BiblioMerge

BiblioMerge: A Python-based automated tool to merge WoS and Scopus bibliographic data, compatible with Biblioshiny, Bibexcel, VOSviewer, SciMAT and ScientoPy

David Diez-Junguitu, Miguel Á. Peña-Cerezo

User Guide
Table of contents

Authors debugging

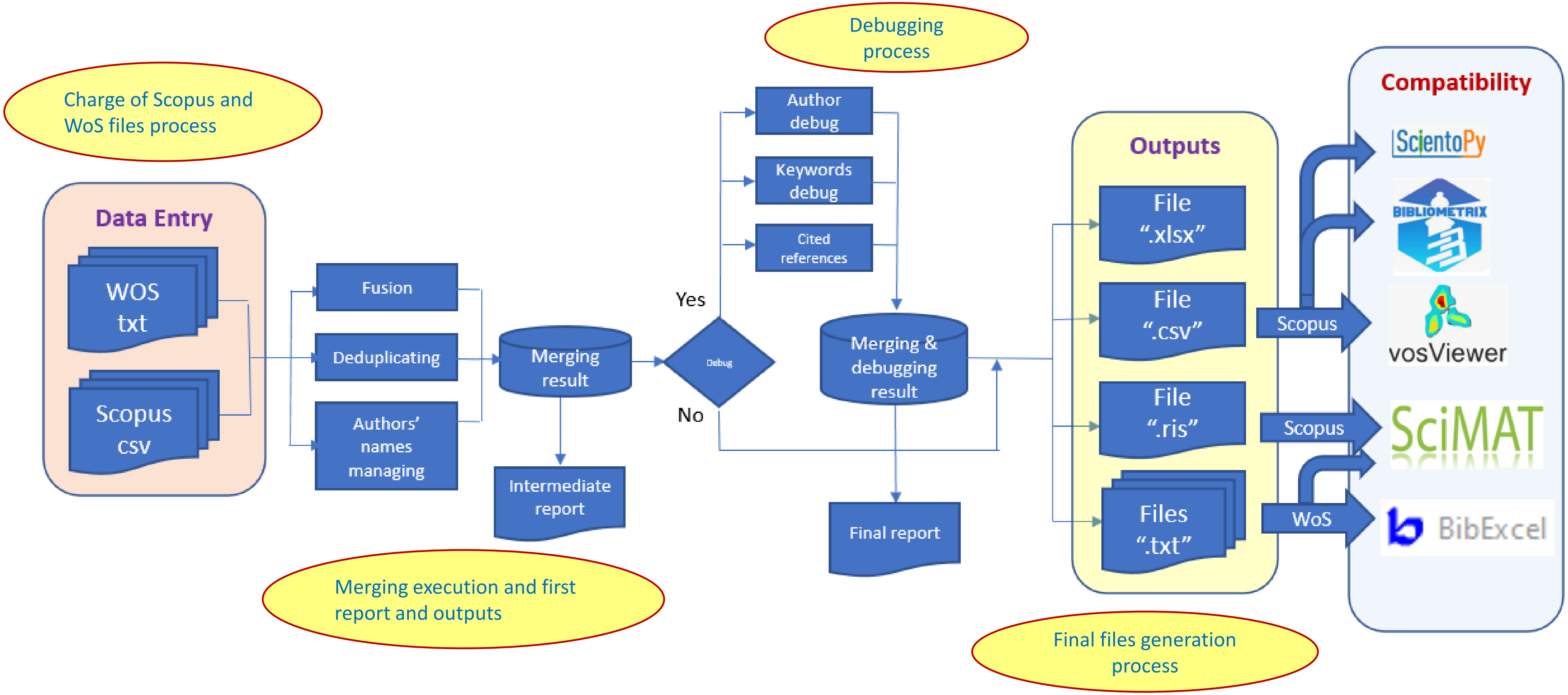
User Guide
Table of contents

- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
  - ❑ Authors debugging
  - ❑ Author and Index Keywords debugging
  - ❑ Cited References debugging
  - ❑ Final Step
- Final files and reports generation

- Authors debugging

User Guide
General process view

Back to index



Charge of Scopus and WoS files process

Data Entry

WOS txt

Scopus csv

Fusion

Deduplicating

Authors' names managing

Merging result

Intermediate report

Merging execution and first report and outputs

Debug

Yes

No

Debugging process

Author debug

Keywords debug

Cited references

Merging & debugging result

Final report

Outputs

File ".xlsx"

File ".csv"

File ".ris"

Files ".txt"

Final files generation process

Scopus

Scopus

WoS

Compatibility

ScientoPy

BIBLIOMETRIX

vosViewer

SciMAT

BibExcel

User Guide
Table of contents

- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
  - □ Authors debugging
  - □ Author and Index Keywords debugging
  - □ Cited References debugging
  - □ Final Step
- Final files and reports generation

- Authors debugging

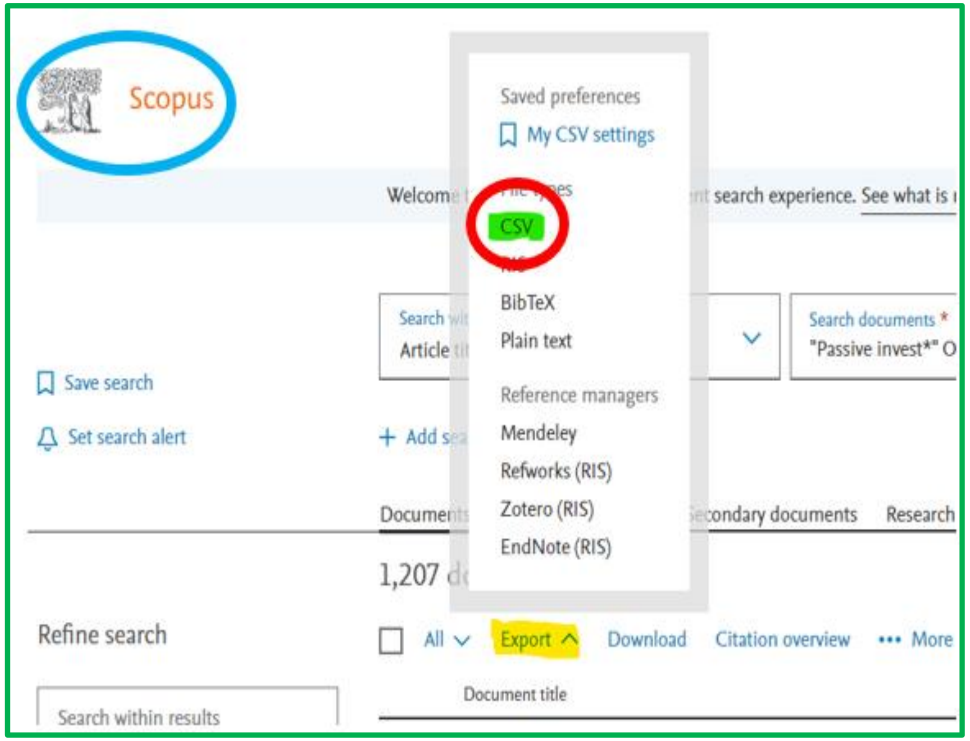
User Guide

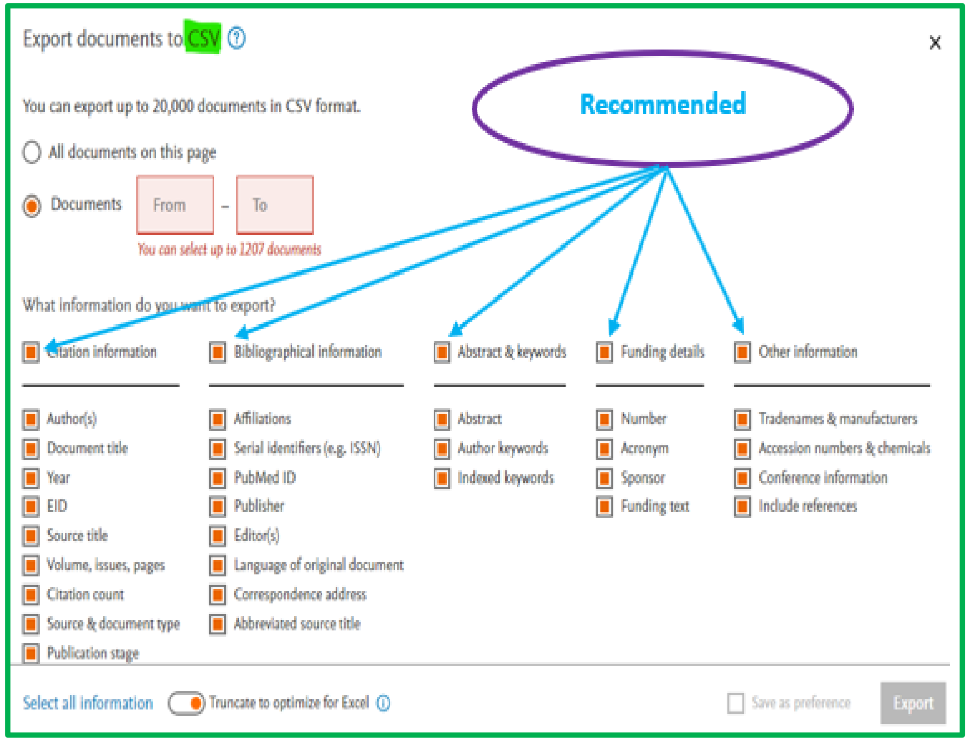
Preliminary steps

Back to index

1. We need to export Scopus documents in CSV format

and selecting all available information recommended





Export documents to CSV ⓘ ×

Recommended

You can export up to 20,000 documents in CSV format.

○ All documents on this page

● Documents From – To

You can select up to 1207 documents

What information do you want to export?

☑ Citation information
- ☑ Author(s)
- ☑ Document title
- ☑ Year
- ☑ EID
- ☑ Source title
- ☑ Volume, issues, pages
- ☑ Citation count
- ☑ Source & document type
- ☑ Publication stage

☑ 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
- ☑ Indexed keywords

☑ Funding details
- ☑ Number
- ☑ Acronym
- ☑ Sponsor
- ☑ Funding text

☑ Other information
- ☑ Tradenames & manufacturers
- ☑ Accession numbers & chemicals
- ☑ Conference information
- ☑ Include references

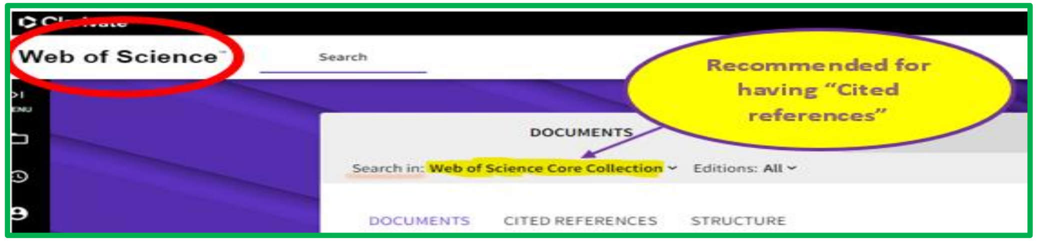
Select all information Truncate to optimize for Excel ⓘ □ Save as preference Export

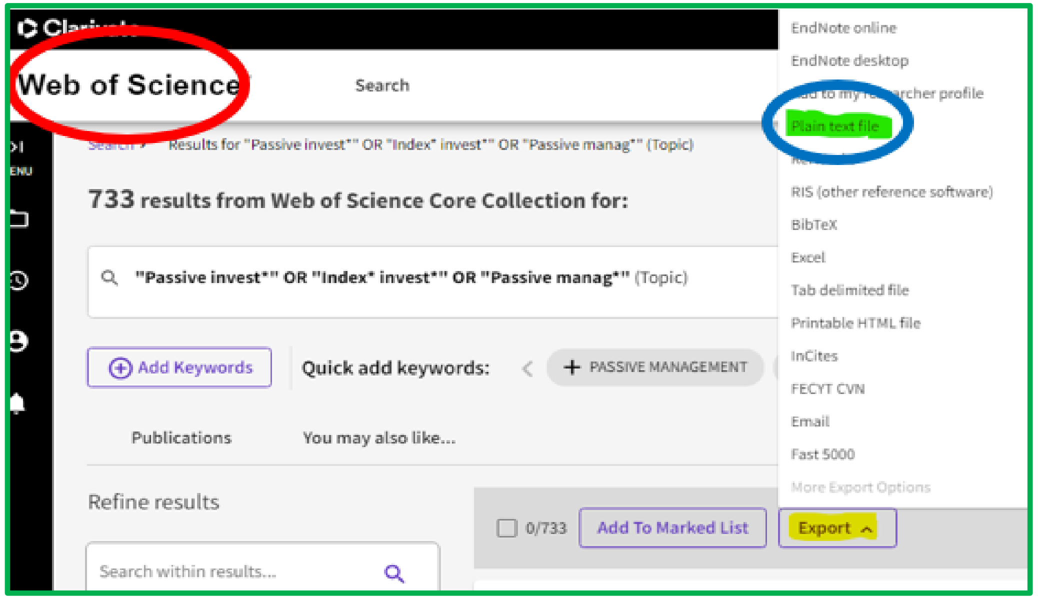
User Guide

Preliminary steps

Back to index

2. We need to export **Web of Science** records in Plain text format



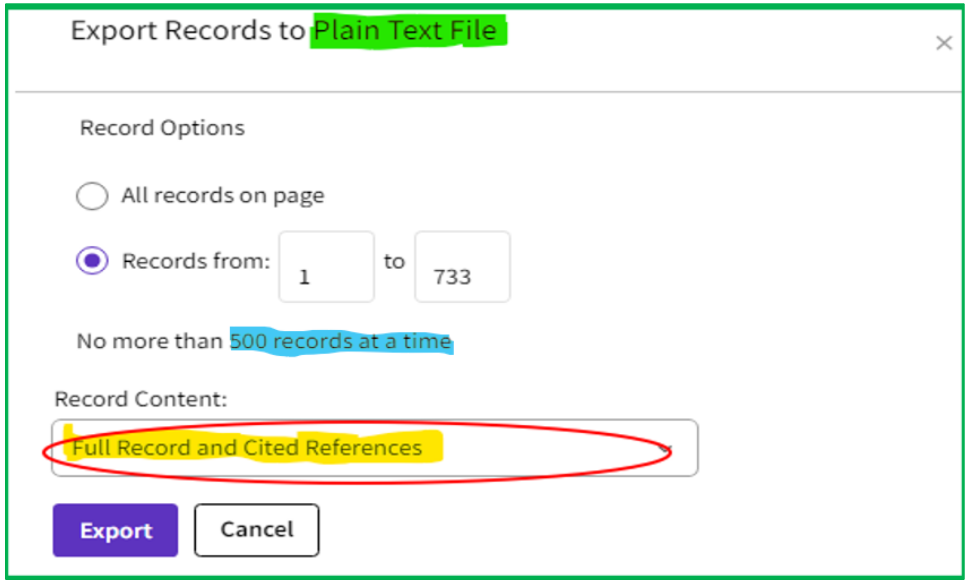


Results for "Passive invest*" OR "Index* invest*" OR "Passive manag*" (Topic)

733 results from Web of Science Core Collection for:

"Passive invest*" OR "Index* invest*" OR "Passive manag*" (Topic)

Selecting all available information and exporting it in batches of 500 records each



Export Records to Plain Text File

Record Options

○ All records on page

● Records from: 1 to 733

No more than 500 records at a time

Record Content: Full Record and Cited References

User Guide
Table of contents

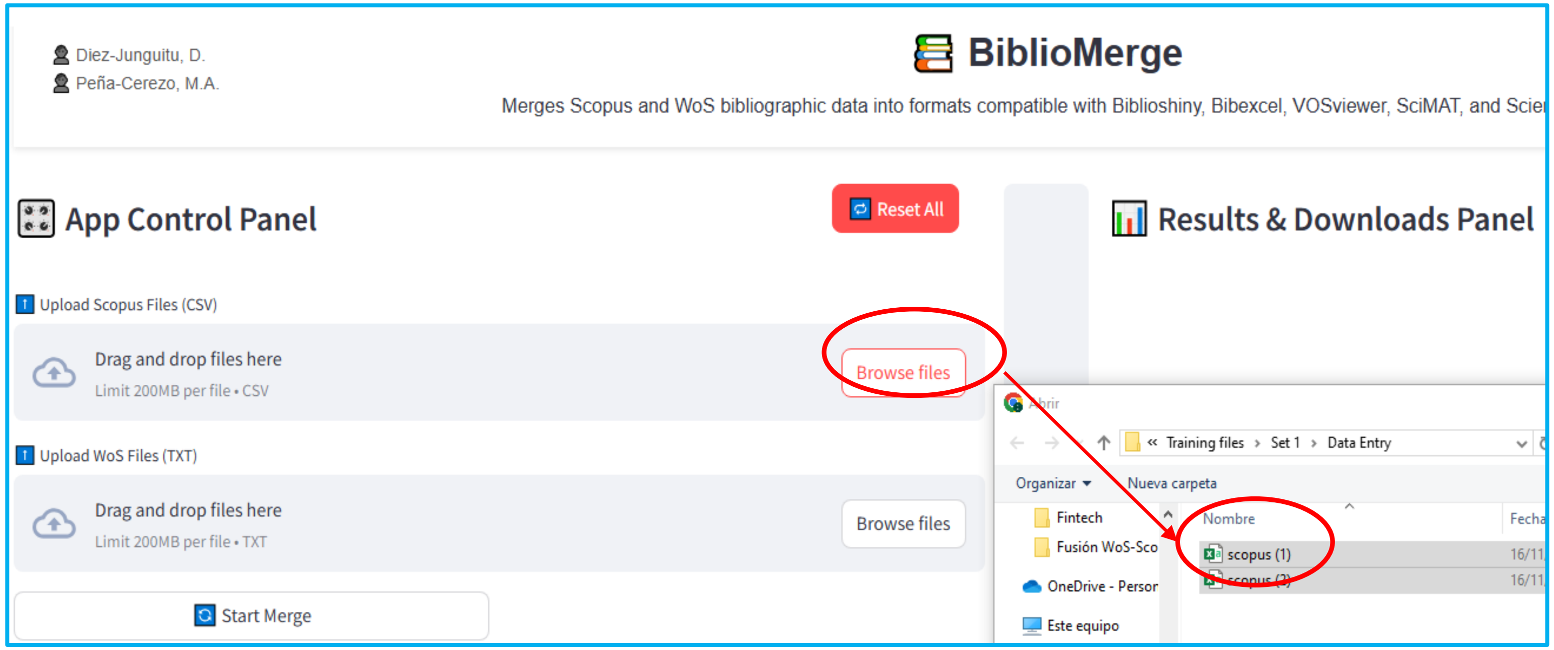
- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
  - ❑ Authors debugging
  - ❑ Author and Index Keywords debugging
  - ❑ Cited References debugging
  - ❑ Final Step
- Final files and reports generation

- Authors debugging

User Guide
Charge of Scopus and WoS files process

[Back to index](#)

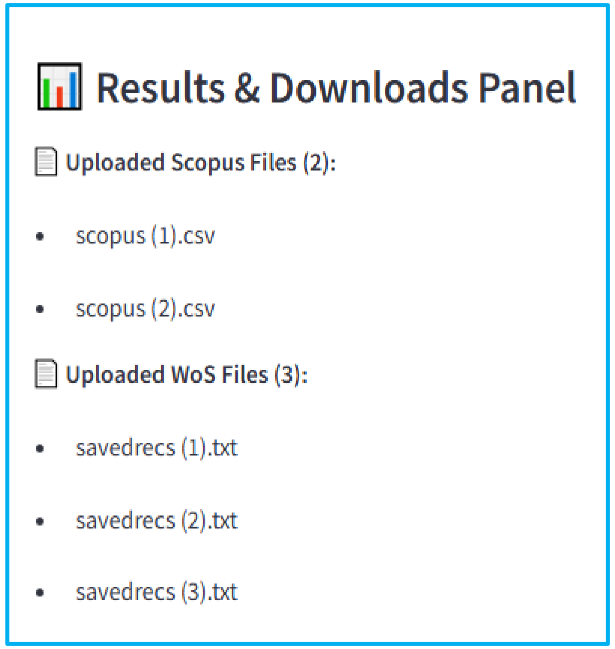
Clicking **Browse files** options the Explorer will help to guide you to the entry files


Diez-Junguitu, D.
Peña-Cerezo, M.A.

BiblioMerge

Merges Scopus and WoS bibliographic data into formats compatible with Biblioshiny, Bibexcel, VOSviewer, SciMAT, and Scie

You will have a **notice** when the upload is completed


Results & Downloads Panel

Uploaded Scopus Files (2):

- scopus (1).csv
- scopus (2).csv

Uploaded WoS Files (3):

- savedrecs (1).txt
- savedrecs (2).txt
- savedrecs (3).txt

User Guide
Table of contents

- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
    - ❑ Authors debugging
    - ❑ Author and Index Keywords debugging
    - ❑ Cited References debugging
    - ❑ Final Step
- Final files and reports generation

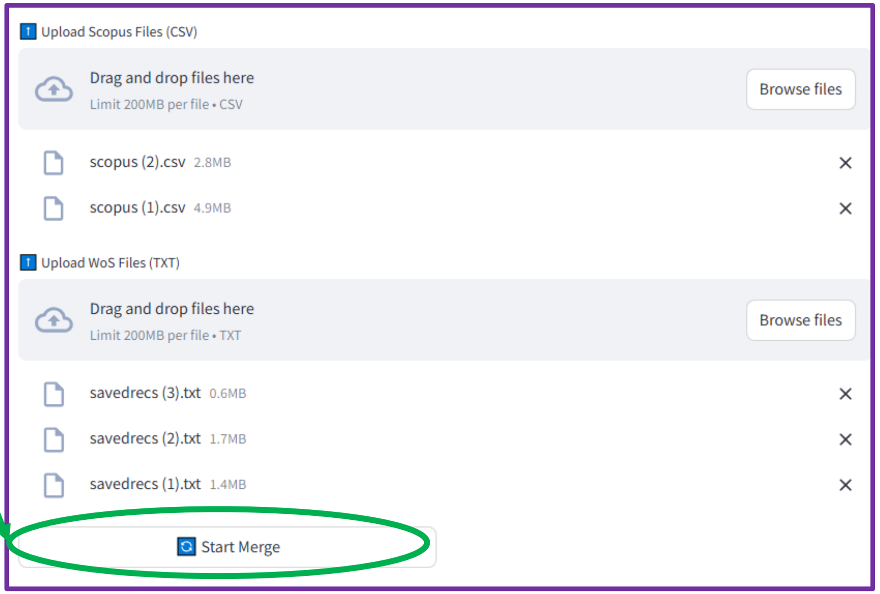
- Authors debugging

User Guide

Merging execution and first report and outputs

Back to index

After uploading Scopus and WoS records, **Start Merge** should be clicked to continue the merging process,



Upload Scopus Files (CSV)

Drag and drop files here

Limit 200MB per file • CSV

Browse files

scopus (2).csv 2.8MB ×

scopus (1).csv 4.9MB ×

Upload WoS Files (TXT)

Drag and drop files here

Limit 200MB per file • TXT

Browse files

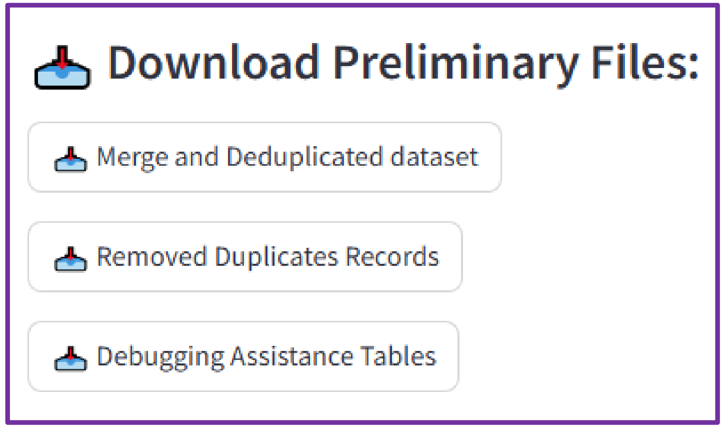
savedrecs (3).txt 0.6MB ×

savedrecs (2).txt 1.7MB ×

savedrecs (1).txt 1.4MB ×

Start Merge

and after a few seconds, the process will have been concluded and generated 3 preliminary files that can be downloaded into your device…..



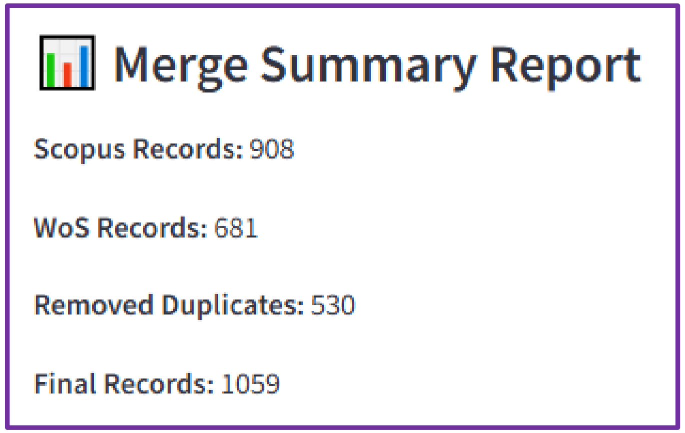
Download Preliminary Files:

Merge and Deduplicated dataset

Removed Duplicates Records

Debugging Assistance Tables

…and a summary report of the merge and some grahps of key parameters



Merge Summary Report

Scopus Records: 908

WoS Records: 681

Removed Duplicates: 530

Final Records: 1059

User Guide
Table of contents

- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
  - Authors debugging
  - Author and Index Keywords debugging
  - Cited References debugging
  - Final Step
- Final files and reports generation

- Authors debugging

User Guide

Debugging of Authors, Keywords and Cited References process

Back to index

Meaning of Debugging:

- We have a database of scientific documents over which we plan to perform bibliometric analysis
- For the analysis it will be needed to account the **authors, keywords, and cited references** in their respective fields in the database, and stablish metrics and connections
- Frequently, these concepts in the data base are expressed in different orthographies or in synonymous words. As for example:
  - For Authors: “Cheng Y” could be the same orthography for two different Authors
  - For Keywords: “Trade”, “Trading”, “trading strategy”, “trading strategies”, could be synonymous to “trading” from a particular point of view
- The object of Debugging is to disambiguate, harmonize o group these situations in order to get the clearest interpretation of the analysis results

Debugging process:

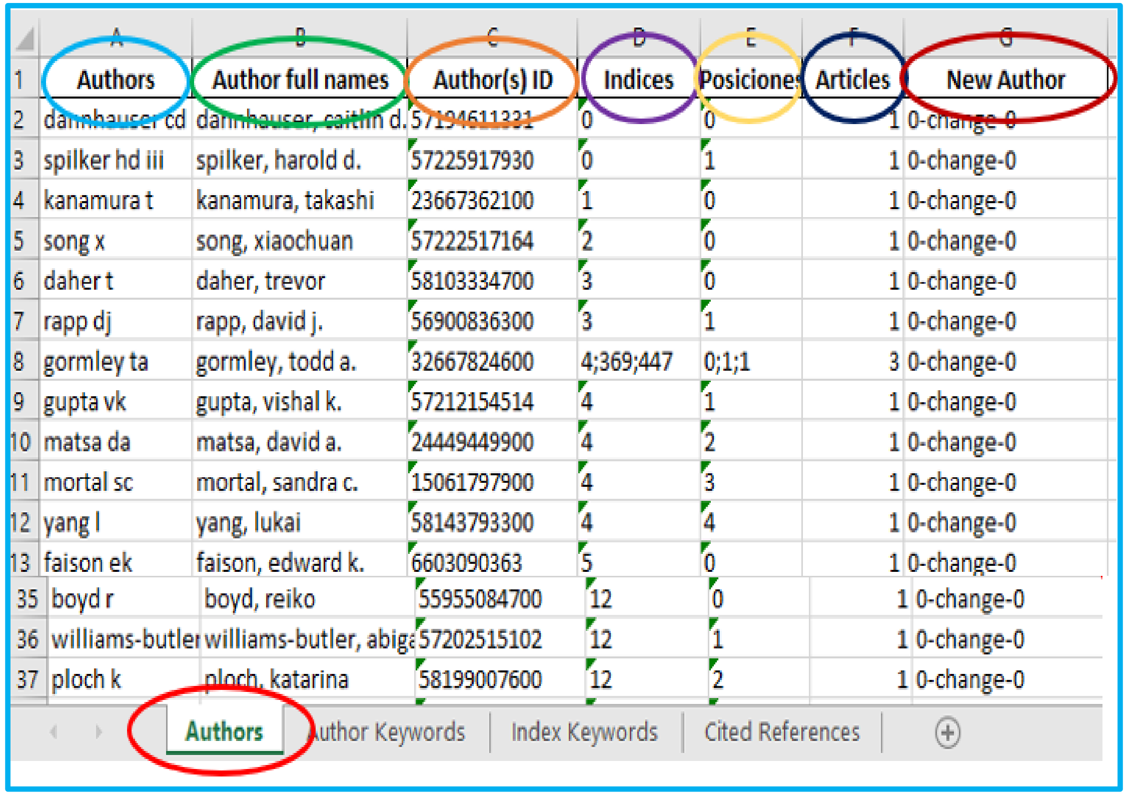
- The process consist in three steps:
  1. **Identify** the situations to be disambiguated, harmonized or grouped
  2. **Establish the association** of these situation with their alternative, decided by the researcher, in an Excel file
  3. Incorporate this association in the App to **replace** the old words (group of words) by the new ones in the database
- For facilitate the three steps, in the Merging process, it was generated the Excel file “**Debugging-Tables.xlsx**”. It has four sheets, for the managing of **Authors**, **Author Keywords**, **Index Keywords**, and **Cited References** debugging
- Although the general process is the same for the four concepts, it would be explained one by one as some particularities could be worth to focus the identification step in a different way

User Guide

Debugging of Authors, Keywords and Cited References process

Back to index

Authors debugging:



	A	B	C	D	E	F	G
1	Authors	Author full names	Author(s) ID	Indices	Posiciones	Articles	New Author
2	dannhauser cd	dannhauser, caitlin d.	57194611331	0	0	1	0-change-0
3	spilker hd iii	spilker, harold d.	57225917930	0	1	1	0-change-0
4	kanamura t	kanamura, takashi	23667362100	1	0	1	0-change-0
5	song x	song, xiaochuan	57222517164	2	0	1	0-change-0
6	daher t	daher, trevor	58103334700	3	0	1	0-change-0
7	rapp dj	rapp, david j.	56900836300	3	1	1	0-change-0
8	gormley ta	gormley, todd a.	32667824600	4;369;447	0;1;1	3	0-change-0
9	gupta vk	gupta, vishal k.	57212154514	4	1	1	0-change-0
10	matsa da	matsa, david a.	24449449900	4	2	1	0-change-0
11	mortal sc	mortal, sandra c.	15061797900	4	3	1	0-change-0
12	yang l	yang, lukai	58143793300	4	4	1	0-change-0
13	faison ek	faison, edward k.	6603090363	5	0	1	0-change-0
35	boyd r	boyd, reiko	55955084700	12	0	1	0-change-0
36	williams-butler	williams-butler, abiga	57202515102	12	1	1	0-change-0
37	ploch k	ploch, katarina	58199007600	12	2	1	0-change-0

Fields explanation:

- **Authors**: correspond to the list of different authors in the Authors field of the database
- **Authors full names**: correspond to the long name of the author. It will help to distinguish different authors with the same short name
- **Author(s) ID**: it is an identification code for each individual author in Scopus databases. In base to this ID, previous debugging was already performed in merging execution, so no need for debugging in the "Scopus part" of the merged database, and only need to check the WoS origin records
- **Indices**: it points out in which records of the database appears the corresponding author
- **Posiciones**: it points out in which position within the Author field of the database appears the corresponding author
- **Articles**: number of articles in which it appears the corresponding author
- **New Author**: it is the field where the alternative word or group of words should be place, and that will be the one/s that will replace the word or group of words appearing in the "Author" field. This field should be filled if necessary or the whole line eliminated (if the application find the message **"0–change-0"** any debugging would be performed)

User Guide

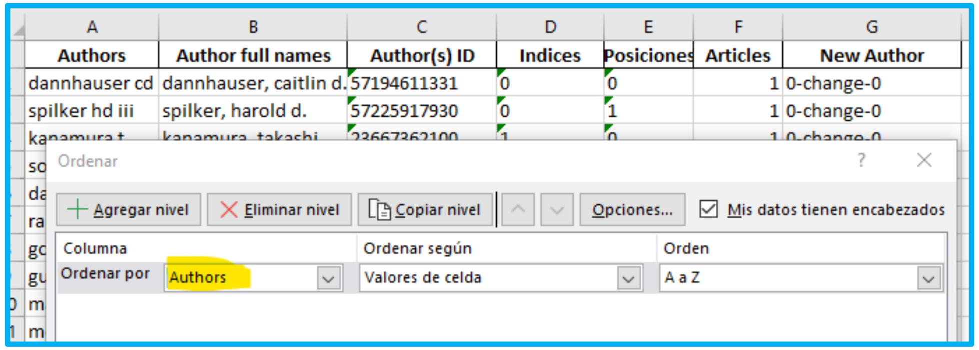
Debugging of Authors, Keywords and Cited References process

Back to index

Authors debugging:

Recommended procedure:

1. Order the Excel table by alphabetically by Author



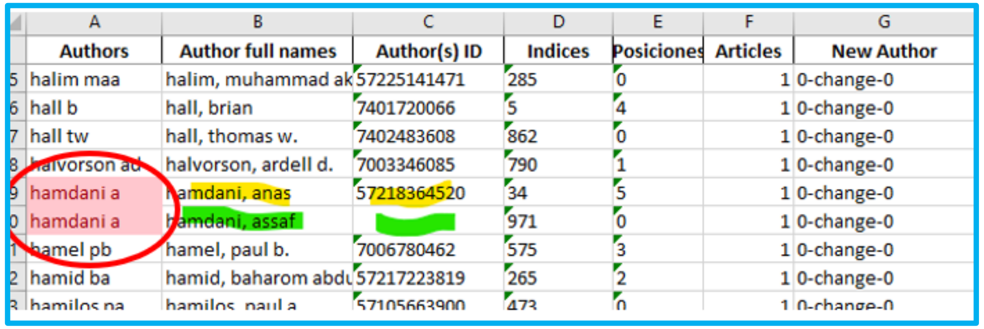
Authors	Author full names	Author(s) ID	Indices	Posiciones	Articles	New Author
dannhauser cd	dannhauser, caitlin d.	57194611331	0	0	1	0-change-0
spilker hd iii	spilker, harold d.	57225917930	0	1	1	0-change-0
kanamura t	kanamura, takashi	23667362100	1	0	1	0-change-0

Ordenar

☑ Mis datos tienen encabezados

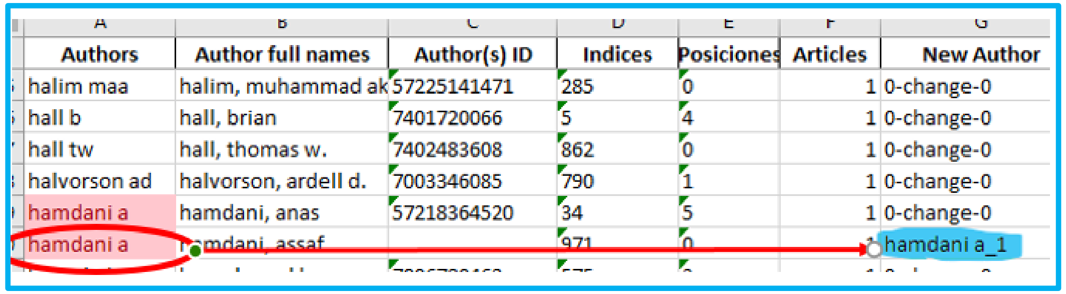
Columna		Ordenar según	Orden
Ordenar por	Authors	Valores de celda	A a Z

2. Pay attention to lines with Author's(ID) empty field, and check if in the rows before or after there are authors with identical orthographies. Help yourself with Excel utility for identifying duplicates (Conditional Formats) in Authors column



Authors	Author full names	Author(s) ID	Indices	Posiciones	Articles	New Author
halim maa	halim, muhammad ak	57225141471	285	0	1	0-change-0
hall b	hall, brian	7401720066	5	4	1	0-change-0
hall tw	hall, thomas w.	7402483608	862	0	1	0-change-0
halvorson ad	halvorson, ardell d.	7003346085	790	1	1	0-change-0
hamdani a	hamdani, anas	57218364520	34	5	1	0-change-0
hamdani a	hamdani, assaf		971	0	1	0-change-0
hamel pb	hamel, paul b.	7006780462	575	3	1	0-change-0
hamid ba	hamid, baharom abdu	57217223819	265	2	1	0-change-0
hamilos pa	hamilos, paul a	57105663900	473	0	1	0-change-0

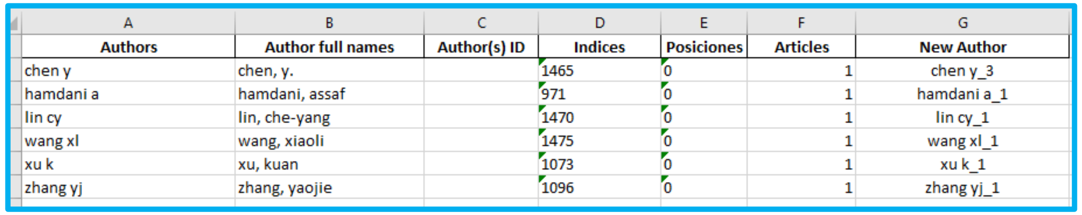
3. Choose an alternative name to disambiguate one of them, and assign it in the 'New Author' field



Authors	Author full names	Author(s) ID	Indices	Posiciones	Articles	New Author
halim maa	halim, muhammad ak	57225141471	285	0	1	0-change-0
hall b	hall, brian	7401720066	5	4	1	0-change-0
hall tw	hall, thomas w.	7402483608	862	0	1	0-change-0
halvorson ad	halvorson, ardell d.	7003346085	790	1	1	0-change-0
hamdani a	hamdani, anas	57218364520	34	5	1	0-change-0
hamdani a	hamdani, assaf		971	0	1	hamdani a_1

4. Erase all the rows that have not been associated, that are the ones that still have the original legend "o-change-0"

The final aspect of the table should be like that:

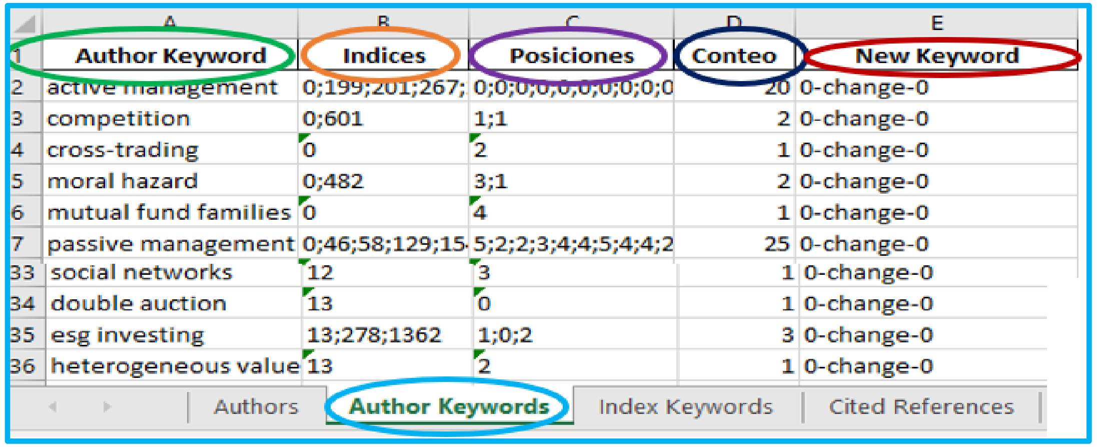


Authors	Author full names	Author(s) ID	Indices	Posiciones	Articles	New Author
chen y	chen, y.		1465	0	1	chen y_3
hamdani a	hamdani, assaf		971	0	1	hamdani a_1
lin cy	lin, che-yang		1470	0	1	lin cy_1
wang xl	wang, xiaoli		1475	0	1	wang xl_1
xu k	xu, kuan		1073	0	1	xu k_1
zhang yj	zhang, yaojie		1096	0	1	zhang yj_1

User Guide

Debugging of Authors, Keywords and Cited References process

Author Keywords debugging (identical procedure for Index Keywords):



	A	B	C	D	E
1	Author Keyword	Indices	Posiciones	Conteo	New Keyword
2	active management	0;199;201;267;	0;0;0;0;0;0;0;0;0;0	20	0-change-0
3	competition	0;601	1;1	2	0-change-0
4	cross-trading	0	2	1	0-change-0
5	moral hazard	0;482	3;1	2	0-change-0
6	mutual fund families	0	4	1	0-change-0
7	passive management	0;46;58;129;15	5;2;2;3;4;4;5;4;4;2	25	0-change-0
33	social networks	12	3	1	0-change-0
34	double auction	13	0	1	0-change-0
35	esg investing	13;278;1362	1;0;2	3	0-change-0
36	heterogeneous value	13	2	1	0-change-0

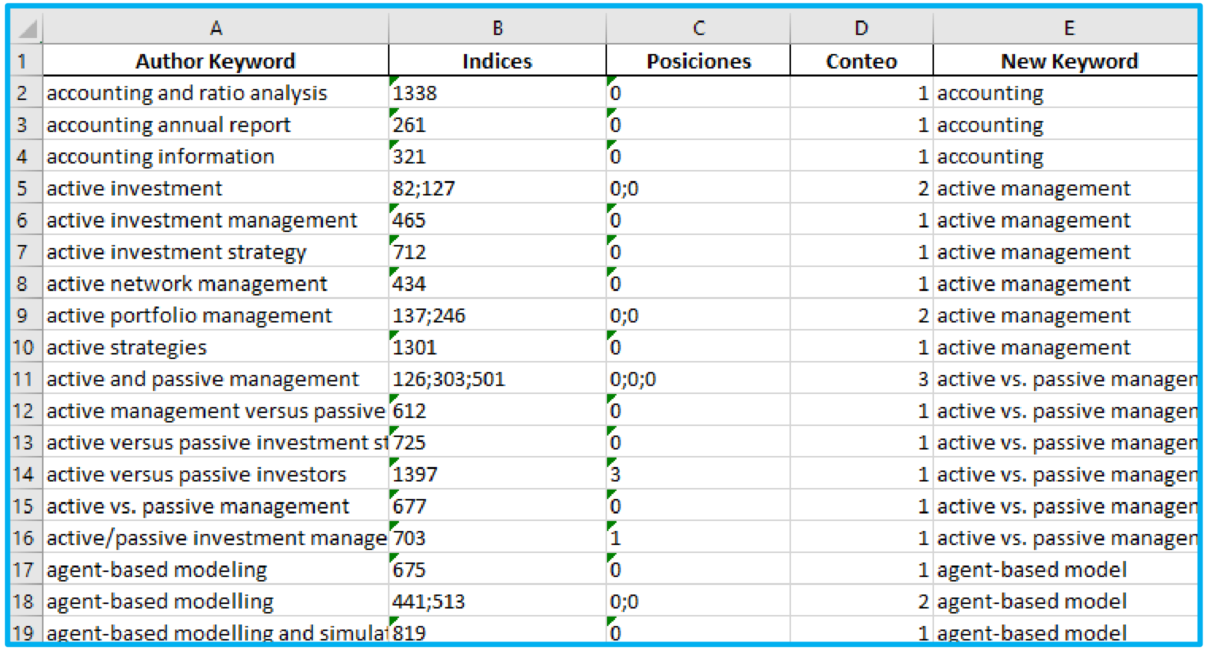
Fields explanation:

- **Author Keyword**: correspond to the list of different author keywords in the Author Keywords field of the database
- **Indices**: The concept is the same as in authors debugging
- **Posiciones**: The concept is the same as in authors debugging
- **Count**: number of timer the corresponding author keyword appears in the database
- **New Keyword**: it is the field where the alternative word or group of words should be place, and that will be the one/s that will replace the word or group of words appearing in the "Author Keyword" field. This field should be filled if necessary or the whole line eliminated (if the application find the message "o– change-0" any debugging would be performed)

Recommended procedure:

1. Use ordering by "Count" (number of appearances) or/and by "Author Keyword" alphabetically to find similar orthographies or synonymous
2. Chose one of the orthographies and associate it to the other one in the field "New Keyword"
3. Erase all the rows that have not been associated, that are the ones that still have the original legend "0–change-0"

The final aspect of the table should be like that:

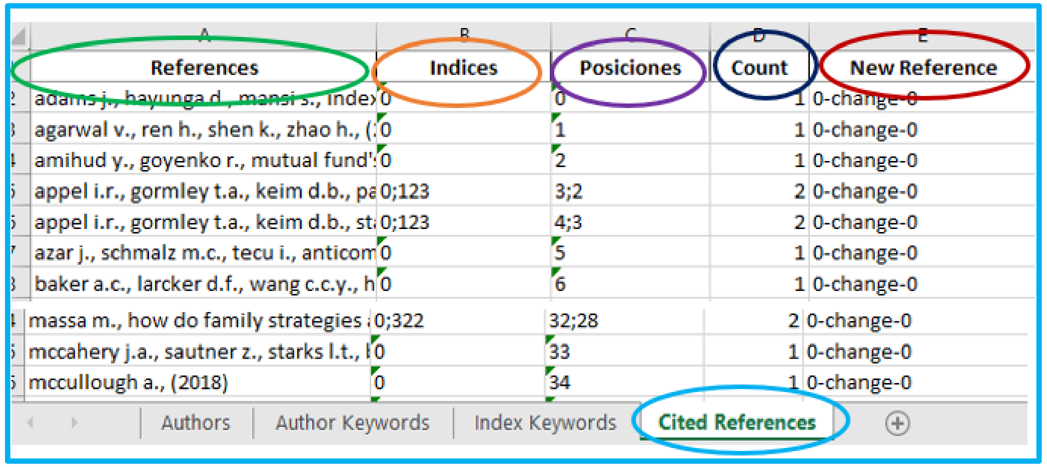


	A	B	C	D	E
1	Author Keyword	Indices	Posiciones	Conteo	New Keyword
2	accounting and ratio analysis	1338	0	1	accounting
3	accounting annual report	261	0	1	accounting
4	accounting information	321	0	1	accounting
5	active investment	82;127	0;0	2	active management
6	active investment management	465	0	1	active management
7	active investment strategy	712	0	1	active management
8	active network management	434	0	1	active management
9	active portfolio management	137;246	0;0	2	active management
10	active strategies	1301	0	1	active management
11	active and passive management	126;303;501	0;0;0	3	active vs. passive managen
12	active management versus passive	612	0	1	active vs. passive managen
13	active versus passive investment st	725	0	1	active vs. passive managen
14	active versus passive investors	1397	3	1	active vs. passive managen
15	active vs. passive management	677	0	1	active vs. passive managen
16	active/passive investment manage	703	1	1	active vs. passive managen
17	agent-based modeling	675	0	1	agent-based model
18	agent-based modelling	441;513	0;0	2	agent-based model
19	agent-based modelling and simulat	819	0	1	agent-based model

User Guide

Debugging of Authors, Keywords and Cited References process

Cited References debugging



References	Indices	Posiciones	Count	New Reference
adams j., hayunga d., mansi s., index	0	0	1	0-change-0
agarwal v., ren h., shen k., zhao h., (	0	1	1	0-change-0
amihud y., goyenko r., mutual fund'	0	2	1	0-change-0
appel i.r., gormley t.a., keim d.b., pa	0;123	3;2	2	0-change-0
appel i.r., gormley t.a., keim d.b., st	0;123	4;3	2	0-change-0
azar j., schmalz m.c., tecu i., anticom	0	5	1	0-change-0
baker a.c., larcker d.f., wang c.c.y., h	0	6	1	0-change-0
massa m., how do family strategies	0;322	32;28	2	0-change-0
mccahery j.a., sautner z., starks l.t.,	0	33	1	0-change-0
mccullough a., (2018)	0	34	1	0-change-0

Fields explanation:

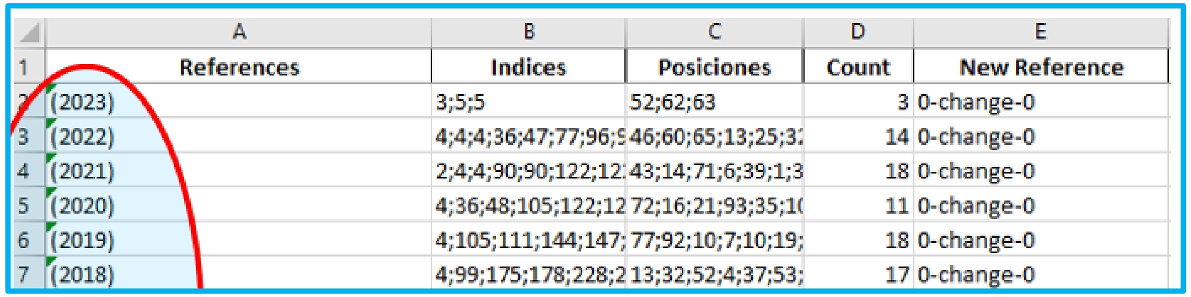
- **References**: correspond to the list of different cited references in the correspondent field of the database. Typical structure: "authors, title, journal, etc"
- **Rests of fields**: similar structure and meaning than in previous cases
- **New Reference**: This field should be filled if necessary or the whole line eliminated (if the application find the message "aaa – add equivalence or erase the row" any debugging would be performed)

Special characteristics:

1. It uses to be a large file, with around 50 K rows
2. It uses to contain "trash" row that have nothing to do with bibliographical references

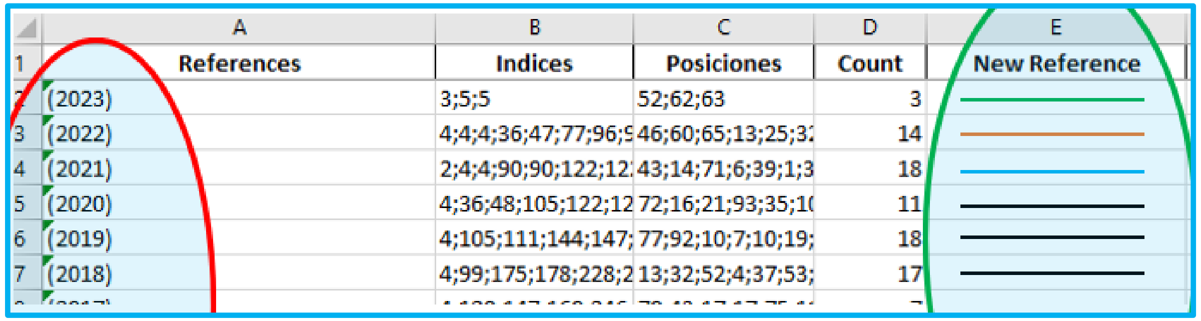
Recommended procedure:

1. Order by "References" alphabetically to find this "trash" rows, for example:



	A	B	C	D	E
1	References	Indices	Posiciones	Count	New Reference
2	(2023)	3;5;5	52;62;63	3	0-change-0
3	(2022)	4;4;4;36;47;77;96;9	46;60;65;13;25;3	14	0-change-0
4	(2021)	2;4;4;90;90;122;12	43;14;71;6;39;1;3	18	0-change-0
5	(2020)	4;36;48;105;122;12	72;16;21;93;35;1	11	0-change-0
6	(2019)	4;105;111;144;147;	77;92;10;7;10;19;	18	0-change-0
7	(2018)	4;99;175;178;228;2	13;32;52;4;37;53;	17	0-change-0

2. Empty the field "New Reference", and so, the "trash" data will be erased in the data base



	A	B	C	D	E
1	References	Indices	Posiciones	Count	New Reference
2	(2023)	3;5;5	52;62;63	3	
3	(2022)	4;4;4;36;47;77;96;9	46;60;65;13;25;3	14	
4	(2021)	2;4;4;90;90;122;12	43;14;71;6;39;1;3	18	
5	(2020)	4;36;48;105;122;12	72;16;21;93;35;1	11	
6	(2019)	4;105;111;144;147;	77;92;10;7;10;19;	18	
7	(2018)	4;99;175;178;228;2	13;32;52;4;37;53;	17	

User Guide

Debugging of Authors, Keywords and Cited References process

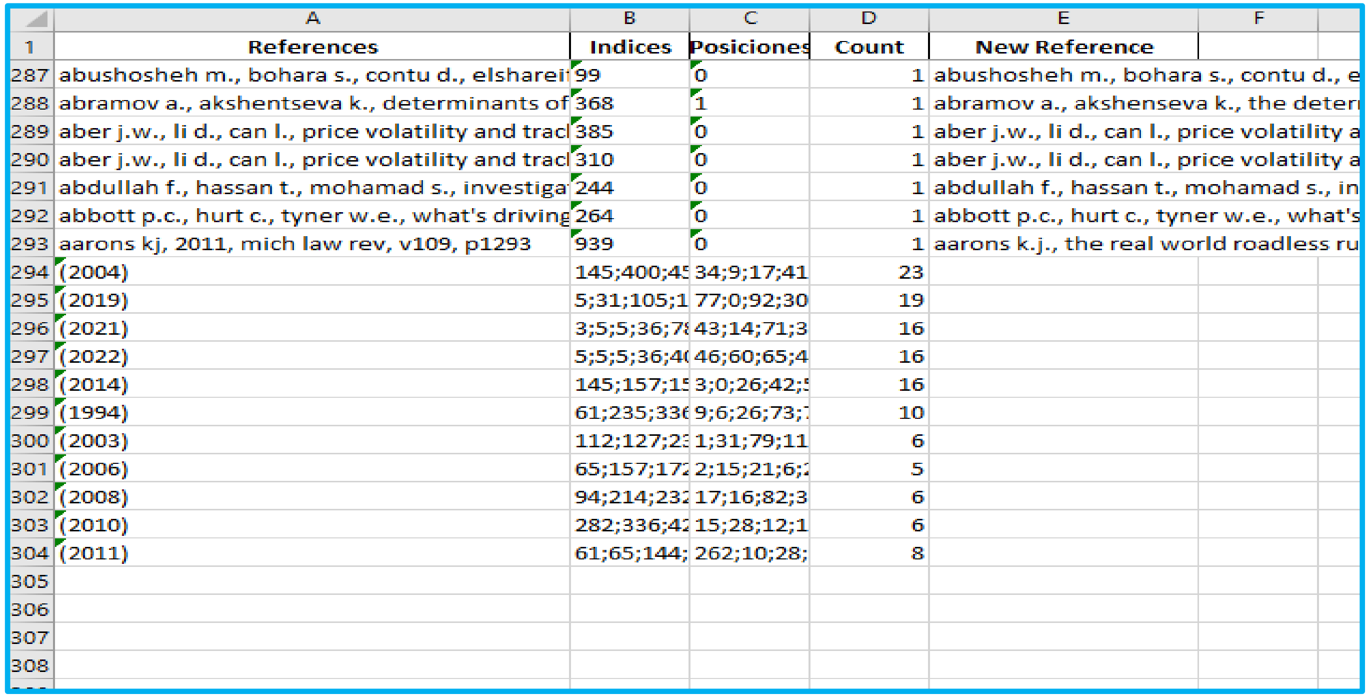
Back to index

Cited References debugging

Recommended procedure:

3. Putting apart these treated rows, it could be ordered the rest by "count" to try to identify the relevant references and get a workable number of rows
4. Order alphabetically by "References" and identify the different orthographies for a same reference
5. Chose one of the orthographies and associate it to the other one in the field "New Reference"
6. Join this part with the one worked in the point number 2

The final aspect of the table should be like that



	A	B	C	D	E	F
1	References	Indices	Posiciones	Count	New Reference	
287	abushosheh m., bohara s., contu d., elsharei	99	0	1	abushosheh m., bohara s., contu d., e	
288	abramov a., akshentseva k., determinants of	368	1	1	abramov a., akshenseva k., the deter	
289	aber j.w., li d., can l., price volatility and trac	385	0	1	aber j.w., li d., can l., price volatility a	
290	aber j.w., li d., can l., price volatility and trac	310	0	1	aber j.w., li d., can l., price volatility a	
291	abdullah f., hassan t., mohamad s., investiga	244	0	1	abdullah f., hassan t., mohamad s., in	
292	abbott p.c., hurt c., tyner w.e., what's driving	264	0	1	abbott p.c., hurt c., tyner w.e., what's	
293	aarons kj, 2011, mich law rev, v109, p1293	939	0	1	aarons k.j., the real world roadless ru	
294	(2004)	145;400;45	34;9;17;41	23		
295	(2019)	5;31;105;1	77;0;92;30	19		
296	(2021)	3;5;5;36;78	43;14;71;3	16		
297	(2022)	5;5;5;36;40	46;60;65;4	16		
298	(2014)	145;157;15	3;0;26;42;5	16		
299	(1994)	61;235;336	9;6;26;73;7	10		
300	(2003)	112;127;23	1;31;79;11	6		
301	(2006)	65;157;172	2;15;21;6;2	5		
302	(2008)	94;214;232	17;16;82;3	6		
303	(2010)	282;336;42	15;28;12;1	6		
304	(2011)	61;65;144;	262;10;28;	8		
305						
306						
307						
308						

User Guide

Debugging of Authors, Keywords and Cited References process

Back to index

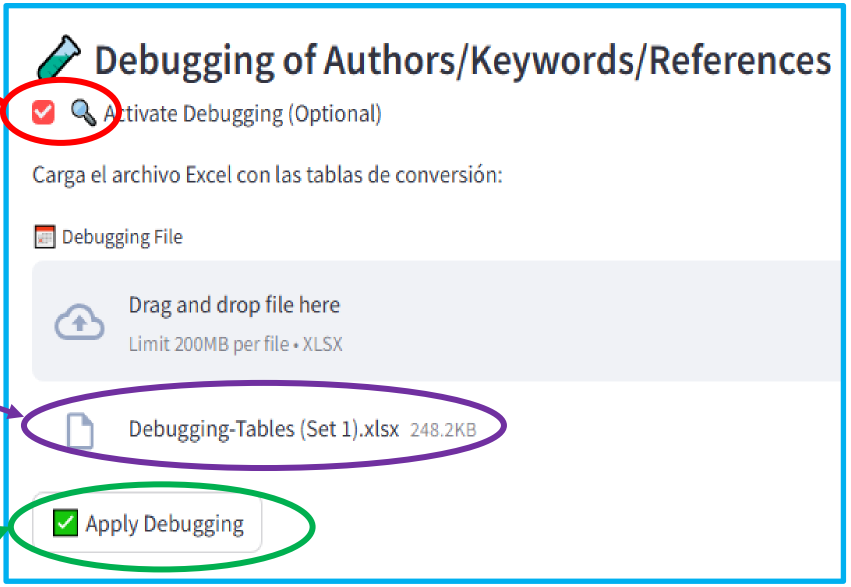
Final Step:

- Once established the associations on any or all the four sheets, the file "Debugging-Tables.xlsx" used for this task should be saved with either this name or a different one.

- Next, we will go to the App, clicking "**Activate Debugging**"…

- Then we will upload the file

- …and clicking the "**Apply Debugging**" the process will be executed



Debugging of Authors/Keywords/References

☑ 🔍 Activate Debugging (Optional)

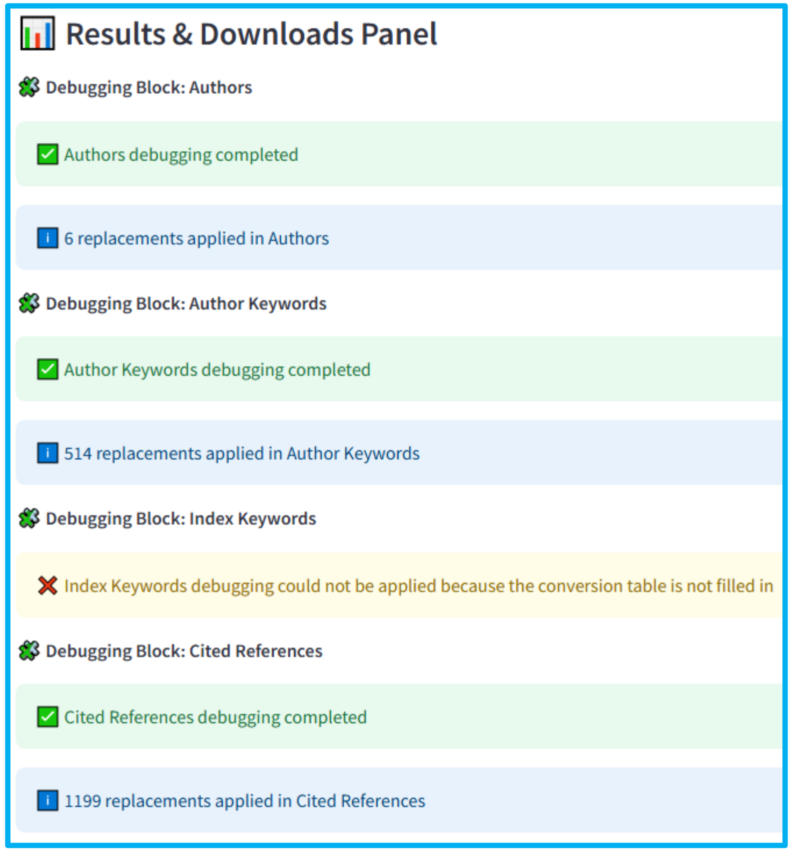
Carga el archivo Excel con las tablas de conversión:

Debugging File

Drag and drop file here
Limit 200MB per file • XLSX

Debugging-Tables (Set 1).xlsx 248.2KB

✅ Apply Debugging



Results & Downloads Panel

Debugging Block: Authors

✅ Authors debugging completed

ℹ 6 replacements applied in Authors

Debugging Block: Author Keywords

✅ Author Keywords debugging completed

ℹ 514 replacements applied in Author Keywords

Debugging Block: Index Keywords

❌ Index Keywords debugging could not be applied because the conversion table is not filled in

Debugging Block: Cited References

✅ Cited References debugging completed

ℹ 1199 replacements applied in Cited References

User Guide
Table of contents

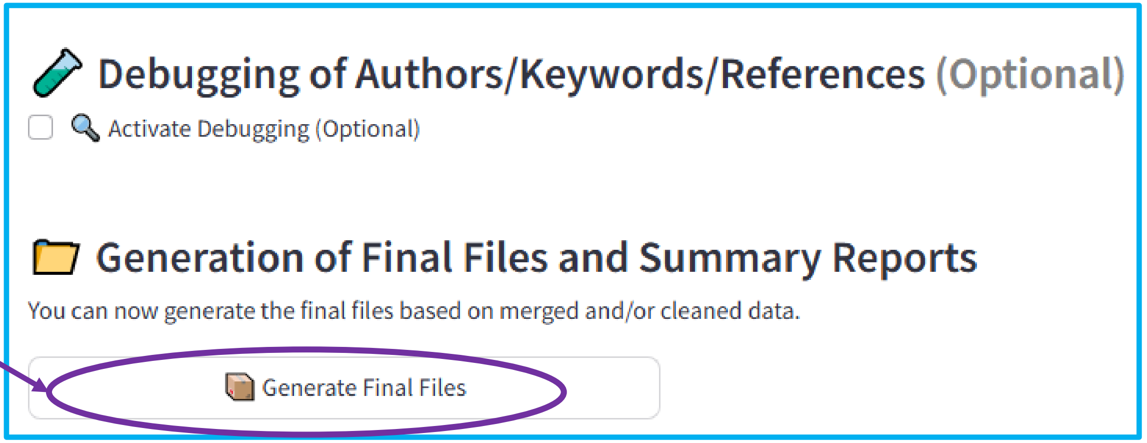
- General process view
- Preliminary steps
- Charge of Scopus and WoS files process
- Merging execution and first report and outputs
- Debugging of Authors, Keywords and Cited References
    - Authors debugging
    - Author and Index Keywords debugging
    - Cited References debugging
    - Final Step
- Final files and reports generation

- Authors debugging

User Guide
Final files and reports generation

Back to index

- By activating this "**Generate Final Files**", **with or without debugging** …



🧪 Debugging of Authors/Keywords/References (Optional)

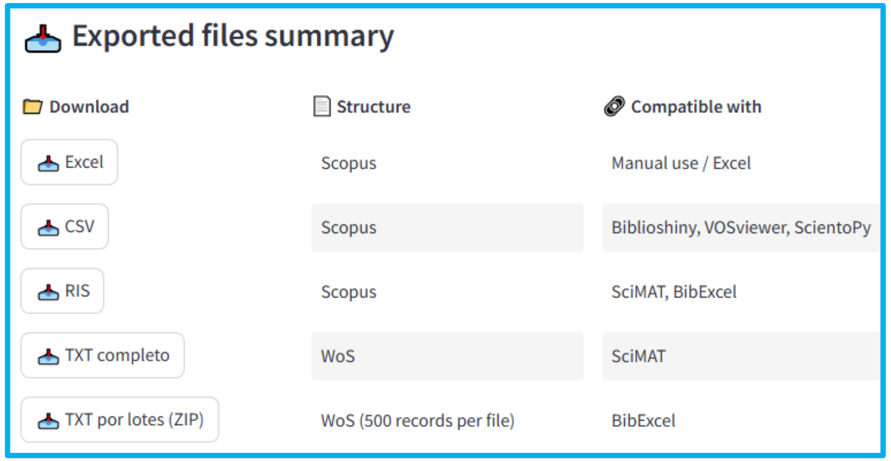
□ 🔍 Activate Debugging (Optional)

📁 Generation of Final Files and Summary Reports

You can now generate the final files based on merged and/or cleaned data.

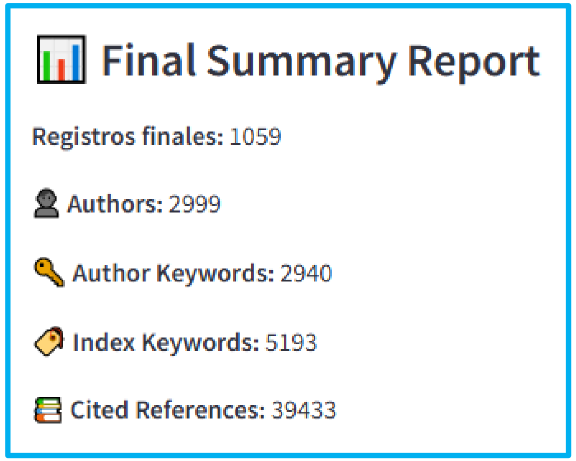
…the following files will be generated and can be downloaded in your device:

…and some general information:



📥 Exported files summary

📁 Download	📄 Structure	🔗 Compatible with
📥 Excel	Scopus	Manual use / Excel
📥 CSV	Scopus	Biblioshiny, VOSviewer, ScientoPy
📥 RIS	Scopus	SciMAT, BibExcel
📥 TXT completo	WoS	SciMAT
📥 TXT por lotes (ZIP)	WoS (500 records per file)	BibExcel



📊 Final Summary Report

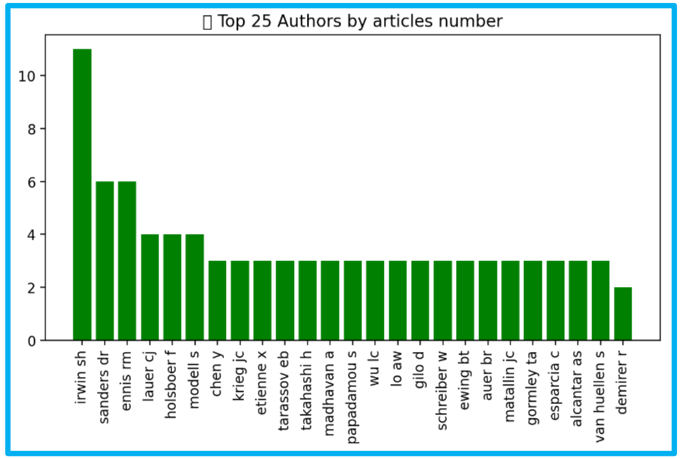
Registros finales: 1059

👤 **Authors:** 2999

🔑 **Author Keywords:** 2940

🏷 **Index Keywords:** 5193

📚 **Cited References:** 39433



The End

Thanks for your attention

BiblioMerge

BiblioMerge: A Python-based automated tool to merge WoS and Scopus bibliographic data, compatible with Biblioshiny, Bibexcel, VOSviewer, SciMAT and ScientoPy

David Diez-Junguitu, Miguel Á. Peña-Cerezo

BiblioMergeApp@gmail.com