Cleaning Patent Data with Open Refine (Google Refine)

Paul Oldham

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By **STEVE LOHR** AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

The field known as "big data" offers a contemporary case study. The catchphrase stands for the modern abundance of digital data from many sources — the web, sensors, smartphones and corporate databases — that can be mined with clever software for discoveries and insights. Its promise is smarter, data-driven decision-making in every field. That is why data scientist is the economy's hot new job.

Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

"Data wrangling is a huge — and surprisingly so — part of the job," said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. "It's something that is not appreciated by data civilians. At times, it feels like everything we do."

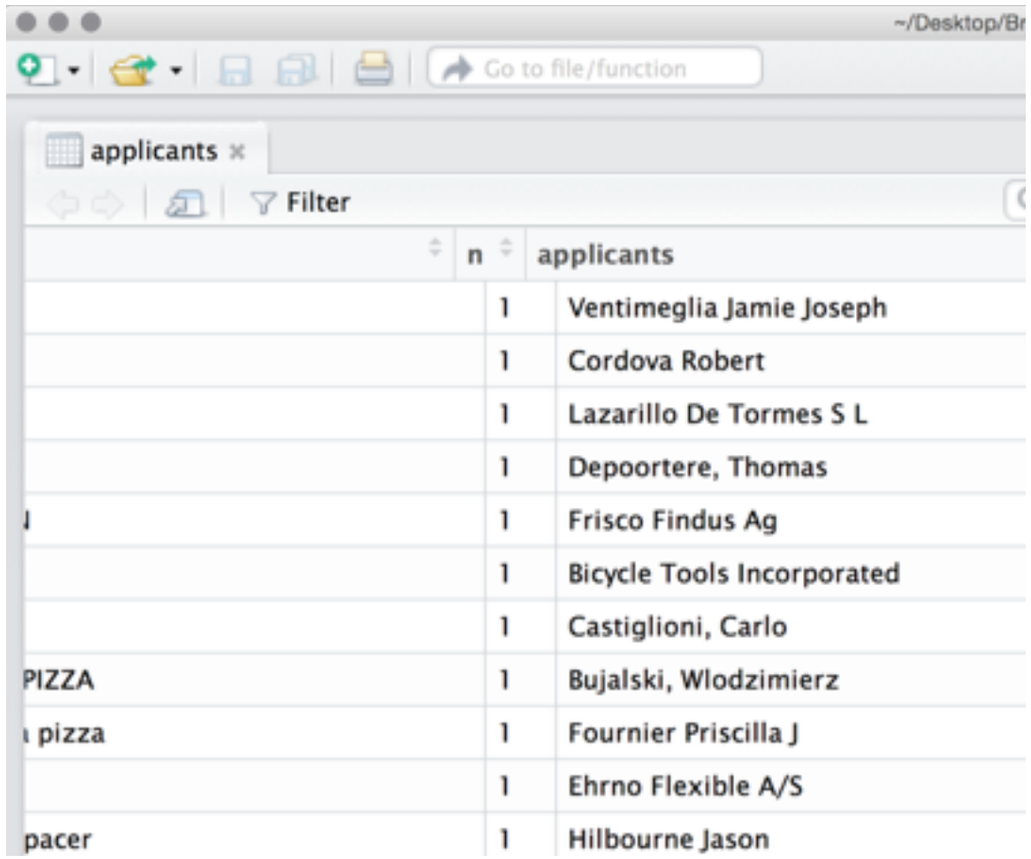
Two Main Tasks

1. Data Cleaning. This mainly involves cleaning up patent inventor and applicant names and cleaning up data fields.

2. Tidying Data. Transforming patent data into a form that can be used for statistics, mapping and visualisation. This normally involves splitting data into new columns & tables.

Cleaning problems

Inventor(s)	Applicant(s)
FAHEY-BURKE SAMUEL Jâ€,[US];	ELWHA LLCâ€,[US]
FAHEY-BURKE SAMUEL Jâ€,[US];	ELWHA LLCâ€,[US]; ELWHA LLC
ARICO MARIA BEATRICEâ€,[IT];	NOVARTIS VACCINES & DIAGNOSTICâ€,[IT]
RAPPUOLI RINOâ€,[IT]; PIZZA M	NOVARTIS VACCINES & DIAGNOSTICâ€,[IT]
OSTER PHILIPP; RAPPUOLI RINO	NOVARTIS VACCINES & DIAGNOSTICâ€,[IT]
	PAN PACIFIC PLASTICS MFG INCâ€,[US]
FINDLAY MICHAEL Câ€,[US]	FINDLAY MICHAEL Câ€,[US]
CONTARINO JR ALFRED Fâ€,[US]	CONTARINO JR ALFRED Fâ€,[US]
MCGUYER CARTER Wâ€,[US]; R	ROBBINS IND INCâ€,[US]



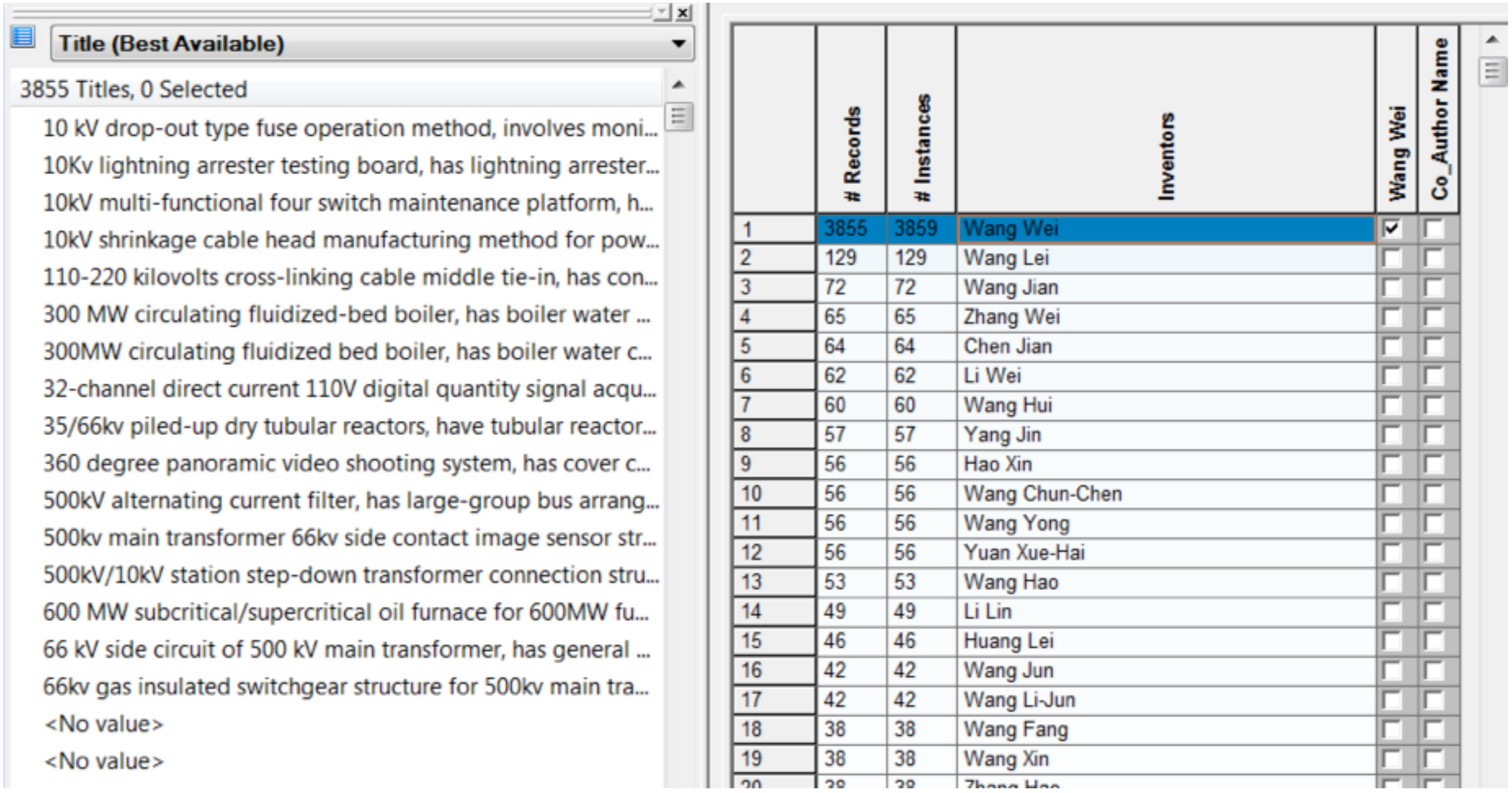
Splitting Applicant Names in R

What's in a name? Lumps and Splits in Patent Data

- Cleaning Applicant and Inventor Names is one of the biggest and hardest challenges in patent analytics.
- There are two big problems:
- 1. Splitting of names (synonyms)

  (<Kirk, James, T.>, <Kirk, T., James>, <James, T, Kirk>, <James, Tiberius, Kird) including variations in punctuation.
- 2. Lumping of names (homonyms), for example, <Garcia, Carlos> or <Silva, Jose> or <Smith, John> or <Wang Wei>.
- The most important of these problems is not splitting or synonyms. It is lumping.
- See Fegley & Torvik 2013 Has Large-Scale Named-Entity Network Analysis Been Resting on a Flawed Assumption. PLOS ONE.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Wang Wei - Super Inventor?

Appears in the scientific literature and patent literature for synthetic biology. His name also co-occurs with other inventors. But he is actually multiple persons.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Basic Cleaning Principles

1. **Rubbish in = Rubbish out**. Cleaning cannot be avoided if you want to produce meaningful results.
2. **Accuracy is best achieved by using other fields as match criteria**. Use whatever you have available. The INPADOC First Family Member (or priority number) is preferred (but use what you have in your dataset).
3. **Cleaning involves multiple steps and is an iterative process requiring patience**. Use different match criteria along the way.
4. **Document the steps you take and their limitations**. Place this in a “code book” at the front of an Excel workbook or in a text file.

Open Refine

- An open source project originally known as Google Refine
- Open Refine is a programme that runs on your computer but uses the browser as your interface.
- It does not need an internet connection.
- It includes a scripting language called General Refine Expression Language (GREL) to write cleaning scripts.

http://openrefine.org/

Download OpenRefine

Home

Download

Documentation

Community

Post archive

Mapping OpenRefine Ecosystem

2014 survey results

A Governance Model for OpenRefine

Using OpenRefine: a

OpenRefine Core

Google Refine 2.5 - Stable version

- **Windows kit**, Download, unzip, and double-click on *google-refine.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. **NOTE:** If you have issues installing Refine on Mac, please refer to issue 590 - OpenRefine 2.5 for mac support java 6 and 7 only
- **Linux kit**, Download, extract, then type *./refine* to start. **NOTE:** OpenRefine 2.5 for linux support java 6 and 7 only

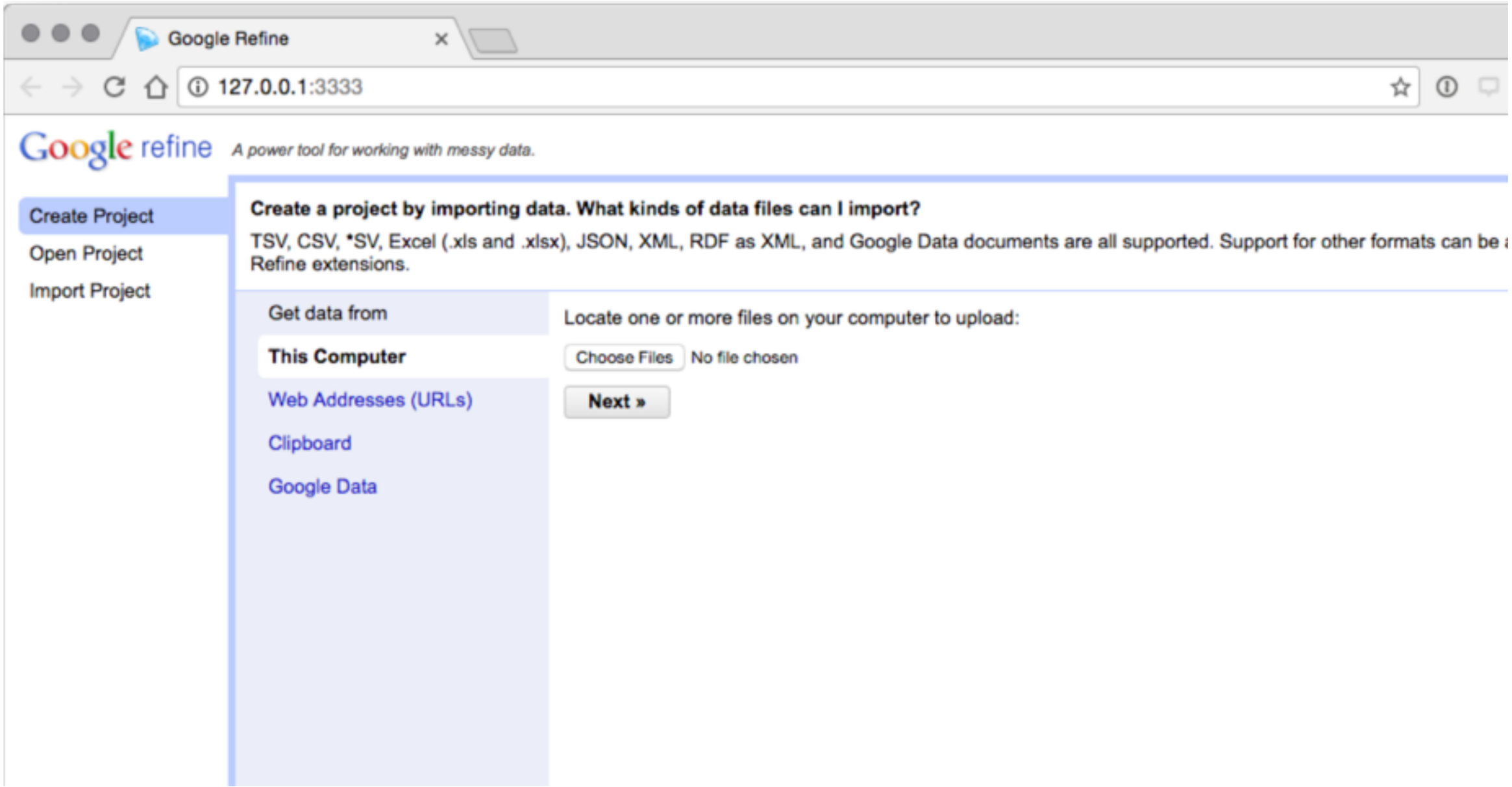
See also the installation instruction

OpenRefine 2.6 - Development version

Download OpenRefine 2.6-beta1 (link for all releases).

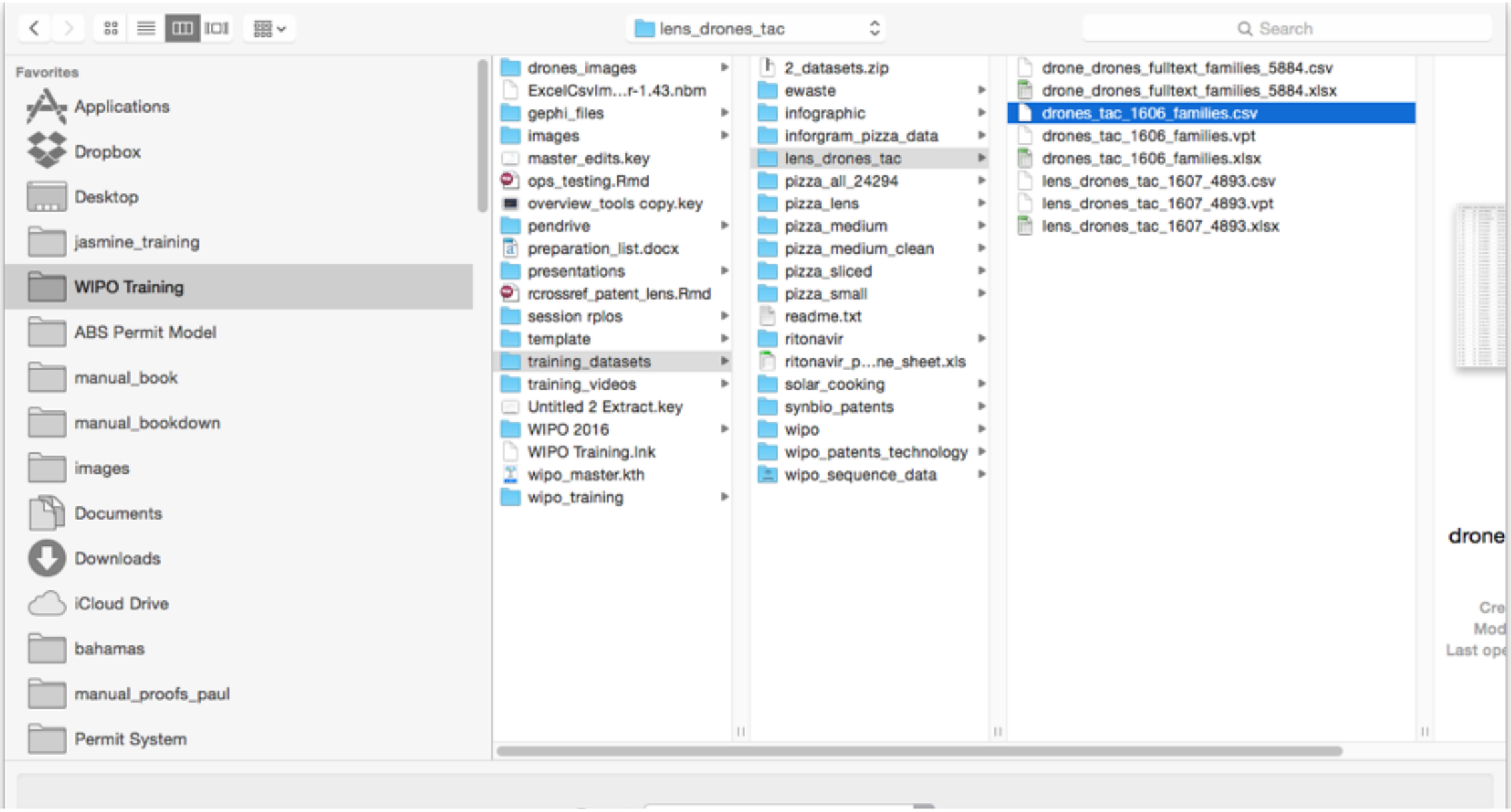
All releases

All previous releases are available here.



Choose Files

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Open Open Refine: Choose drones_tac_1606_families dataset

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Google Refine

127.0.0.1:3333

Google refine *A power tool for working with messy data.*

Create Project
Open Project
Import Project

« Start Over | Configure Parsing Options | Project name drones_tac_1606_f

	#	Jurisdiction	Kind	Publication Number	Lens ID	Publication Date	Publication Year	Application Number	Applicat		
1.	1	AR	A1	AR 021856 A1	189-466-043-698-920	07/08/02	2002	AR P990105416 A	27/10/99		Un Espaci Volumen C Producir U Eficacia D Particular .
2.	2	AR	A1	AR 075121 A1	143-256-917-407-318	09/03/11	2011	AR P100100007 A	04/01/10	FR 0950012 A 20090105	Metodo Y Mejorados Sismica A
3.	3	AT	A	AT A12022004 A	013-820-289-119-542	15/10/05	2005	AT 12022004 A	15/07/04	AT 12022004 A 20040715	Unbemanr
4.	4	AT	A	AT A33585 A	080-780-325-	15/01/95	1995	AT 33585 A	06/02/85	AT 33585 A 19850206	Doppelsto ZweivÃ¶lk GÃ nzlich

Parse data as

CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
RDF/N3 files
XML files
Open Document Format spreadsheets (.ods)
RDF/XML files

Character encoding

Update Preview

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

Escape special characters with \

- [ ] 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] Parse cell text into numbers, dates, ...
- [x] Quotation marks are used to enclose cells containing column separators
- [x] Store blank rows
- [x] Store blank cells as nulls
- [ ] Store file source (file names, URLs) in each row

Version 2.5 [r2407]

Help
About

Choose commas (the default) as separator

Basic Cleaning Steps

- Regularising the Case (titles, publication numbers etc.)
- Remove leading and trailing whitespace (repeat later)
- Add columns based on other columns (e.g. publication country)
- Deal with character encoding and related problems
- Reformat dates
- Fill blank cells with NA
- **The objective is to create a new clean core dataset**. Then you can split the data into new tables on applicants, inventors, IPCs to clean up and visualise.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

drones_tac_1606_families csv

127.0.0.1:3333/project?project=1971363208037

Google refine drones_tac_1606_families csv Permalink

Open... Export Help

Facet / Filter | Undo / Redo 0

Refresh | Reset All | Remove All

Title change

1534 choices Sort by: name count Cluster

"huangpeijing" And Its Processing Method 1

14alpha,17alpha-c2-Ä_berbrÄ_ckte 19-nor-Progesteronderivate 1

2-substituted-4-1-thiomorpholino-iminoalkyl- Thiazoles And Derivatives Thereof 1

1606 rows

Extensions: Freebase

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

« first ‹ previous 1 - 10 next › last »

Publication Date	Publication Year	Application Num	Application Date	Priority Number	Title	Applicants	Inventors	URL	Type	Full Text
/08/02	2002	AR P990105416 A	27/10/99	IT RE980117 A 19981123				https://lens.org/189-466-043-696-920	unknown	no
9/03/11	2011	AR P100100007 A	04/01/10	FR 0950012 A 20090105				https://lens.org/143-256-917-407-318	unknown	no
5/10/05	2005	AT 12022004 A	15/07/04	AT 12022004 A 20040715	Unbemannter Hubschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEORG	https://lens.org/013-820-289-119-542	unknown	no
5/01/95	1995	AT 33585 A	06/02/85	AT 33585 A 19850206	Doppelstocktrogbeute FÃ_r Die ZweivÃ¶lkerbetriebsweise Zur GÃ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no
1/09/72	1972	AT 91271 A	03/02/71	DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072-084-291-460-544	unknown	no
5/01/74	1974	AT 1005171 A	22/11/71	DE 2064859 A 19701230	Verfahren Zur Herstellung Von Neuen 6Î±-fluor-16Î±,18-dimethyl-1,4-pregnadien-3,20-dionderivaten	SCHERING AG		https://lens.org/034-658-237-174-32X	unknown	no
5/01/97	1997	AT 94100711 T	19/01/94	DE 4301763 A 19930123	Wagenkastenaufbau, Insbesondere FÃ_r Eisenbahnwagen Zur PersonenbefÃ¶rderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN;;ELSNER OLAF	https://lens.org/180-373-684-761-176	unknown	no
5/01/00	2000	AT 97902456 T	03/02/97	US 59560596 A 19960202	Stabile Herbizide Zusammensetzungen, Die Metalchelate Von Herbiziden Dionen Enthalten	ZENECA LTD	SCHER HERBERT BENSON;;CHEN JINLING	https://lens.org/102-558-221-531-261	unknown	no
5/11/01	2001	AT 96401865 T	30/08/96	FR 9510385 A 19950905	Fernbedienungssystem FÃ_r Eine Mit Einem	AEROSPATIALE MATRA	BARNIER JEAN-FRANCOIS	https://lens.org/179-736-376-179-488	unknown	no

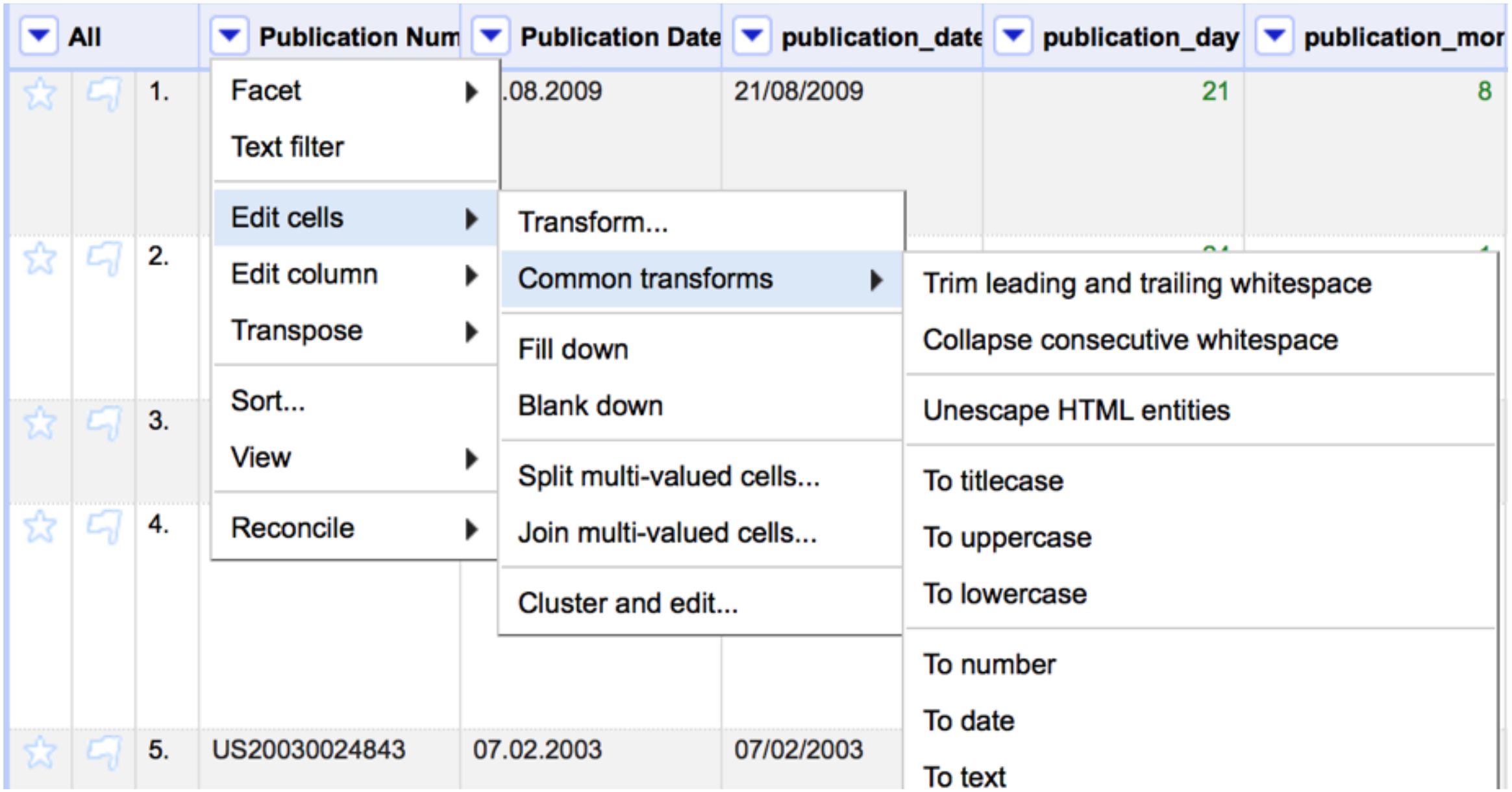
Facet ▸ | Text filter | Edit cells ▸ | Edit column ▸ | Transpose ▸ | Sort... | View ▸ | Reconcile ▸

Text facet | Numeric facet | Timeline facet | Scatterplot facet | Custom text facet... | Custom numeric facet... | Customized facets ▸

Title - Bring up Text Facet - Inspect for problems

A facet basically brings up a window with the column that we can then act on to clean it up (remove <i>). Usually this involves clustering. Choose Title Case from the edit menu.

WIPO WORLD INTELLECTUAL PROPERTY ORGANIZATION



Open Refine works on columns

Most functions are found in the column pull down menu. Choose the Publication Number - then to Uppercase (line 693 and others are lowercase)

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

drones_tac_1606_families csv

127.0.0.1:3333/project?project=1971363208037

Google refine drones_tac_1606_families csv Permalink

Open... Export Help

Facet / Filter | Undo / Redo 0

Refresh | Reset All | Remove All

Title change

1534 choices Sort by: name count Cluster

"huangpeijing" And Its Processing Method 1

14alpha,17alpha-c2-Ä_berbrÄ_ckte 19-nor-Progesteronderivate 1

2-substituted-4-1-thiomorpholino-iminoalkyl- Thiazoles And Derivatives Thereof 1

1606 rows

Extensions: Freebase

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

« first ‹ previous 1 - 10 next › last »

Publication Date	Publication Year	Application Num	Application Date	Priority Number	Title	Applicants	Inventors	URL	Type	Full Text
/08/02	2002	AR P990105416 A	27/10/99	IT RE980117 A 19981123				https://lens.org/189-466-043-696-920	unknown	no
9/03/11	2011	AR P100100007 A	04/01/10	FR 0950012 A 20090105				https://lens.org/143-256-917-407-318	unknown	no
5/10/05	2005	AT 12022004 A	15/07/04	AT 12022004 A 20040715	Unbemannter Hubschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEORG	https://lens.org/013-820-289-119-542	unknown	no
5/01/95	1995	AT 33585 A	06/02/85	AT 33585 A 19850206	Doppelstocktrogbeute FÃ_r Die ZweivÃ¶lkerbetriebsweise Zur GÃ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no
1/09/72	1972	AT 91271 A	03/02/71	DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072-084-291-460-544	unknown	no
5/01/74	1974	AT 1005171 A	22/11/71	DE 2064859 A 19701230	Verfahren Zur Herstellung Von Neuen 6Î±-fluor-16Î ±,18-dimethyl-1,4-pregnadien-3,20-dionderivaten	SCHERING AG		https://lens.org/034-658-237-174-32X	unknown	no
5/01/97	1997	AT 94100711 T	19/01/94	DE 4301763 A 19930123	Wagenkastenaufbau, Insbesondere FÃ_r Eisenbahnwagen Zur PersonenbefÃ¶rderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN;;ELSNER OLAF	https://lens.org/180-373-684-761-176	unknown	no
5/01/00	2000	AT 97902456 T	03/02/97	US 59560596 A 19960202	Stabile Herbizide Zusammensetzungen, Die Metalchelate Von Herbiziden Dionen Enthalten	ZENECA LTD	SCHER HERBERT BENSON;;CHEN JINLING	https://lens.org/102-558-221-531-261	unknown	no
5/11/01	2001	AT 96401865 T	30/08/96	FR 9510385 A 19950905	Fernbedienungssystem FÃ_r Eine Mit Einem	AEROSPATIALE MATRA	BARNIER JEAN-FRANCOIS	https://lens.org/179-736-376-179-488	unknown	no

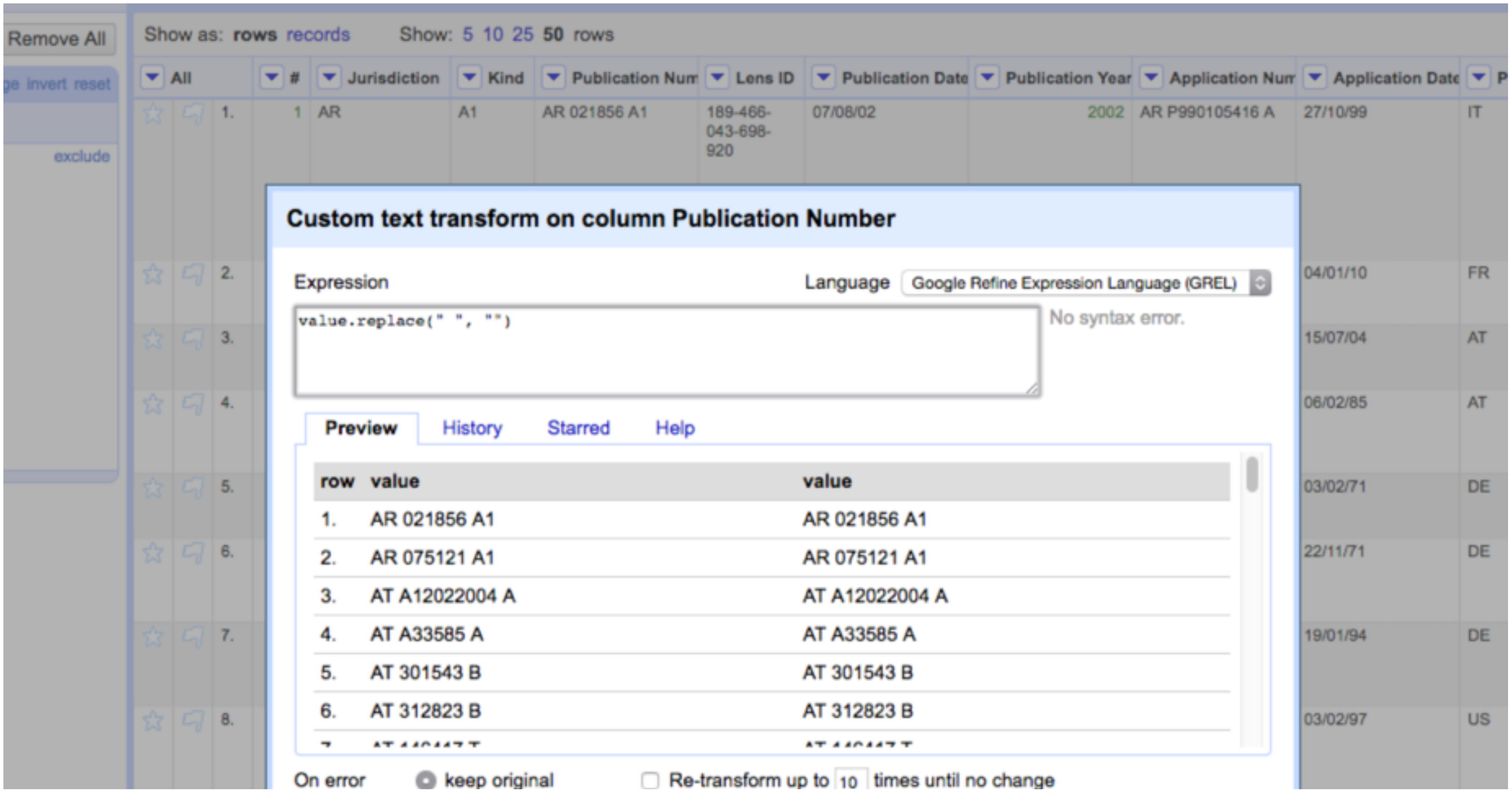
Facet ▸ | Text filter | Edit cells ▸ | Edit column ▸ | Transpose ▸ | Sort... | View ▸ | Reconcile ▸

Text facet | Numeric facet | Timeline facet | Scatterplot facet | Custom text facet... | Custom numeric facet... | Customized facets ▸

Harmonize Title to Titlecase

A facet basically brings up a window with the column that we can then act on to clean it up (remove <i>). Usually this involves clustering. Choose Title Case from the edit menu.

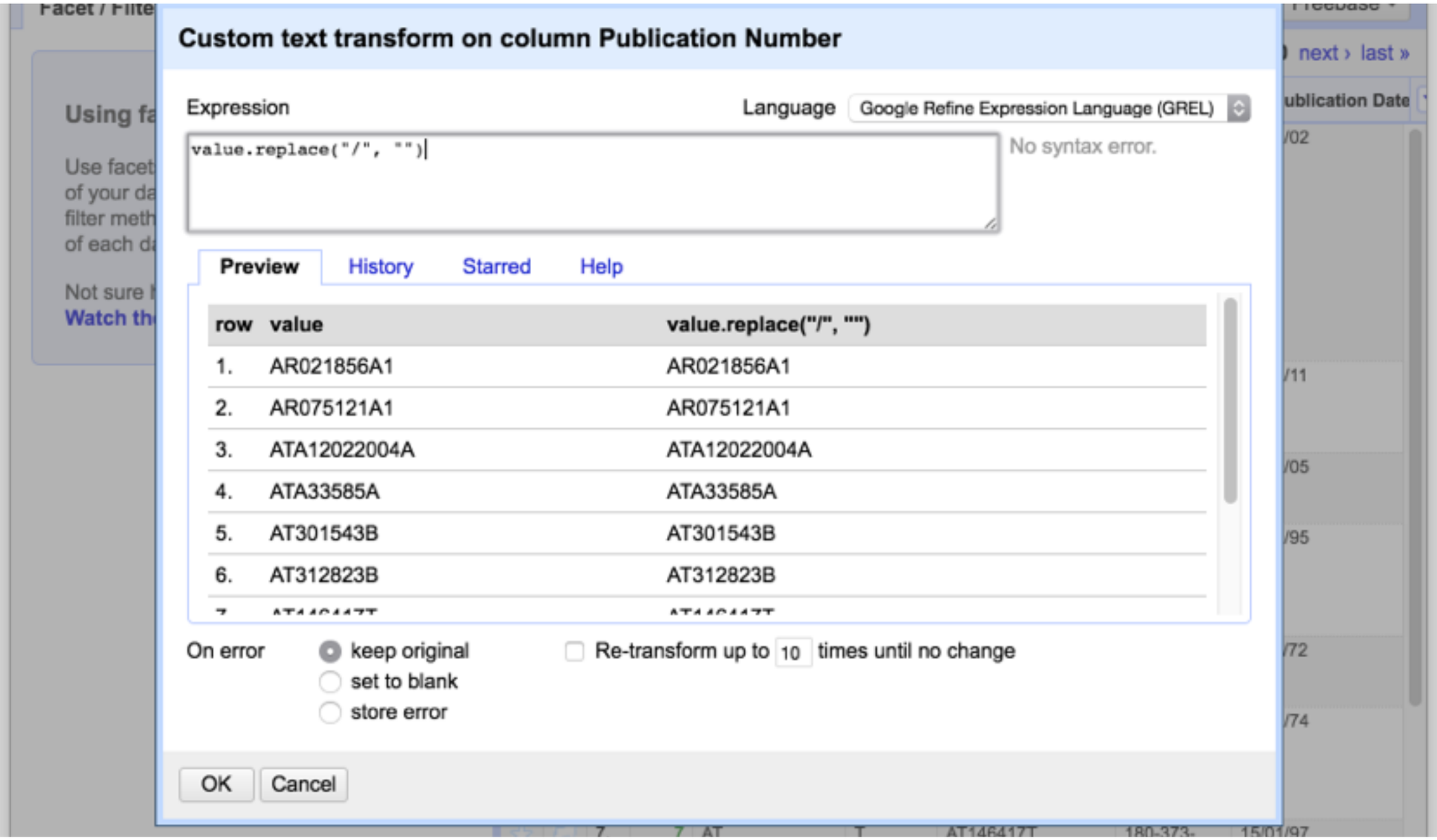
WIPO WORLD INTELLECTUAL PROPERTY ORGANIZATION



Remove spaces in publication numbers

Edit Cells > Transform . Enter **value.replace(" ", "")**. **Then OK**

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Remove Slash in publication numbers

Edit Cells > Transform . Enter **value.replace("/", ""). Then OK**
(cleans up 372 cells for using numbers in other databases)

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Google refine drones_tac_1606_families csv Permalink

Open... Export Help

Facet / Filter | Undo / Redo 1

Refresh | Reset All | Remove All

Title change

1534 choices Sort by: name count | Cluster

*huangpeijing" And Its Processing Method 1

14alpha,17alpha-c2-Ã_berbrÃ_ckte 19-nor-Progesteronderivate 1

2-substituted-4-1-thiomorpholino-iminoalkyl- Thiazoles And Derivatives Thereof 1

3 Drone Stereo Camera 3d Modeling System Using An Image 1

3d Model And Beacon For Automatic Delivery Of Goods 1

1925 rows

Extensions: Freebase

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

« first ‹ previous **1 - 10** next › last »

Priority Numbers	Title	Applicants	Inventors	URL	Type	Full Text	Cited Count	Simple Family S	Extended Family	Sequence Count	C
	bmento Pasivo Que nciona En mbinacion Con Un rtante Activo Y Alojado	ASK IND SPA		https://lens.org/189-466-043-698-920	unknown	no	0	16	16	0	HC
	ICHEL			https://lens.org/143-256-917-407-318	unknown	no	0	22	22	0	GC
AT 12022004 A 20040715	L GES	SCHIEBEL HANS-GEORG		https://lens.org/013-820-289-119-542	unknown	no	0	23	24	0	BE
AT 33585 A 19850206	Doppelstocktrogbeute FÃ_r Die ZweivÃ¶lkerbetriebsweise Zur GÃ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no	0	2	2	0	AC
DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072-084-291-460-544	unknown	no	0	14	14	0	CC
DE 2064859 A 19701230	Verfahren Zur Herstellung Von Neuen 6Î±-fluor-16Î±,18-dimethyl-1,4-pregnadien-3,20-dionderivaten	SCHERING AG		https://lens.org/034-658-237-174-32X	unknown	no	0	29	29	0	CC
DE 4301763 A 19930123	Wagenkastenaufbau, Insbesondere FÃ_r Eisenbahnwagen Zur PersonenbefÃ¶rderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN;;ELSNER OLAF	https://lens.org/180-373-684-761-176	unknown	no	0	8	8	0	BE
US 59560596 A 19960202	Stabile Herbizide Zusammensetzungen, Die Metallchelate Von Herbiziden Dionen Enthalten	ZENECA LTD	SCHER HERBERT BENSON;;CHEN JINLING	https://lens.org/102-558-221-531-261	unknown	no	0	30	32	0	CC
FR 9510385 A 19950905	Fernbedienungssystem FÃ_r Eine Mit Einem Draht Gesteuerte Rakete	AEROSPATIALE MATRA	BARNIER JEAN-FRANCOIS	https://lens.org/179-736-376-179-488	unknown	no	0	8	8	0	GC
DE 4447401 A 19941223	14alpha,17alpha-c2-Ã_berbrÃ_ckte 19-nor-Progesteronderivate	SCHERING AG	SCHOELLKOPF KLAUS;;HALFBRODT WOLFGANG;;KUHNKE JOACHIM;;SCHWEDE	https://lens.org/119-053-067-033-961	unknown	no	0	49	50	0	CC

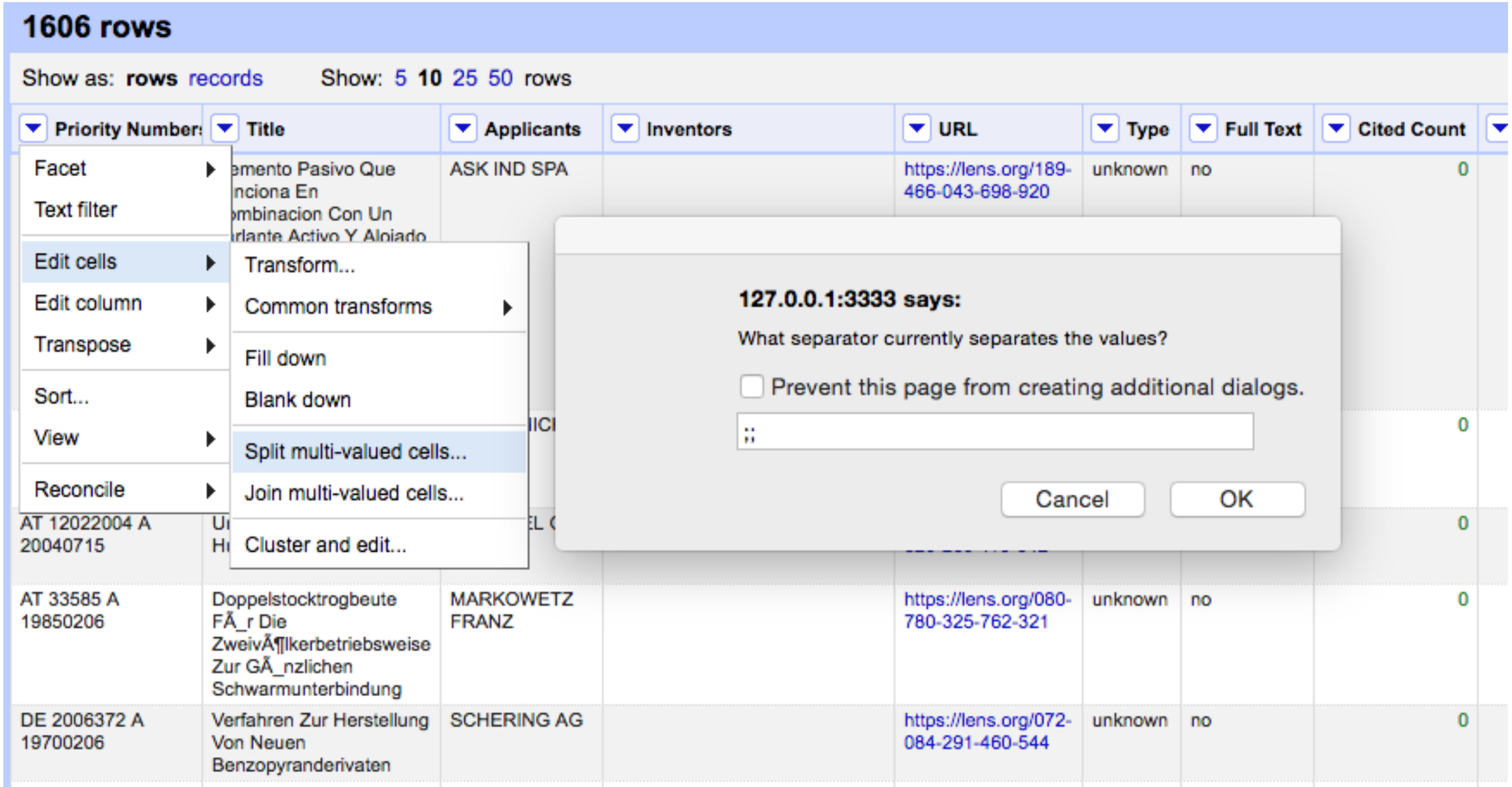
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile

Transform...
Common transforms
Fill down
Blank down
Split multi-valued cells...
Join multi-valued cells...
Cluster and edit...

Extract more information (priority numbers)

We want the priority year. This information is concatenated. So we need to **Edit Cells > Split Multivalued Cells**.

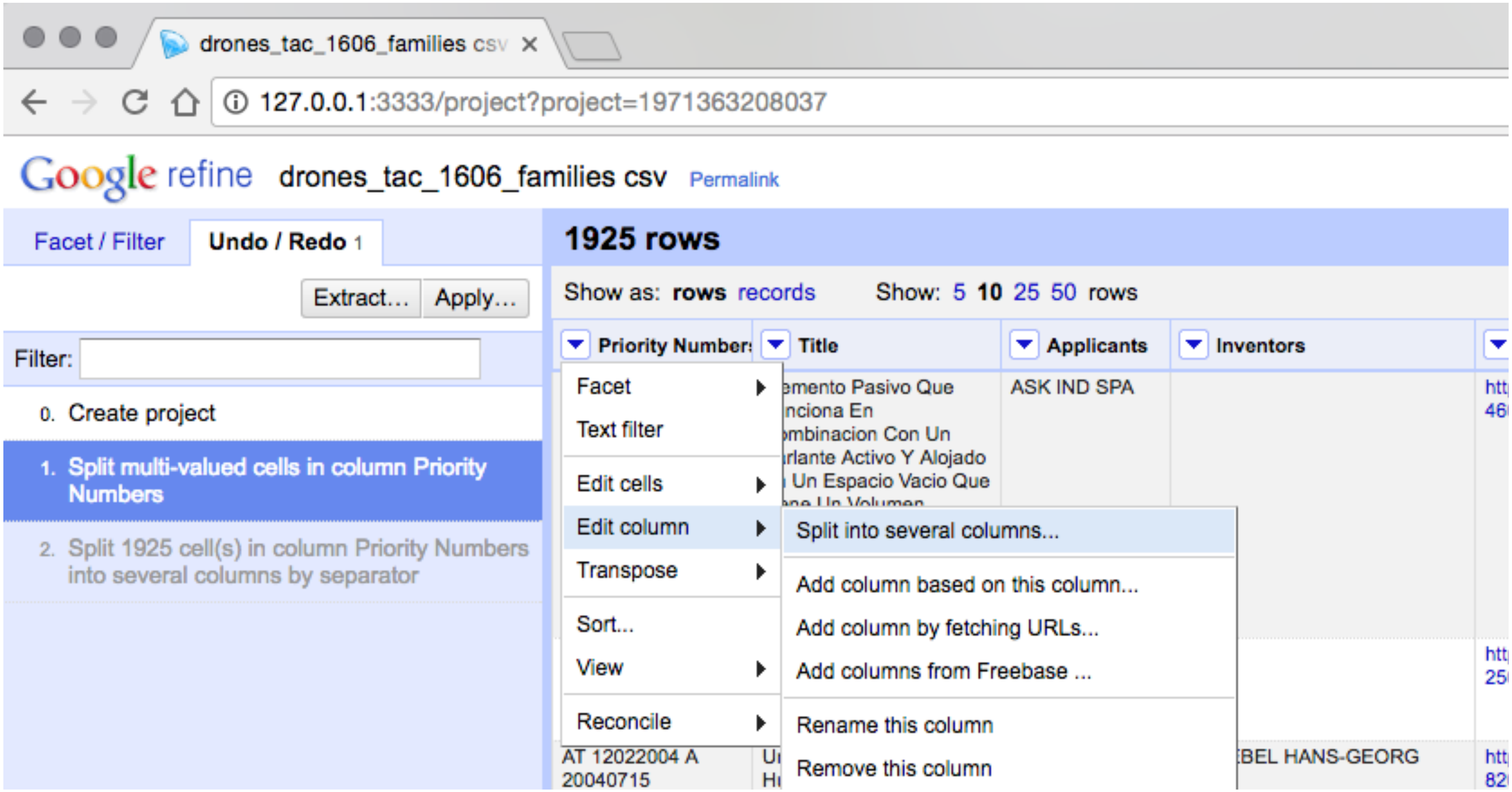
WIPO WORLD INTELLECTUAL PROPERTY ORGANIZATION



Priority Numbers: Split on Double Semi-colon

Note that the Lens is unusual in separating on double semi-colons. Normally it is a single semi-colon.

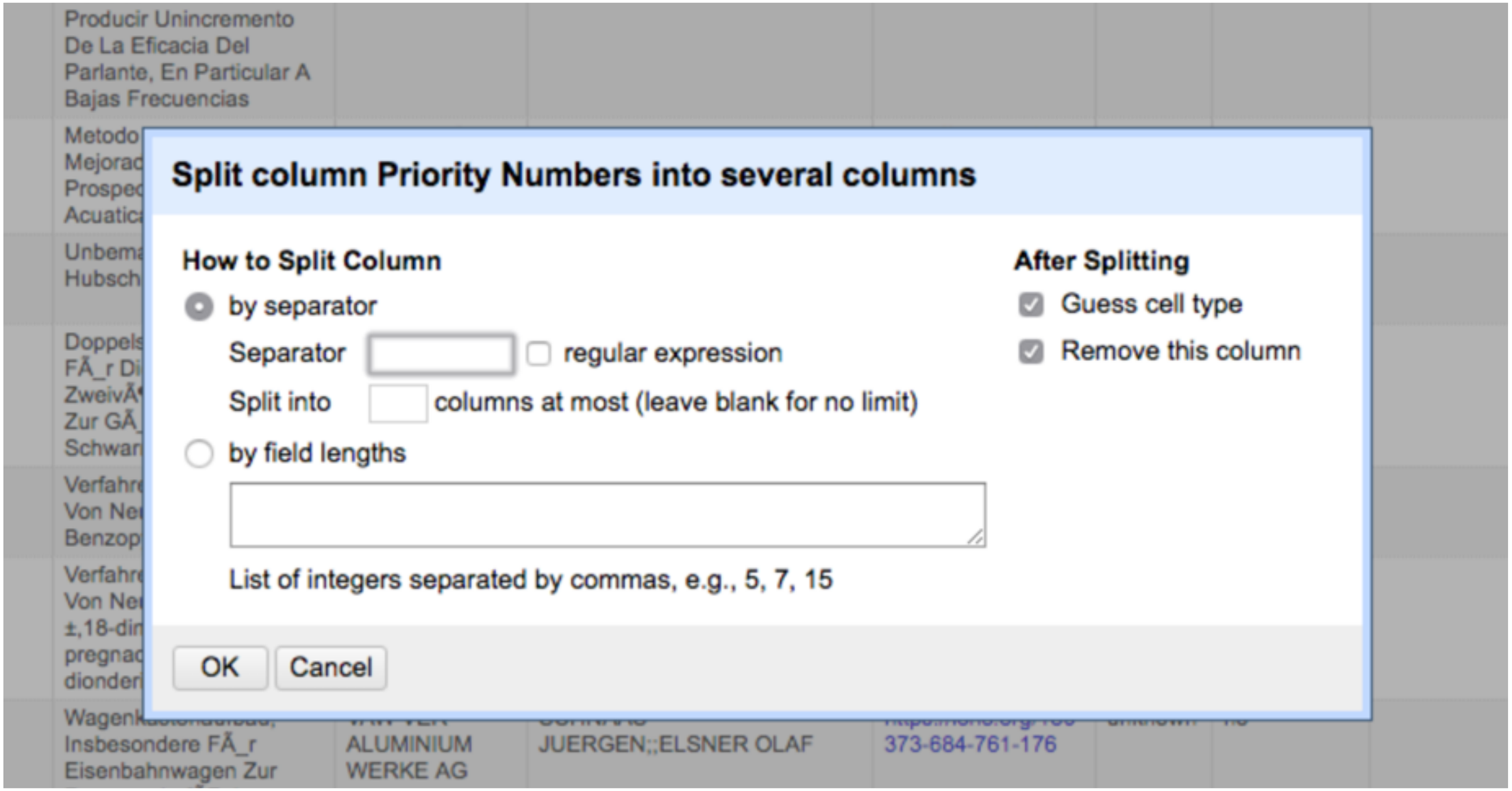
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Next Choose Edit Column - Split into several columns

Note that this will increase the number of rows because some records have more than one priority number.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Enter a single space

The default is a comma. delete the comma and press the Space bar once.
IMPORTANT: UNCHECK REMOVE THIS COLUMN to keep original priority data field.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

1925 rows

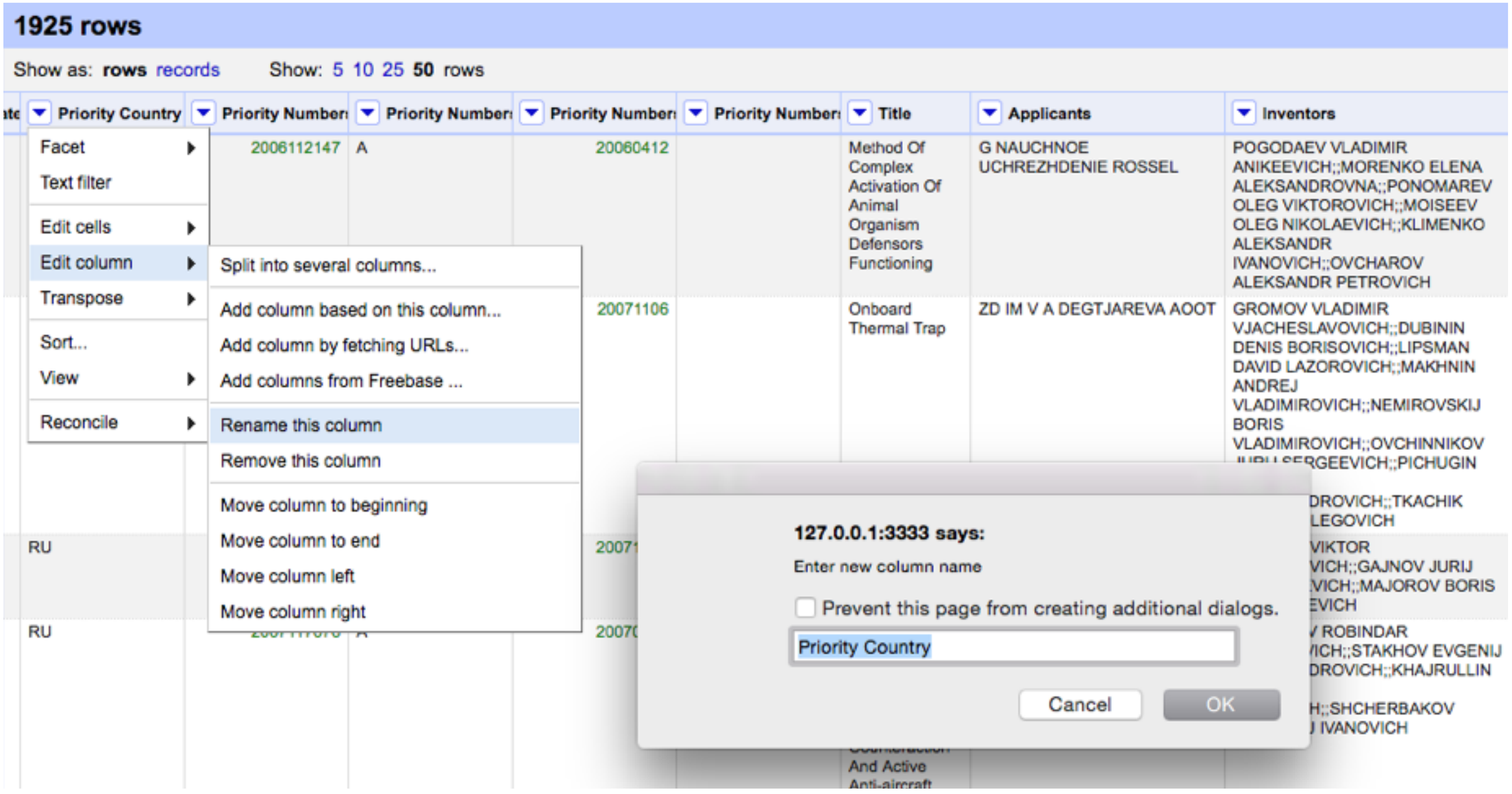
Show as: **rows** records Show: 5 **10** 25 50 rows

Priority Number:	Priority Number:	Priority Number:	Priority Number:	Priority Number:	Title
IT	RE980117	A	19981123		Elemento Pasivo Que Funciona En Combinacion Con Un Parlante Activo Y Aloj En Un Espacio Vacio Tiene Un Volumen Cerrado De Aire Para Producir Unincrement De La Eficacia Del Parlante, En Particula Bajas Frecuencias
FR	950012	A	20090105		Metodo Y Dispositivo Mejorados Para La Prospeccion Sismica Acuatica
AT	12022004	A	20040715		Unbemannter Hubschrauber
AT	33585	A	19850206		Doppelstocktrogbeute

Note we have more rows, and new columns

We had 1606 rows, we now have 1925 because some documents have more than one priority number.

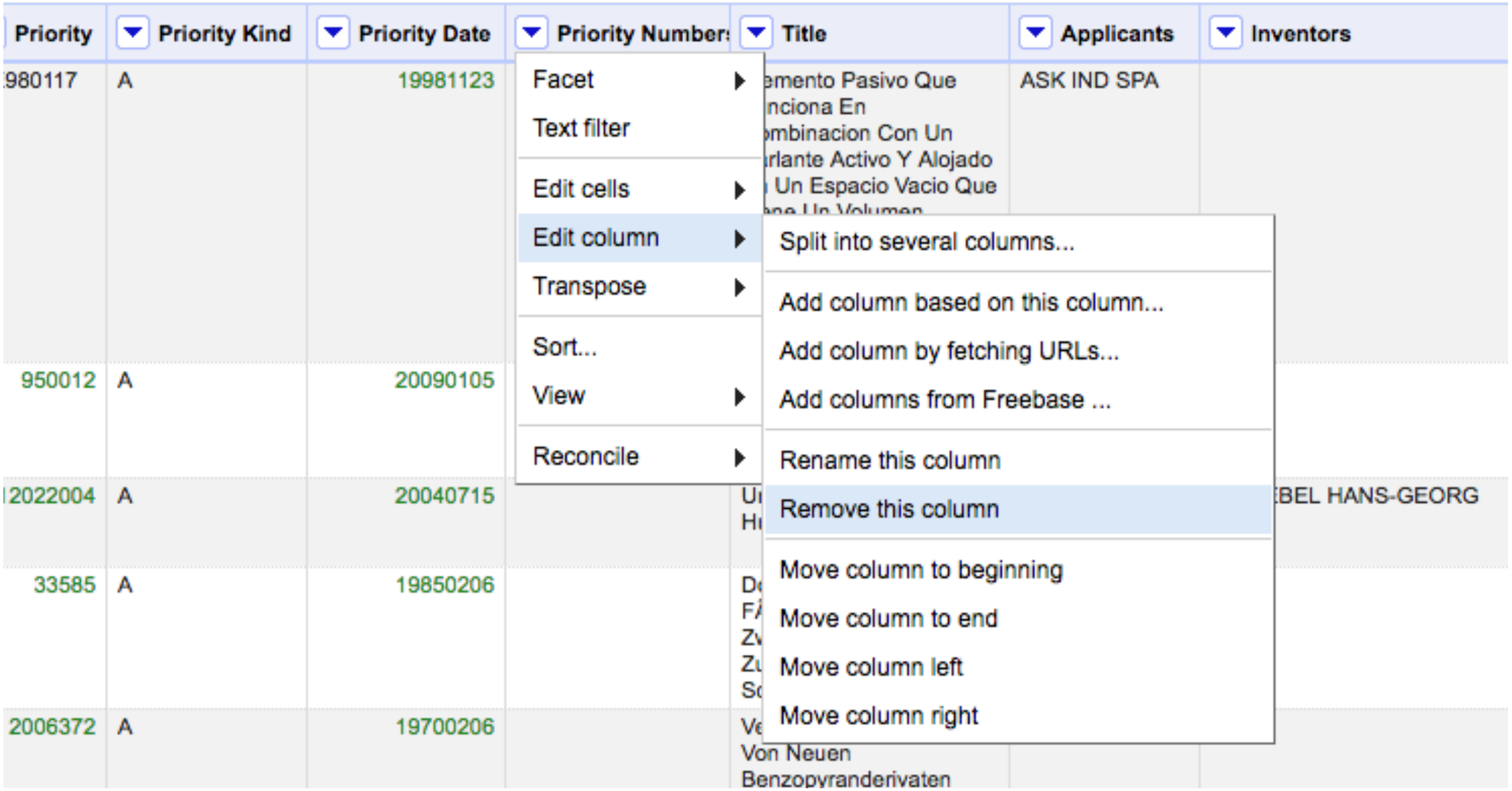
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Rename Columns

1. Priority Country, 2. Priority. 3 Priority Kind, 4. Priority Date (leave last empty column as is)

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



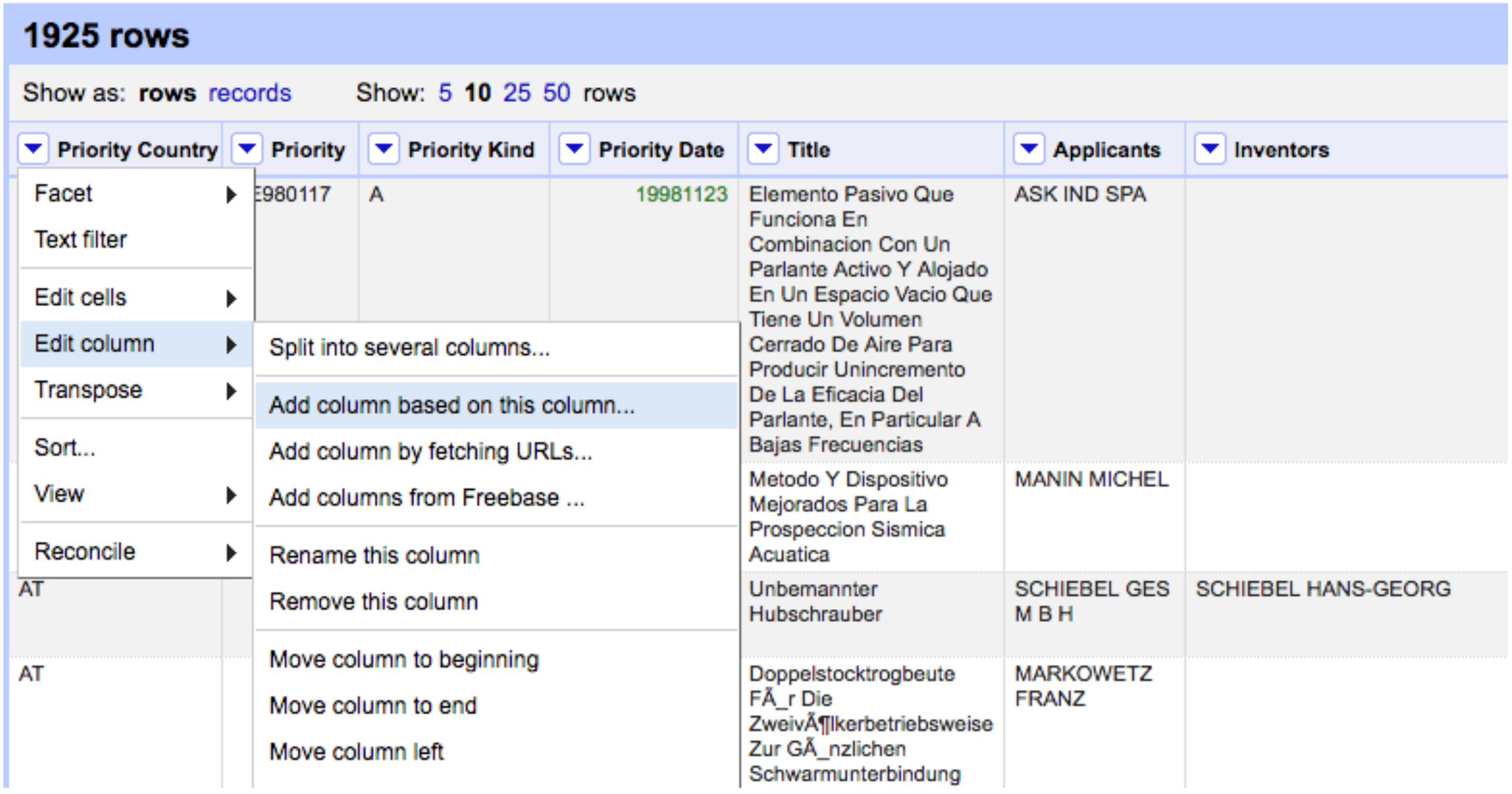
Remove Blank Column

Appears to be nothing in this column (probably created from a trailing blank space) so Edit Column > Remove this column.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

CREATE A NEW COLUMN TO JOIN PRIORITY DATA.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Add Column based on Columns

Choose Priority Country

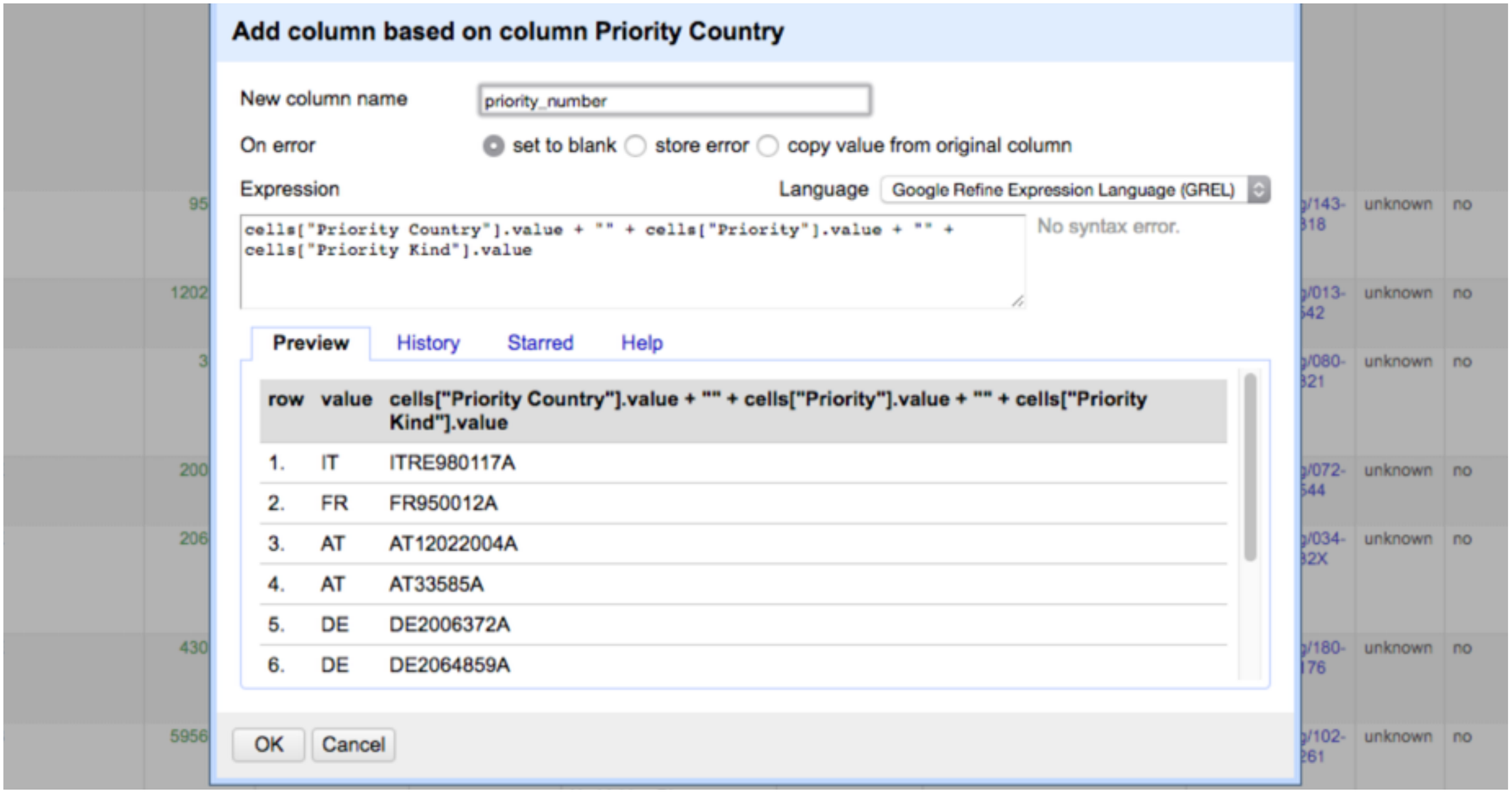
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

We now need to use some of the GREL language to combine the fields

```
cells["Priority Country"].value + "" +
cells["Priority"].value + "" + cells["Priority Kind"].value
```

See the open refine recipes: https://github.com/OpenRefine/OpenRefine/wiki/Recipes . GREL is General Regular Expression Language.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Paste the code

If nothing happens in the code try deleting Country in Priority Country and then retyping. Note the “”. This is the separator. Try putting in “-” to see the effect.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

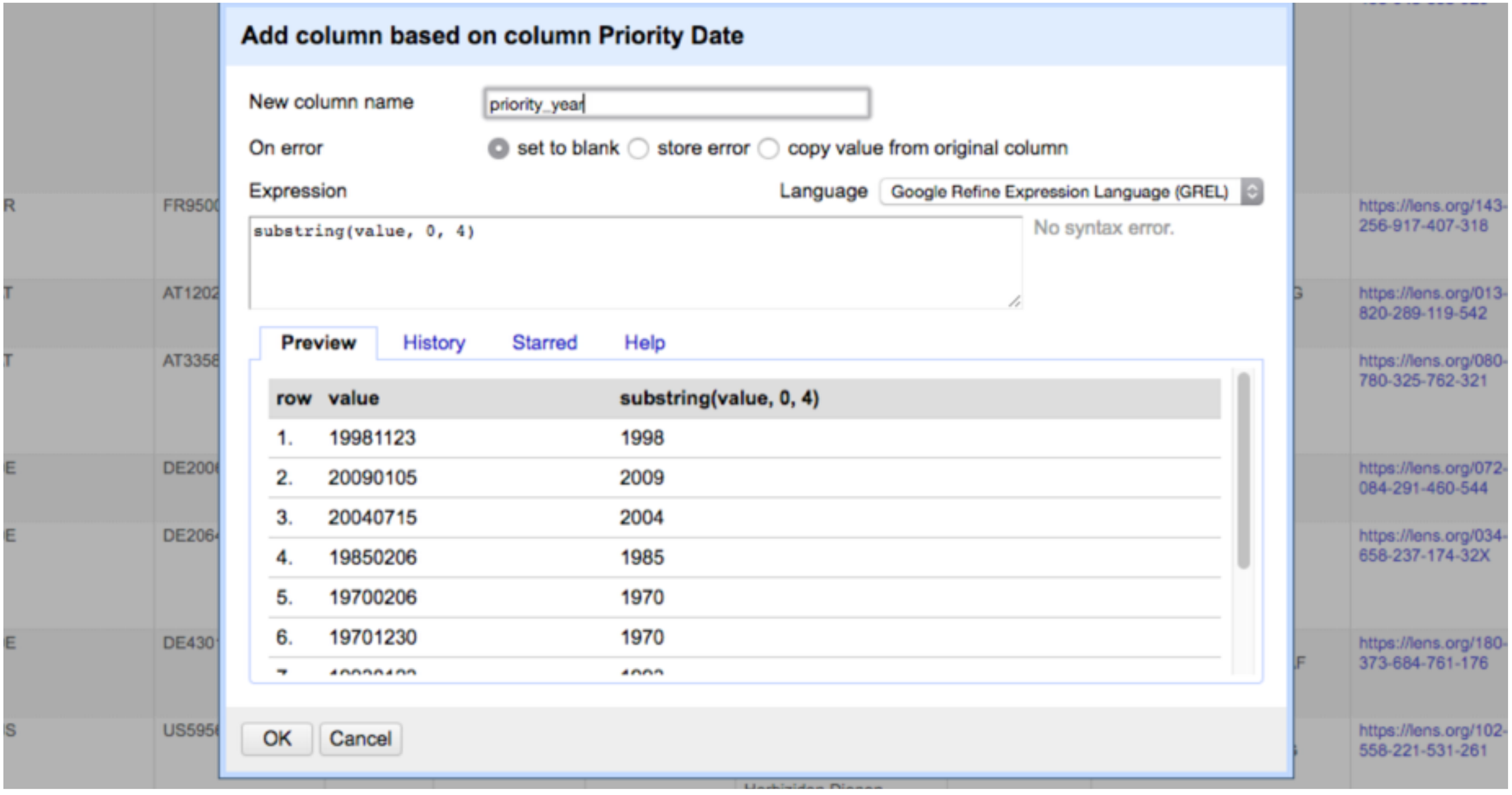
1925 rows

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

Priority Country	priority_number	Priority	Priority Kind	Priority Date	Title	Applicants	Inventors
IT	ITRE980117A	RE980117	A	19981123	Elemento Pasivo Que Funciona En Combinacion Con Un Parlante Activo Y Alojado En Un Espacio Vacio Que Tiene Un Volumen Cerrado De Aire Para Producir Unincremento De La Eficacia Del Parlante, En Particular A Bajas Frecuencias	ASK IND SPA	
FR	FR950012A	950012	A	20090105	Metodo Y Dispositivo Mejorados Para La Prospeccion Sismica Acuatica	MANIN MICHEL	
AT	AT12022004A	12022004	A	20040715	Unbemannter Hubschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEO
AT	AT33585A	33585	A	19850206	Doppelstocktrogbeute FÃ_r Die ZweivÃ¶lkerbetriebsweise Zur GÃ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ	
DE	DE2006372A	2006372	A	19700206	Verfahren Zur Herstellung Von Neuen	SCHERING AG	

The New Priority Number Column

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

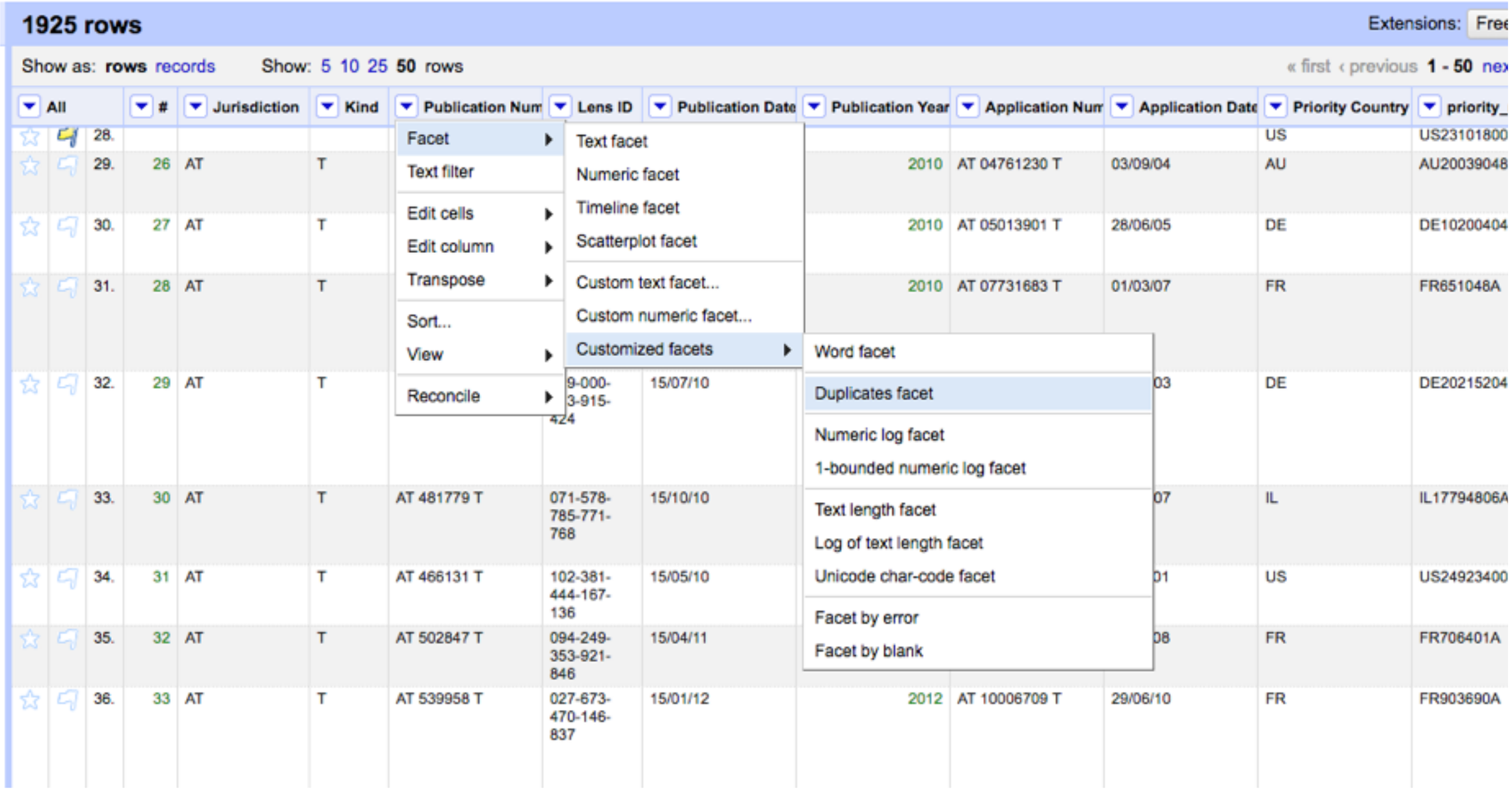


Extract the Priority Year.

Choose Priority Date > Edit Column > Add Column based on this Column

Enter **substring(value, 0, 4). This reads the first four characters.**

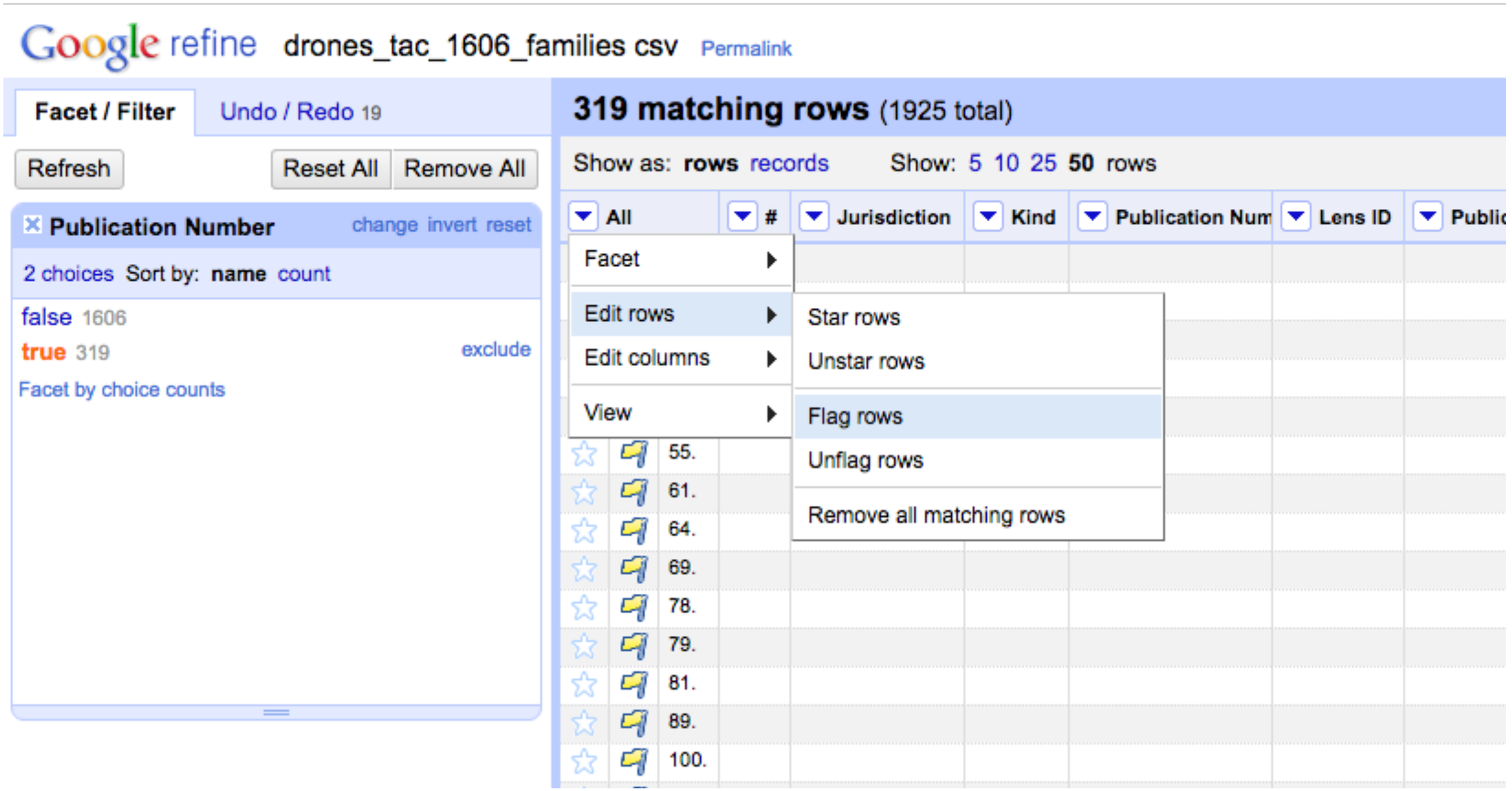
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Identifying Duplicates

We have 1925 rows in our priority tables and 1606 in the original set. Go **Publication Number > Facet > Duplicates facet**

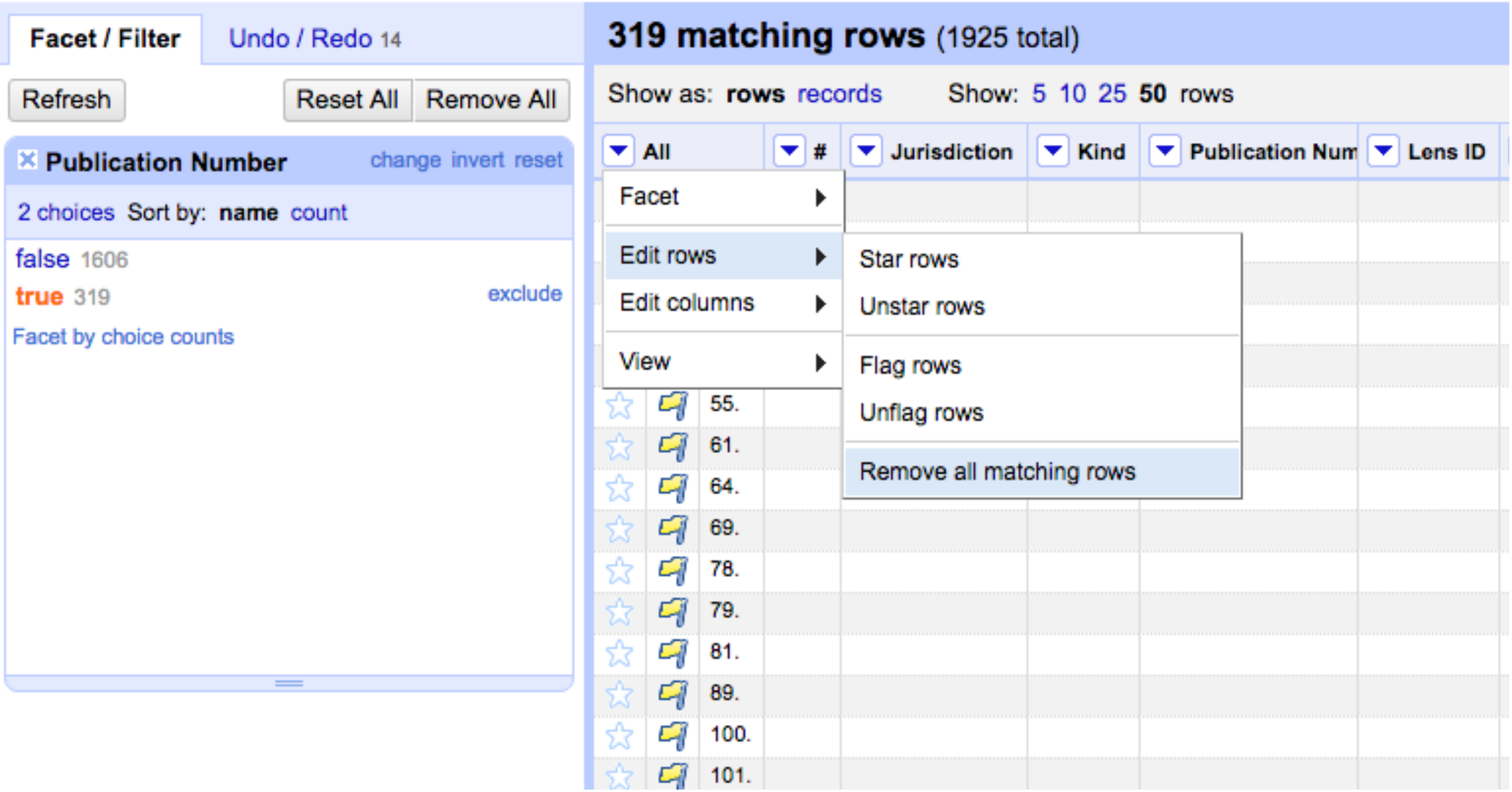
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Adding a Flag

We will flag the duplicates because we want to see whether the dates are earlier or later. Select true in Panel > Edit rows > Flag rows.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Remove Flagged Duplicates

First Select **true** in the facet panel. Then **All > Edit rows > Remove all matching rows**. Then click on **false** to see the 1606.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Export the cleaned dataset as a new core dataset.

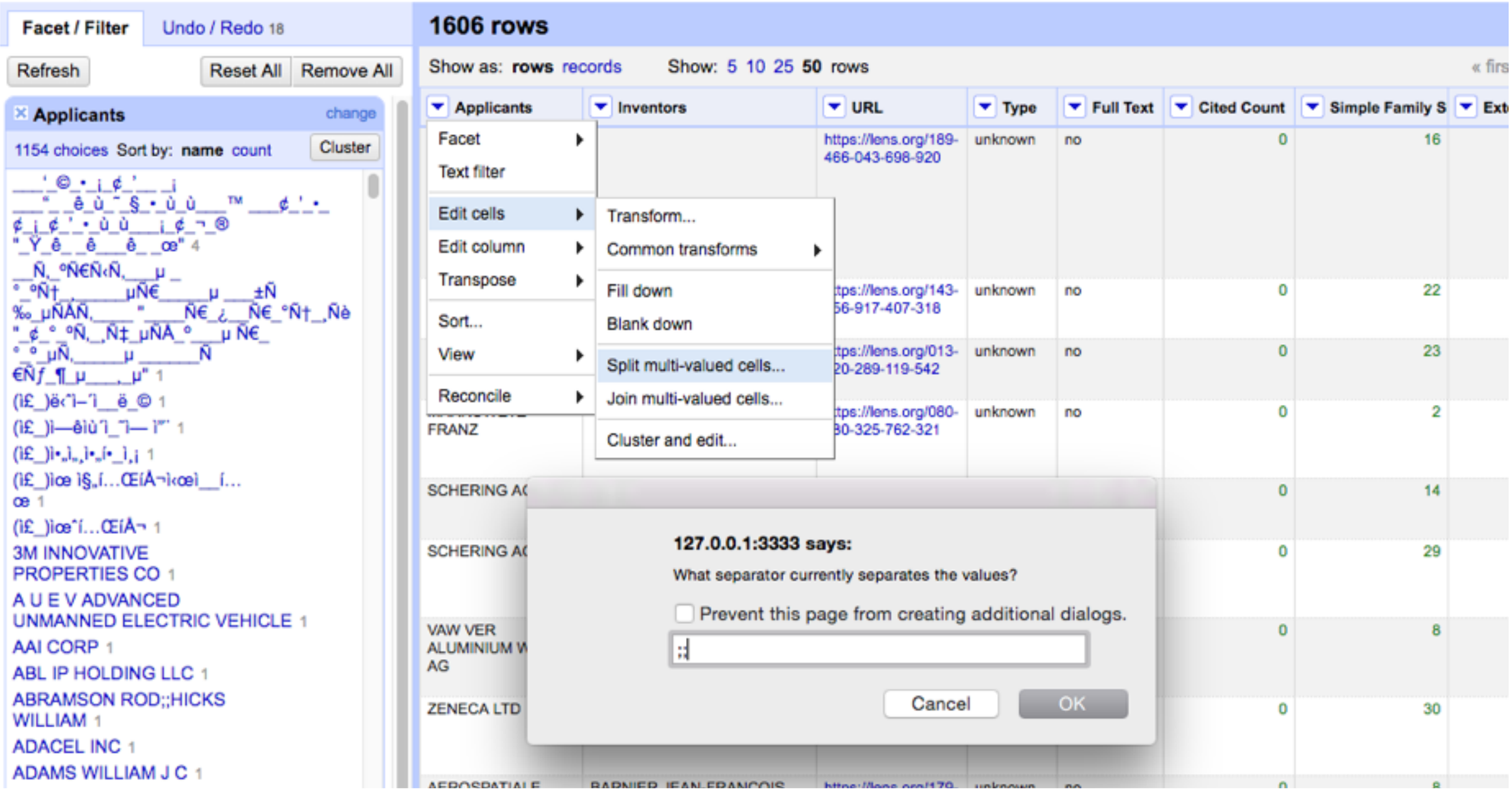
When you have finished the basic clean up you export this dataset as a new core dataset (mark accordingly).

In the next steps you will split the dataset into applicants, inventors, IPCs etc as new tables based on the core.

So, mistakes in the core will cause you problems. Bear this in mind.

Cleaning names

- We will keep going back to our core dataset so keep it somewhere safe.
- In the next steps we need to split, clean and save new data tables for applicants, inventors, IPCs etc.
- Cleaning names is a multi-step process
  - 1. Split names (each name goes on its own row)
  - 2. Create a text facet using Applicants > Facet > Text facet
  - 3. In the Text facet panel use the Cluster button to bring up the cleaning algorithm panels.
  - 4. Carefully review the names, cluster and reclusters and apply the algorithms in the list one by one.
  - 5. browser clusters - to check (and lookup on google)
  - 6. Export as an applicants table.

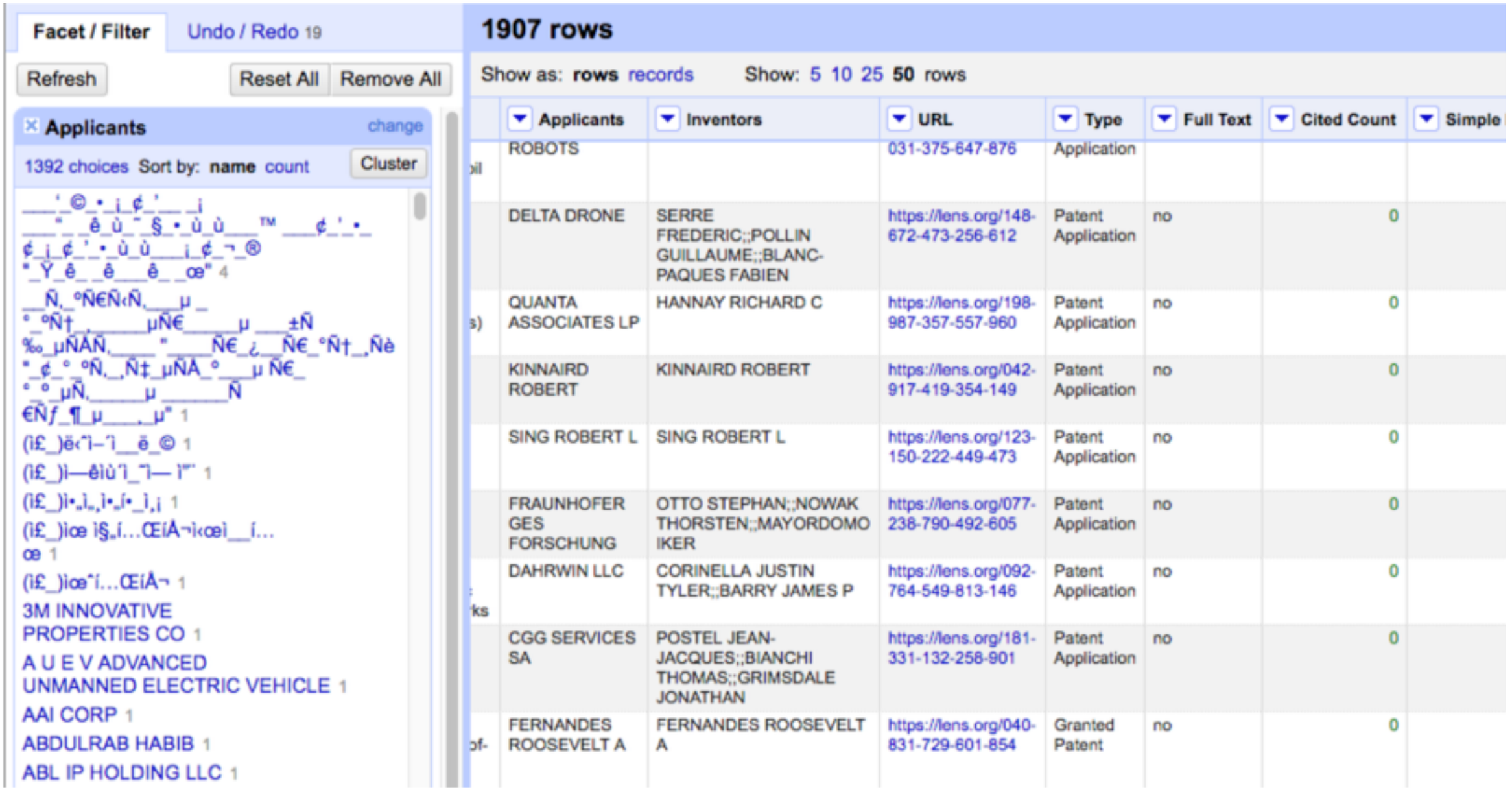


Step 1: Edit cells -> Split multivalued -> ;;

Lens data is separated on “;;”. Make sure you use “;;

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

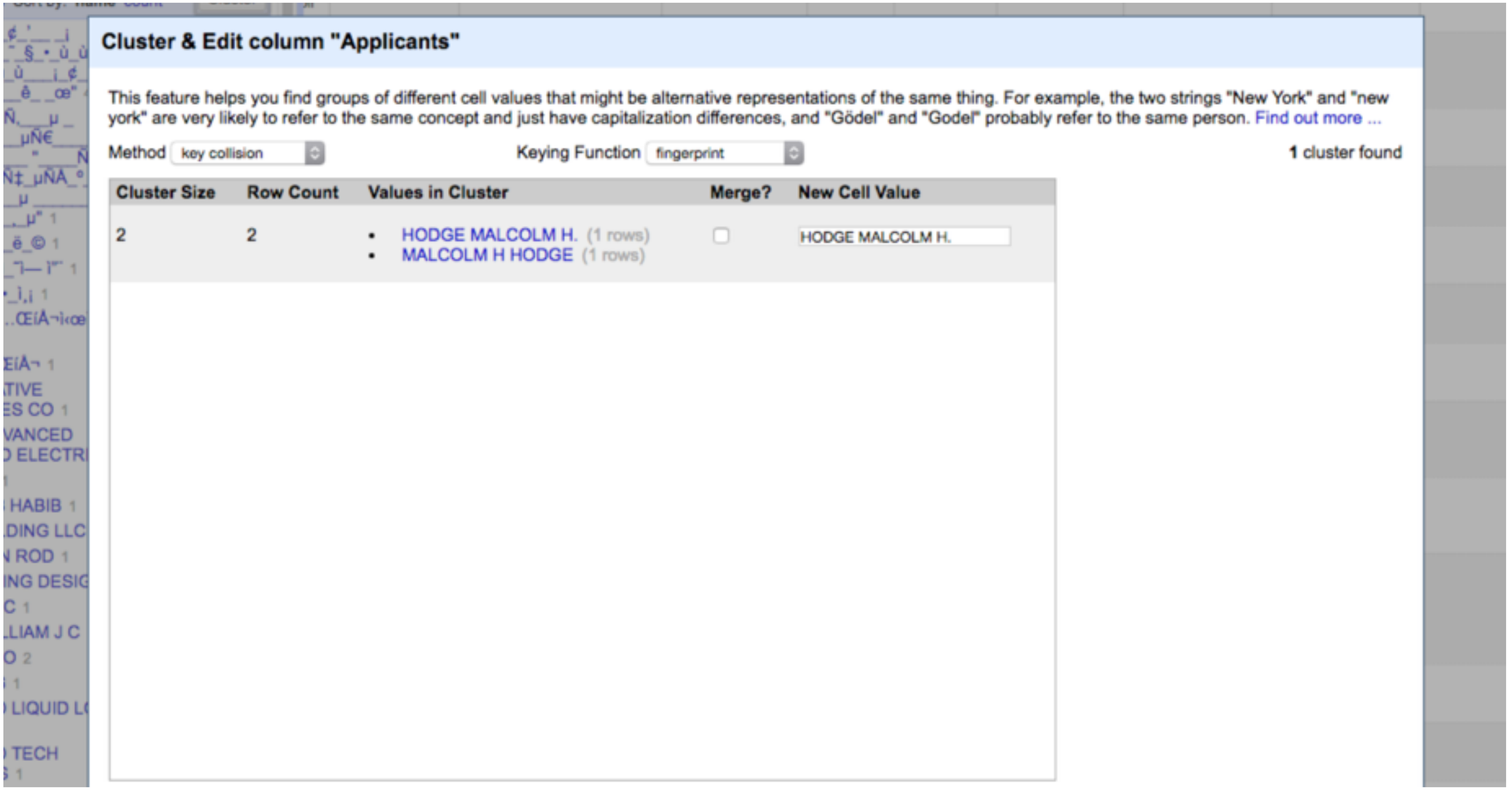
note increase in number of rows



To Start Cleaning use the Cluster Button

Open Refine contains six cleaning algorithms. To start cleaning use cluster.

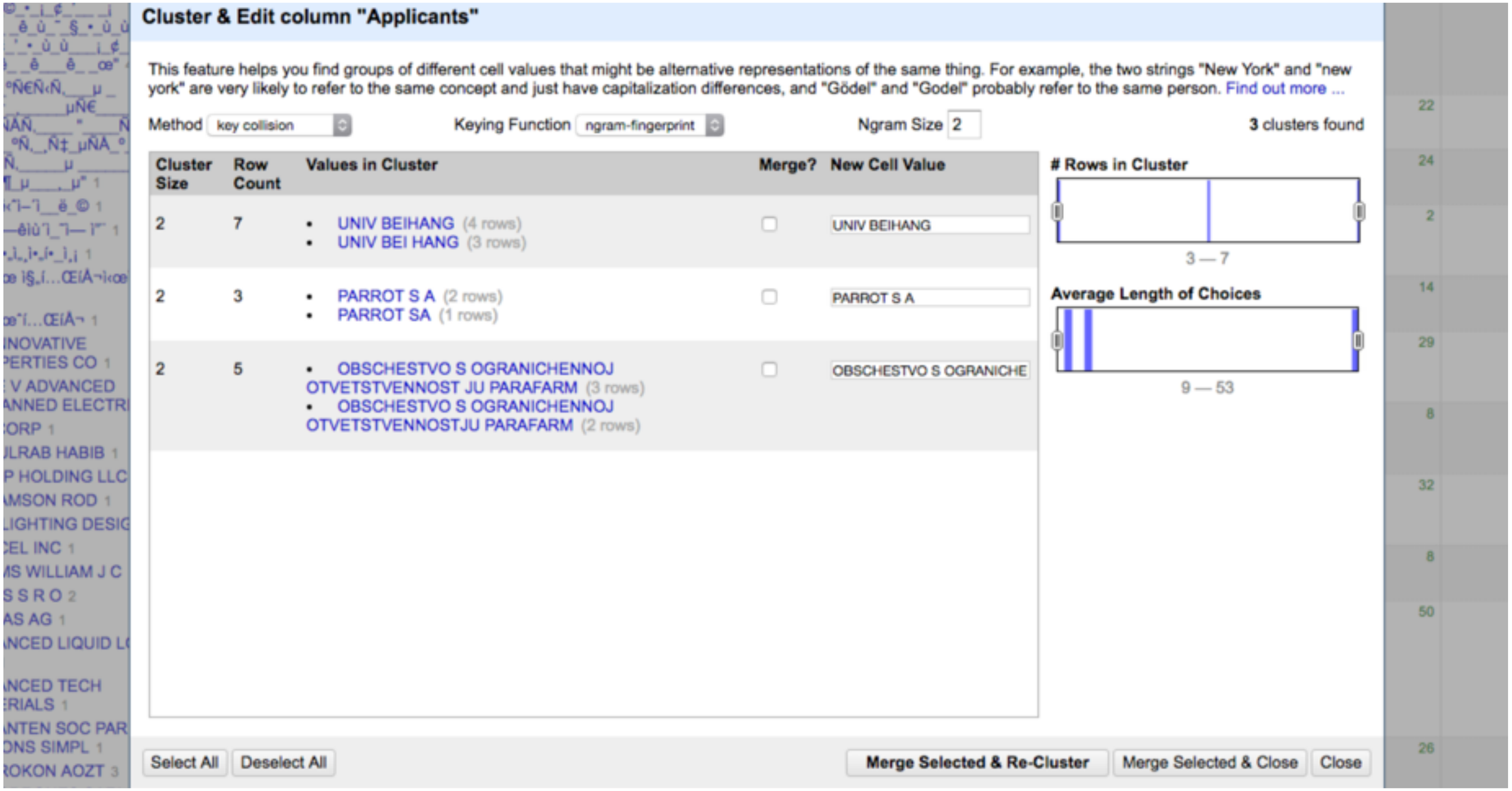
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Select Merge Selected & Re-Cluster

Check the box to merge

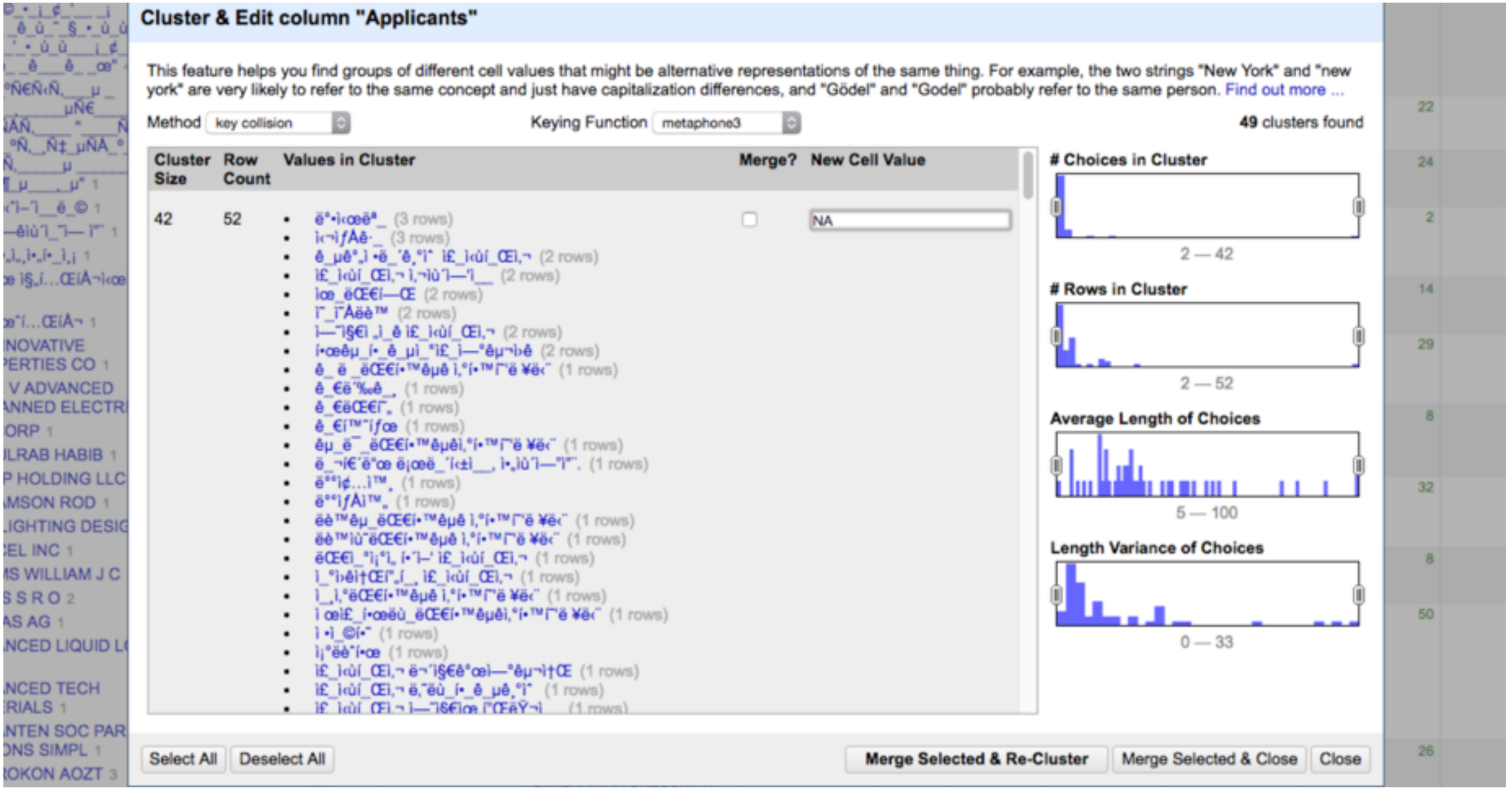
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Choose ngram_fingerprint

Looks OK, Check boxes, Merge Selected and recluster

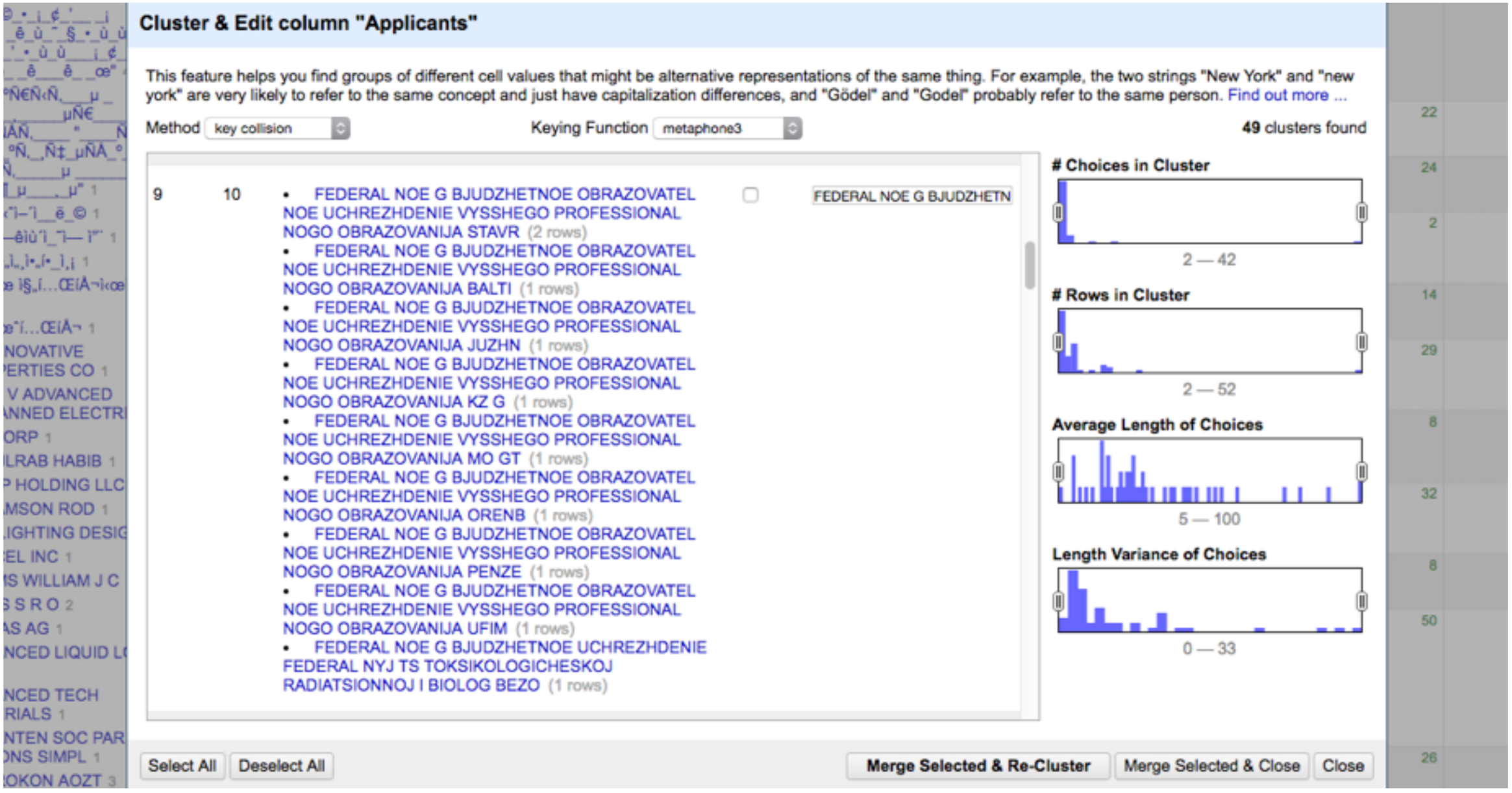
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Choose Metaphone3 and review

Arises from problems with character sets not being recognised (e.g. Mandarin or other). Either fix character sets or classify as NA.

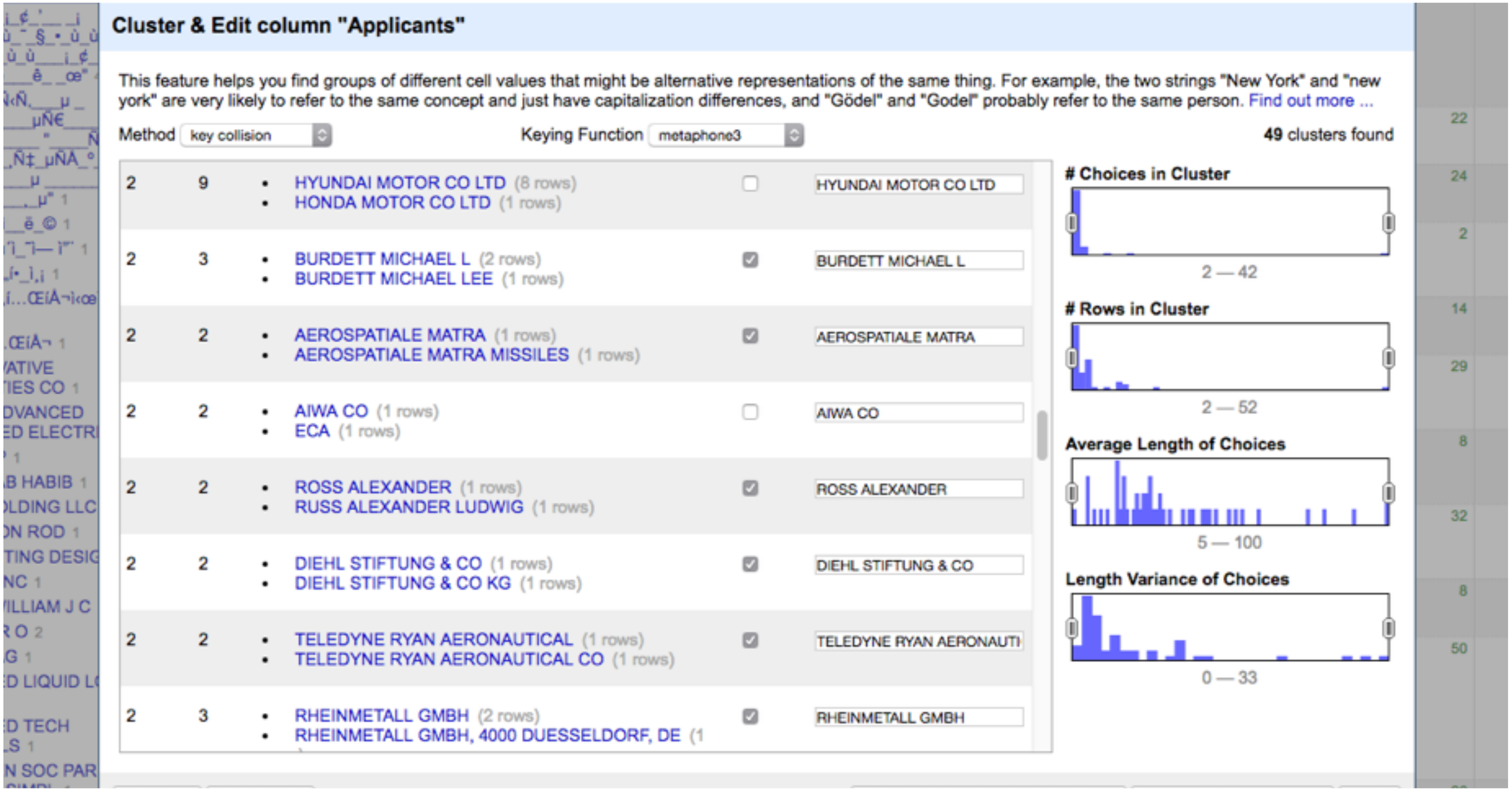
WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Cleaning can present difficult choices

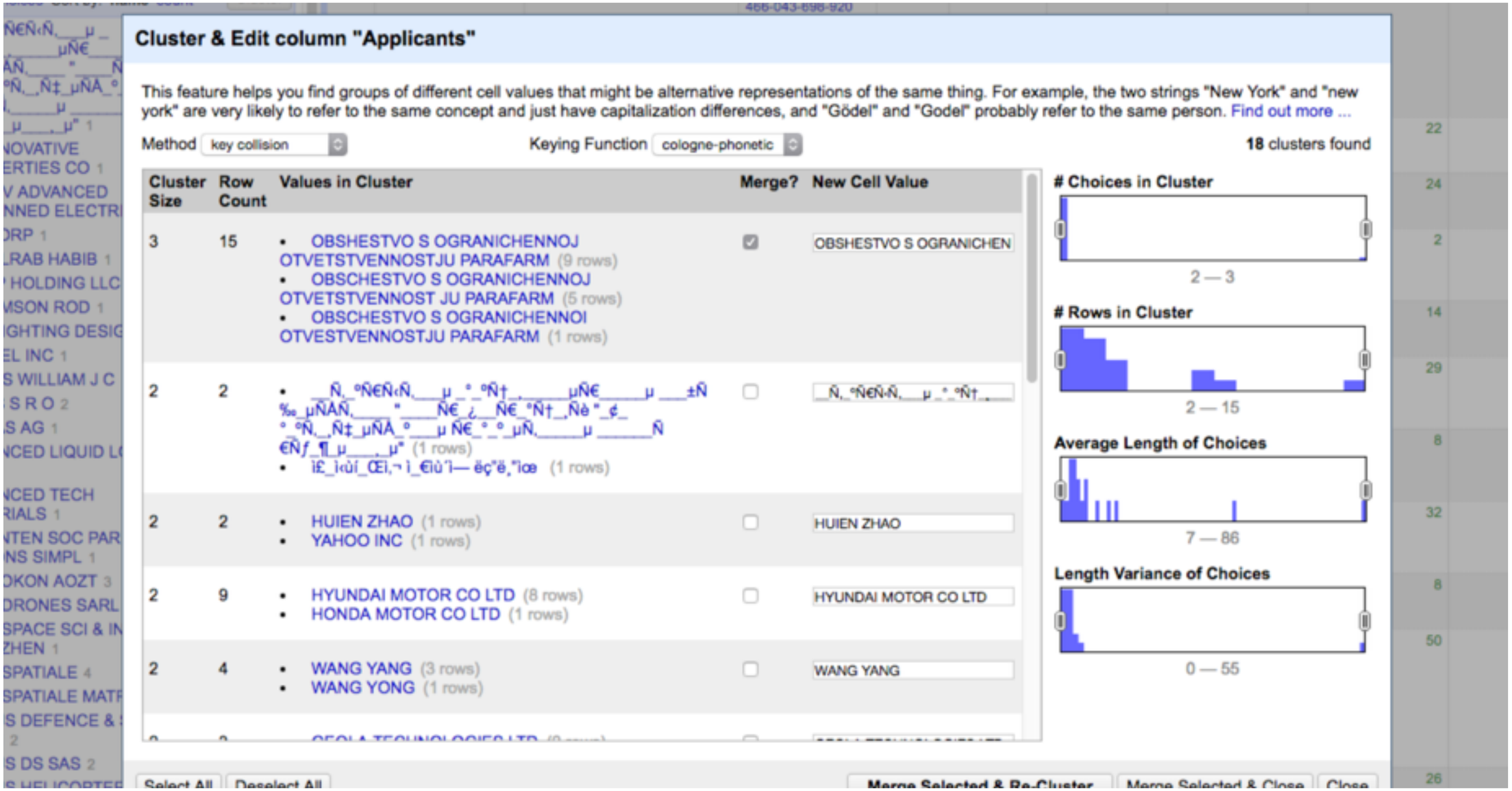
It is better to use other match criteria (such as priority numbers, family numbers or inventors) to improve accuracy.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Choose Metaphone3 and review

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Cologne Phonetic

Select those that capture accurately then Merge Selected and Close.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

before

Applicants change

1392 choices Sort by: name count Cluster

PARROT 29
THALES SA 27
IDEAL TOY CORP 14
DCNS 13
MATSUSHITA ELECTRIC IND CO LTD 10
US NAVY 10
OBSHESTVO S OGRANICHENNOJ OTVETSTVENNOSTJU PARAFARM 9
BOEING CO 8
G MASHINOSTROITEL NOE KB RADUGA IM A JA BEREZNJAKA AOOT 8
HYUNDAI MOTOR CO LTD 8
ALMUHAIRBI EIDA 7
ALSHDAIFAT WASFI 7
BAODING LONG TENG SPORTS GOODS CO LTD 7
DIEHL GMBH & CO 7
EUROCOPTER FRANCE 7
FLIR SYSTEMS 7
KASSAB FARAH AFIF 7
BOSE CORP 6
DELTA DRONE 6
DUROV DMITRIJ SERGEEVICH 6

Applicants	Inventors
ASK IND SPA	
MANIN MICHEL	
SCHIEBEL GES M B H	SCHIEBEL HA
MARKOWETZ FRANZ	
SCHERING AG	
SCHERING AG	
VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN;;EL
ZENECA LTD	SCHER HERE BENSON;;CH
AEROSPATIALE MATRA	BARNIER JEA

after

Applicants change

1312 choices Sort by: name count Cluster

NA 63
PARROT 29
THALES SA 27
OBSHESTVO S OGRANICHENNOJ OTVETSTVENNOSTJU PARAFARM 15
IDEAL TOY CORP 14
DCNS 13
BAODING LONG TENG SPORTS GOODS CO LTD 10
MATSUSHITA ELECTRIC IND CO LTD 10
US NAVY 10
BOEING CO 8
G MASHINOSTROITEL NOE KB RADUGA IM A JA BEREZNJAKA AOOT 8
HYUNDAI MOTOR CO LTD 8
ALMUHAIRBI EIDA 7
ALSHDAIFAT WASFI 7
DIEHL GMBH & CO 7
EUROCOPTER FRANCE 7
FLIR SYSTEMS 7
KASSAB FARAH AFIF 7
UNIV BEIHANG 7
BOSE CORP 6
DELTA DRONE 6
DUROV DMITRIJ SERGEEVICH 6

Applicants	Inventors
ASK IND SPA	
MANIN MICHEL	
SCHIEBEL GES M B H	SCHIEBEL HANS-GE
MARKOWETZ FRANZ	
SCHERING AG	
SCHERING AG	
VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN;;ELSNER
ZENECA LTD	SCHER HERBERT BENSON;;CHEN JIN
AEROSPATIALE MATRA	BARNIER JEAN-FRA
SCHERING AG	SCHOELLKOPF KLAUS;;HALFBRODT

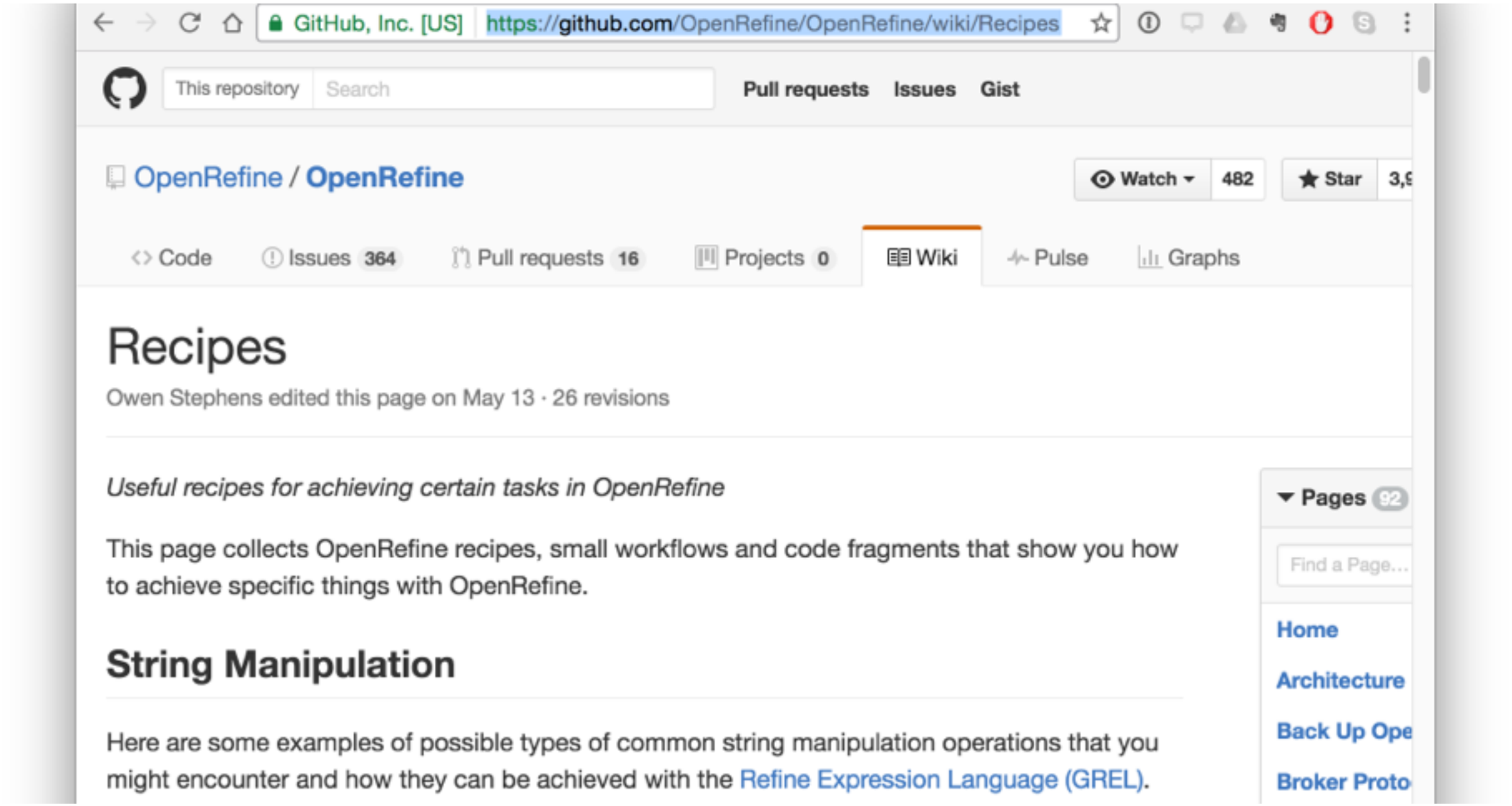
Differences may not be radical but are important in terms of accuracy (NA = added, not available)

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Next steps

- When you have cleaned the applicants - export and label as applicants.
- Go back to your clean core dataset and this time repeat the steps above for inventors. Then export and label.
- Finally, depending on your objectives, you will probably want to go back to the core dataset and split on IPCs/ CPCs for ranking technology areas. Export and label.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



Going Further

The Open Refine recipes will help you to perform common tasks

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

https://github.com/OpenRefine/OpenRefine/wiki/Recipes

Open Refine

- Open refine is free and you can do a lot with it to clean up patent data. It is important to read the recipes and other documentation.
- One problem with cleaning names in Open refine is that it does not appear to be possible to match names based on sharing another field (e.g. priority numbers and applicants) For small datasets this is fine. For larger datasets it will be a problem.
- An alternative used by researchers, companies and a growing number of patent offices is Vantage Point. However, this is a tool that needs a paid for licence so bear this in mind.