



Cleaning Patent Data with Open Refine (Google Refine)

Paul Oldham

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

✉ f t MORE

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

By STEVE LOHR AUG. 17, 2014

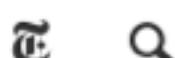


Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

✉ Email

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.

f Share



TECHNOLOGY

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



More

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.

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

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.”

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]
RAPPOLI RINOâ€,[IT]; PIZZA M	NOVARTIS VACCINES & DIAGNOSTICâ€,[IT]
OSTER PHILIPP; RAPPOLI 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]
MCGLYNN CARTER WAE [US]	McGlynn Carter Wae [US]

~/Desktop/Br		
applicants x		
	n	applicants
	1	Ventimiglia Jamie Joseph
	1	Cordova Robert
	1	Lazarillo De Tormes S L
	1	Depoortere, Thomas
	1	Frisco Findus Ag
	1	Bicycle Tools Incorporated
	1	Castiglioni, Carlo
PIZZA	1	Bujalski, Wladzimirz
pizza	1	Fournier Priscilla J
	1	Ehrno Flexible A/S
pacer	1	Hilbourne Jason

Splitting Applicant Names in R

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.

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.

Title (Best Available)

3855 Titles, 0 Selected

	# Records	# Instances	Inventors	Wang Wei	Co_Author Name
1	3855	3859	Wang Wei	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	129	129	Wang Lei	<input type="checkbox"/>	<input type="checkbox"/>
3	72	72	Wang Jian	<input type="checkbox"/>	<input type="checkbox"/>
4	65	65	Zhang Wei	<input type="checkbox"/>	<input type="checkbox"/>
5	64	64	Chen Jian	<input type="checkbox"/>	<input type="checkbox"/>
6	62	62	Li Wei	<input type="checkbox"/>	<input type="checkbox"/>
7	60	60	Wang Hui	<input type="checkbox"/>	<input type="checkbox"/>
8	57	57	Yang Jin	<input type="checkbox"/>	<input type="checkbox"/>
9	56	56	Hao Xin	<input type="checkbox"/>	<input type="checkbox"/>
10	56	56	Wang Chun-Chen	<input type="checkbox"/>	<input type="checkbox"/>
11	56	56	Wang Yong	<input type="checkbox"/>	<input type="checkbox"/>
12	56	56	Yuan Xue-Hai	<input type="checkbox"/>	<input type="checkbox"/>
13	53	53	Wang Hao	<input type="checkbox"/>	<input type="checkbox"/>
14	49	49	Li Lin	<input type="checkbox"/>	<input type="checkbox"/>
15	46	46	Huang Lei	<input type="checkbox"/>	<input type="checkbox"/>
16	42	42	Wang Jun	<input type="checkbox"/>	<input type="checkbox"/>
17	42	42	Wang Li-Jun	<input type="checkbox"/>	<input type="checkbox"/>
18	38	38	Wang Fang	<input type="checkbox"/>	<input type="checkbox"/>
19	38	38	Wang Xin	<input type="checkbox"/>	<input type="checkbox"/>
20	29	29	Zhang Hao	<input type="checkbox"/>	<input type="checkbox"/>

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.

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.

The screenshot shows the official website for OpenRefine. At the top, there's a header with tabs for "Home", "Download", "Documentation", "Community", and "Post archive". Below the header, the main content area features a large "OPEN Refine" logo with a blue diamond icon and the text "A free, open source, powerful tool for working with messy data". A "Welcome!" section provides an overview of the tool. To the right, there's a section titled "Using OpenRefine - The Book" featuring a thumbnail of the book cover and a list of topics covered. Further down, there's an "Introduction to OpenRefine" section and a "1. Explore Data" article. On the far left, there's a sidebar titled "Tweets" displaying several tweets from users like Tom Morris and Francis T. O'Donovan. The bottom of the page shows a video player for a "Google Refine 2.0 - Introduction (1 of 3) (video version 2)" video.

<http://openrefine.org/>

Download OpenRefine

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

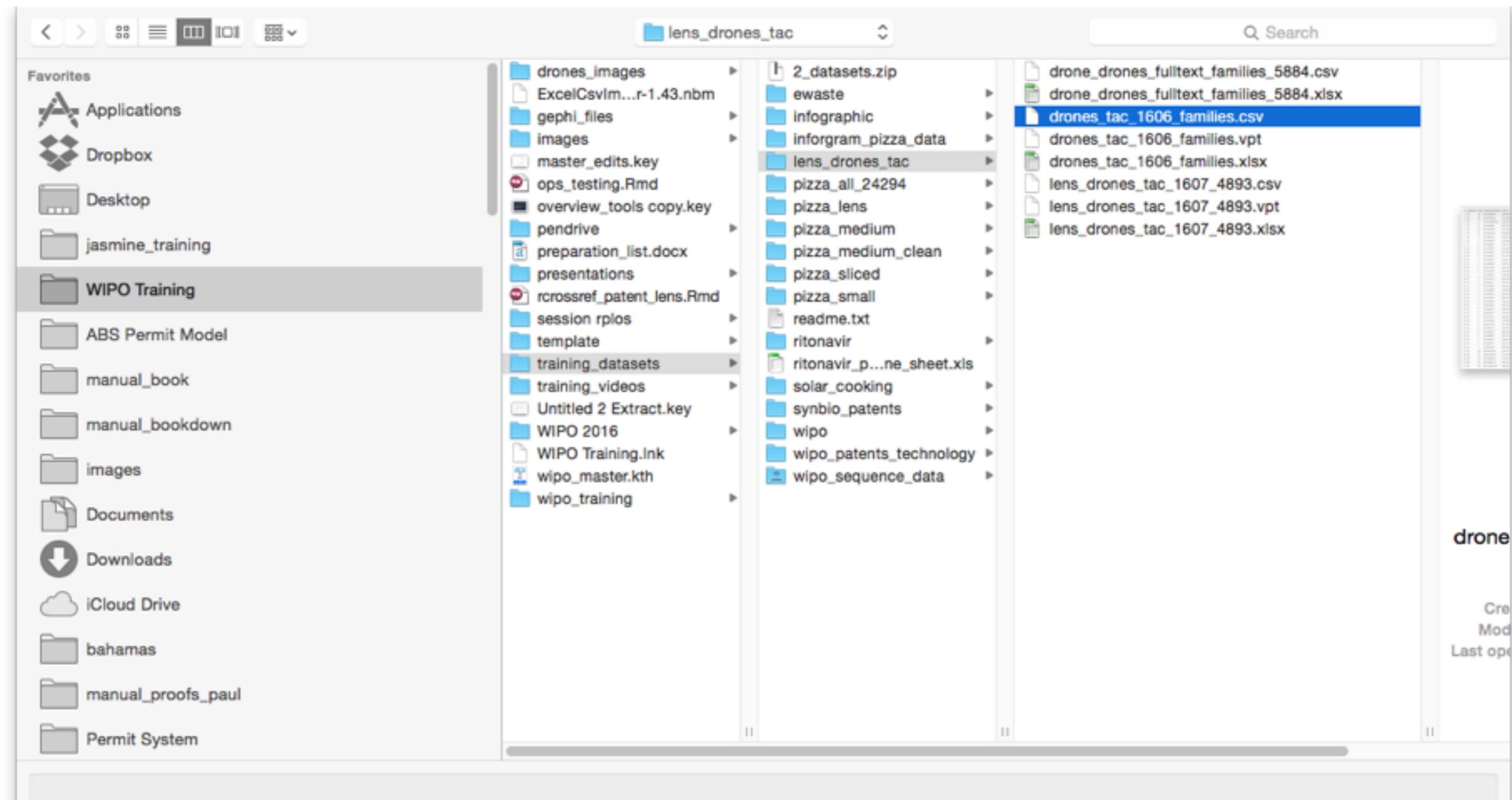
Download OpenRefine 2.6-beta1 ([link for all releases](#)).

All releases

All previous releases are available [here](#).

The screenshot shows the Google Refine web application running at 127.0.0.1:3333. The title bar says "Google Refine". The left sidebar has three options: "Create Project" (selected), "Open Project", and "Import Project". The main content area has a heading "Create a project by importing data. What kinds of data files can I import?". It lists supported formats: TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents. Below this, it says "Refine extensions." A section titled "Get data from" lists four options: "This Computer" (selected), "Web Addresses (URLs)", "Clipboard", and "Google Data". To the right, there is a file upload field with "Choose Files" and "No file chosen" buttons, and a "Next »" button.

Choose Files



Open Open Refine: Choose `drones_tac_1606_families` dataset

Choose commas (the default) as separator

The screenshot shows the Google Refine interface with a data grid and configuration options.

Data Grid:

#	Jurisdiction	Kind	Publication Number	Lens ID	Publication Date	Publication Year	Application Number	Applicat
1.	AR	A1	AR 021856 A1	189-466-043-698-920	07/08/02	2002	AR P990105416 A	27/10/99
2.	AR	A1	AR 075121 A1	143-256-917-407-318	09/03/11	2011	AR P100100007 A	04/01/10
3.	AT	A	AT A12022004 A	013-820-289-119-542	15/10/05	2005	AT 12022004 A	15/07/04
4.	AT	A	AT A33585 A	080-780-325-	15/01/95	1995	AT 33585 A	06/02/85

Parse data as:

- CSV / TSV / separator-based files** (selected)
- 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:

Columns are separated by:

- commas (CSV)
- tabs (TSV)
- custom ,

Escape special characters with: \

Parsing Options:

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

Help | **About**

Version 2.5 [r2407]

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.

Google refine drones_tac_1606_families csv X ① 127.0.0.1:3333/project?project=1971363208037

1606 rows

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
7/08/02	2002	AR P990105416 A	27/10/99	IT RE980117 A 19981123	Facet	Text facet		https://lens.org/189-466-043-698-920	unknown	no
					Text filter	Numeric facet				
					Edit cells	Timeline facet				
					Edit column	Scatterplot facet				
					Transpose	Custom text facet...				
					Sort...	Custom numeric facet...				
					View	Customized facets		https://lens.org/143-256-917-407-318	unknown	no
					Reconcile					
9/03/11	2011	AR P100100007 A	04/01/10	FR 0950012 A 20090105	Unbemannter Hubeschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEORG	https://lens.org/013-820-289-119-542	unknown	no
5/10/05	2005	AT 12022004 A	15/07/04	AT 12022004 A 20040715	Doppelstocktrogbeute FÄ_r Die ZweiMjkerbetriebsweise Zur GÄ_nzlichen Schwamunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no
5/01/95	1995	AT 33585 A	06/02/85	AT 33585 A 19850206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072-084-291-460-544	unknown	no
1/09/72	1972	AT 91271 A	03/02/71	DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen 6Ä-fluor-16Ä ±,18-dimethyl-1,4- pregnadien-3,20- diomerivaten	SCHERING AG		https://lens.org/034-658-237-174-32X	unknown	no
5/01/74	1974	AT 1005171 A	22/11/71	DE 2064859 A 19701230	Wagenkastenaufbau, Insbesondere FÄ_r Eisenbahnwagen Zur PersonenbefÄ_rderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN; ELSNER OLAF	https://lens.org/180-373-684-761-176	unknown	no
5/01/97	1997	AT 94100711 T	19/01/94	DE 4301763 A 19930123	Stabile Herbizide Zusammensetzungen, Die Metalchelate Von Herbiziden Dienen Enthalten	ZENECA LTD	SCHER HERBERT BENSON; CHEN JINLING	https://lens.org/102-558-221-531-261	unknown	no
5/01/00	2000	AT 97902456 T	03/02/97	US 59560596 A 19960202	Fernbedienungssystem FÄ_r Eine Mit Einem	AEROSPATIALE MATRA	BARNIER JEAN-FRANCOIS	https://lens.org/179-736-376-179-468	unknown	no
5/11/01	2001	AT 96401865 T	30/08/96	FR 9510385 A 19950905						

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.

All	Publication Num	Publication Date	publication_date	publication_day	publication_mor
1.	Facet	.08.2009	21/08/2009	21	8
2.	Edit cells	Transform...			
3.	Edit column	Common transforms		Trim leading and trailing whitespace	
4.	Transpose	Fill down		Collapse consecutive whitespace	
5.	Sort...	Blank down		Unescape HTML entities	
	View	Split multi-valued cells...		To titlecase	
	Reconcile	Join multi-valued cells...		To uppercase	
		Cluster and edit...		To lowercase	
				To number	
				To date	
				To text	
5.	US20030024843	07.02.2003	07/02/2003		

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)

Google refine drones_tac_1606_families csv X ① 127.0.0.1:3333/project?project=1971363208037

Facet / Filter Undo / Redo ▾ 1606 rows Extensions: Freebase ▾

Refresh Reset All Remove All

Show as: rows records Show: 5 10 25 50 rows ▾ first ▾ previous 1 - 10 next ▾ last ▾

Title change

1534 choices Sort by: name count Cluster

Publication Date	Publication Year	Application Num	Application Date	Priority Number	Title	Applicants	Inventors	URL	Type	Full Text
7/08/02	2002	AR P990105416 A	27/10/99	IT RE980117 A 19981123	Facet ▾	Text facet				
					Text filter	Numeric facet				
					Edit cells ▾	Timeline facet				
					Edit column ▾	Scatterplot facet				
					Transpose ▾	Custom text facet...				
					Sort... ▾	Custom numeric facet...				
					View ▾	Customized facets ▾				
					Reconcile ▾					
9/03/11	2011	AR P100100007 A	04/01/10	FR 0950012 A 20090105	Unbemannter Hubeschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEORG	https://lens.org/013- 820-289-119-542	unknown	no
5/10/05	2005	AT 12022004 A	15/07/04	AT 12022004 A 20040715	Doppelstocktrogbeute FÄ_r Die ZweiMjkerbetriebsweise Zur GÄ_nzlichen Schwamunterbindung	MARKOWETZ FRANZ		https://lens.org/080- 780-325-762-321	unknown	no
5/01/95	1995	AT 33585 A	06/02/85	AT 33585 A 19850206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072- 084-291-460-544	unknown	no
1/09/72	1972	AT 91271 A	03/02/71	DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen 6fz-fluor-16f ±,18-dimethyl-1,4- pregnadien-3,20- dionderivaten	SCHERING AG		https://lens.org/034- 658-237-174-32X	unknown	no
5/01/74	1974	AT 1005171 A	22/11/71	DE 2064859 A 19701230	Wagenkastenaufbau, Insbesondere FÄ_r Eisenbahnwagen Zur Personenbeförderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN; ELSNER OLAF	https://lens.org/180- 373-684-761-176	unknown	no
5/01/97	1997	AT 94100711 T	19/01/94	DE 4301763 A 19930123	Stabile Herbizide Zusammensetzungen, Die Metalchelate Von Herbiziden Dienen Enthalten	ZENECA LTD	SCHER HERBERT BENSON; CHEN JINLING	https://lens.org/102- 558-221-531-261	unknown	no
5/01/00	2000	AT 97902456 T	03/02/97	US 59560596 A 19960202	Fernbedienungssystem FÄ_r Eine Mit Einem	AEROSPATIALE MATRA	BARNIER JEAN-FRANCOIS	https://lens.org/179- 736-376-179-468	unknown	no
5/11/01	2001	AT 96401865 T	30/08/96	FR 9510385 A 19950905						

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.

Remove All

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

All	#	Jurisdiction	Kind	Publication Num	Lens ID	Publication Date	Publication Year	Application Num	Application Date	P
1.	1	AR	A1	AR 021856 A1	189-466-043-698-920	07/06/02	2002	AR P990105416 A	27/10/99	IT
2.										04/01/10 FR
3.										15/07/04 AT
4.										06/02/85 AT
5.										03/02/71 DE
6.										22/11/71 DE
7.										19/01/94 DE
8.										03/02/97 US

Custom text transform on column Publication Number

Expression Language

No syntax error.

row	value	value
1.	AR 021856 A1	AR 021856 A1
2.	AR 075121 A1	AR 075121 A1
3.	AT A12022004 A	AT A12022004 A
4.	AT A33585 A	AT A33585 A
5.	AT 301543 B	AT 301543 B
6.	AT 312823 B	AT 312823 B
7.	AT A44444 T	AT A44444 T

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

Remove spaces in publication numbers

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

Facet / Filter

Using facets

Use facets of your data to filter methods of each data source.

Not sure how to use facets? Watch the video!

Custom text transform on column Publication Number

Expression: value.replace("/", "")

Language: Google Refine Expression Language (GREL)

No syntax error.

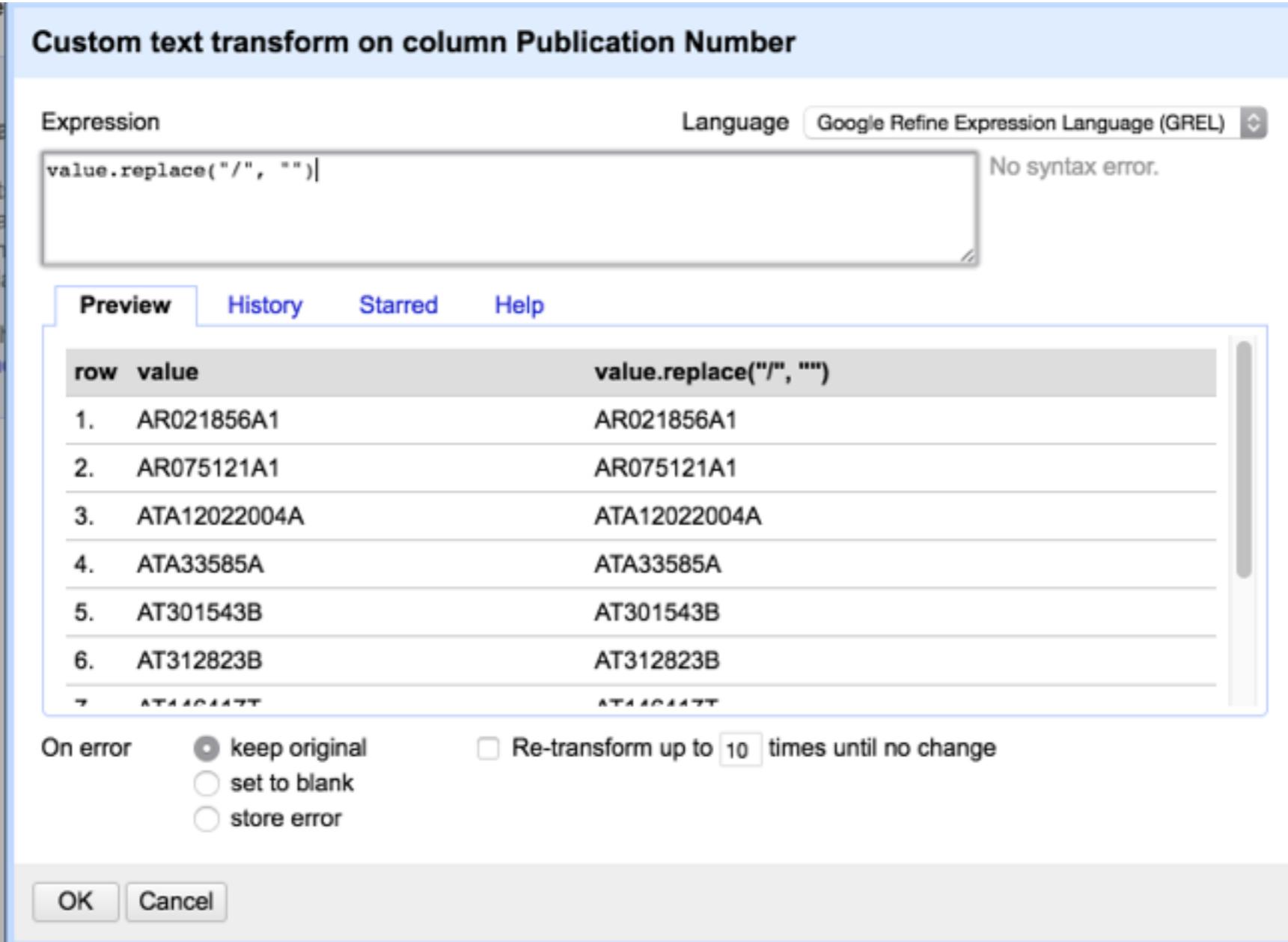
Preview History Starred Help

row	value	value.replace("/", "")
1.	AR021856A1	AR021856A1
2.	AR075121A1	AR075121A1
3.	ATA12022004A	ATA12022004A
4.	ATA33585A	ATA33585A
5.	AT301543B	AT301543B
6.	AT312823B	AT312823B
7.	ATA12022004A	ATA12022004A

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel



Remove Slash in publication numbers

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

Google refine drones_tac_1606_families.csv Permalink

Facet / Filter Undo / Redo 1 1925 rows Extensions: Freebase ▾

Refresh Reset All Remove All Show as: rows records Show: 5 10 25 50 rows < first < previous 1 - 10 next > last ▾

Title	Priority Number	Title	Applicants	Inventors	URL	Type	Full Text	Cited Count	Simple Family S	Extended Family	Sequence Count	C
1534 choices Sort by: name count Cluster		amento Pasivo Que Inclina En combinación Con Una Variante Activa Y Aliviada	ASK IND SPA		https://lens.org/189-466-043-698-920	unknown	no	0	16	16	0	HC
"NEN/N_ _N_ _N..._Nt_		Transform...										
"_N_ _NAN_ _N_		Common transforms										
N _NAN_ _N_		Transpose										
"N_ _NAN_ _N_		Fill down										
"N_ _NAN_ _N_		Sort...										
"N_ _NAN_ _N_		Blank down										
"N_ _NAN_ _N_		View										
"N_ _NAN_ _N_		Split multi-valued cells...										
"N_ _NAN_ _N_		Reconcile										
"N_ _NAN_ _N_		Join multi-valued cells...										
"N_ _NAN_ _N_	AT 12022004 A 20040715	Ui Hi Cluster and edit...	EL GES	SCHIEBEL HANS-GEORG	https://lens.org/013-620-289-119-542	unknown	no	0	23	24	0	BE
"N_ _NAN_ _N_	AT 33585 A 19850206	Doppelstocktrogbeute FÄ_r Die ZweivÄ_lkerbetriebeweise Zur GÄ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no	0	2	2	0	AC
"N_ _NAN_ _N_	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
"N_ _NAN_ _N_	DE 2064859 A 19701230	Verfahren Zur Herstellung Von Neuen 6Ä-fluor-16Ä±,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
"N_ _NAN_ _N_	DE 4301763 A 19930123	Wagenkastenaufbau, Insbesondere FÄ_r Eisenbahnwagen Zur Personenbeförderung	VAW VER ALUMINIUM WERKE AG	SCHNAAS JUERGEN; ELSNER OLAF	https://lens.org/180-373-684-761-176	unknown	no	0	8	8	0	BE
"N_ _NAN_ _N_	US 59560598 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	0	30	32	0	CC
"N_ _NAN_ _N_	FR 9510385 A 19950905	Fernbedienungssystem FÄ_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
"N_ _NAN_ _N_	DE 4447401 A 19941223	14alpha,17alpha-c2Ä_berbrÄ_cke 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

Extract more information (priority numbers)

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

1606 rows

Show as: rows records

Show: 5 10 25 50 rows

Priority Number	Title	Applicants	Inventors	URL	Type	Full Text	Cited Count
Facet Text filter	lemento Pasivo Que Funciona En Combinacion Con Un Irrlante Activo Y Aliojado	ASK IND SPA		https://lens.org/189-466-043-698-920	unknown	no	0
Edit cells	Transform...						
Edit column	Common transforms						
Transpose	Fill down						
Sort...	Blank down						
View	Split multi-valued cells...						0
Reconcile	Join multi-valued cells...						
AT 12022004 A 20040715	Ur Hi	EL C	IICI				0
AT 33585 A 19850206	Doppelstocktrogbeute FÄ_r Die ZweivÄ¶lkerbetriebsweise Zur GÄ_nzlichen Schwamunterbindung	MARKOWETZ FRANZ		https://lens.org/080-780-325-762-321	unknown	no	0
DE 2006372 A 19700206	Verfahren Zur Herstellung Von Neuen Benzopyranderivaten	SCHERING AG		https://lens.org/072-084-291-460-544	unknown	no	0

127.0.0.1:3333 says:

What separator currently separates the values?

Prevent this page from creating additional dialogs.

::

Cancel

OK

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.

• • • drones_tac_1606_families csv X

← → C ⌂ 127.0.0.1:3333/project?project=1971363208037

Google refine drones_tac_1606_families csv Permalink

Facet / Filter Undo / Redo 1 Extract... Apply...

Filter:

0. Create project

1. Split multi-valued cells in column Priority Numbers

2. Split 1925 cell(s) in column Priority Numbers into several columns by separator

1925 rows

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

Priority Number	Title	Applicants	Inventors
Facet	emento Pasivo Que Inciona En Combinacion Con Un Elemento Activo Y Alojado En Un Espacio Vacio Que no Lleve Volumen	ASK IND SPA	
Text filter			
Edit cells			
Edit column	Split into several columns...		
Transpose	Add column based on this column...		
Sort...	Add column by fetching URLs...		
View	Add columns from Freebase ...		
Reconcile	Rename this column		
AT 12022004 A 20040715	Remove this column		BEL HANS-GEORG

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.

Producir Unincremento
De La Eficacia Del
Parlante, En Particular A
Bajas Frecuencias

Metodo
Mejorad
Prospect
Acutatica

Unbema
Hubsch

Doppels
FÄ_r Di
ZweivÄ
Zur GÄ
Schwarz

Verfahre
Von Ne
Benzop

Verfahre
Von Ne
±,18-din
pregnac
dionder

Wagen
Insbesondere FÄ_r
Eisenbahnwagen Zur

ALUMINIUM
WERKE AG

JUERGEN;;ELSNER OLAF

373-684-761-176

Split column Priority Numbers into several columns

How to Split Column

- by separator

Separator regular expression

Split into columns at most (leave blank for no limit)

- by field lengths

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

After Splitting

Guess cell type

Remove this column

OK

Cancel

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.

1925 rows

Show as: **rows** records

Show: **5 10 25 50** rows

<input type="button" value="▼"/> Priority Number:	<input type="button" value="▼"/> Title				
IT	RE980117	A	19981123		Elemento Pasivo Que Funciona En Combinacion Con Un Parlante Activo Y Aloh En Un Espacio Vacio Tiene Un Volumen Cerrado De Aire Para Producir Unincremento 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		Doppelstocktrochheute

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.

1925 rows

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

Priority Country	Priority Number	Priority Number	Priority Number	Priority Number	Title	Applicants	Inventors
Facet Text filter	2006112147	A	20060412		Method Of Complex Activation Of Animal Organism Defensors Functioning	G NAUCHNOE UCHREZHDENIE ROSSEL	POGODAEV VLADIMIR ANIKEEVICH;;MORENKO ELENA ALEKSANDROVNA;;PONOMAREV OLEG VIKTOROVICH;;MOISEEV OLEG NIKOLAEVICH;;KLIMENKO ALEKSANDR IVANOVICH;;OVCHAROV ALEKSANDR PETROVICH
Edit cells							
Edit column	Split into several columns...						
Transpose	Add column based on this column...			20071106	Onboard Thermal Trap	ZD IM V A DEGTJAREVA AOOT	GROMOV VLADIMIR VJACHESLAVOVICH;;DUBININ DENIS BORISOVICH;;LIPSMAN DAVID LAZOROVICH;;MAKHNIN ANDREJ VLADIMIROVICH;;NEMIROVSKIY BORIS VLADIMIROVICH;;OVCHINNIKOV IURII SERGEEVICH;;PICHUGIN
Sort...	Add column by fetching URLs...						
View	Add columns from Freebase ...						
Reconcile	Rename this column						
	Remove this column						
	Move column to beginning						DROVICH;;TKACHIK LEGOVICH
RU	Move column to end						VIKTOR VICH;;GAJNOV JURIJ VICH;;MAJOROV BORIS EVICH
	Move column left						V ROBINDAR VICH;;STAKHOV EVGENIJ DROVICH;;KHAJRULLIN
RU	Move column right						H;;SHCHERBAKOV J IVANOVICH

127.0.0.1:3333 says:

Enter new column name

Prevent this page from creating additional dialogs.

Priority Country

Cancel OK

Rename Columns

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

Priority	Priority Kind	Priority Date	Priority Number	Title	Applicants	Inventors
980117	A	19981123	Facet	emento Pasivo Que nciona En ombinacion Con Un rante Activo Y Alojado n Un Espacio Vacio Que no Tiene Volumen	ASK IND SPA	
			Text filter			
			Edit cells			
			Edit column	Split into several columns...		
			Transpose	Add column based on this column...		
			Sort...	Add column by fetching URLs...		
950012	A	20090105	View	Add columns from Freebase ...		
12022004	A	20040715	Reconcile	Rename this column		
33585	A	19850206		Remove this column		
2006372	A	19700206		Move column to beginning		
				Move column to end		
				Move column left		
				Move column right		
				Von Neuen Benzopyranderivaten		

Remove Blank Column

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

**CREATE A NEW COLUMN TO
JOIN PRIORITY DATA.**

1925 rows

Show as: **rows** records

Show: 5 10 25 50 rows

Priority Country	Priority	Priority Kind	Priority Date	Title	Applicants	Inventors
Facet	► 980117	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	
Text filter						
Edit cells	►					
Edit column	►	Split into several columns...				
Transpose	►	Add column based on this column...				
Sort...		Add column by fetching URLs...				
View	►	Add columns from Freebase ...			MANIN MICHEL	
Reconcile	►	Rename this column				
AT		Remove this column		Unbemannter Hubschrauber	SCHIEBEL GES M B H	SCHIEBEL HANS-GEORG
AT		Move column to beginning		Doppelstocktrogbeute FÄ_r Die ZweivÄ¶lkerbetriebsweise Zur GÄ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ	
		Move column to end				
		Move column left				

Add Column based on Columns

Choose Priority Country

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.

Add column based on column Priority Country

New column name priority_number

On error set to blank store error copy value from original column

Expression Language Google Refine Expression Language (GREL)

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

No syntax error.

Preview History Starred Help

row	value	cells["Priority Country"].value + "" + cells["Priority"].value + "" + cells["Priority Kind"].value
1.	IT	ITRE980117A
2.	FR	FR950012A
3.	AT	AT12022004A
4.	AT	AT33585A
5.	DE	DE2006372A
6.	DE	DE2064859A

OK Cancel

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.

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¶kerbetriebsweise Zur GÃ_nzlichen Schwarmunterbindung	MARKOWETZ FRANZ	
DE	DE2006372A	2006372	A	19700206	Verfahren Zur Herstellung Von Neuen	SCHERING AG	

The New Priority Number Column

Add column based on column Priority Date

New column name

On error set to blank store error copy value from original column

Expression

```
substring(value, 0, 4)
```

No syntax error.

Preview History Starred Help

row	value	substring(value, 0, 4)
1.	19981123	1998
2.	20090105	2009
3.	20040715	2004
4.	19850206	1985
5.	19700206	1970
6.	19701230	1970
-	40000000	4000

OK Cancel

The screenshot shows the 'Add column based on column Priority Date' dialog in Google Refine. It includes fields for 'New column name' (priority_year), 'On error' (set to blank), 'Expression' (substring(value, 0, 4)), and 'Language' (Google Refine Expression Language (GREL)). Below the dialog is a preview table showing the transformation of various priority dates into their corresponding years. The table has columns for row number, original value, and the result of the expression. The preview shows examples like '19981123' becoming '1998', '20090105' becoming '2009', and '19700206' becoming '1970'. At the bottom are 'OK' and 'Cancel' buttons.

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.

1925 rows

Extensions: Free

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next »

All	#	Jurisdiction	Kind	Publication Num	Lens ID	Publication Date	Publication Year	Application Num	Application Date	Priority Country	priority_
28.						Facet	Text facet			US	US23101800
29.	26	AT	T			Text filter	Numeric facet	2010 AT 04761230 T	03/09/04	AU	AU20039048
30.	27	AT	T			Edit cells	Timeline facet	2010 AT 05013901 T	28/06/05	DE	DE10200404
31.	28	AT	T			Edit column	Scatterplot facet	2010 AT 07731683 T	01/03/07	FR	FR651048A
32.	29	AT	T			Transpose	Custom text facet...				
						Sort...	Custom numeric facet...				
						View	Customized facets				
33.	30	AT	T	AT 481779 T	071-578-785-771-768	15/10/10		Duplicates facet	03	DE	DE20215204
34.	31	AT	T	AT 466131 T	102-381-444-167-136	15/05/10		Numeric log facet			
35.	32	AT	T	AT 502847 T	094-249-353-921-846	15/04/11		1-bounded numeric log facet			
36.	33	AT	T	AT 539958 T	027-673-470-146-837	15/01/12		Text length facet	07	IL	IL17794806A
								Log of text length facet			
								Unicode char-code facet	01	US	US24923400
								Facet by error	08	FR	FR706401A
								Facet by blank			

Identifying Duplicates

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

Facet / Filter Undo / Redo 19

Refresh Reset All Remove All

Publication Number change invert reset

2 choices Sort by: name count

false 1606
true 319

Facet by choice counts

exclude

319 matching rows (1925 total)

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

All # Jurisdiction Kind Publication Num Lens ID Public

Facet ►

Edit rows ► Star rows

Edit columns ► Unstar rows

View ► Flag rows

Unflag rows

Remove all matching rows

Star	Flag	#	Jurisdiction	Kind	Publication Num	Lens ID	Public
☆	!	55.					
☆	!	61.					
☆	!	64.					
☆	!	69.					
☆	!	78.					
☆	!	79.					
☆	!	81.					
☆	!	89.					
☆	!	100.					

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.

Facet / Filter Undo / Redo 14

Refresh Reset All Remove All

Publication Number change invert reset

2 choices Sort by: name count

false 1606 true 319 Facet by choice counts

exclude

319 matching rows (1925 total)

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

All	#	Jurisdiction	Kind	Publication Num	Lens ID
Facet					
Edit rows				Star rows	
Edit columns				Unstar rows	
View				Flag rows	
				Unflag rows	
				Remove all matching rows	
55.					
61.					
64.					
69.					
78.					
79.					
81.					
89.					
100.					
101.					

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.

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.

The screenshot shows a web-based dataset interface with a list of 1606 records. The table has columns for ID, Jurisdiction, and Kind. The 'Jurisdiction' column contains values like AR, AT, and T. The 'Kind' column contains values like A1, A, and B. The 'Jurisdiction' header is currently sorted. A context menu is open on the right side of the table, listing various export options: Export project, Tab-separated value (which is highlighted), Comma-separated value, HTML table, Excel, ODF spreadsheet, Triple loader, MQLWrite, Custom tabular exporter..., and Templating... The 'Comma-separated value' option is currently selected.

ID	Jurisdiction	Kind					
1.	AR	A1					
2.	AR	A1				917-407-318	
3.	AT	A	ATA12022004A	013-820-289-119-542	15/10/05		
4.	AT	A	ATA33585A	080-780-325-762-321	15/01/95		
5.	AT	B	AT301543B	072-084-291-460-544	11/09/72		
6.	AT	B	AT312823B	034-658-237-174-32X	25/01/74		
7.	AT	T	AT146417T	180-373-684-761-176	15/01/97		

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.

Facet / Filter Undo / Redo 18

Refresh Reset All Remove All

Applicants change

1154 choices Sort by: name count Cluster

Applicants	Inventors	URL	Type	Full Text	Cited Count	Simple Family S	Ext
Facet		https://lens.org/189-466-043-698-920	unknown	no	0	16	
Text filter							
Edit cells	Transform...						
Edit column	Common transforms						
Transpose	Fill down	https://lens.org/143-56-917-407-318	unknown	no	0	22	
Sort...	Blank down						
View	Split multi-valued cells...	https://lens.org/013-20-289-119-542	unknown	no	0	23	
Reconcile	Join multi-valued cells...	https://lens.org/080-30-325-762-321	unknown	no	0	2	
FRANZ	Cluster and edit...						
SCHERING AG					0	14	
SCHERING AG	127.0.0.1:3333 says: What separator currently separates the values? <input type="checkbox"/> Prevent this page from creating additional dialogs. ;;				0	29	
VAW VER ALUMINIUM WAG					0	8	
ZENECA LTD					0	30	
AEROSPATIALE	RAPNIER JEANFRANCOIS	https://lens.org/170...	unknown	no	0	0	

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

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

note increase in number of rows

Facet / Filter Undo / Redo 19

Refresh Reset All Remove All

Applicants change

1392 choices Sort by: name count Cluster

	Applicants	Inventors	URL	Type	Full Text	Cited Count	Simple
xil	ROBOTS		031-375-647-876	Application			
s)	DELTA DRONE	SERRE FREDERIC;;POLLIN GUILLAUME;;BLANC-PAQUES FABIEN	https://lens.org/148-672-473-256-612	Patent Application	no	0	
ks	QUANTA ASSOCIATES LP	HANNAY RICHARD C	https://lens.org/198-987-357-557-960	Patent Application	no	0	
of-	KINNAIRD ROBERT	KINNAIRD ROBERT	https://lens.org/042-917-419-354-149	Patent Application	no	0	
	SING ROBERT L	SING ROBERT L	https://lens.org/123-150-222-449-473	Patent Application	no	0	
	FRAUNHOFER GES FORSCHUNG	OTTO STEPHAN;;NOWAK THORSTEN;;MAYORDOMO IKER	https://lens.org/077-238-790-492-605	Patent Application	no	0	
	DAHRWIN LLC	CORINELLA JUSTIN TYLER;;BARRY JAMES P	https://lens.org/092-764-549-813-146	Patent Application	no	0	
	CGG SERVICES SA	POSTEL JEAN-JACQUES;;BIANCHI THOMAS;;GRIMSDALE JONATHAN	https://lens.org/181-331-132-258-901	Patent Application	no	0	
	FERNANDES ROOSEVELT A	FERNANDES ROOSEVELT A	https://lens.org/040-831-729-601-854	Granted Patent	no	0	

To Start Cleaning use the Cluster Button

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

Cluster & Edit column "Applicants"

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 1 cluster found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none">HODGE MALCOLM H. (1 rows)MALCOLM H HODGE (1 rows)	<input type="checkbox"/>	HODGE MALCOLM H.

Select Merge Selected & Re-Cluster

Check the box to merge

Cluster & Edit column "Applicants"

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: ngram-fingerprint Ngram Size: 2 3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Rows in Cluster
2	7	<ul style="list-style-type: none"> UNIV BEIHANG (4 rows) UNIV BEI HANG (3 rows) 	<input type="checkbox"/>	UNIV BEIHANG	 3 — 7
2	3	<ul style="list-style-type: none"> PARROT S A (2 rows) PARROT SA (1 rows) 	<input type="checkbox"/>	PARROT S A	 2
2	5	<ul style="list-style-type: none"> OBSCHESTVO S OGRANICHENNOJ OTVETSTVENNOST JU PARAFARM (3 rows) OBSCHESTVO S OGRANICHENNOJ OTVETSTVENNOSTJU PARAFARM (2 rows) 	<input type="checkbox"/>	OBSCHESTVO S OGRANICHE	 9 — 53

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Choose ngram_fingerprint

Looks OK, Check boxes, Merge Selected and recluster

Cluster & Edit column "Applicants"

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 metaphone3 49 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? <input type="checkbox"/>	New Cell Value
42	52	<ul style="list-style-type: none">• ë®·íœëë_ (3 rows)• íœifAéë_ (3 rows)• é_µë®·í·ë_ë·í·í£_íùí_ŒI,· (2 rows)• í£_íùí_ŒI,·í·íùí_í·í (2 rows)• íœ_ëŒÉI—Œ (2 rows)• í·í£Aéë™ (2 rows)• í—í§ÉI_í·ë_í£_íùí_ŒI,· (2 rows)• í·œëµ_í·ë_µ_í£_í—ëµ·íë (2 rows)• é_ë_ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• é_€ë%ë_ (1 rows)• é_€ëŒÉEΓ_ (1 rows)• é_ŒI™í·íœ (1 rows)• éµ_ë_ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• é_í·ë_ë_ë_ëjœë_í·í·í,·í·íùí_í·í (1 rows)• é**í¢...í™_ (1 rows)• é**ífAí™_ (1 rows)• éé™ëµ_ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• éé™íù_ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• éŒÉI_í·í·í,·í·í·í£_íùí_ŒI,· (1 rows)• í·í·íëí_ŒI,·í·í·í£_íùí_ŒI,· (1 rows)• í·í·í,·ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• í·í·íœí_í·í·íœí_ëŒÉI·™ëµë_í·í·™í·ë_¥ëç (1 rows)• í·í·í·í·í (1 rows)• í·í·í·í·í·í (1 rows)• í£_íùí_ŒI,·í·ë_í£_íùí_ŒI,·í·í (1 rows)• í£_íùí_ŒI,·í·ë_í£_íùí_ŒI,·í·í (1 rows)		NA

Choices in Cluster

2 — 42

Rows in Cluster

2 — 52

Average Length of Choices

5 — 100

Length Variance of Choices

0 — 50

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

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.

Cluster & Edit column "Applicants"

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: metaphone3 49 clusters found

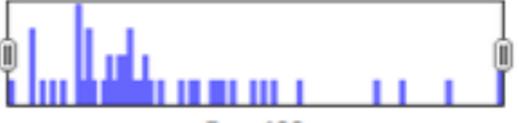
Priority	Count	Choices
9	10	<ul style="list-style-type: none"> FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA STAVR (2 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA BALTI (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA JUZHN (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA KZ G (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA MO GT (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA ORENB (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA PENZE (1 rows) FEDERAL NOE G BJUDZHENOE OBRAZOVATEL NOE UCHREZHDENIE VYSSHEGO PROFESSIONAL NOGO OBRAZOVANIJA UFIM (1 rows) FEDERAL NYJ TS TOKSIKOLOGICHESKOJ RADIATSIONNOJ I BIOLOG BEZO (1 rows)

Choices in Cluster

 2 — 42

Rows in Cluster

 2 — 52

Average Length of Choices

 5 — 100

Length Variance of Choices

 0 — 33

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Cleaning can present difficult choices

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

Cluster & Edit column "Applicants"

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 **metaphone3**

49 clusters found

2	9	• HYUNDAI MOTOR CO LTD (8 rows)	<input type="checkbox"/>	HYUNDAI MOTOR CO LTD
		• HONDA MOTOR CO LTD (1 rows)		
2	3	• BURDETT MICHAEL L (2 rows)	<input checked="" type="checkbox"/>	BURDETT MICHAEL L
		• BURDETT MICHAEL LEE (1 rows)		
2	2	• AEROSPATIALE MATRA (1 rows)	<input checked="" type="checkbox"/>	AEROSPATIALE MATRA
		• AEROSPATIALE MATRA MISSILES (1 rows)		
2	2	• AIWA CO (1 rows)	<input type="checkbox"/>	AIWA CO
		• ECA (1 rows)		
2	2	• ROSS ALEXANDER (1 rows)	<input checked="" type="checkbox"/>	ROSS ALEXANDER
		• RUSS ALEXANDER LUDWIG (1 rows)		
2	2	• DIEHL STIFTUNG & CO (1 rows)	<input checked="" type="checkbox"/>	DIEHL STIFTUNG & CO
		• DIEHL STIFTUNG & CO KG (1 rows)		
2	2	• TELEDYNE RYAN AERONAUTICAL (1 rows)	<input checked="" type="checkbox"/>	TELEDYNE RYAN AERONAUTICAL
		• TELEDYNE RYAN AERONAUTICAL CO (1 rows)		
2	3	• RHEINMETALL GMBH (2 rows)	<input checked="" type="checkbox"/>	RHEINMETALL GMBH
		• RHEINMETALL GMBH, 4000 DUESSELDORF, DE (1		

Choices in Cluster



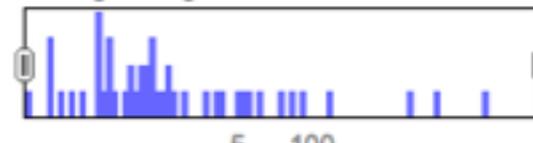
2 — 42

Rows in Cluster



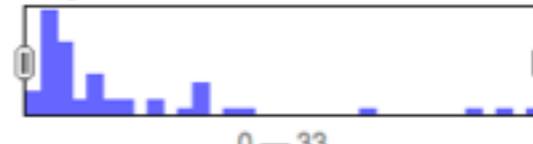
2 — 52

Average Length of Choices



5 — 100

Length Variance of Choices



0 — 33

Choose Metaphone3 and review

Cologne Phonetic

Select those that capture accurately then Merge Selected and Close.

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	

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	
ASK IND SPA		
MANIN MICHEL		
SCHIEBEL GES M B H		
MARKOWETZ FRANZ		
SCHERING AG		
SCHERING AG		
VAV VER ALUMINIUM WERKE AG		
ZENECA LTD		
AEROSPATIALE MATRA		
SCHERING AG		

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

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.

[GitHub, Inc. \[US\]](#) <https://github.com/OpenRefine/OpenRefine/wiki/Recipes>

This repository Search Pull requests Issues Gist

OpenRefine / OpenRefine Watch 482 Star 3,9

Code Issues 364 Pull requests 16 Projects 0 Wiki Pulse Graphs

Recipes

Owen Stephens edited this page on May 13 · 26 revisions

Useful recipes for achieving certain tasks in OpenRefine

This page collects OpenRefine recipes, small workflows and code fragments that show you how to achieve specific things with OpenRefine.

String Manipulation

Here are some examples of possible types of common string manipulation operations that you might encounter and how they can be achieved with the Refine Expression Language (GREL).

Pages 92

Find a Page...
Home
Architecture
Back Up Ope
Broker Proto

Going Further

The Open Refine recipes will help you to perform common tasks

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.