Curating data with OpenRefine to visualize them in VOSviewer

If you choose to create a map based on bibliographic data to visualize keywords cooccurrences in VOSviewer you will visualize keywords as they are indexed in the Web of Science or Scopus. If written in a different manner, occurrences of the same term will be considered as different items: e.g. "climate-change" and "climate change", or "city" and "cities", while they are the same concept.

A small number of this kind of errors should not affect too much your visualization in VOSviewer but too many of them could blur the picture and diminish the relevance of the map. Most of these issues can be solved in few steps thanks to OpenRefine. This tutorial explains how.

Curating data from Web of Science

Exporting data from the Web of Science

Export data from the WoS as you usually do when working with VOSviewer. As file format, select the **Tab-delimited format (Win, UTF-8)**.

Send to File		×
Number of Re	ecords: All records on page Records 1 to 500	
Record Content:	Full Record and Cited References	•
File Format	Tab-delimited (Win, UTF-8)	•
	Send Cancel	

Creating a project in OpenRefine

Download OpenRefine on your computer and launch the software. Then upload the file you want to clean in OpenRefine and click on **Next**.



Then you must make sure that the different fields of the file are well recognized. Select the following options: CSV/TSV/separator-based files; **separated by tabs** (TSV); then choose **UTF-8** as the Encoding character and Parse next **1 line** as column header and create your project.



Refiñe ⋗	A po	wer too	ol for v	vorking with mes	sy da	ta.										
Create Project	«	Start	Over	Configure Pa	arsin	g Op	tions	Project name	e example_wos					Create Project »		
Open Project	Т	PT	AU		BA	BE	GP	AF	BF	CA	п		so		P	
Import Project Language Settings	1.	L	Oha	inian, D				Ohanian, Daniel			Sympathy ar Exclusion: TI Migration of and Women Survivors of Armenian Gé from the Eas Mediterranea Canada, 192	he Child the enocide tern an to		OCIDE STUDIES	3	
	•	III													•	
		Pars	e da	ata as			Char	acter encoding UT	'F-8				U	pdate Preview	ŕ	
		CSV / files	TSV	/ separator-b	ased			mns are separated commas (CSV)	by			Igno first		0 line(s) at beginning of	=	
		Line-t	based	text files				abs (TSV)						file		
		Fixed	-width	n field text files				ustom \t				Par nex		1 line(s) as column		
		PC-A	xis te:	kt files			Esca	pe special charact	ers w	vith \				headers		
Marsian 0.0 FTDI IMI/1		JSON	l files									Dise		0 row(s)		

Once your project is created, each row contains information about a publication and each column is a metadata. The first line of the WoS file is the column header.

In the column "**DE**", go to "Edit cells", "Split multi-valued cells".

Refine wos Perma	link Oper	n Export - He
Facet / Filter Undo / Redo 0	200 rows	Extensio
Using facets and filters Use facets and filters to select subsets of your data to act on. Choose facet and	Fa	DE 🔽
filter methods from the menus at the top of each data column.		it cells
Not sure how to get started? Watch these screencasts	Fill down Tra Blank down So	nspose
	Split multi-valued cells Vie	•w •
	Join multi-valued cells Re	concile 🕨
	Cluster and edit	

As separator, type : "; ". Beware not forgetting the space after the semi-colon.

Split multi-valu			
 by separator Separator by field lengths 	;	regular expression	
List of integers	separated by com	mas, e.g., 5, 7, 15	đ
OK Cancel			



From now you have a difference between the number of rows (keywords) and the nmber of records (publications).



Then, still in the column "DE", go to "Edit cells" and "Cluster and edit".

Refiñe	example_wos Per	malink							
Facet / Filter	Undo / Redo 1	10	15 row	s					
	0	Sho	was: ro	ws recor	ds Sho	w: 5 10	25 50 ro	ows	
Using face	ets and filters	- CL	. 💌 SP	💌 но	💌 DE	💌 ID		💌 АВ	C 1
Use facets a	nd filters to select subsets to act on. Choose facet and				Facet Text filter	•		In 1918, some 500,000 Ottoman Armenians	[Ohanian, I Univ Calif L Angeles, H Los Angele
· · · · · · · · · · · · · · · · · · ·	s from the menus at the top				Edit cells	•	Transfo		4 USA
Not sure hov	v to get started?				Edit columr Transpose	י ו ווווווווווווווווווווווווווווווווווו		on transforms	•
Watch these	escreencasts				Sort		Fill dow Blank d		
					View	•	Split mu	ulti-valued cells	
					Reconcile	•	Join mu	Ilti-valued cells	
							Cluster	and edit	

A new window opens. You can select which terms you want to merge. First click on "Select all", then uncheck any words whose merging is irrelevant. By default, the new cell value is the value that occurs the most often but you may desire to change it. Once you made your choices, click on "**Merge selected & re-cluster**".

lethod key co	llision •	Keying Function fi	ingerprint	•	16 clusters foun
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Choices in Cluster
3	5	 Familial Mediterranean Fever (2 rows) familial Mediterranean fever (2 rows) Familial Mediterranean fever (1 rows) 	V	Familial Mediterranean Feve	2-3
2	2	Cardiovascular disease (1 rows) cardiovascular disease (1 rows)	V	Cardiovascular disease	# Rows in Cluster
2	5	 Mediterranean Sea (3 rows) Mediterranean sea (2 rows) 	V	Mediterranean Sea	2 — 13 Average Length of Choices
2	3	Biogeography (2 rows)biogeography (1 rows)	V	Biogeography	
2	2	 Microsatellite (1 rows) microsatellite (1 rows) 	V	Microsatellite	4 — 28
2	4	Nutrition (2 rows) nutrition (2 rows)	V	Nutrition	
2	4	Diet (2 rows) diet (2 rows)	V	Diet	



A second method of clustering is available: "nearest neighbor", you may try it. Then close this window.

Once this is done, in the column "DE", go to "Edit cells", "Join multi-valued cells".

Refine wos Permal	ink							
Facet / Filter Undo / Redo 3	101	5 row	s					
Extract Apply	Shov	v as: ro	ws recor	ds Sho	w: 5 10	25 50 ra	ws	
Filter:	💌 CL	💌 SP	💌 но	💌 DE	💌 ID	1	💌 АВ	C 1
0. Create project				Facet Text filter	•		In 1918, some 500,000 Ottoman Armenians	[Ohanian, Da Univ Calif Los Angeles, Hist
1. Mass edit 0 cells in column DE				Edit cells	•	Transfo		Los Angeles, 4 USA
2. Split multi-valued cells in column DE				Edit colum	n ▶	Commo	n transforms	•
3. Mass edit 55 cells in column DE				Transpose	•	Fill dow	n	
				Sort		Blank d	own	
				View	•	Split mu	ilti-valued cells	
				Reconcile	•	Join mu	lti-valued cells	
						Cluster	and edit	

Enter "; " as separator.

Enter separator to use between values

;		
ſ	OK	Annuler

You should now have as many rows as records. You can now save your project and export the file in TSV (Tab-separated value).



This file should be readable in VOSviewer as an original Web of Science file. The same cleaning may be done with Scopus files, in a slightly different way.



Curating data from Scopus

Exporting data from Scopus

Export data from Scopus as you usually do when working with VOSviewer in a commaseparated value format (CSV).

i have chosen to export 44 ielect your method of exp		CSV BibTeX Excel	O Plain Text ASCII in HTML	
 What information do you value Citation information Author(s) Document title Year EID Source title volume, issue, pages Citation count Source & document type DOI 	ant to export? Bibliographical information Affiliations Serial identifiers (e.g. ISSN) PubMed ID Publisher Editor(s) Language of original document Correspondence address Abbreviated source title	 Abstract & keywords Abstract Author keywords Index keywords 	Funding details Number Acronym Sponsor Funding text	Other information Tradenames & manufacturers Accession numbers & chemicals Conference information Include references

Creating a project in OpenRefine

Upload the file you want to clean in OpenRefine and click on **Next**. Choose UTF-8 as character encoding and CSV separator-based file and create the project.

Refiñe 🤛	Α ροι	ver tool for worl	ang with	messy data.							
Create Project	«		Configu	ire Pi Options	roject n	ame scopus 4	5 csv		Crea	te Project »	•
Open Project		Authors	Title		Year	Source title	Volume	Issue	Art. No.	Page start	
Import Project	1.	Eberhart		optimizer	1995	Proceedings				39	-
Language Settings		Russell, Kennedy James		g particle rm theory		of the International Symposium on Micro Machine and Human Science					
	2.	Swain M.J., Ballard D.H.	Cold	or indexing	1991	International Journal of	7	1		11	
		Parse data	aas	Characte	r encoc	ling UTF-8			Updat	e Preview	^
	:	CSV / TSV / separator-ba files	sed	Columns comn tabs (nas (CS	oarated by SV)		Igno Igno Igno		line(s) eginning of	
		Line-based te files	xt	custo	m ,	characters with	i V	Pars next		line(s) olumn ders	
Version 2.8 [TRUNK]		Fixed-width fi text files	eld					Disc initia		row(s) ata	
		PC-Axis text JSON files	files					Load mos		row(s) ata	
Preferences Help											

In the column « Author keywords », « Edit cells », and click on « **Split multi-valued cells** », write "; " as separator. Then merge the terms in the same way than with the Web of Science file.



Refine	scopus 45 csv Perma	alink			C	pen	Export
Facet / Filter	Undo / Redo 0	20 rows					E
		Show as: rows	records Show	v: 5	5 10 25 150t rows	evious	1 - 10 n
Using face	ets and filters		Author Keyword	•	Document Type	💌 Acc	ess Type
	nd filters to select subsets	ward 9517385&	Facet Text filter	•	nference Paper		
	to act on. Choose facet and s from the menus at the top column	4c8b2fd9275856b	Edit cells	Þ	Transform		
Not sure hov	v to get started?	ward 6246210& 87&partnerID=40&	Edit column Transpose	► ►	Common transfo	rms	•
Watch these	e screencasts	ba518f11c5c599e ward	Sort		Blank down		
		2688781& 1087032&	View	۲	Split multi-valued	d cells	
		5f93e399f150d96	Reconcile	►	Join multi-valued	I cells	
		ward 44399896&	Cooperative control; Graph theory;	Ar	Cluster and edit.		

Finally join the multi-valued cells with "; " as separator and export the file as a csv file. This file should be readable in VOSviewer.

Clean several files with the same rules

If you want to view several files in VOSviewer, the files curation in OpenRefine could take some time, and the merging choices could not always be the same since the maximum number of occurrences for a concept may change from one file to the other.

To be sure to apply always the same rules you can go in the "Undo/Redo" tab and click on "Extract".



Then the different operations you have just made appear in JSON format. You may copypaste it as text.





Then, when you have created a new project for another file, in the "Undo/Redo" tab, click on "Apply".



Paste the text with the history of operations you have done with the previous file and click on "**Perform operations**".

"to": "Sustainable agriculture"	
3,	
{	
"fromBlank": false,	
"fromError": false,	
"from": [
"Epidemiology",	
"epidemiology"	
1,	
"to": "Epidemiology"	
(
"fromBlank": false,	
"fromError": false,	
"from": [
"Children",	
"children"	
1,	
"to": "Children"	
},	
"op": "core/multivalued-cell-join",	
"description": "Join multi-valued cells in column DE",	
"columnName": "DE",	
"keyColumnName": "PT",	
"separator": "; "	

The file has been curated in the same way than the previous one. You may now export it in the format you want to read it in VOSviewer.

