

Bibliominer

Cleaning guide

Before and after, step by step, on a real corpus

A Scopus export is not a dataset. DOIs are missing, journals are empty, authors are shifted from one column to another, affiliations are written ten ways for the same institution, and the references are free text that no co-citation analysis can read.

This guide follows a real file of 128 articles, from the moment it is imported to the delivered corpus. For each step it shows two screens: the problem as it is in the file (**before**), then the result (**after**). In between, it says what Bibliominer does by itself (which API it queries, which rule it applies) and what is left to your judgement.

Contents

1	The corpus of this guide	2
1.1	What the cleaning changed	2
1.2	The role of Bibliominer, on one page	2
1.3	The five principles	3
2	Import and corpus review	3
2.1	The problem in the file	3
3	Step 1: DOI	6
3.1	The problem in the file	6
4	Step 2: Metadata	8
4.1	The problem in the file	8
5	Step 3: Sources	10
5.1	The problem in the file	10
6	Step 4: Authors	11
6.1	What Bibliominer completes by itself	11
6.2	Authors with affiliations (AWA)	12
6.2.1	The problem	12
7	Step 5: Affiliations	14
7.1	The problem in the file	14
8	Step 6: Text	21
8.1	The problem in the file	21
9	Step 7: References	22
9.1	The problem in the file	22
10	Correcting any value: All documents	24
11	Final: the deliverable file	26
11.1	What you get	27

1 The corpus of this guide

All the screenshots and all the examples come from the same file: `merged.csv`, a Scopus export of **128 articles** (2010–2026), cleaned end to end with Bibliominer. The **before** screenshots show this file right after the import; the **after** screenshots, the same file once each step has been applied. Nothing has been retouched: these are the screens as the application displays them.

1.1 What the cleaning changed

The end-of-cleaning report (section 11) sums up the whole workflow. These are the figures quoted in the Methods section of an article.

	Imported file	Analysed corpus	Cleaned file
Articles	128	126	126
Distinct authors	141	140	138
Affiliations (distinct spellings)	222	216	67
Articles without DOI	7	5	1
Articles without keywords	16	14	0
Author keywords (total)	538	538	572
References (total)	6224	6215	6520

Analysed corpus = the imported file without its documents that are not research (an editorial here, section 2) and without the article the user removed during the corpus review.

1.2 The role of Bibliominer, on one page

Step	The problem	What Bibliominer does	Result
Import	Non-research documents, duplicates, incomplete articles	Rule 0 (research types), duplicate detection, corpus review with incomplete articles first	1 editorial removed
DOI	5 articles without an identifier	Search by title in Crossref and OpenAlex , acceptance of verified matches only	5 → 1
Metadata	2 missing years	Lists them and locks the step as long as a year is missing	2 → 0
Sources	2 journals without a name	Matching by ISSN with the SCImago index	2 → 0
Authors	Empty or shifted author columns, misassigned affiliations (AWA)	Completes from the corpus itself (Scopus identifiers), suggests each author's affiliation	0 incomplete author
Affiliations	216 spellings for 67 organisations	City and country via GeoNames , organisation via ROR , splitting into blocks, merge suggestions	216 → 67
Text	14 articles without keywords	Suggests keywords drawn from the corpus vocabulary	14 → 0
References	Free text unreadable for co-citation	Reconciliation with Crossref , OpenAlex (and Semantic Scholar)	6065 → 6534

1.3 The five principles

1. **Nothing is decided in your place.** Each suggestion shows what it relies on (an API, an identifier, a rule), and you accept or refuse it. Only what is *already written in the file* is taken over without asking, and the screen says so.
2. **Nothing disappears silently.** A removed document is written to `removed_articles.csv` with its reason; a document removed by mistake is put back in one click.
3. **The file is the backup.** The corpus can be downloaded at any time, and imported again later to resume.
4. **Your keys, your quota.** GeoNames, OpenAlex and Semantic Scholar are called with your own credentials, entered on the import screen.
5. **Every figure is traceable.** The final report says where the corpus starts from, what was removed, recovered, and with which tool.

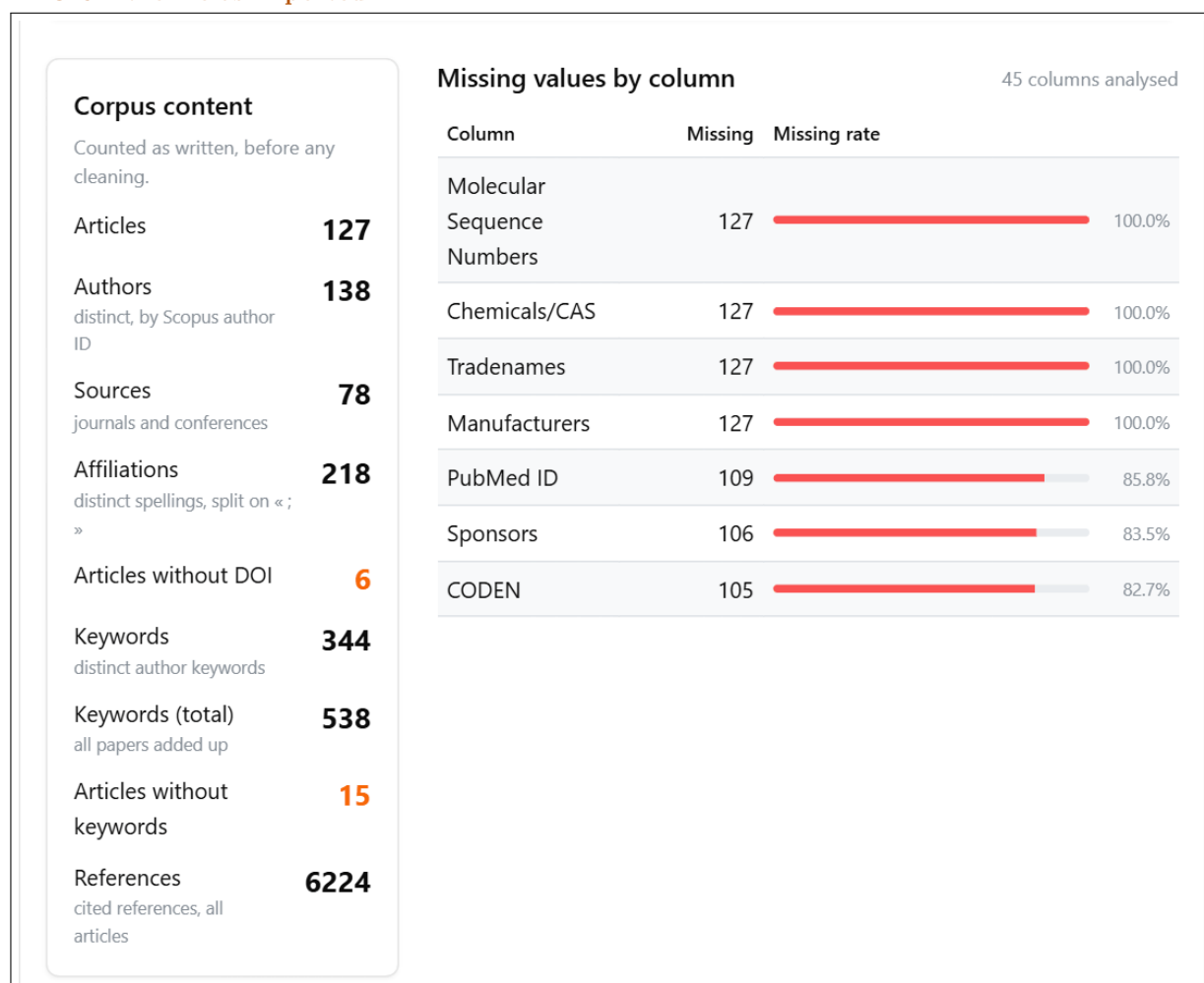
Why

A corpus that cannot be accounted for is a corpus that cannot be defended. An automatic merge of affiliations would confuse two institutions that only share a name, and the error would be invisible in the final corpus.

2 Import and corpus review

2.1 The problem in the file

The import screen reads the file without changing anything and says what it contains: number of articles, authors, sources, affiliations, articles without DOI, without keywords, total number of references, and the missing values column by column.

BEFORE: the file as imported

The imported file: 127 research articles (the 128th, an editorial, is already set aside), 218 affiliation spellings, 6 articles without DOI, 15 without keywords. The right-hand column gives the missing values per column.

What Bibliominer does

- **Rule 0:** only the research types are kept (article, review, conference paper, book, book chapter). The editorial of the file is removed at import, and recorded in `removed_articles.csv`.
- **Duplicates:** same EID, or failing that same DOI. The number is announced before any deletion; here, none.
- **Corpus review:** before the first step, the list of articles is displayed, *incomplete articles first* (without DOI, without authors, without abstract, without keywords), each with its badges.

BEFORE: the file as imported

Rule 0: 1 non-research document(s) removed at import (Editorial 1). Counts below are out of 127: the rows are listed in removed_articles.csv.

Corpus

The papers this study is based on. Delete any paper that does not belong to your study before cleaning starts: nothing is lost, a deleted paper can be returned.

Cleaning has started
Papers can only be deleted or returned before the first step is applied: later steps have already saved their work paper by paper.

127 PAPER(S) IN THE STUDY
20 incomplete paper(s) first — no DOI, authors, abstract or keywords — then the complete ones.

Estudio de la sostenibilidad energética en los portales de Carpeta Personal de Salud
Actas de las 24th Jornadas de Ingeniería del Software y Bases de Datos, JISBD 2019 - 2019.0 CONFERENCE PAPER No DOI No abstract No keywords
Authors: García-Berná J.A.; Fernández-Alemán J.L.; Carrillo-De-Gea J.M.; Nicolás J.; Moros B.; Toval A.; Mancebo J.; Calero C.; García F.
Affiliations: Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, Spain; Instituto de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha, Ciudad Real, Spain
[View on Scopus](#)

Mixed finite element method of a seawater intrusion problem in confined aquifers
E3S Web of Conferences - 2021.0 CONFERENCE PAPER No keywords
Authors: Medarhi I.; Farhloul M.; Najib K.; Slimani S.; Zine A.
Affiliations: LERMA, Ecole Mohammadia d'Ingénieurs, Université Mohammed V, Agdal-Rabat, Morocco; LMAID, Ecole Nationale Supérieure des Mines de Rabat, Agdal, 1000, Morocco; Département de Mathématiques et de Statistique, Université de Moncton, Moncton, E1A 3E9, NB, Canada; Institut Camille Jordan, Ecole centrale Lyon, Ecully, 69134, France
[View on Scopus](#)

Calibration of volatility: Coupling between alternative regularization strategy and Dupire's equation
International Journal of Applied Mathematics and Statistics - 2013.0 ARTICLE No DOI
Authors: Aboulaich R.; Medarhi I.
Affiliations: Laboratory of Study and Research in Mathematics Applied (LERMA), Mohammadia School of Engineering (EMI), Mohammed V-Agdal University, Rabat, Morocco; LIRIMA, International Laboratory for Research in Computer Science and Applied Mathematics, Morocco
[View on Scopus](#)

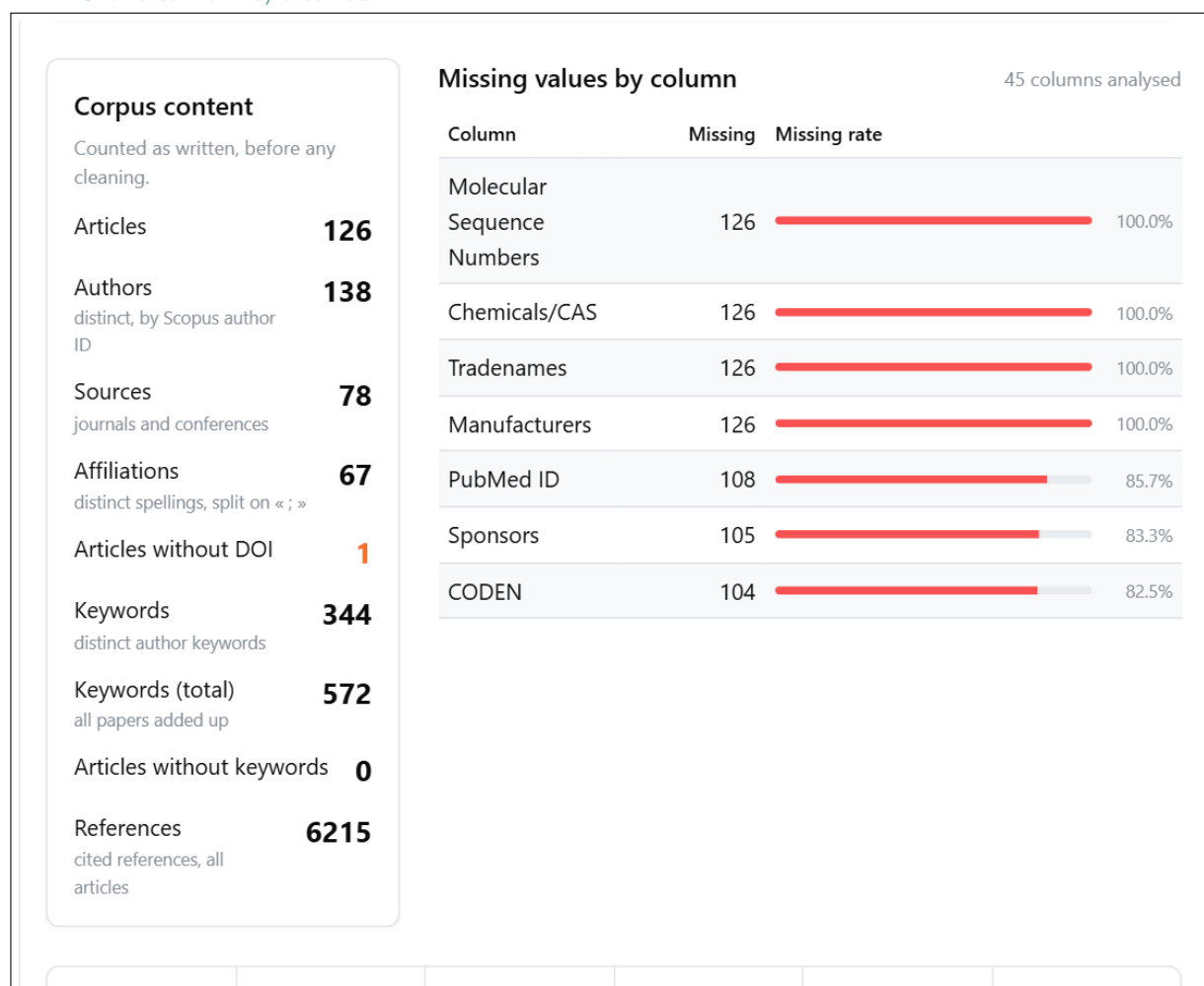
Improving Accessibility for People with Disabilities: A Case Study on Inclusive Beach Tourism*
Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS - 2019.0 CONFERENCE PAPER No keywords
Authors: Mayordomo-Martínez D.; Sanchez-Aarnoutse J.-C.; Merzoukid K.; García-Hernández M.; Carrillo-De-Gea J.M.; García-Berna J.A.; Fernández-Aleman J.L.; Idri A.; García-Mateos G.
Affiliations: Department of Structures and Construction and Graph Expression, Technical University of Cartagena, Cartagena, 30202, Spain; Department of Information and Communications Technologies, Technical University of Cartagena, Cartagena, 30302, Spain; FAMDIF, Murcia, 30005, Spain; Dept. of Computer Science and Systems, University of Murcia, Murcia, 30100, Spain; ENSIAS, Mohammed v University, Rabat, Morocco
[View on Scopus](#)

Evaluation and neuronal network-based classification of the PHRs privacy policies
Proceedings of the Annual Hawaii International Conference on System Sciences - 2012.0 CONFERENCE PAPER No keywords
Authors: Carrión I.; Fernández-Alemán J.L.; Jayne C.; Palmer-Brown D.; Toval A.; Carrillo-De-Gea J.M.

The corpus review: 21 incomplete articles at the top of the list, each with what it is missing (No DOI, No authors, No abstract, No keywords). An article that does not belong to the study is removed here; it is put back in one click in “Deleted papers”.

What you decide

Remove the articles that do not belong to your study. In this corpus, the user removed one: it is the difference between 127 and 126 articles.

AFTER: the same file, cleaned

The same screen after the cleaning: 126 articles, 67 affiliations instead of 218, a single article without DOI, no article without keywords any more.

3 Step 1: DOI

3.1 The problem in the file

The DOI is the identifier that links an article to everything else: its references, the citations it receives, the databases that describe it. Without it, an article cannot be matched with the others. In this file, **6 articles** did not carry one.

BEFORE: the file as imported

Rule 0: 1 non-research document(s) removed at import (Editorial 1). Counts below are out of 127; the rows are listed in removed_articles.csv.

DOI

Every document below carries no identifier. Crossref and OpenAlex are searched by title, and only verified matches are accepted on their own. The rest is yours to judge.

127 Documents	6 Without DOI	6 Not searched	121 With a DOI
-------------------------	-------------------------	--------------------------	--------------------------

Batch search
Queries Crossref and OpenAlex, then accepts only the verified matches, below the threshold you decide yourself.

[Search and accept verified](#)

0 / 6 documents resolved **6 to go**

To review (6) **Resolved (0)** 6 documents still need a decision.

Calibration of volatility: Coupling between alternative regularization strategy and Dupire's equation
Aboulaich R.; Medarhi I. - 2013.0 - International Journal of Applied Mathematics and Statistics - Article

Not searched yet. [Search this one](#)

DOI
Paste it yourself if you found it elsewhere
 [Save](#)

Requirements engineering tools
Carrillo De Gea J.M.; Nicols J.; Alemin J.L.F.; Toval A.; Ebert C.; Vizcano A. - 2011.0 - Article

Not searched yet. [Search this one](#)

DOI
Paste it yourself if you found it elsewhere
 [Save](#)

Six articles without DOI, none searched yet. “Search and accept verified” starts the batch search; “Search this one” starts it for a single article.

What Bibliominer does

For each article without DOI, Bibliominer queries **Crossref** and **OpenAlex** with its title, and compares the returned title with the title of the article, after normalisation (case, accents, punctuation). A match is called *verified* if

$$\text{similarity}(\text{title}, \text{found title}) \geq 0.80 \quad \text{and} \quad |\text{year} - \text{found year}| \leq 1$$

(a conference paper often appears the year after the conference). Only verified matches are accepted by the batch search; the others stay proposed, with their score, the database that found them and the year.

Homogeneous ensemble based support vector machine in breast cancer diagnosis
El Ouassif B.; Idri A.; Hosni M. - 2021.0 - HEALTHINF 2021 - 14th International Conference on Health Informatics; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021 - Conference paper

[100% title match](#) [CROSSREF](#) [2021](#)

Homogeneous Ensemble based Support Vector Machine in Breast Cancer Diagnosis
10.5220/0010230403520360 [open](#) [Use this DOI](#)

DOI
Paste it yourself if you found it elsewhere
 [Save](#)

A search: Crossref returns the same title (100 % title match), the same year (2021) and the DOI 10.5220/0010230403520360. A click on “Use this DOI” accepts it; a DOI found elsewhere can also be pasted.

Real example

Homogeneous ensemble based support vector machine in breast cancer diagnosis (2021)

BEFORE imported file

DOI: (empty)

AFTER cleaned file

DOI: 10.5220/0010230403520360 (Crossref, title 100% identical)

AFTER: the same file, cleaned

The screenshot shows the 'Resolved' tab in the Bibliominer interface. It displays three documents that have been resolved by another file. Each document entry includes the title, authors, year, database, and a 100% title match score. The DOI field is populated for each document.

Title	Authors	Year	Database	DOI
Requirements engineering tools	Carrillo De Gea J.M.; Nicols J.; Alemin J.L.F.; Toval A.; Ebert C.; Vizcano A.	2011	OPENALEX	10.1109/ms.2011.81
Detection of normality/pathology on chest radiographs using LBP	Carrillo-De-Gea J.M.; Garcia-Mateos G.	2010	CROSSREF	10.5220/0002747001670172
Impact of threshold values for filter-based univariate feature selection in heart disease classification	Benhar H.; Idri A.; Hosni M.	2020	CROSSREF	10.5220/0008947403910398

The recovered DOIs, in the “Resolved” tab: each keeps the database that found it (Crossref or OpenAlex) and the title score. A single article remains without DOI: no database knows it.

Why

A wrongly accepted DOI attaches the references and citations of another article to this one. This is why an approximate match is never accepted on its own: it is shown to you.

4 Step 2: Metadata

4.1 The problem in the file

Year, number of citations, document type, language. A missing year takes the article out of every time series: annual production, growth, trends. Here, **2 articles** had no year.

BEFORE: the file as imported

Metadata
Year, citations, document type and language. This step is locked until every one is filled: a missing year makes every trend that follows meaningless.

127 Documents	0 Missing citations	2 Missing year	0 Missing type	0 Missing language
-------------------------	-------------------------------	--------------------------	--------------------------	------------------------------

125 / 127 documents resolved **2 to go**

To review (2) Resolved (0) 2 documents still need a decision.

A survey of the vehicle routing problem and its variants: formulations and solutions
10.23939/mmc2024.01.333 2-52.0-85188461431 [open in Scopus](#)

Missing: Year

Year
Without it the document leaves every time series

Save

Reviewing Machine Learning Techniques in Credit Card Fraud Detection
10.5220/0013072500003838 2-52.0-85215282075 [open in Scopus](#)

Missing: Year

Year
Without it the document leaves every time series

Save

Two articles without a year (“Missing: Year”). The step is locked as long as any remain: a missing year distorts everything that follows.

What Bibliominer does

Bibliominer spots, column by column, the four fields the indicators depend on (*Cited by*, *Year*, *Document Type*, *Language*), shows for each article its DOI and its Scopus link to go and check, and **refuses to move on** as long as a year is missing.

What you decide

Enter the value read on the Scopus record or on the publisher’s page. Zero citations is a valid answer, not an absence.

Real example

A survey of the vehicle routing problem and its variants (DOI 10.23939/mmc2024.01.333)

BEFORE imported file

Year: (empty)

AFTER cleaned file

Year: 2024

AFTER: the same file, cleaned

To review (0) Resolved (2) 2 documents you already settled. Open one to change your mind.

You settled this - 27/09/2026 00:08

Reviewing Machine Learning Techniques in Credit Card Fraud Detection
10.5220/0013072500003838 2-s2.0-85215282075

Citations: 0 is a valid answer, not an absence. Year: Without it the document leaves every time series.

Document type: Conference paper Language: English

You settled this - 27/09/2026 00:08

A survey of the vehicle routing problem and its variants: formulations and solutions
10.23939/immc.2024.01.333 2-s2.0-85188461431

Citations: Year:

The two completed articles, in “Resolved”: what was entered stays visible and editable.

5 Step 3: Sources

5.1 The problem in the file

Two articles had **no journal name**. They then disappear from every analysis by source: rankings, Bradford’s law, SCImago quartiles.

BEFORE: the file as imported

Sources

Name the publication venues Scopus left blank. Naming a journal settles every article that carries it at once: you work on groups, not on rows.

127 Documents 2 Without a source 2 SCImago can resolve 0 To name by hand

0 / 2 source groups resolved 2 to go

SCImago (2) Name by hand (2)

Matched through the ISSN, an identifier, not a resemblance. The provenance is recorded, so a title coming from a reference file is never confused with one you typed.

Select all Clear 0 of 2 selected Accept 0 selected Accept all 2

EID	Title	ISSN	Document Type	Year	scimago_issn	scimago_title	scimago_publisher	scimago_type	scimago_found
2-s2.0-105018113129	A Fuzzy Logic-Based Decision Support System for Predicting Entrepreneurial Intention Among Textile Students: [Un sistema de apoyo a la toma de decisiones basado en la lógica difusa para predecir la intención empresarial entre los estudiantes del sector textil]	29534917	Article	2025.0	29534917	Data and Metadata (discontinued)	—	journal	True
2-s2.0-79959765268	Requirements engineering tools	07407459	Article	2011.0	07407459	IEEE Software	IEEE Computer Society	journal	True

Two documents without a source; SCImago can resolve both. The table shows, for each, the ISSN of the file and the journal SCImago associates with it.

What Bibliominer does

Bibliominer looks up the **ISSN** of the article in the **SCImago** index (bundled with the application). The match is made by identifier, never by title similarity. The provenance of each title is recorded: a title coming from SCImago is never confused with a title typed by hand. The articles are handled by *group*: naming a journal settles at once every article that carries it.

Real example

Requirements engineering tools (2011)

BEFORE imported file

Source title: (empty) ISSN: 07407459

AFTER cleaned file

Source title: IEEE Software (SCImago, through ISSN 07407459)

AFTER: the same file, cleaned

Source titles filled during cleaning
What Bibliominer wrote, and where each title came from. Open one to change it: the new title applies to every article that carries it.

Group	Articles	ISSN	Source Title	Action
Data and Metadata (discontinued)	1 article	ISSN 29534917	A Fuzzy Logic-Based Decision Support System for Predicting Entrepreneurial Intention Among Textile Students; [Un sistema de apoyo a la toma de decisiones basado en la lógica difusa para predecir la intención empresarial entre los estudiantes del secto...	Change
IEEE Software	1 article	ISSN 07407459	Requirements engineering tools	Change

The titles written during the cleaning, with their provenance (“From SCImago”) and the ISSN that found them. “Change” corrects the title for every article of the group.

6 Step 4: Authors

Three columns describe the authors of an article in Scopus: **Authors** (short names), **Author full names** (full names, with the Scopus identifier in parentheses) and **Author(s) ID**. A fourth one, **Authors with affiliations** (AWA), links each author to their affiliation. When they do not say the same thing, the whole analysis by author, institution and country is wrong.

6.1 What Bibliominer completes by itself

Before showing you anything, Bibliominer takes over what the file already contains, in another place. It guesses nothing: it copies.

BEFORE: the file as imported

Authors
Restore missing authors, then say which affiliation each author belongs to. Nothing is written to the corpus until you apply.

123 / 127 documents resolved 4 to go

Documents	Without authors	Without affiliation	Invalid AWA
127	0	0	4

Filled from this corpus
2 author cell(s) had stray spaces, now tidied.

- "Predicting Stock Market Price Movement using Machine Learning Techniques": 1 Scopus ID(s) read from its own "Author full names", where Scopus writes them in brackets.
- "Constructing Ensembles: A Diversity-Driven Approach with Correlation and Q-Statistics": 4 Scopus ID(s) read from its own "Author full names", where Scopus writes them in brackets.
- "Artificial Intelligence in Hydrogen Energy: A Comprehensive Bibliometric Analysis": 1 full name(s) found in other articles, under the same Scopus author ID.
- "Meta-Ensemble Learning for Predicting NASDAQ Stock Movements: An Empirical Study": 1 short name(s) missing from "Authors", computed from the full names.
- "Mixed finite element method of a seawater intrusion problem in confined aquifers": 5 short name(s) missing from "Authors", computed from the full names.
- "Predicting Academic Performance Using Machine Learning: Moroccan Middle School Data": 3 short name(s) missing from "Authors", computed from the full names.

Missing authors (0) Affiliations (0) Authors with affiliations (4)

The “Filled from this corpus” box: what was completed without asking anything, article by article. 4 AWA articles remain to be settled.

What Bibliominer does

- **Identifiers read in the full names.** Scopus writes “Najib, Khalid (6505786280)”: if the identifier is missing in the Author(s) ID column, it is taken from there.
- **Full names recovered by identifier.** A Scopus identifier designates a person: their full name, written on another article of the corpus, is theirs. Safeguard: the family name must match the short name of the article.
- **Short names computed.** Authors is only an abbreviation of the full names: a forgotten short name can be computed.
- **Stray spaces removed** (“Hosni, Mohamed (5718...)”).

What cannot be recovered goes to the [Missing authors](#) panel, author by author, with the right author designated.

Real example

Artificial Intelligence in Hydrogen Energy (2025): 3 authors, 3 identifiers, 2 full names.

BEFORE imported file

Author full names: ; Hosni, Mohamed (57189341317); Medarhri, Ibtissam (55818093100)
 Author(s) ID: 59962429900; 57189341317; 55818093100

AFTER cleaned file

Author full names: Zerhouni, Jaouad (59962429900); Hosni, Mohamed (57189341317); Medarhri, Ibtissam (55818093100)
(recovered from three other articles of the corpus, under the same identifier)

6.2 Authors with affiliations (AWA)

The problem

Each block of the AWA column starts with the short name of an author, followed by their affiliation. Scopus leaves authors out there, or attaches an affiliation to the wrong author. Here, **4 articles** were in this situation.

BEFORE: the file as imported

Requirements engineering tools 2011.0 [open in Scopus](#)

Some authors are absent from "Authors with affiliations". Carrillo De Gea, Juan M.: listed as authors, but with no block in "Authors with affiliations", so Scopus gives no affiliation for them. Tick their affiliation below, or add it.

Carrillo De Gea, Juan M. Carrillo De Gea J.M. absent from AWA
 No block for this author in "Authors with affiliations": Scopus gives no affiliation. Tick the right one, or add it.
 36462477200 [open account in Scopus](#)
☐ suggested Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain
[+ Add new affiliation](#)

Nicols, Joaquin Nicols J. locked
 7401657472 [open account in Scopus](#)
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Alemn, Jos L. Fernndez Alemn J.L.F. locked
 41868972600 [open account in Scopus](#)
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Toval, Ambrosio Toval A. locked
 22982044800 [open account in Scopus](#)
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Ebert, Christof Ebert C. no affiliation given
 This author's block in "Authors with affiliations" carries no affiliation. Tick the right one, or add it.
 56406144100 [open account in Scopus](#)
☐ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain
[+ Add new affiliation](#)

Vizcano, Aurora Vizcano A. locked
 7003393897 [open account in Scopus](#)
☐ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☒ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

[+ Insert a missing author](#) [Save this article](#)

"Requirements engineering tools": Carrillo De Gea has no block in AWA ("absent from AWA") and Ebert has no affiliation ("no affiliation given"). The safe authors are locked ("locked"). The suggestion comes from the other articles of the same author, through their Scopus identifier.

BEFORE: the file as imported

Software development effort estimation using feature selection techniques 2018.0 [open in Scopus](#)

The affiliations look misassigned. 1 author(s) match several affiliations while 1 match none, and the article has exactly one affiliation per author (2). Scopus probably attached an affiliation to the wrong author: check each author below.

Hosni, Mohamed Hosni M. no affiliation given
 This author's block in "Authors with affiliations" carries no affiliation. Tick the right one, or add it.
 57189341317 [open account in Scopus](#)
☐ suggested Software Project Management Research Team ENSIAS, Mohammed V University of Rabat, Morocco
☐ suggested Department of Web and Mobile Engineering, Software Project Management Research Team, Mohammed V University Rabat, Morocco
[+ Add new affiliation](#)

Idri, Ali Idri A. several affiliations match
 This author's text matches 2 of the article's affiliations, so it is ambiguous. Tick only those that are really theirs.
 6602789810 [open account in Scopus](#)
 ▶ Original AWA text for this author
☐ suggested Software Project Management Research Team ENSIAS, Mohammed V University of Rabat, Morocco
☐ suggested Department of Web and Mobile Engineering, Software Project Management Research Team, Mohammed V University Rabat, Morocco
[+ Add new affiliation](#)

[+ Insert a missing author](#) [Save this article](#)

Another case: one author matches two affiliations, another matches none, while the article has exactly one affiliation per author. Scopus probably attached an affiliation to the wrong author.

What Bibliominer does

Bibliominer matches each author with their AWA block by name, detects the four cases (author absent from AWA, block without affiliation, ambiguous affiliation, swapped affiliations) and **suggests** an author's affiliation from their other articles in the corpus, found through their Scopus identifier. An article that still has an author or affiliation problem does not appear in AWA: each problem is settled in its own panel, and AWA only re-examines the article afterwards.

What you decide

Tick the affiliation of each flagged author, or add one. The authors whose affiliation is certain are locked.

AFTER: the same file, cleaned

Requirements engineering tools
2011 open in Scopus
SAVED

Some authors are absent from "Authors with affiliations". Carrillo De Gea, Juan M.: listed as authors, but with no block in "Authors with affiliations", so Scopus gives no affiliation for them. Tick their affiliation below, or add it.

Carrillo De Gea, Juan M. Carrillo De Gea J.M. absent from AWA
No block for this author in "Authors with affiliations". Scopus gives no affiliation. Tick the right one, or add it.
36462477200 open account in Scopus
☒ suggested Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain
Add new affiliation

Nicols, Joaquin Nicols J. locked
7481657472 open account in Scopus
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Alezn, Jos L. Fernndez Alezn J.L.F. locked
41860972600 open account in Scopus
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Toval, Ambrosio Toval A. locked
22982044000 open account in Scopus
☒ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☐ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

Ebert, Christof Ebert C. no affiliation given
This author's block in "Authors with affiliations" carries no affiliation. Tick the right one, or add it.
56486144100 open account in Scopus
☐ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☒ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain
Add new affiliation

Vizcano, Aurora Vizcano A. locked
7003393897 open account in Scopus
☐ Software Engineering Research Group, Faculty of Computer Science, University of Murcia, Spain
☒ Alarcos Research Group, Escuela Superior de Informática, University of Castilla-La Mancha, Spain

+ Insert a missing author
Saved Save again

The same article settled: Carrillo De Gea receives "Software Engineering Research Group, University of Murcia" (suggestion accepted), Ebert receives "Alarcos Research Group". "Saved": the decision is kept and stays editable.

7 Step 5: Affiliations

7.1 The problem in the file

A Scopus affiliation is a free string. The same laboratory is written in several ways, in several languages, with or without an abbreviation, with or without its parent university. This file had **218 distinct spellings** for what is, in the end, **67 organisations**.

Real example: one organisation, three spellings

BEFORE imported file

MMCS Research Team, LMAID, ENSMR, Rabat, Morocco
 MMCS Research Team, LMAID ENSMR, Rabat, Morocco
 MMCS Research Team, LMAID, Mines Rabat, Rabat, Morocco

AFTER cleaned file

parent 1: National School of Mineral Industry, city: Rabat, country: Morocco

BEFORE: the file as imported

Affiliations

Country, then city, then the affiliations themselves. An affiliation is only readable in its geographic context, and that same context is what lets two spellings be compared, so standardisation happens city by city.

218 Distinct affiliations	218 Not resolved	0 Resolved
-------------------------------------	----------------------------	----------------------

DETECTION
Country and city through GeoNames, then institution, locally, in the same pass.

ROR: OFFLINE DUMP
Run detection

Run the detection first

Review, standardisation and the organisation view all work on detected values: there is nothing to arbitrate before that.

Before the detection: 218 distinct affiliations, none resolved. All the arbitration work is on detected values: there is nothing to decide before.

What Bibliominer does

The detection, in a single pass:

- Country and city via GeoNames.** The city of a school prevails over the seat of its university (ENSA Safi is in Safi, not at the seat of the university); “Sale” after a dash is recognised as the city of Salé.
 - Splitting into blocks:** each affiliation becomes **sub-unit** (faculty, school, laboratory), **parent** (the organisation it belongs to), **city**, **country**.
 - Organisation** via the **ROR** index (offline) and via the other affiliations of the corpus: translation (“Facultad de Informática” / “Faculty of Computer Science”), abbreviation + city, shared laboratory.
- Each value keeps its provenance: *from text* (read in the string), *deduced*, *ROR*, *unresolved*.

BEFORE: the file as imported

Affiliations

Country, then city, then the affiliations themselves. An affiliation is only readable in its geographic context, and that same context is what lets two spellings be compared, so standardisation happens city by city.

218 Distinct affiliations

41 Not resolved

177 Resolved

DETECTION
Country and city through GeoNames, then institution, locally, in the same pass.

ROR: OFFLINE DUMP

Re-run geo check Institution check

Review Organisations

All countries

Country	Affiliation(s)	Article(s)	Location	Institution	Merge review needed
Morocco	112	162	1	27	8
France	6	6	1	5	0
Spain	68	102	4	0	1
Germany	2	4	0	2	0
Latvia	2	2	0	2	0
Tunisia	1	1	0	1	0
Canada	9	17	0	0	2
United Kingdom	6	6	0	0	0
Poland	3	3	0	0	0
United Arab Emirates	3	5	0	0	1
Egypt	2	3	0	0	0
Lithuania	1	1	0	0	0

After the detection: 177 of the 218 affiliations are resolved automatically; 41 remain to review. The work is done country by country, then city by city.

BEFORE: the file as imported

Review Organisations

All countries > Morocco

City	Affiliation(s)	Article(s)	Location	Institution	Merge review needed
Rabat	50	79	0	19	0
Casablanca	3	3	0	3	0
Ben Guerir	8	9	0	2	0
No city detected	1	1	1	1	0
Tétouan	4	4	0	1	0
Martil	1	1	0	1	0
Meknès	22	34	0	0	4
Oujda	11	18	0	0	1
Fes	4	4	0	0	0
El Jadida	2	2	0	0	1
Ifrane	2	3	0	0	0
Safi	2	2	0	0	1
Larache	1	1	0	0	0
Sale	1	1	0	0	0

Morocco, city by city. Each card says what remains to be done: an institution to confirm, a merge to review (“merge review needed”).

BEFORE: the file as imported

Review Organisations

All countries > Morocco > Rabat

31 / 50 affiliations resolved 19 to go

Detected rows are read-only; press **Edit** to change them. Rows needing review are already open.

Finish this city first
Complete and approve every affiliation in **Rabat**. 19 still need review. Standardization and cross-language review unlock once this city is fully resolved.

ALL - 50 NEEDS REVIEW - 19 FROM TEXT - 29 DEDUCED - 2 ROR - 7 UNRESOLVED - 12

Search an affiliation To do first Most used A-Z

MMCS Research Team, LMAID, ENSMR, Rabat, Morocco
6 articles location institution pending ROR More information

Institution: ROR National School of Mineral Industry
education - autonomous body Not approve Approve

Country Morocco City Rabat Region Optional

Parent organisation Sub-unit
ENSMT

Save

MMCS Research Team, LMAID ENSMR, Rabat, Morocco
2 articles location institution pending ROR More information

Institution: this corpus National School of Mineral Industry
education - autonomous body Not approve Approve

Country Morocco City Rabat Region Optional

Parent organisation Sub-unit
ENSMT

Save

MMCS Research Team, LMAID, Mines Rabat, Rabat, Morocco
2 articles location institution pending Unresolved More information

Country Morocco City Rabat Region Optional

Parent organisation Sub-unit
ENSMT

Save

No parent found, remove the article(s) instead

MMCS Team, LMAID Laboratory, ENSMR-Rabat, Morocco
2 articles location institution pending Unresolved More information

Rabat: 50 affiliations, 19 to review. For “MMCS Research Team, LMAID, ENSMR”, ROR proposes “National School of Mineral Industry”; the abbreviation ENSMR is placed as a sub-unit. The FROM TEXT / DEDUCED / ROR / UNRESOLVED counters say where each value comes from.

What you decide

Approve or refuse each organisation suggestion, complete what remains *unresolved*, then validate the proposed merges. A city must be entirely resolved before its merges are unlocked: affiliations are not merged before it is known what they are.

BEFORE: the file as imported

Review Organisations			
All organisations, every country.			
218	55	100	
Raw spellings	Organisations	Sub-units	
Organisation	City	Country	Sub-units
> University of Murcia	Murcia	Spain	6
> Moulay Ismail University	Meknès	Morocco	3
> Mohammed V University	Rabat	Morocco	2
> –university of Québec	Montréal	Canada	3
> University of Castilla-La Mancha	2 cities	Spain	3
> Mohammed V University in Rabat	2 cities	Morocco	3
> Mohammed 1st University	Oujda	Morocco	2
> Mohamed 1st University	Oujda	Morocco	1
> Mohammed VI Polytechnic University	Ben Guerir	Morocco	1
> University Moulay Ismail of Meknes	Meknès	Morocco	1
> Moulay Ismail University of Meknes	Meknès	Morocco	1
> Universidad de Murcia	Murcia	Spain	2
> Abdelmalek Essaadi University	2 cities	Morocco	3
> African Disaster Mitigation Research Center	Cairo	Egypt	1
> Al Akhawayn University in Ifrane	Ifrane	Morocco	1
> Sidi Mohamed Ben Abdellah University	Fes	Morocco	2
> Technical University of Cartagena	Cartagena	Spain	1
> University Mohammed v	Rabat	Morocco	1
> University Mohammed V of Rabat	Rabat	Morocco	1
> Université de Moncton	Moncton	Canada	1
> Université Mohammed V	Rabat	Morocco	1
> Abulcasis International University of Health Sciences	Rabat	Morocco	1
> Cheikh Zaid International University Hospital	Rabat	Morocco	1
> Edinburgh Napier University	Edinburgh	United Kingdom	1
> Mohammed V University of Rabat	Rabat	Morocco	1
> Mohammed V University Rabat	Rabat	Morocco	1
> Mohammed V-Agdal University	Rabat	Morocco	2
> United Arab Emirates University	Al Ain	United Arab Emirates	1

The Organisations tab before arbitration: 218 spellings, 55 organisations, 100 sub-units.

AFTER: the same file, cleaned

Review Organisations

All countries > Morocco > Rabat

50 / 50 affiliations resolved Nothing left here Detected rows are read-only; press **Edit** to change them. Rows needing review are already open.

↔ Standardize organisation names · review merges View organisations Save progress

ALL · 50 FROM TEXT · 29 DEDUCED · 2 MANUAL · 19

Search an affiliation To do first Most used A-Z

Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco More information

8 articles location institution From text

Country: Morocco City: Rabat Region: Optional

Parent organisation: Mohammed V University Sub-unit: ENSIAS Edit

MMCS Research Team, LMAID, ENSMR, Rabat, Morocco More information

6 articles location institution Manual

Country: Morocco City: Rabat Region: Optional

Parent organisation: National School of Mineral Industry Sub-unit: Edit

Software Project Management Research Team, ENSIAS, Mohammed V University in Rabat, Rabat, Morocco More information

4 articles location institution From text

Country: Morocco City: Rabat Region: Optional

Parent organisation: Mohammed V University Sub-unit: ENSIAS Edit

ENSIAS, Mohammed v University, Rabat, Morocco More information

3 articles location institution From text

Country: Morocco City: Rabat Region: Optional

Parent organisation: Mohammed V University Sub-unit: ENSIAS Edit

Rabat once resolved: each affiliation carries its organisation, its city and its country.

AFTER: the same file, cleaned

Review Organisations			
All organisations, every country.			
216	48	66	
Raw spellings	Organisations	Sub-units	
Organisation	City	Country	Sub-units
> Universidad de Murcia	Murcia	Spain	6
> Mohammed V University	2 cities	Morocco	4
> Moulay Ismail University	Meknès	Morocco	1
> National School of Mineral Industry	Rabat	Morocco	1
> University of Québec	Montréal	Canada	1
> University of Castilla-La Mancha	2 cities	Spain	3
> Mohammed 1st University	Oujda	Morocco	2
> Mohammed VI Polytechnic University	Ben Guerir	Morocco	1
> Abdelmalek Essaadi University	3 cities	Morocco	4
> Université Jean Monnet	Écully	France	1
> Vector Stiftung	Stuttgart	Germany	1
> African Disaster Mitigation Research Center	Cairo	Egypt	1
> Al Akhawayn University in Ifrane	Ifrane	Morocco	1
> Sidi Mohamed Ben Abdellah University	Fes	Morocco	2
> Technical University of Cartagena	Cartagena	Spain	1
> UAE University	Al Ain	United Arab Emirates	1
> Université de Moncton	Moncton	Canada	1
> Abulcasis International University of Health Sciences	Rabat	Morocco	1
> Cadi Ayyad University	Safi	Morocco	1
> Cheikh Zaid International University Hospital	Rabat	Morocco	1
> Edinburgh Napier University	Edinburgh	United Kingdom	1
> FAMDIF	Murcia	Spain	1
> Green Energy Park	Ben Guerir	Morocco	1
> Higher School of Textile and Clothing industries.	Casablanca	Morocco	1
> International Laboratory for Research in Computer Science and Applied Mathematics	—	2 countries	2
> Transport and Telecommunication Institute	Riga	Latvia	1
> University of Coventry	Coventry	United Kingdom	2
> University of Sharjah	Sharjah	United Arab Emirates	1

The organisations after standardisation: 48 organisations and 66 sub-units. The spelling kept is the one most used in the corpus.

Real example**Artificial Intelligence in Hydrogen Energy**

BEFORE imported file

Moulay Ismail University of Meknes, IEST Research Team, LAIDTM, ENSAM, Meknes, Morocco; MMCS Research Team, LMAID, ENSMR, Rabat, Morocco

AFTER cleaned file

subparent: ENSAM, parent 1: Moulay Ismail University, city: Meknès, country: Morocco; parent 1: National School of Mineral Industry, city: Rabat, country: Morocco

Why

The labelled form (subparent: ..., parent 1: ..., city: ..., country: ...) is what lets the analysis count by institution, by city and by country without rereading free text. A comma in an organisation name can no longer be taken for a separator.

8 Step 6: Text

8.1 The problem in the file

15 articles had no author keywords, one had no abstract. Without keywords, an article is absent from the thematic map, the co-word networks and Zipf's law.

BEFORE: the file as imported

The screenshot shows the 'Text' interface in Bibliominer. At the top, it says 'Text' and 'Restore titles, abstracts and keywords. Articles are grouped by what is missing, so you always know which case you are filling.' Below this, a progress bar shows '112 / 127 documents resolved' and '15 to go'. A search bar is on the right. The main area is divided into three columns: '127 Documents', '15 To fix', and '112 Complete'. Below the columns, it says 'K - 14 AK - 1'. A section titled 'To review (15) Resolved (0)' indicates '15 documents still need a decision.' A filter 'K Only Author Keywords' is selected, showing '14 article(s)'. A note states 'Title and abstract are already there — only the keywords are missing.' Two article entries are shown. The first article is 'Mixed finite element method of a seawater intrusion problem in confined aquifers' (2021.0 - Conference paper - E3S Web of Conferences). It has a text input for 'Author keywords' and a list of suggested keywords: '+ CONFINED AQUIFERS', '+ SEAWATER INTRUSION', and '+ SEMI-IMPLICIT SCHEME'. The second article is 'Improving Accessibility for People with Disabilities: A Case Study on Inclusive Beach Tourism*' (2019.0 - Conference paper - Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS). It also has a text input for 'Author keywords' and suggested keywords: '+ ACCESSIBILITY', '+ INCLUSIVE BEACHES', and '+ MOTOR DISABILITY'. Both articles have 'Approve all 3' buttons and a 'Save' button at the bottom right of each entry.

The articles grouped by what they are missing: here “Only Author Keywords”, 14 articles. Under each field, the keywords suggested from the corpus vocabulary.

What Bibliominer does

The *controlled vocabulary* is the set of author keywords already used in the corpus. For an article without keywords, Bibliominer extracts the expressions of its title and abstract, matches them with this vocabulary (up to the plural: “neural networks” = “neural network”) and ranks them by a TF-IDF-like score:

$$\text{score}(t) = (\text{tf}(t) + 2 \mathbf{1}_{\text{title}}(t)) \times \left(\ln \frac{N + 1}{\text{df}(t) + 1} + 0.5 \right) \times w_{\ell}(t)$$

where tf is the number of occurrences of the term in the title and the abstract, $\mathbf{1}_{\text{title}}$ is 1 if the term appears in the title, df is the number of articles of the corpus that already use it as a keyword, N the number of articles, and $w_{\ell} = 1.5$ for a multi-word expression (1 otherwise). A single-word term is only kept if it is in the title or repeated. At most three suggestions, and none rather than a doubtful one. An article without any term of the vocabulary receives a *weak* suggestion: the expressions of its title repeated at least twice in the text.

Real example

Mixed finite element method of a seawater intrusion problem in confined aquifers

BEFORE imported file

Author Keywords: (empty)

AFTER cleaned file

Author Keywords: Confined aquifers; Seawater intrusion; Semi-implicit scheme

AFTER: the same file, cleaned

The screenshot shows the Bibliominer interface for reviewing and settling documents. At the top, it indicates 'To review (0)' and 'Resolved (14)', with a note that 14 documents have already been settled. Two article entries are visible:

- Article 1:** 'Improving Accessibility for People with Disabilities: A Case Study on Inclusive Beach Tourism*'. It is a 2019 conference paper. The author keywords are 'Accessibility', 'Inclusive beaches', and 'Motor disability'. Suggested corpus vocabulary includes '+ ACCESSIBILITY', '+ INCLUSIVE BEACHES', and '+ MOTOR DISABILITY'. There are buttons to 'Approve all 3' and 'Save'.
- Article 2:** 'Neural network-based data analysis for medical-surgical nursing learning'. It is a 2012 conference paper. The author keywords are 'Medical-surgical nursing' and 'Knowledge'. Suggested corpus vocabulary includes '+ MEDICAL-SURGICAL NURSING' and '+ KNOWLEDGE'. There are buttons to 'Approve all 2' and 'Save'.

The completed articles: the keywords approved or entered. In total, 34 keywords added to the corpus (538 → 572).

9 Step 7: References

9.1 The problem in the file

Scopus delivers the references as **free text**, without an identifier. Two articles citing the same work often write it differently: no co-citation analysis can match them.

Real example

Artificial Intelligence in Hydrogen Energy: 20 references in free text.

BEFORE imported file

Petroleum, bp Energy Outlook 2023 edition ..., Stat. Rev. World Energy, pp. 1-53, (2023); Zhang F.; Zhao P.; Niu M.; Maddy J., The survey of key technologies in hydrogen energy storage, Int. J. Hydrogen Energy, 41, pp. 14535-14552, (2016); ...

AFTER cleaned file

```
ref1 | 10.1016/0306-4573(86)90076-2 | 1986 | Paul Nicholls | Empirical validation of Lotka's law
; ref2 | 10.1007/bf02020148 | 1990 | Bluma C. Peritz | A bradford distribution for bibliometrics
; ...
```

BEFORE: the file as imported

References

Scopus gives references as raw text. Crossref and OpenAlex are queried for every article, and only corroborated results enter the global list.

Reconciliation

☐ Also query Semantic Scholar
One request per second, and it misses most large publishers. It roughly doubles the run for a marginal gain.

[Run reconciliation](#)

Nothing reconciled yet

Run the reconciliation to see, article by article, how many references each source returned. Until then the References column holds raw Scopus text, which no co-citation analysis can read.

Applying builds the deliverable: the References column is replaced by the reconciled global list and the author columns are indexed. Both happen only there.

[Apply and continue →](#)

Before the reconciliation: the References column contains the raw Scopus text.

What Bibliominer does

For each article, Bibliominer requests its list of references from **Crossref** and **OpenAlex** (and, if you tick it, from **Semantic Scholar**), through its DOI. The lists are matched with each other and with the Scopus one, and form a *global list* where each reference carries, when it has them, its DOI, its year, its authors and its title. A reference only enters the list if it is corroborated.

AFTER: the same file, cleaned

References

Scopus gives references as raw text. Crossref and OpenAlex are queried for every article, and only corroborated results enter the global list.

Reconciliation

☐ Also query Semantic Scholar
One request per second, and it misses most large publishers. It roughly doubles the run for a marginal gain.

[Run again](#)

Articles
126

Scopus refs
6065

Global refs
6534
×1.1 vs Scopus

Crossref / OpenAlex
5416 / 5102

Filter by title or DOI

Most refs Biggest gain Title

Analysis tables: [references_reconciliation.csv](#) [references_long.csv](#)

Article	Scopus	Crossref	OpenAlex	Sem. Scholar	Global
Reviewing ensemble classification methods in breast cancer 10.1016/j.cmpb.2019.05.019	227	226	234	0	234
Classification techniques in breast cancer diagnosis: A systematic literature review 10.1080/21681163.2020.1811159	214	224	184	0	227
A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard Mapping 10.1007/s11269-024-03940-7	145	144	141	0	146
A method and validation for auditing e-Health applications based on reusable software security requirements specifications 10.1016/j.ijmedinf.2024.105699	119	134	102	92	139
On the risks and safeguards for requirements engineering in global software development: Systematic literature review and quantitative assessment 10.1109/access.2018.2674096	136	128	120	122	136
A systematic mapping study for ensemble classification methods in cardiovascular disease 10.1007/s10462-020-09914-6	121	127	127	0	132
On the value of parameter tuning in heterogeneous ensembles effort estimation 10.1007/s00500-017-2945-4	115	115	106	0	131
Software project management approaches for global software development: A systematic mapping study 10.26599/tst.2018.9010029	123	0	110	98	123

After the reconciliation: for each article, the number of references seen by Scopus, Crossref, OpenAlex and Semantic Scholar, and the global list. Here, 6065 Scopus references become 6534 identified references: 469 more.

Why

Each source misses references that the others have. The corroborated union is more complete than each of them, and each reference in it carries an identifier: this is what makes co-citation, bibliographic coupling, local citations and spectroscopy possible.

10 Correcting any value: All documents

Each step shows what is *missing*. A value that Scopus provided therefore never appears there, even a wrong one: a DOI pointing to another article, a shifted year (published in 2024, indexed in 2025), a misspelled journal title, a wrong author identifier.

At the top of each step, the **Review | All documents** selector opens the table of *all* the articles (those of the part, if the corpus is split), with the fields of the step. An article can be searched by its title, its EID, its DOI, an author or a source; the **Scopus** link opens its record to check the value at the source.

be reverted.						
Q 2024		☐ Only edited (1)		9 of 50 documents Log of manual edits (CSV)		
Document	Year	Citations	Document Type	Language	EID (Scopus ID)	Scopus link
Comparative Analysis of Single and Ensemble Support Vector Regression Methods for... Hosni M. - 2024 2-s2.0-85215279324 Scopus	2024	3	Conference paper	English	2-s2.0-85215279324	scopus.com/pages/publications/85...origin:resultlist
Edit						
A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard... El baida M. - 2024 2-s2.0-85208413949 Scopus	2024	29	Article	English	2-s2.0-85200413949	scopus.com/pages/publications/85...origin:resultlist
Edit						

Metadata step, All documents view: each article with its year, its citations, its type, its language, its EID and its Scopus link. The [Edit](#) button is under the title.

Step	Editable fields
DOI	DOI
Metadata	year, citations, type, language, EID, Scopus link
Sources	source title, ISSN
Authors	names, full names, Scopus identifiers of the authors
Text	title, abstract, author and index keywords
Affiliations, References	read only (see below)

What you decide

Edit opens the record of the article with all the fields of the step; correct, then **Save**. Clicking a value opens the same record on that field. For a journal title or an ISSN, a checkbox offers to correct at once every article of the same journal.

2024

2024

Document

Comparative Analysis of Single and Ensemble Support Vector Regression Methods for Software Development Effort Estimation

Hosni M. · 2024 · 2-s2.0-8-85215279324 · Scopus 



2024

2024

Document

A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Flood risk decomposed: Optimized machine learning hazard mapping and multi-criteria...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Software Testing Effort Estimation Based on Machine Learning Techniques: Single and...

Hosni M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and...

O. Iqbal M. · 2024 · 2-s2.0-49-85298952996 · Scopus 



2024

2024

Document

Real-Time

The record of an article: the expected format is recalled under each field.

What Bibliominer does

Each entry is checked before it enters the corpus: a DOI starts with 10. (a <https://doi.org/> link is accepted and shortened), a year has four plausible digits, citations are an integer, an EID has the form 2-s2.0-... and belongs to no other article. Changing the EID updates the Scopus link and follows the article wherever it is referenced (parts of the split, decisions already taken).

The screenshot shows a document entry in the Bibliominer interface. The document title is "Comparative Analysis of Single and Ensemble Support Vector Regression Methods for...". The Year field is highlighted with a blue "EDITED" badge. The Citations field shows a value of 3 with a small arrow icon next to it. The Scopus link is provided. The interface also shows a search bar with the same document title and a toggle for "Only edited (2)".

The corrected value carries the **edited** badge; on hover, it recalls the original value. The arrow restores it.

Why

A correction is never silent. Each value modified by hand is written to the log (original value, new value, date), reversible in one click, and downloadable from the final screen (**Manual edits (CSV)**) to be attached to the corpus.

Caution

In **Affiliations** and **References**, the view is read only: these steps rewrite their columns when they are applied, and a touch-up of the raw text would be overwritten. An affiliation is corrected in **Review** (all of them are listed there, resolved ones included), a reference in the record of its article.

11 Final: the deliverable file

The final screen brings the whole workflow together: what was missing, what was filled in, and with which tool. It is the Methods section of your article, already written in figures.

The screenshot shows the final screen of the Bibliominer workflow. It includes a summary of cleaning results and a table of missing values.

Before / after cleaning

- Articles: **128 → 126** (2 removed (non-research))
- Missing values (tracked columns): **23 → 1** (22 filled / fixed)
- Affiliations standardized: **216 → 67** (149 variants consolidated)
- References: Scopus → Global: **6065 → 6534** (+469 added via Crossref / OpenAlex)

Column	Missing before	Missing after	Fixed
DOI	5	1	-4
Cited by	0	0	—
Year	2	0	-2
Document Type	0	0	—
Language	0	0	—
Source title	2	0	-2
Title	0	0	—
Abstract	0	0	—
Author Keywords	14	0	-14
Affiliations	0	0	—
Authors w/ affil.	0	0	—
References	0	0	—

Before / after: 128 → 126 articles (2 removed), missing values 23 → 1, affiliations 216 → 67, references 6065 → 6534. The table gives, column by column, the missing values before and after.

Corpus content: before / after

	Imported file	Analysed corpus	Cleaned file	Change
Articles	128	126	126	—
Authors (distinct)	141	140	138	−2
Sources	78	77	78	+1
Affiliations (distinct spellings)	222	216	67	−149
Articles without DOI	7	5	1	−4
Author keywords (distinct)	344	344	344	—
Author keywords (total, all papers)	538	538	572	+34
Articles without keywords	16	14	0	−14
References (total)	6224	6215	6520	+305

« Imported file » is the file exactly as you uploaded it. « Analysed corpus » is what the cleaning actually worked on: the imported file without its non-research documents (rule 0), as shown on the import screen. « Change » compares the cleaned file with the analysed corpus: articles drop when you delete sources or remove duplicates; affiliations drop as spellings of the same organisation are merged.

The content of the corpus at each stage: imported file, analysed corpus, cleaned file.

What this file contains

- Missing **metadata** filled (Cited by, Year, Document Type, Language)
- **Source titles** completed and standardized
- **Text** cleaned (titles, abstracts, keywords)
- **Authors** and author-affiliation links validated
- **Affiliations** rewritten to subparent, parent, city, country
- **References** replaced by the reconciled *global* list, ready for co-citation

merged_cleaned.csv READY

Download the cleaned file Cleaning report

← Back to References Rebuild the export

What the delivered file contains, and the two downloads: the cleaned corpus and the cleaning report.

11.1 What you get

- `merged_cleaned.csv`: the corpus, in Scopus format, with the labelled affiliations, the authors indexed by position and the global list of references. It opens as is in Bibliominer Analysis ([Start your analysis](#) button), or in Python with the `bibliominer-analysis` package (`pip install bibliominer-analysis`).
- `merged_cleaning_report.txt`: the report, with the versions of the tools and the APIs used (here Crossref, OpenAlex, Semantic Scholar, ROR, GeoNames; no API error).
- `removed_articles.csv`, `recovered_dois.csv`, and the reconciliation tables of the references, per article and per cited reference.
- `manual_edits.csv`: each value corrected by hand in [All documents](#), with its original value and its date.

Caution

Without an account, a study is kept for one week from its import; with a Google account, 30 days. The cleaned file remains the backup: download it before leaving, and import it again to resume. Splitting into parts works the same way with or without an account.