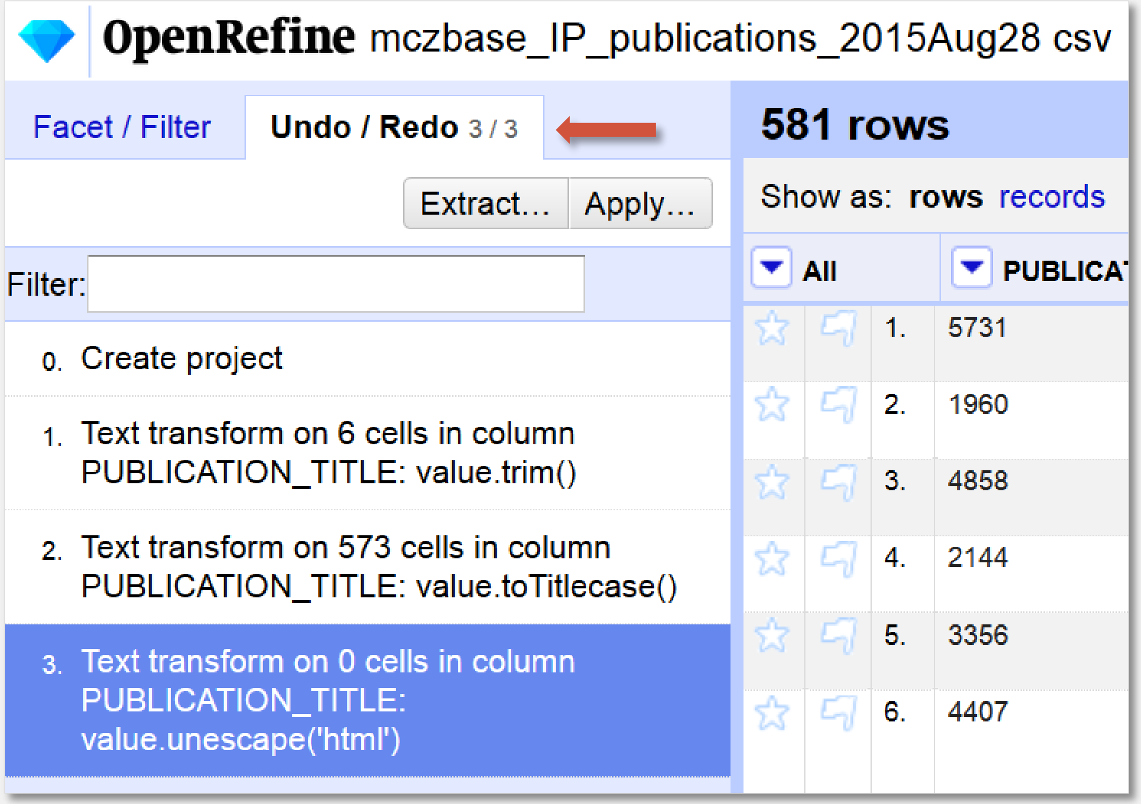
○ by field lengths

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

A note on tracking project history

Don't be afraid to make mistakes!

...you can always use the Undo/Redo tab to navigate to earlier stages in your project.



OpenRefine mczbase_IP_publications_2015Aug28 csv

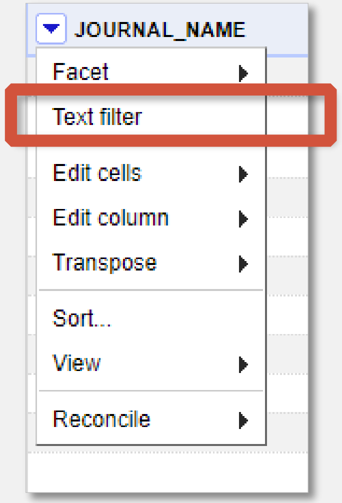
0. Create project
1. Text transform on 6 cells in column PUBLICATION_TITLE: value.trim()
2. Text transform on 573 cells in column PUBLICATION_TITLE: value.toTitlecase()
3. Text transform on 0 cells in column PUBLICATION_TITLE: value.unescape('html')

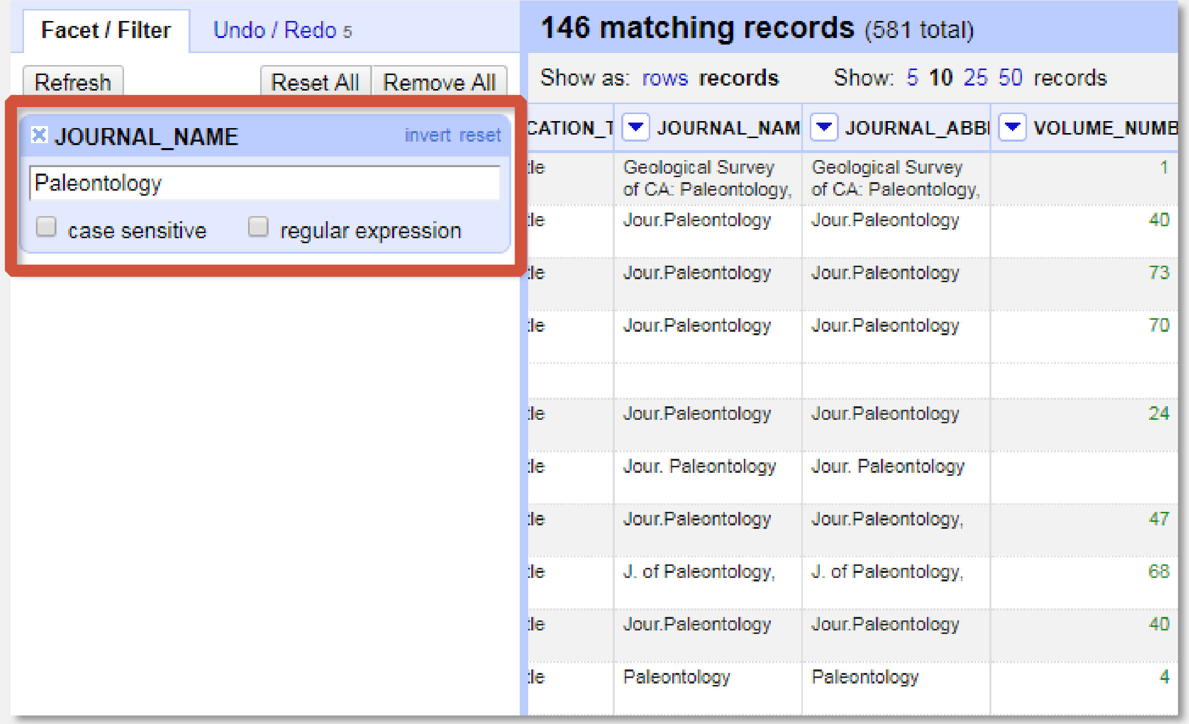
581 rows

Show as: **rows** records

	All	PUBLICA
☆	1.	5731
☆	2.	1960
☆	3.	4858
☆	4.	2144
☆	5.	3356
☆	6.	4407

Text filtering





☒ JOURNAL_NAME invert reset

Paleontology

□ case sensitive □ regular expression

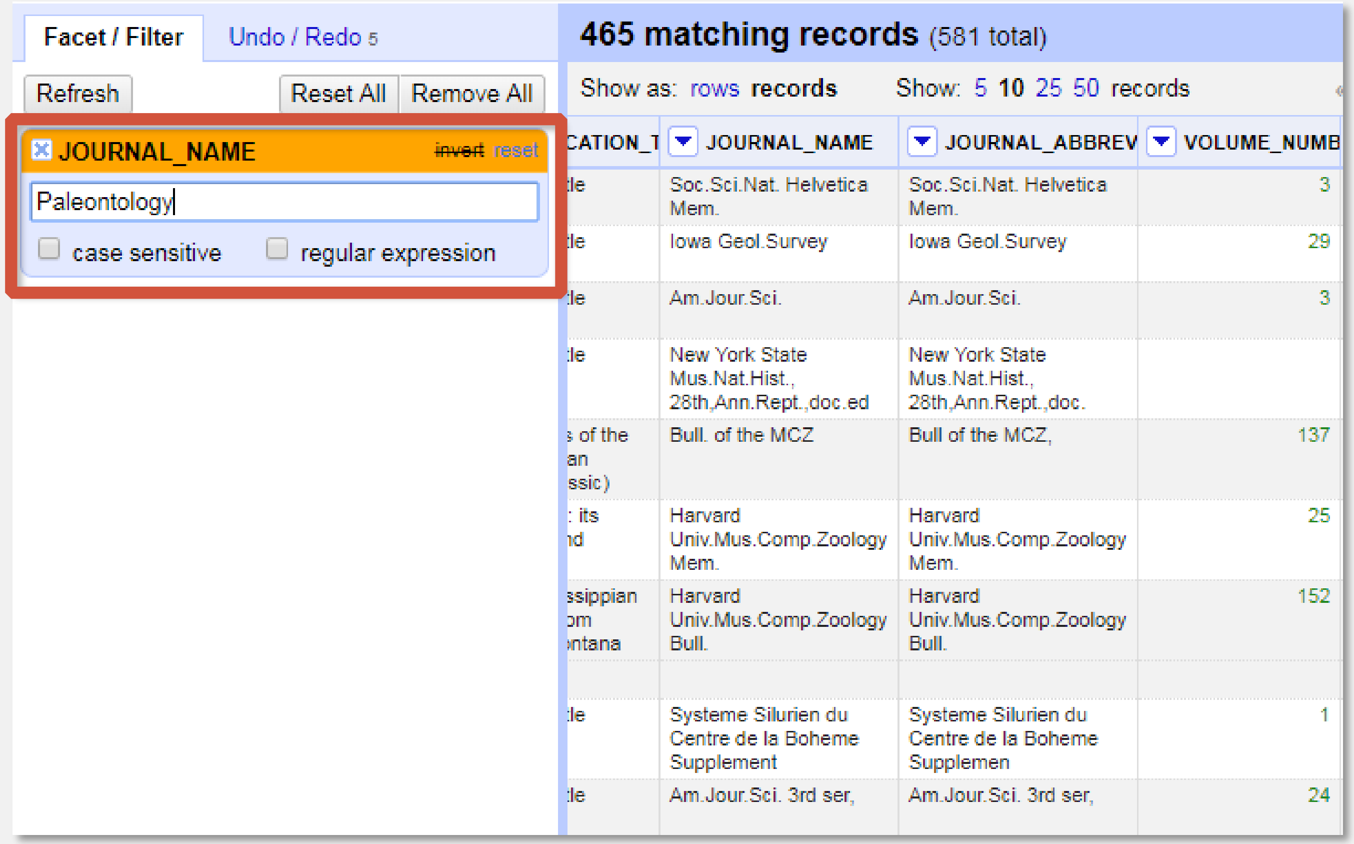
146 matching records (581 total)

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

CATION_T	JOURNAL_NAM	JOURNAL_ABBI	VOLUME_NUMB
le	Geological Survey of CA: Paleontology,	Geological Survey of CA: Paleontology,	1
le	Jour.Paleontology	Jour.Paleontology	40
le	Jour.Paleontology	Jour.Paleontology	73
le	Jour.Paleontology	Jour.Paleontology	70
le	Jour.Paleontology	Jour.Paleontology	24
le	Jour. Paleontology	Jour. Paleontology	
le	Jour.Paleontology	Jour.Paleontology,	47
le	J. of Paleontology,	J. of Paleontology,	68
le	Jour.Paleontology	Jour.Paleontology	40
le	Paleontology	Paleontology	4

*You can also use regular expressions in the filter!

Text filtering



JOURNAL_NAME

Paleontology

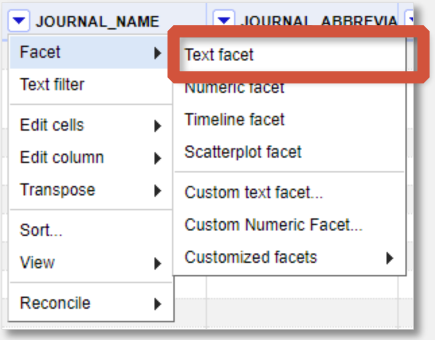
☐ case sensitive ☐ regular expression

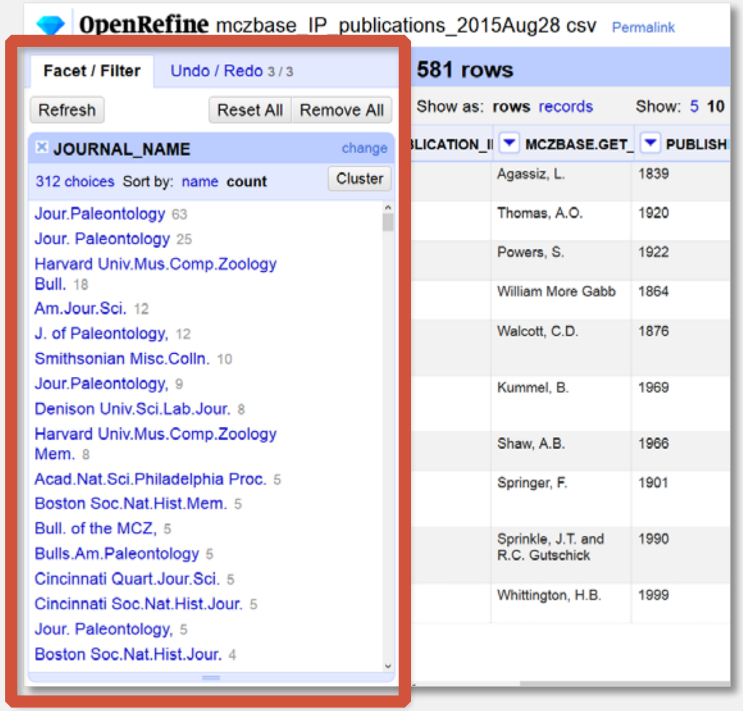
465 matching records (581 total)

CATION_T	JOURNAL_NAME	JOURNAL_ABBREV	VOLUME_NUMB
le	Soc.Sci.Nat. Helvetica Mem.	Soc.Sci.Nat. Helvetica Mem.	3
le	Iowa Geol.Survey	Iowa Geol.Survey	29
le	Am.Jour.Sci.	Am.Jour.Sci.	3
le	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.ed	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.	
s of the an ssic)	Bull. of the MCZ	Bull of the MCZ,	137
: its nd	Harvard Univ.Mus.Comp.Zoology Mem.	Harvard Univ.Mus.Comp.Zoology Mem.	25
ssippian om ntana	Harvard Univ.Mus.Comp.Zoology Bull.	Harvard Univ.Mus.Comp.Zoology Bull.	152
le	Systeme Silurien du Centre de la Boheme Supplement	Systeme Silurien du Centre de la Boheme Supplemen	1
le	Am.Jour.Sci. 3rd ser,	Am.Jour.Sci. 3rd ser,	24

Faceting

- View unique values
- Select set of data based on unique values
- Edit all cells with the same value
- Extract values and counts





OpenRefine mczbase_IP_publications_2015Aug28 csv

581 rows

JOURNAL_NAME

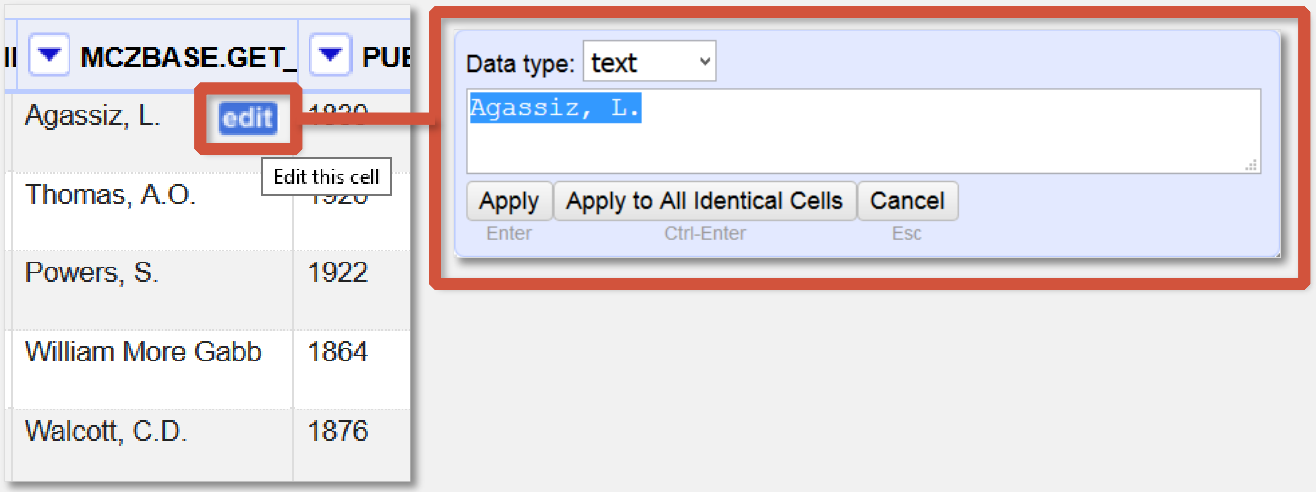
312 choices Sort by: name count

Jour.Paleontology	63
Jour. Paleontology	25
Harvard Univ.Mus.Comp.Zoology Bull.	18
Am.Jour.Sci.	12
J. of Paleontology,	12
Smithsonian Misc.Colln.	10
Jour.Paleontology,	9
Denison Univ.Sci.Lab.Jour.	8
Harvard Univ.Mus.Comp.Zoology Mem.	8
Acad.Nat.Sci.Philadelphia Proc.	5
Boston Soc.Nat.Hist.Mem.	5
Bull. of the MCZ,	5
Bulls.Am.Paleontology	5
Cincinnati Quart.Jour.Sci.	5
Cincinnati Soc.Nat.Hist.Jour.	5
Jour. Paleontology,	5
Boston Soc.Nat.Hist.Jour.	4

LICATION_I	MCZBASE.GET_	PUBLISH
	Agassiz, L.	1839
	Thomas, A.O.	1920
	Powers, S.	1922
	William More Gabb	1864
	Walcott, C.D.	1876
	Kummel, B.	1969
	Shaw, A.B.	1966
	Springer, F.	1901
	Sprinkle, J.T. and R.C. Gutschick	1990
	Whittington, H.B.	1999

Editing Values

In a Cell

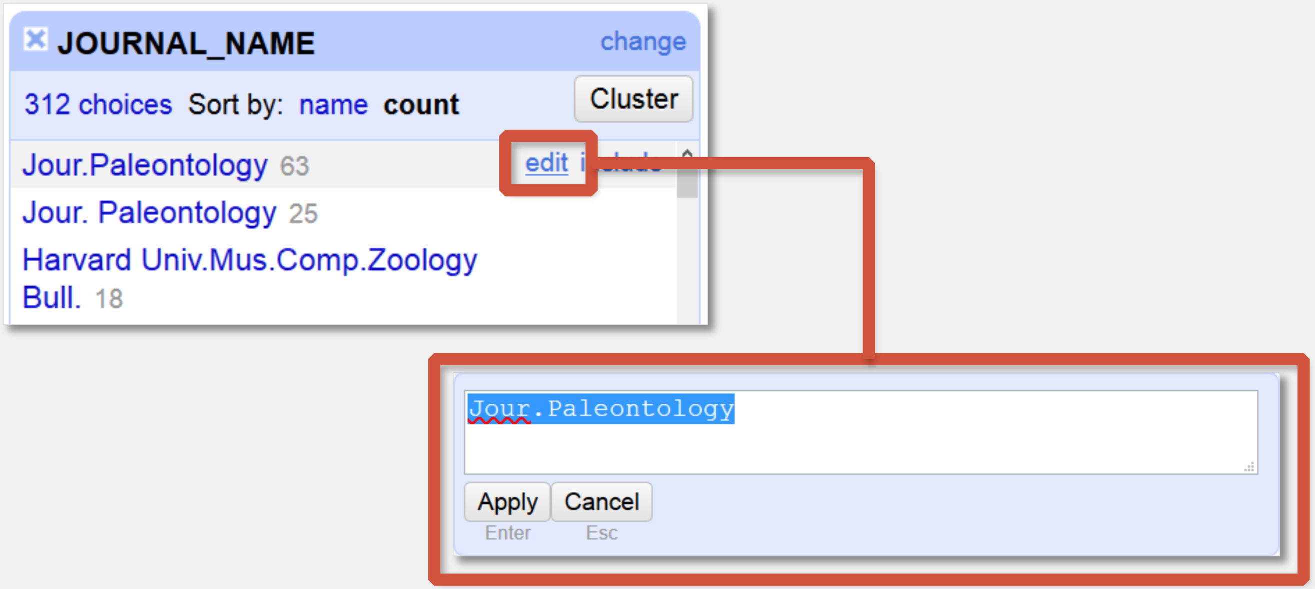


MCZBASE.GET_	PUE
Agassiz, L.	[illegible]
Thomas, A.O.	[illegible]
Powers, S.	1922
William More Gabb	1864
Walcott, C.D.	1876

Data type: text

Agassiz, L.

In a Facet



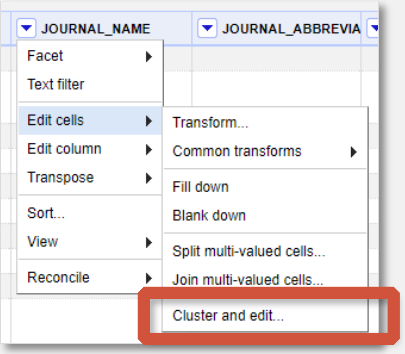
JOURNAL_NAME

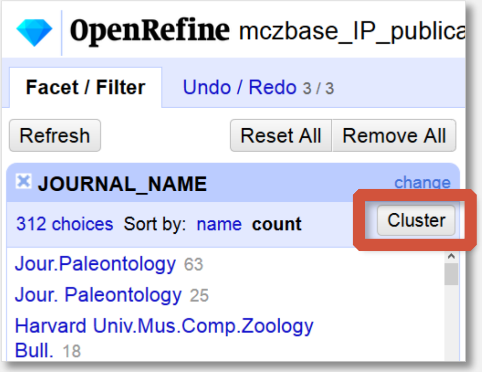
312 choices Sort by: name count

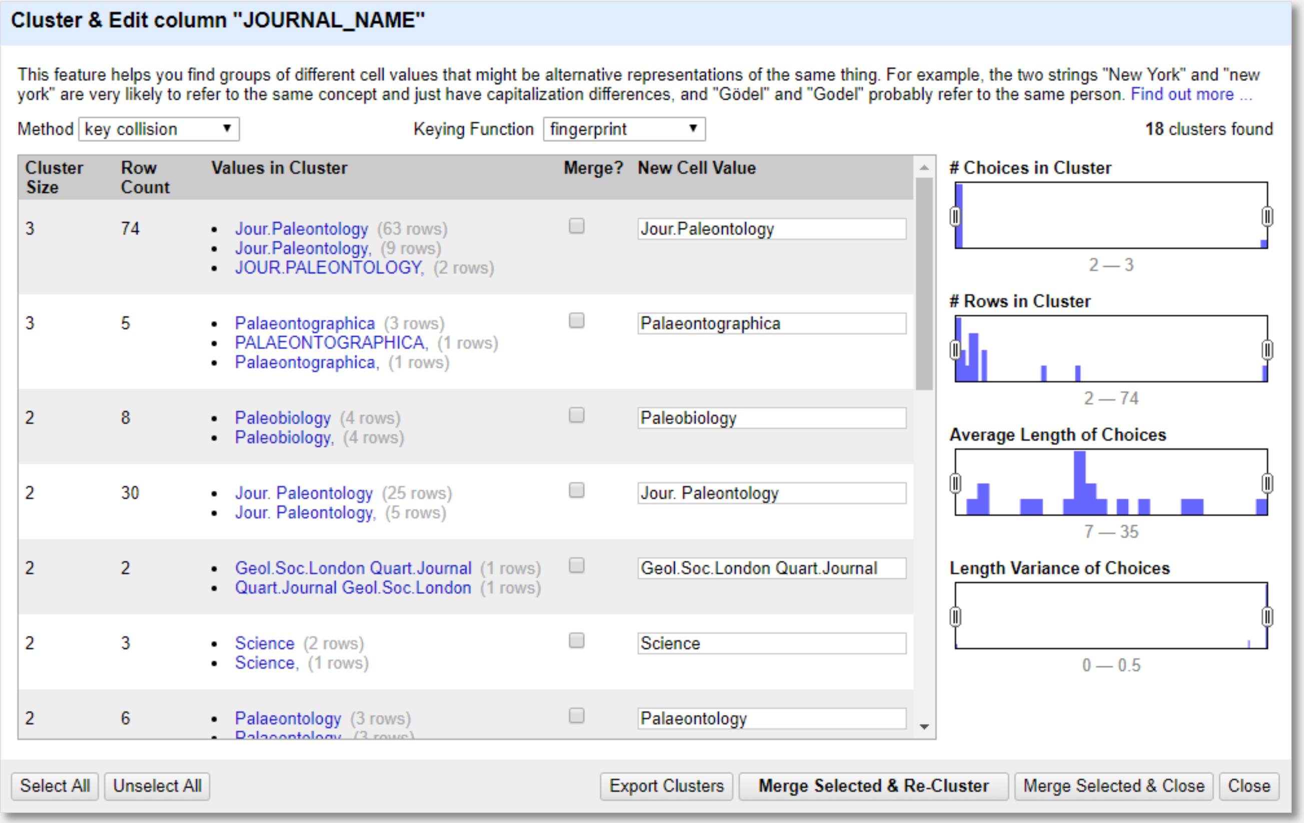
Jour.Paleontology	63
Jour. Paleontology	25
Harvard Univ.Mus.Comp.Zoology Bull.	18

Jour.Paleontology

Clustering






Cluster & Edit column "JOURNAL_NAME"

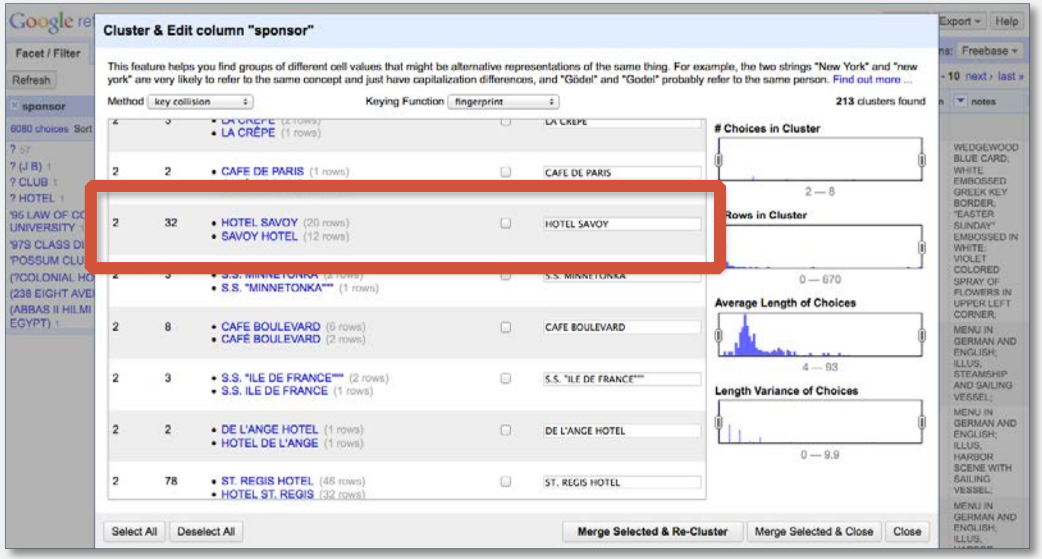
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method key collision ▾ Keying Function fingerprint ▾ 18 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	74	Jour.Paleontology (63 rows) Jour.Paleontology, (9 rows) JOUR.PALEONTOLOGY, (2 rows)	□	Jour.Paleontology
3	5	Palaeontographica (3 rows) PALAEONTOGRAPHICA, (1 rows) Palaeontographica, (1 rows)	□	Palaeontographica
2	8	Paleobiology (4 rows) Paleobiology, (4 rows)	□	Paleobiology
2	30	Jour. Paleontology (25 rows) Jour. Paleontology, (5 rows)	□	Jour. Paleontology
2	2	Geol.Soc.London Quart.Journal (1 rows) Quart.Journal Geol.Soc.London (1 rows)	□	Geol.Soc.London Quart.Journal
2	3	Science (2 rows) Science, (1 rows)	□	Science
2	6	Palaeontology (3 rows) Palaeontology (3 rows)	□	Palaeontology

Clustering

Caveat



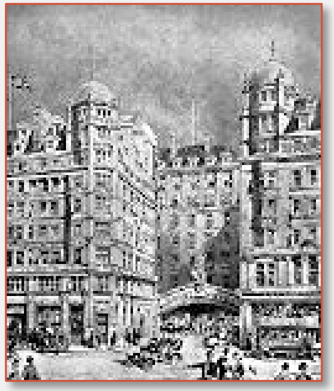
Cluster & Edit column "sponsor"

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   213 clusters found

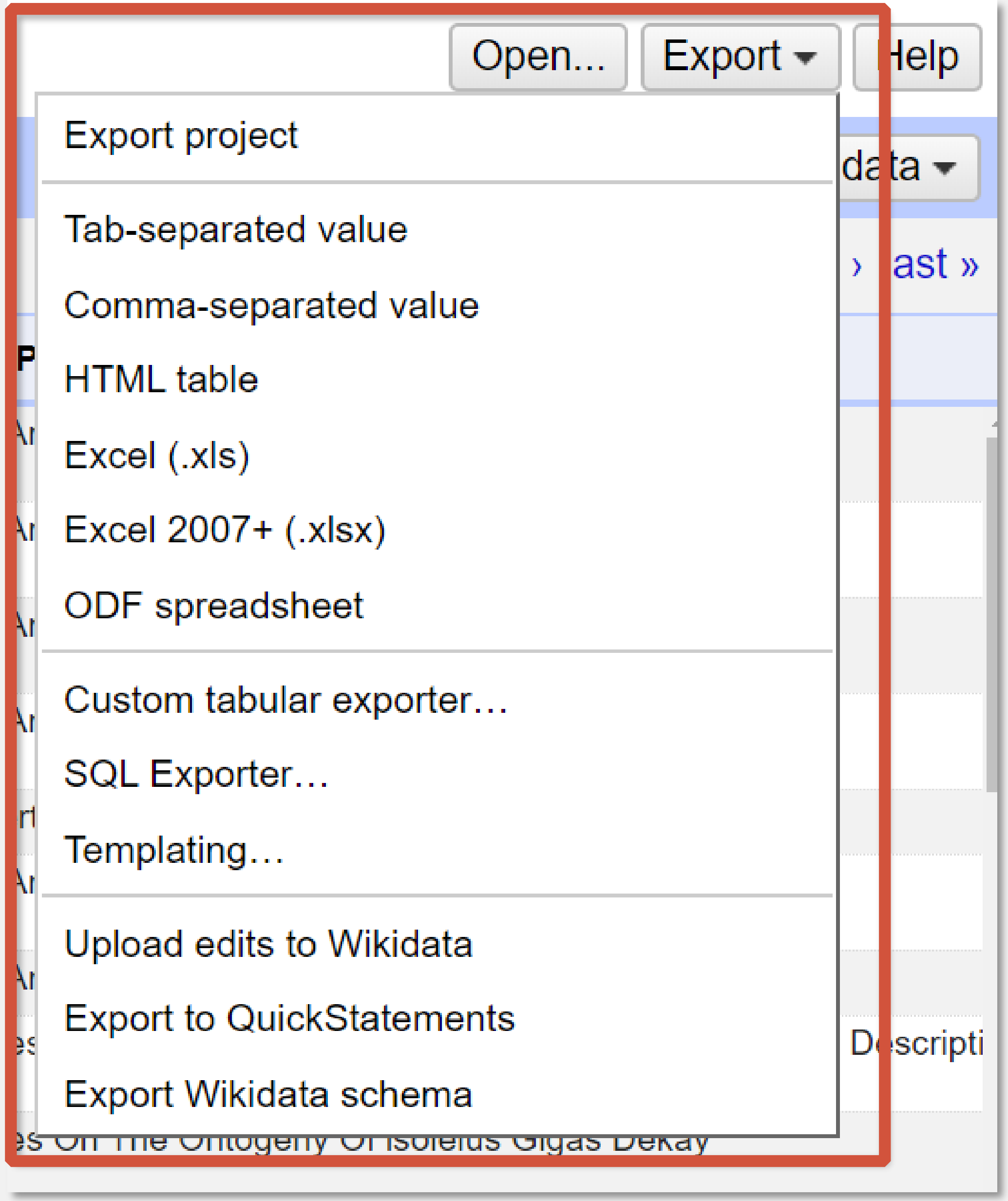
2	3	• LA CRÊPE (1 rows)	□	LA CRÊPE
2	2	• CAFE DE PARIS (1 rows)	□	CAFE DE PARIS
2	32	• HOTEL SAVOY (20 rows) • SAVOY HOTEL (12 rows)	□	HOTEL SAVOY
2	3	• S.S. "MINNETONKA""" (1 rows)	□	S.S. MINNETONKA
2	8	• CAFE BOULEVARD (6 rows) • CAFÉ BOULEVARD (2 rows)	□	CAFE BOULEVARD
2	3	• S.S. "ILE DE FRANCE""" (2 rows) • S.S. ILE DE FRANCE (1 rows)	□	S.S. "ILE DE FRANCE"""
2	2	• DE L'ANGE HOTEL (1 rows) • HOTEL DE L'ANGE (1 rows)	□	DE L'ANGE HOTEL
2	78	• ST. REGIS HOTEL (46 rows) • HOTEL ST. REGIS (32 rows)	□	ST. REGIS HOTEL

Hotel Savoy
59th St. & 5th Ave.
New York, New York

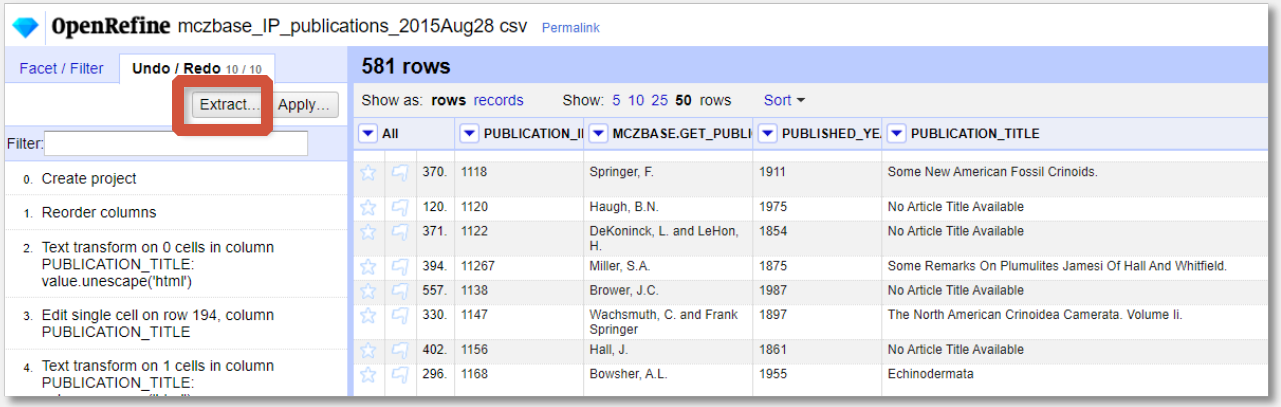


Savoy Hotel
Strand
London WC2R 0EU
United Kingdom

Exporting your data



Extracting history (JSON)



mczbase_IP_publications_2015Aug28 csv Permalink

Facet / Filter | **Undo / Redo** 10 / 10

Extract… Apply…

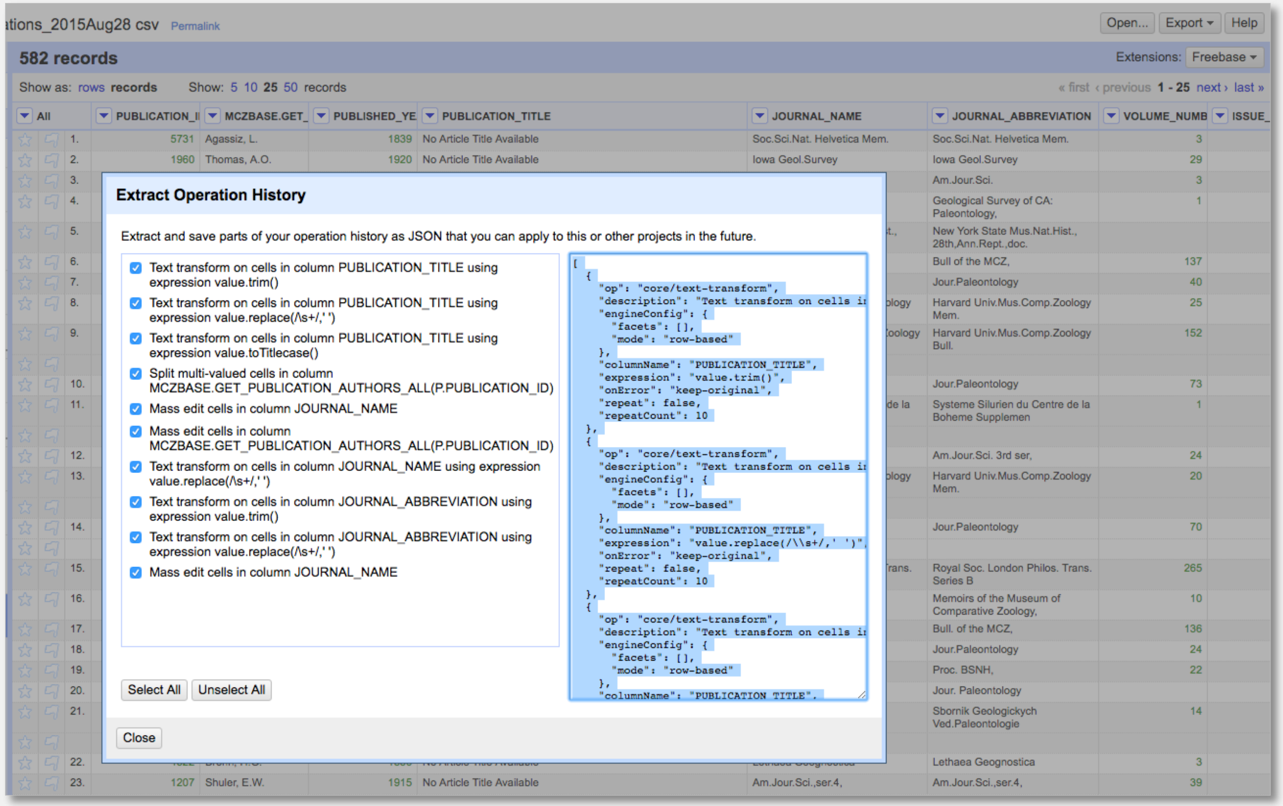
Filter:

0. Create project
1. Reorder columns
2. Text transform on 0 cells in column PUBLICATION_TITLE: value.unescape('html')
3. Edit single cell on row 194, column PUBLICATION_TITLE
4. Text transform on 1 cells in column PUBLICATION_TITLE:

581 rows

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

All	PUBLICATION_II	MCZBASE.GET_PUBLI	PUBLISHED_YE.	PUBLICATION_TITLE
370.	1118	Springer, F.	1911	Some New American Fossil Crinoids.
120.	1120	Haugh, B.N.	1975	No Article Title Available
371.	1122	DeKoninck, L. and LeHon, H.	1854	No Article Title Available
394.	11267	Miller, S.A.	1875	Some Remarks On Plumulites Jamesi Of Hall And Whitfield.
557.	1138	Brower, J.C.	1987	No Article Title Available
330.	1147	Wachsmuth, C. and Frank Springer	1897	The North American Crinoidea Camerata. Volume Ii.
402.	1156	Hall, J.	1861	No Article Title Available
296.	1168	Bowsher, A.L.	1955	Echinodermata



ations_2015Aug28 csv Permalink

582 records

Show as: rows **records**   Show: 5 10 **25** 50 records   « first ‹ previous **1 - 25** next › last »

All	PUBLICATION_I	MCZBASE.GET_	PUBLISHED_YE	PUBLICATION_TITLE	JOURNAL_NAME	JOURNAL_ABBREVIATION	VOLUME_NUMB	ISSUE_
1.	5731	Agassiz, L.	1839	No Article Title Available	Soc.Sci.Nat. Helvetica Mem.	Soc.Sci.Nat. Helvetica Mem.	3	
2.	1960	Thomas, A.O.	1920	No Article Title Available	Iowa Geol.Survey	Iowa Geol.Survey	29	
3.						Am.Jour.Sci.	3	
4.						Geological Survey of CA: Paleontology,	1	
5.					t.,	New York State Mus.Nat.Hist., 28th,Ann.Rept.,doc.		
6.						Bull of the MCZ,	137	
7.						Jour.Paleontology	40	
8.					ology	Harvard Univ.Mus.Comp.Zoology Mem.	25	
9.					oology	Harvard Univ.Mus.Comp.Zoology Bull.	152	
10.						Jour.Paleontology	73	
11.					de la	Systeme Silurien du Centre de la Boheme Supplemen	1	
12.						Am.Jour.Sci. 3rd ser,	24	
13.					ology	Harvard Univ.Mus.Comp.Zoology Mem.	20	
14.						Jour.Paleontology	70	
15.					rans.	Royal Soc. London Philos. Trans. Series B	265	
16.						Memoirs of the Museum of Comparative Zoology,	10	
17.						Bull. of the MCZ,	136	
18.						Jour.Paleontology	24	
19.						Proc. BSNH,	22	
20.						Jour. Paleontology		
21.						Sbornik Geologickych Ved.Paleontologie	14	
22.						Lethaea Geognostica	3	
23.	1207	Shuler, E.W.	1915	No Article Title Available	Am.Jour.Sci.,ser.4,	Am.Jour.Sci.,ser.4,	39	

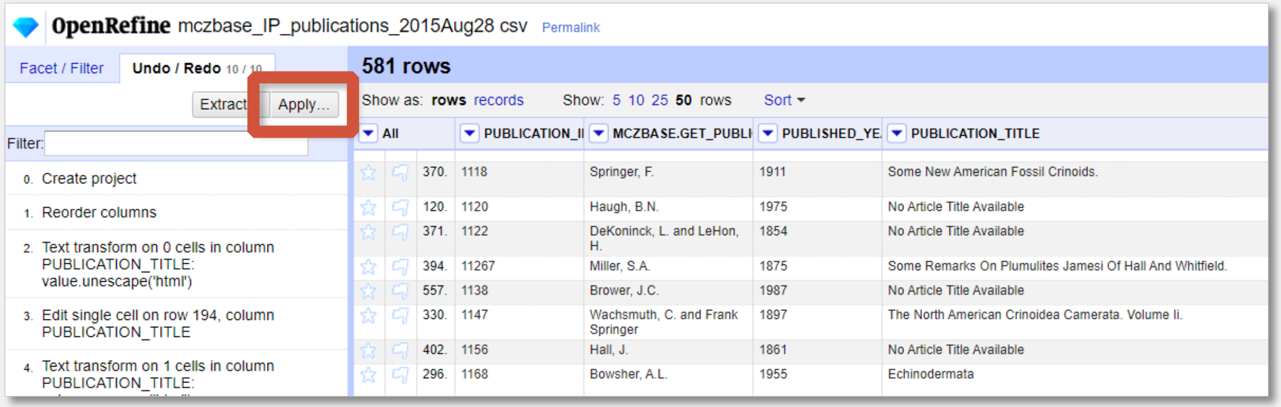
Extract Operation History

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

- ☑ Text transform on cells in column PUBLICATION_TITLE using expression value.trim()
- ☑ Text transform on cells in column PUBLICATION_TITLE using expression value.replace(/\s+/,' ')
- ☑ Text transform on cells in column PUBLICATION_TITLE using expression value.toTitlecase()
- ☑ Split multi-valued cells in column MCZBASE.GET_PUBLICATION_AUTHORS_ALL(P.PUBLICATION_ID)
- ☑ Mass edit cells in column JOURNAL_NAME
- ☑ Mass edit cells in column MCZBASE.GET_PUBLICATION_AUTHORS_ALL(P.PUBLICATION_ID)
- ☑ Text transform on cells in column JOURNAL_NAME using expression value.replace(/\s+/,' ')
- ☑ Text transform on cells in column JOURNAL_ABBREVIATION using expression value.trim()
- ☑ Text transform on cells in column JOURNAL_ABBREVIATION using expression value.replace(/\s+/,' ')
- ☑ Mass edit cells in column JOURNAL_NAME

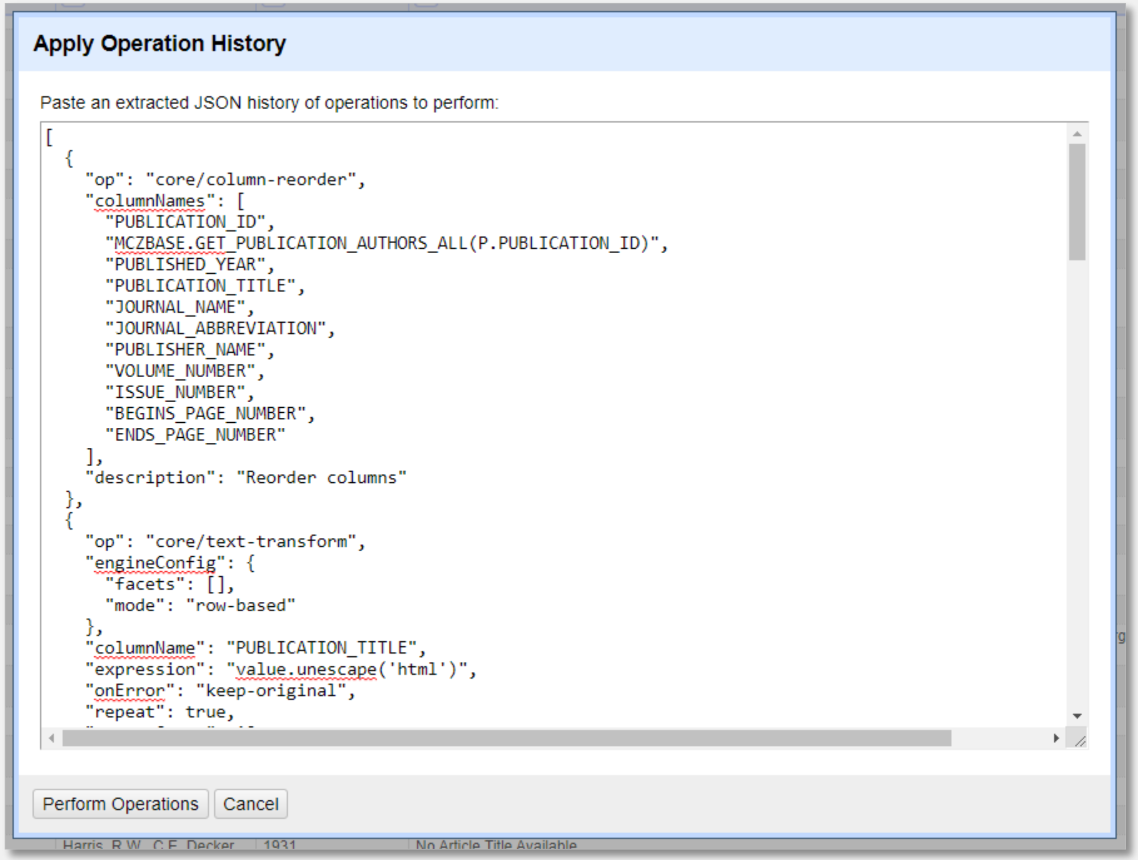
```
[
  {
    "op": "core/text-transform",
    "description": "Text transform on cells i
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "PUBLICATION_TITLE",
    "expression": "value.trim()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells i
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "PUBLICATION_TITLE",
    "expression": "value.replace(/\\s+/,' ')",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells i
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "PUBLICATION_TITLE",
```

Applying previous history (JSON) to new project



0. Create project
1. Reorder columns
2. Text transform on 0 cells in column PUBLICATION_TITLE: value.unescape('html')
3. Edit single cell on row 194, column PUBLICATION_TITLE
4. Text transform on 1 cells in column PUBLICATION_TITLE:

All	PUBLICATION_II	MCZBASE.GET_PUBLI	PUBLISHED_YE.	PUBLICATION_TITLE
370.	1118	Springer, F.	1911	Some New American Fossil Crinoids.
120.	1120	Haugh, B.N.	1975	No Article Title Available
371.	1122	DeKoninck, L. and LeHon, H.	1854	No Article Title Available
394.	11267	Miller, S.A.	1875	Some Remarks On Plumulites Jamesi Of Hall And Whitfield.
557.	1138	Brower, J.C.	1987	No Article Title Available
330.	1147	Wachsmuth, C. and Frank Springer	1897	The North American Crinoidea Camerata. Volume Ii.
402.	1156	Hall, J.	1861	No Article Title Available
296.	1168	Bowsher, A.L.	1955	Echinodermata



Apply Operation History

Paste an extracted JSON history of operations to perform:

```
[
  {
    "op": "core/column-reorder",
    "columnNames": [
      "PUBLICATION_ID",
      "MCZBASE.GET_PUBLICATION_AUTHORS_ALL(P.PUBLICATION_ID)",
      "PUBLISHED_YEAR",
      "PUBLICATION_TITLE",
      "JOURNAL_NAME",
      "JOURNAL_ABBREVIATION",
      "PUBLISHER_NAME",
      "VOLUME_NUMBER",
      "ISSUE_NUMBER",
      "BEGINS_PAGE_NUMBER",
      "ENDS_PAGE_NUMBER"
    ],
    "description": "Reorder columns"
  },
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "PUBLICATION_TITLE",
    "expression": "value.unescape('html')",
    "onError": "keep-original",
    "repeat": true,
```

Resources guides.library.Illinois.edu/openrefine

- *OpenRefine page*: http://openrefine.org/
- *Github repository*: https://github.com/OpenRefine/OpenRefine
- *Open Refine Documentation:* https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Free Online Course: https://bigdatauniversity.com/courses/introduction-to-openrefine/
- *Additional Reconciliation Services:* https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources
- *More on Google Regular Expression Language:* https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions
- Understanding Expressions: https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions
- GREL Cheat Sheet: http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf

Slides created and/or updated by: Qian Zhang, Kelly Applegate, Brinna Michael, and Tricia Lampron
Data set provided by Qian Zhang

QUESTIONS?

Tricia Lampron

[lampron2@Illinois.edu]