

Bibliominer

Analysis guide

Each screen, each chart, and the formula that produces it

A value that cannot be recomputed cannot be defended in front of a reviewer. This guide goes through the analysis application page by page, on the corpus cleaned in the cleaning guide (126 articles, 2010–2026). For each chart and each table: the screenshot, the exact formula the code applies, and how to read the result, with the values obtained on this corpus.

The formulas are those of the `bibliominer-analysis` package: they are what runs behind the interface, and they give the same figures from Python.

Contents

1	Conventions	3
1.1	Notation	3
1.2	Three rules valid everywhere	3
2	Overview: the identity card	3
3	Production	7
4	Documents	9
5	Data quality: what the corpus allows	11
6	Corpus profile	12
6.1	The Price index and the age of references	13
6.2	Concentration: Gini, Lorenz, CR4, CR10	14
6.3	Self-citations	15
6.4	Interdisciplinarity	16
6.5	Open access	17
7	Authors	17
8	Institutions	20
9	Sources	25
10	Cities	28
11	Countries	32
12	Local impact	34
13	Reference spectroscopy	36
14	Indicators	37
14.1	Growth	38
14.2	Trend and projection	39
14.3	Collaboration	40
14.4	Co-authorship index (CAI)	41
14.5	Sample size (Cochran)	41
15	Laws: Lotka, Bradford, Zipf	42
15.1	Lotka: author productivity	42
15.2	Bradford: the dispersion of sources	44
15.3	Zipf: keyword frequency	45
16	Topics: the keywords	46
17	Text mining: what the abstracts say	48
18	Thematic map: Callon's strategic map	50
19	Thematic evolution	51
20	Conceptual structure	52

21 Three fields	54
22 Networks: the common principle	55
23 Network lab: what the eye does not see	57
24 Citation network	59
25 Coupling clusters	62
26 The report: Add to report	63
27 The same results in Python	64

1 Conventions

1.1 Notation

Symbol	Meaning
N	number of documents of the corpus (here 126)
C_d	global citations of document d : the Scopus <i>Cited by</i> column (GC , global citations)
LC_d	local citations: number of documents <i>of the corpus</i> that cite d
y_d	publication year of d ; $y_{\max}$ the last year of the corpus
n_t	number of documents published in year t
A	number of distinct authors (by Scopus identifier)
A_d	number of distinct authors of document d
$C_{(1)} \geq$	the citations sorted in decreasing order
$C_{(2)} \geq \dots$	

1.2 Three rules valid everywhere

- **Full counting.** A document signed by two countries counts once for each. Totals per country, city or institution therefore exceed the number of documents, on purpose.
- **Authors are identified by their Scopus identifier**, not by their name: two spellings of the same researcher (“Fernández-Alemán J.L.”, “Fernandez-Aleman J.L.”) are one person, and two namesakes stay two people. Keywords are compared case-insensitively.
- **The filter bar is global.** Period, document type: every open chart is recomputed on the filtered sub-corpus. Each block also has its own period setting (calendar icon) and its buttons: add to the report, export the image or the table.

Caution

Start with **Data quality** (section 5): this page says, analysis by analysis, on which share of the corpus each figure is computed. An indicator computed on 40 % of the documents is not quoted as if it covered the whole corpus.

2 Overview: the identity card

DOCUMENTS 126	AUTHORS 138	AFFILIATIONS 67
INSTITUTIONS 48	COUNTRIES 15	SOURCES 78
CITATIONS 1,935	KEYWORDS 344	REFERENCES 6,533

The counters of the corpus: documents, authors, affiliations, institutions, countries, sources, citations, keywords, references.

How it is computed

- **Authors:** distinct Scopus identifiers (138). **Affiliations:** distinct spellings after cleaning (67). **Institutions:** distinct parent organisations (48).
- **Citations:** $\sum_d C_d = 1935$. **Keywords:** distinct author keywords, case-insensitive (344). **References:** rows of the reconciled global list (6 533).

Key figures			
The numbers that open a bibliometric paper.			
Timespan 2010-2026	Annual growth 18.44 %	Mean age 5.32 years	Citations / document 15.36
Authors / document 4.73	Authors / co-authored document 4.82	Single-authored documents 3	International co-authorship 52.4 %
References / document 51.8	Author keywords 344	Index keywords 1,079	Document types 4

Key figures: the card placed at the top of a bibliometric article.

How it is computed

$$\text{Annual growth} = \left[\left(\frac{n_{y_{\max}}}{n_{y_{\min}}} \right)^{1/(y_{\max} - y_{\min})} - 1 \right] \times 100 \quad (= 18.44 \%)$$

$$\text{Mean document age} = \frac{1}{N} \sum_d (y_{\max} - y_d) \quad (= 5.32 \text{ years})$$

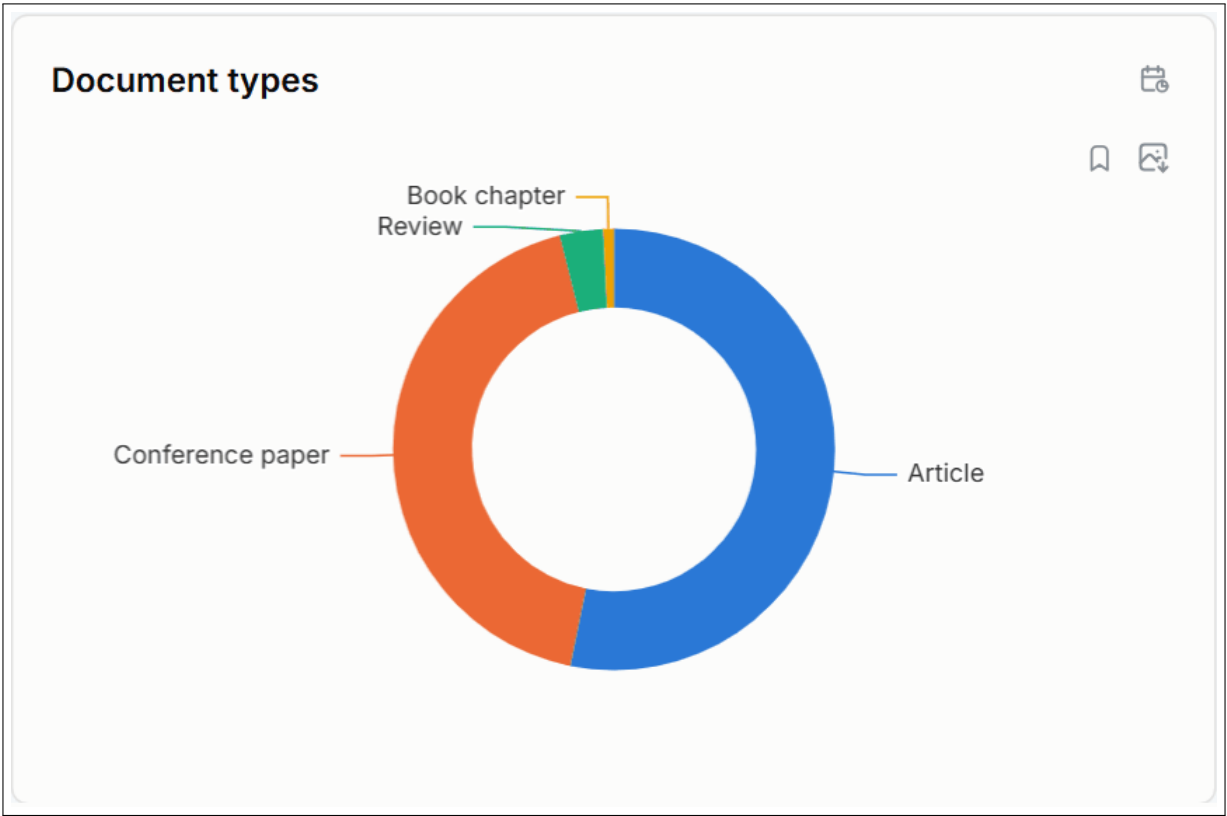
$$\text{Citations per document} = \frac{1}{N} \sum_d C_d \quad (= 15.36)$$

$$\text{Authors per document} = \frac{1}{N} \sum_d A_d \quad (= 4.73)$$

Authors per co-authored document = mean of A_d over co-authored documents only ($A_d > 1$) $(= 4.82)$

$$\text{International co-authorship} = \frac{\#\{d : \text{at least 2 countries}\}}{\#\{d : \text{at least 1 country}\}} \times 100 \quad (= 52.4 \%)$$

Age is counted from the last year *of the corpus*, never from today: otherwise the value would change every 1 January without anything having changed in the data.



Document types: share of each type in the corpus.

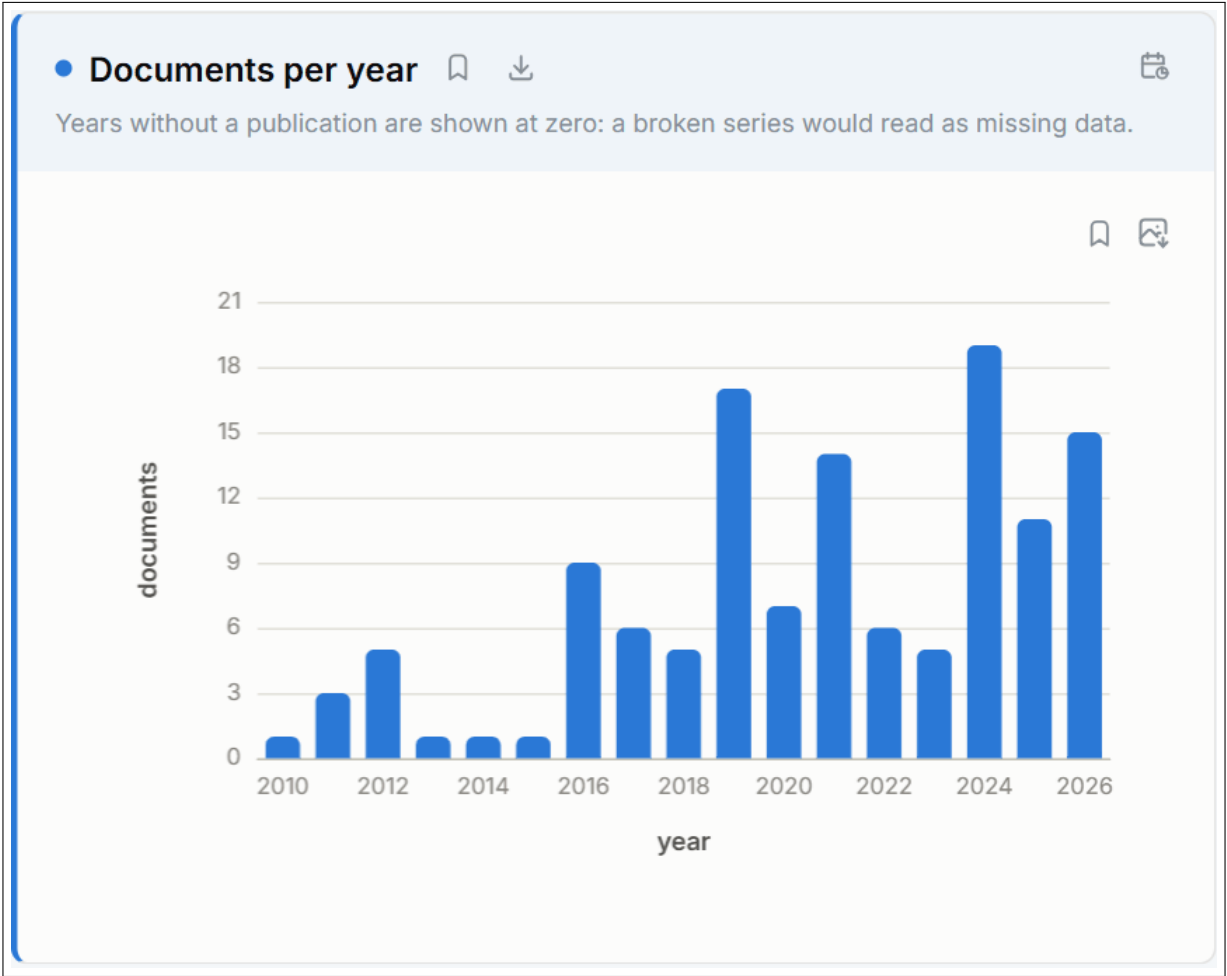
Most cited documents 			
"/yr" relates citations to age: without that column a 2016 paper always crushes a 2024 one.			
Title	Year	Cit.	/yr
Reviewing ensemble classification methods in breast cancer Hosni M.	2019	188	23.5
Systematic literature review of ensemble effort estimation Idri A.	2016	150	13.64
Requirements engineering tools: Capabilities, survey and assessment Carrillo de Gea J.M.	2012	95	6.33
Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles Idri A.	2016	95	8.64
On the value of parameter tuning in heterogeneous ensembles effort estimation Hosni M.	2018	75	8.33
Security vulnerabilities in healthcare: an analysis of medical devices and software Mejía-Granda C.M.	2024	73	24.33
Requirements engineering tools Carrillo de Gea J.M.	2011	55	3.44
Effects of competitive computer-assisted learning versus conventional teaching methods on the... Fernández-Alemán J.L.	2011	55	3.44
Investigating heterogeneous ensembles with filter			

The most cited documents (excerpt): global citations C_d and citations per year.

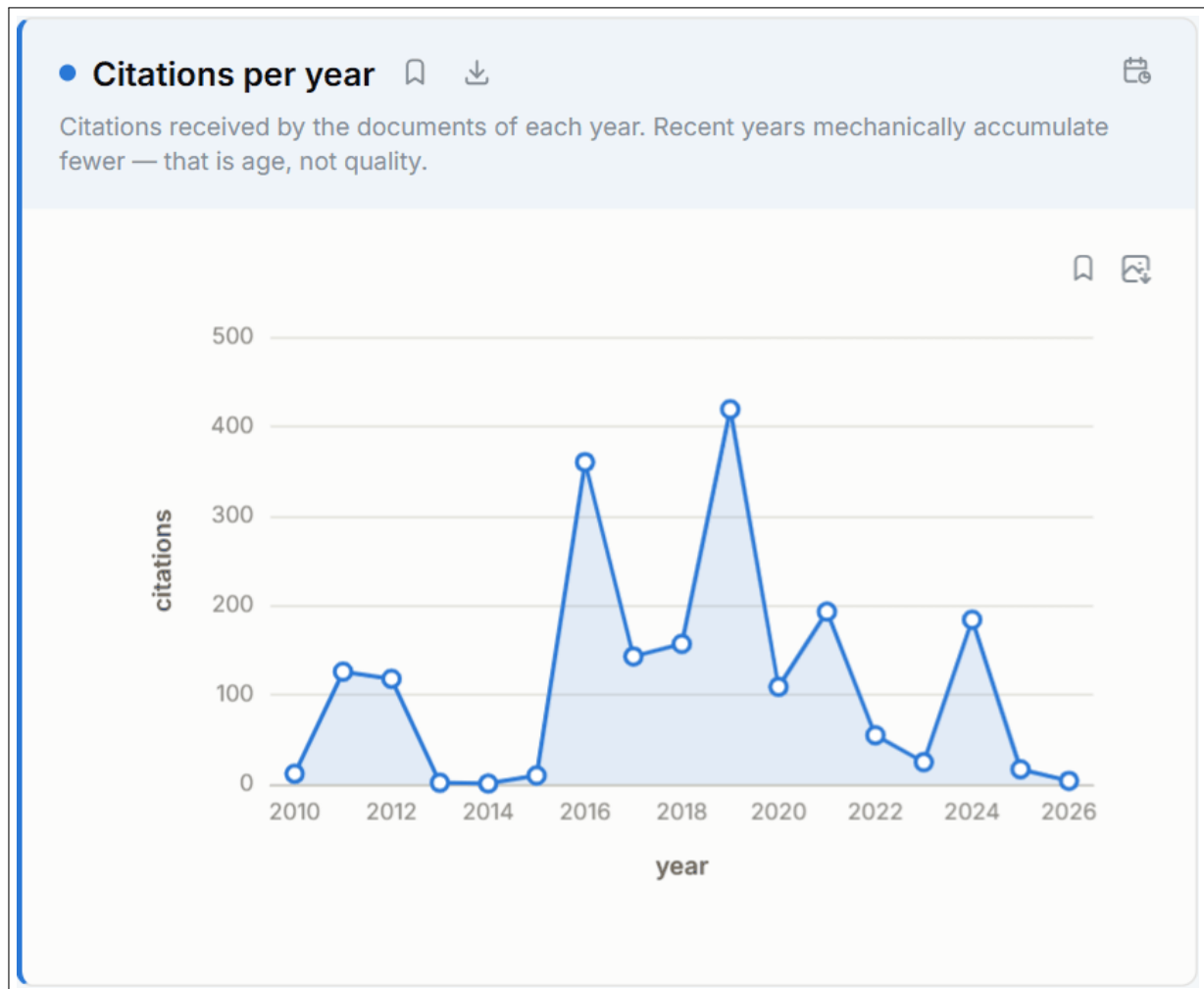
How it is computed

Citations per year = $C_d / (y_{\max} - y_d + 1)$. They correct the advantage of older articles, which have had more time to be cited.

3 Production



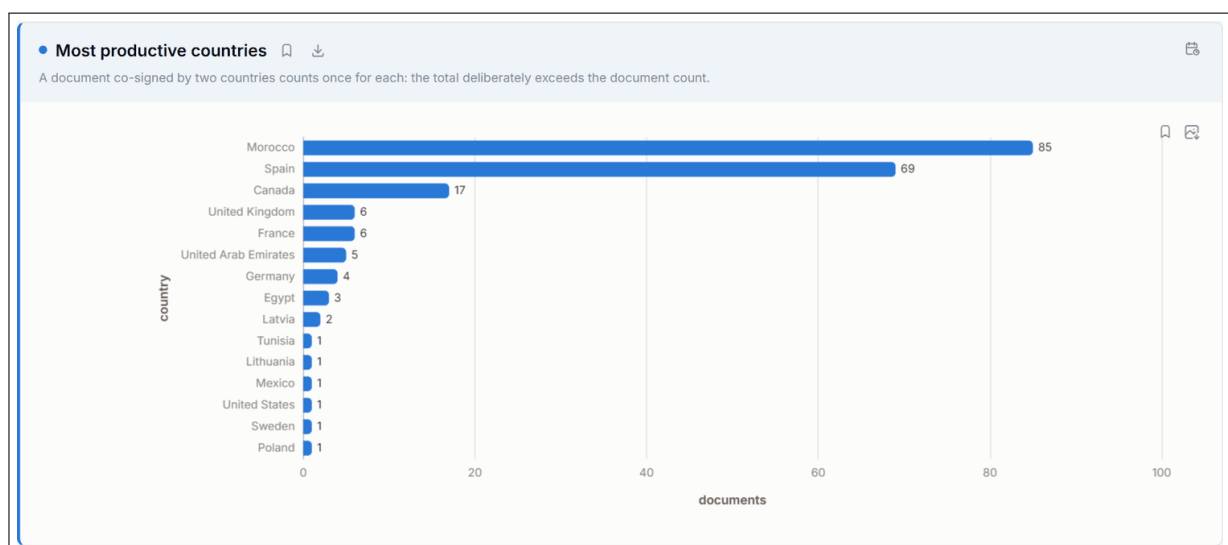
Documents per year, n_t . Years without publications are shown at zero.



Citations per year: $\sum_{d:y_d=t} C_d$, the citations received by the documents published in year t .

How to read it

The last years mechanically have fewer citations: their documents have had less time to receive them. That is age, not quality. To compare years, see the normalised citations (section 4).



Most productive countries, in full counting: a document co-signed by Morocco and Spain counts for each.

4 Documents

Document list  					Citations			Year	Title	Rows
					 Search title, author, source					200
126 of 126 documents										
Title	Source	Year	GC	LC						
Reviewing ensemble classification methods in breast cancer Hosni M. · Review	Computer Methods and Programs in Biomedicine	2019	188							
Systematic literature review of ensemble effort estimation Idri A. · Article	Journal of Systems and Software	2016	150							
Requirements engineering tools: Capabilities, survey and assessment Carrillo de Gea J.M. · Article	Information and Software Technology	2012	95							
Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles Idri A. · Article	Applied Soft Computing Journal	2016	95							
On the value of parameter tuning in heterogeneous ensembles effort estimation Hosni M. · Article	Soft Computing	2018	75							
Security vulnerabilities in healthcare: an analysis of medical devices and software Mejía-Granda C.M. · Article	Medical and Biological Engineering and Computing	2024	73							
Requirements engineering tools Carrillo de Gea J.M. · Article	IEEE Software	2011	55							
Effects of competitive computer-assisted learning versus conventional teaching methods on the acquisition and retention of knowledge in medical surgical nursing students Fernández-Alemán J.L. · Article	Nurse Education Today	2011	55	0						
Investigating heterogeneous ensembles with filter feature selection for software effort estimation Hosni M. · Conference paper	ACM International Conference Proceeding Series	2017	47							
Design and development of a mobile app for accessible beach tourism information for people with disabilities Mayordomo-Martínez D. · Article	International Journal of Environmental Research and Public Health	2019	46							
Systematic mapping study of ensemble effort estimation Idri A. · Conference paper	ENASE 2016 - Proceedings of the 11th International Conference on Evaluation of Novel Software Approaches t...	2016	40							
On the risks and safeguards for requirements engineering in global software development: Systematic literature review and quantitative assessment Nicolás J. · Article	IEEE Access	2018	32	0						
E-health internationalization requirements for audit purposes Ouhbi S. · Article	Computer Methods and Programs in Biomedicine	2017	31							
Effects of gamification on the benefits of student response systems in learning of human anatomy: Three experimental studies López-Jiménez J.J. · Article	International Journal of Environmental Research and Public Health	2021	31	0						
Assessing the impact of parameters tuning in ensemble based breast Cancer classification Idri A. · Article	Health and Technology	2020	31							
Software vulnerabilities overview: A descriptive study Sánchez M.C. · Article	Tsinghua Science and Technology	2020	30							
Evaluating a gamification proposal for learning usability heuristics: Heureka Sobrinho-Duque R. · Article	International Journal of Human Computer Studies	2022	30							
Classification techniques in breast cancer diagnosis: A systematic literature review El Ouassif B. · Article	Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization	2021	30							
Heterogeneous Ensembles for Software Development Effort Estimation Hosni M. · Conference paper	Proceedings - 2016 3rd International Conference on Soft Computing and Machine Intelligence, ISCMi 2016	2017	29							
A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard Mapping El baldi M. · Article	Water Resources Management	2024	29							
Software project management approaches for global software development: A systematic mapping study Baita M.E. · Article	Tsinghua Science and Technology	2018	27	0						

The list of documents (excerpt), with global (GC) and local (LC) citations.

• **Normalized citations**   

Citations divided by the mean of the SAME publication year. 1.0 is exactly average for its cohort, 3.0 is three times better than its contemporaries. It is the only column that lets a 2016 paper and a 2024 one be ranked together.

Title	Year	Citations	Year mean	Normalized
Reviewing ensemble classification methods in breast cancer Hosni M.	2019	188	24.65	7.628
Security vulnerabilities in healthcare: an analysis of medical devices and software Mejia - Granda C.M.	2024	73	9.68	7.538
A watermarking-based multimodal healthcare data protection using blockchain and deep learning Chekira C.	2026	2	0.27	7.5
Requirements engineering tools: Capabilities, survey and assessment Carrillo de Gea J.M.	2012	95	23.6	4.025
A method and validation for auditing e-Health applications based on reusable software security requirements specifications Mejia - Granda C.M.	2025	6	1.55	3.882
Spatio - Temporal Analysis of Urban Expansion and Its Impact on Agricultural Land in the Casablanca Metropolitan Periphery Nakhili B.	2026	1	0.27	3.75
Systematic literature review of ensemble effort estimation Idri A.	2016	150	40	3.75
Predictive Maintenance in Industrial Systems Using Machine Learning : A Review Benmansour O.	2026	1	0.27	3.75
Evaluating a gamification proposal for learning usability heuristics: Heureka Sobrinho - Duque R.	2022	30	9.17	3.273
A theory on human factors in DevOps adoption Pérez - Sánchez J.	2025	5	1.55	3.235
A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard Mapping El baida M.	2024	29	9.68	2.995
On the value of parameter tuning in heterogeneous ensembles effort estimation Hosni M.	2018	75	31.4	2.389
Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles Idri A.	2016	95	40	2.375
Effects of gamification on the benefits of student response systems in learning of human anatomy: Three experimental studies López - Jiménez J.J.	2021	31	13.79	2.249
Classification techniques in breast cancer diagnosis: A systematic literature review El Ouassif B.	2021	30	13.79	2.176
Real - Time Urban Flood Depth Mapping: Convolutional Neural Networks for Pluvial and Fluvial Flood Emulation El baida M.	2024	20	9.68	2.065
Assessing the impact of parameters tuning in ensemble based breast Cancer classification Idri A.	2020	31	15.57	1.991
Investigating heterogeneous ensembles with filter feature selection for software effort estimation Hosni M.	2017	47	23.83	1.972
Software vulnerabilities overview: A descriptive study Sánchez M.C.	2020	30	15.57	1.927
Design and development of a mobile app for accessible beach tourism information for people with disabilities Mayordomo - Martínez D.	2019	46	24.65	1.866

Normalised citations: each document compared with the documents of the same year.

How it is computed

$$NC_d = \frac{C_d}{\bar{C}_{y_d}}, \quad \bar{C}_y = \frac{1}{n_y} \sum_{d': y_{d'}=y} C_{d'}$$

$NC_d = 1$: exactly the mean of its year; $NC_d = 3$: three times better than its contemporaries. It is the only measure that makes it possible to rank an article from 2016 and one from 2024 together. A year without any citation gives 0.

5 Data quality: what the corpus allows

• **What this corpus supports**  

Each row links a family of analyses to the field it depends on, with its real coverage. This is what tells you, before writing a sentence, whether the figure covers the whole corpus or two thirds of it.

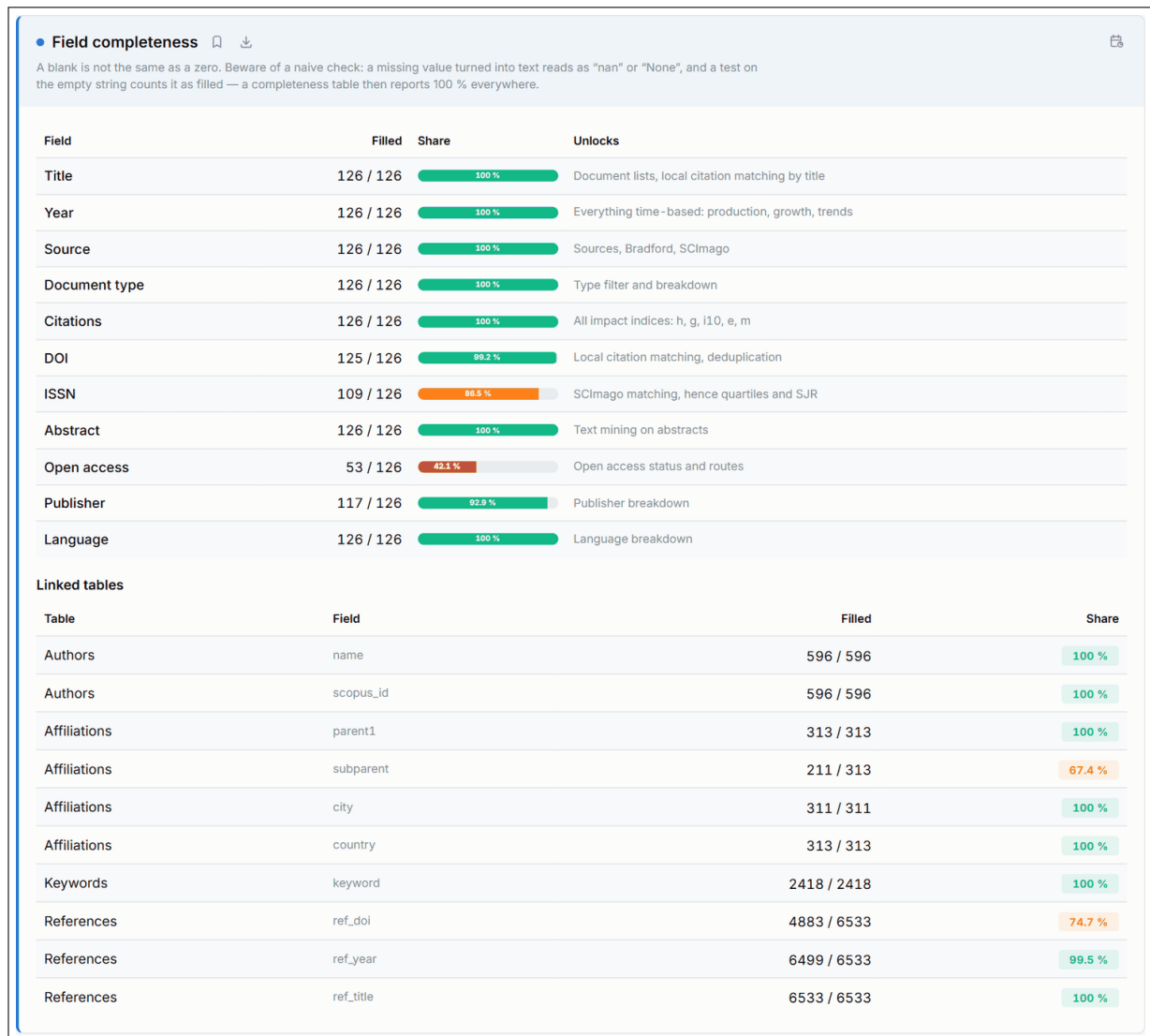
Analysis	Status	Coverage	Depends on	Note
Production and growth	RELIAB...	100 %	Publication year	Documents without a year are excluded from every time series.
Impact indices (h, g, i10, e, m)	RELIAB...	100 %	Citation count	Computed on THIS corpus, not on each person's whole body of work.
Author analysis	RELIAB...	100 %	Author names	Scopus identifier present on 100 % of rows — homonyms are only separable where it is.
Institutions	RELIAB...	100 %	Parent organisation	Internal units filled on 67 % of affiliations.
Cities and geography	RELIAB...	100 %	City	Drives the local / national / international collaboration scale.
Keywords and thematic map	RELIAB...	100 %	Keywords	Co-word, thematic map and evolution all rest on this field.
Text mining	RELIAB...	100 %	Abstract	Terms are extracted from titles and abstracts.
Local citations and historiograph	PARTIAL	74.7 %	Reference DOIs	229 internal citation links actually matched.
SCImago quartiles and SJR	PARTIAL	75.4 %	ISSN matching	59 of 78 sources matched.
Open access	LIMITED	42.1 %	Open access flag	Scopus only flags OPEN articles; a blank is not a closed article.

What this corpus allows: each analysis, the field it depends on, its coverage and its status.

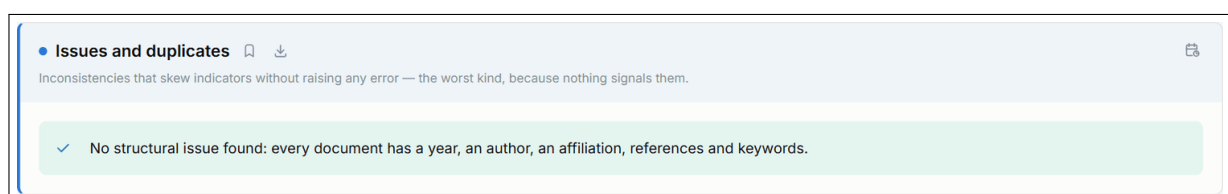
How it is computed

$$\text{coverage}(\text{field}) = \frac{\#\{d : \text{field filled}\}}{N} \times 100$$

Status: **ready** if the coverage is $\geq 90\%$, **partial** between 50 and 90%, **limited** below. A “nan” or “None” value or an empty string counts as missing.



Completeness field by field, then table by table (authors, affiliations, keywords, references).



Anomalies and duplicates: missing or future years, unreadable citations, documents without author or affiliation.

6 Corpus profile

6.1 The Price index and the age of references



Price index per publication year, and distribution of the age of the cited references.

How it is computed

For each dated reference: $\text{age} = y_{\text{citing}} - y_{\text{reference}}$ (an age of -1 or -2 , an “in press” reference, is brought back to 0; beyond 200 years, the year is considered wrong and set aside).

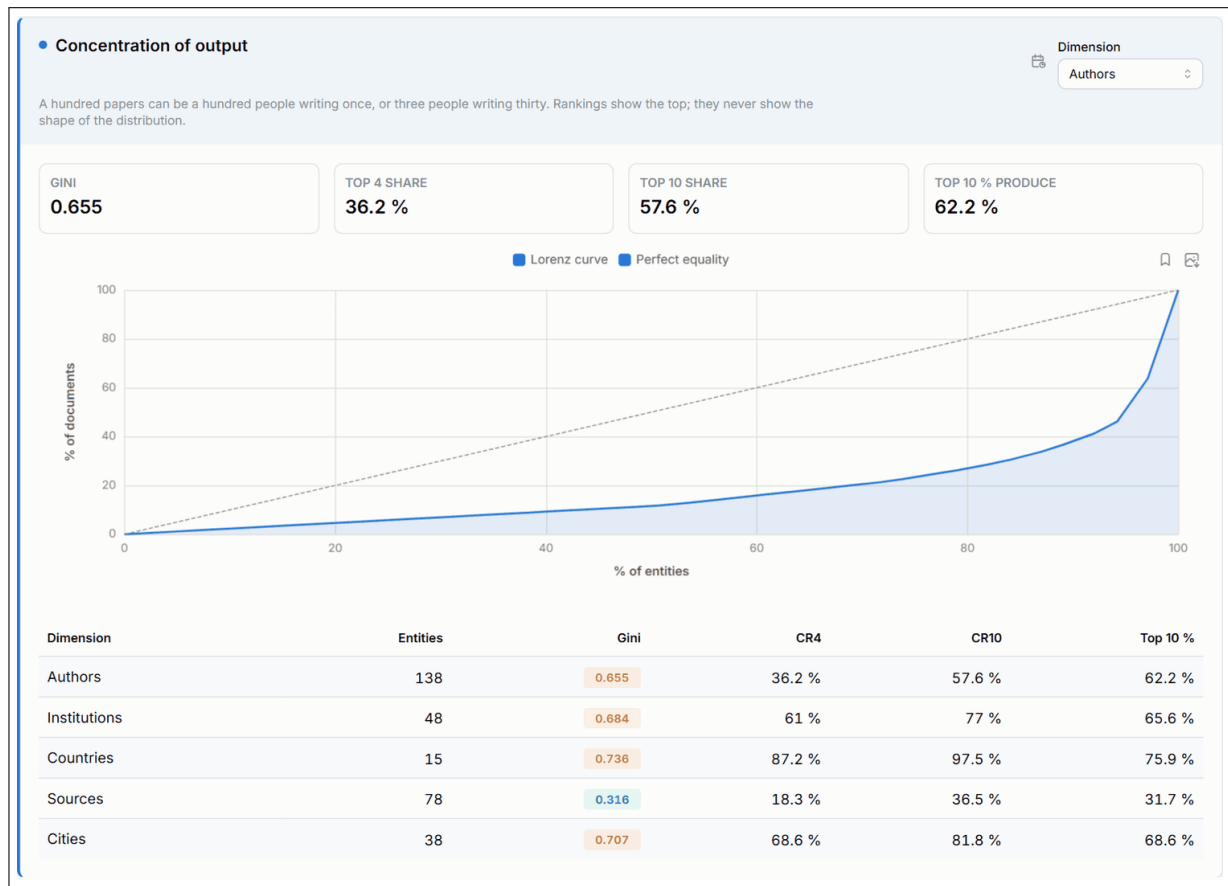
$$\text{Price index} = \frac{\#\{\text{references aged} < 5 \text{ years}\}}{\#\{\text{dated references}\}} \times 100 \quad (= 34.9\%)$$

Median age = cited half-life = 7 years; mean age = 9.6 years, over 6 488 dated references.

How to read it

Above about 50 %, a research front that renews itself quickly; below about 30 %, a field resting on established works. At 34.9 %, this corpus is in between.

6.2 Concentration: Gini, Lorenz, CR4, CR10



Lorenz curve of the authors, and concentration table for each dimension.

How it is computed

For n entities with productions $x_{(1)} \leq \dots \leq x_{(n)}$ (sorted in increasing order):

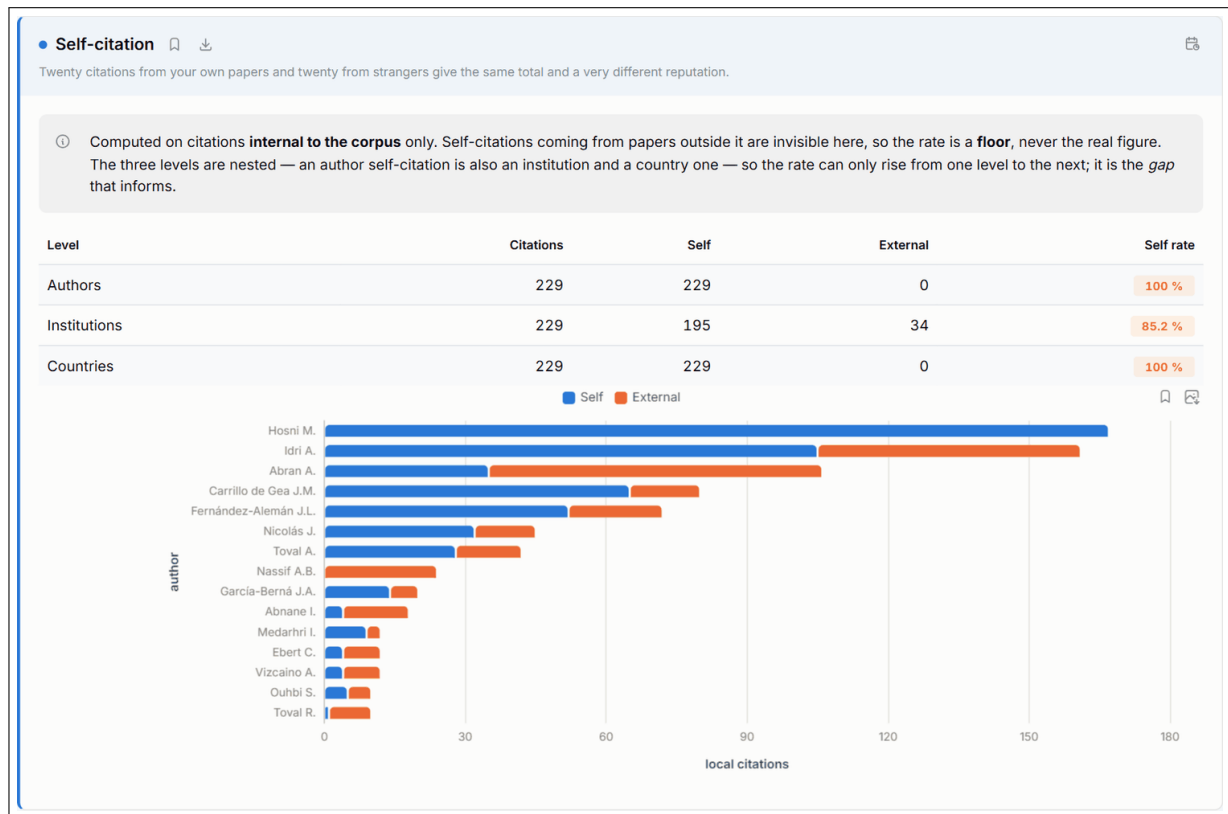
$$G = \frac{2 \sum_{i=1}^n i x_{(i)}}{n \sum_{i=1}^n x_{(i)}} - \frac{n+1}{n}$$

Lorenz curve: cumulative share of the entities (x-axis) against cumulative share of the production (y-axis); the diagonal is perfect equality. CR4 and CR10: share of the production of the top 4 and top 10 entities; “Top 10 %”: share of the 10 % most productive.

How to read it

Authors: Gini 0.655, CR10 57.6 %; countries: Gini 0.736, CR4 87.2 %. The production of countries is highly concentrated (Morocco and Spain), that of sources only slightly (Gini 0.316).

6.3 Self-citations



Self-citation rate at the three levels, and per author.

How it is computed

On the citations *internal* to the corpus (citing → cited pairs, section 12):

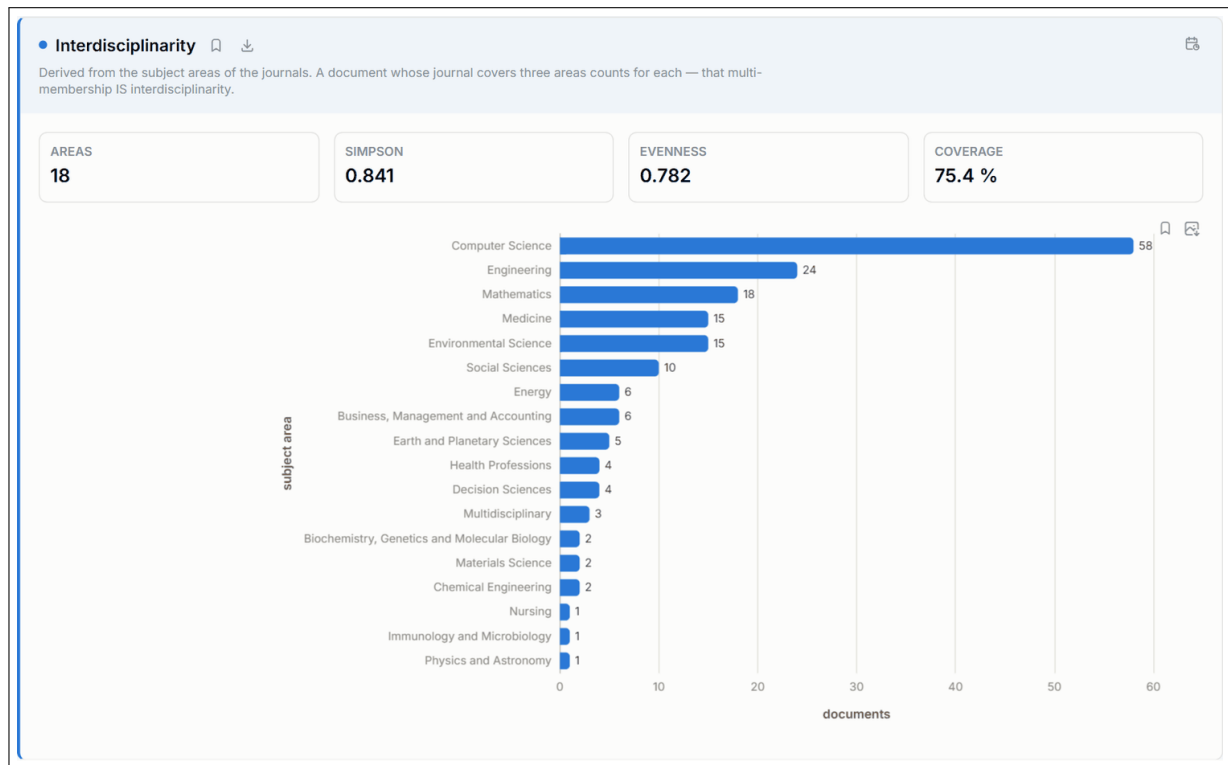
$$\text{rate} = \frac{\#\{\text{pairs whose citing and cited documents share at least one entity}\}}{\#\{\text{pairs}\}} \times 100$$

The entity is an author (Scopus identifier), an institution or a country.

Caution

This rate only sees the citations internal to the corpus: it is a floor, not the real rate. Here, 229 internal citations, all between documents that share an author (100%): the corpus is that of a single research group.

6.4 Interdisciplinarity



SCImago areas of the corpus journals.

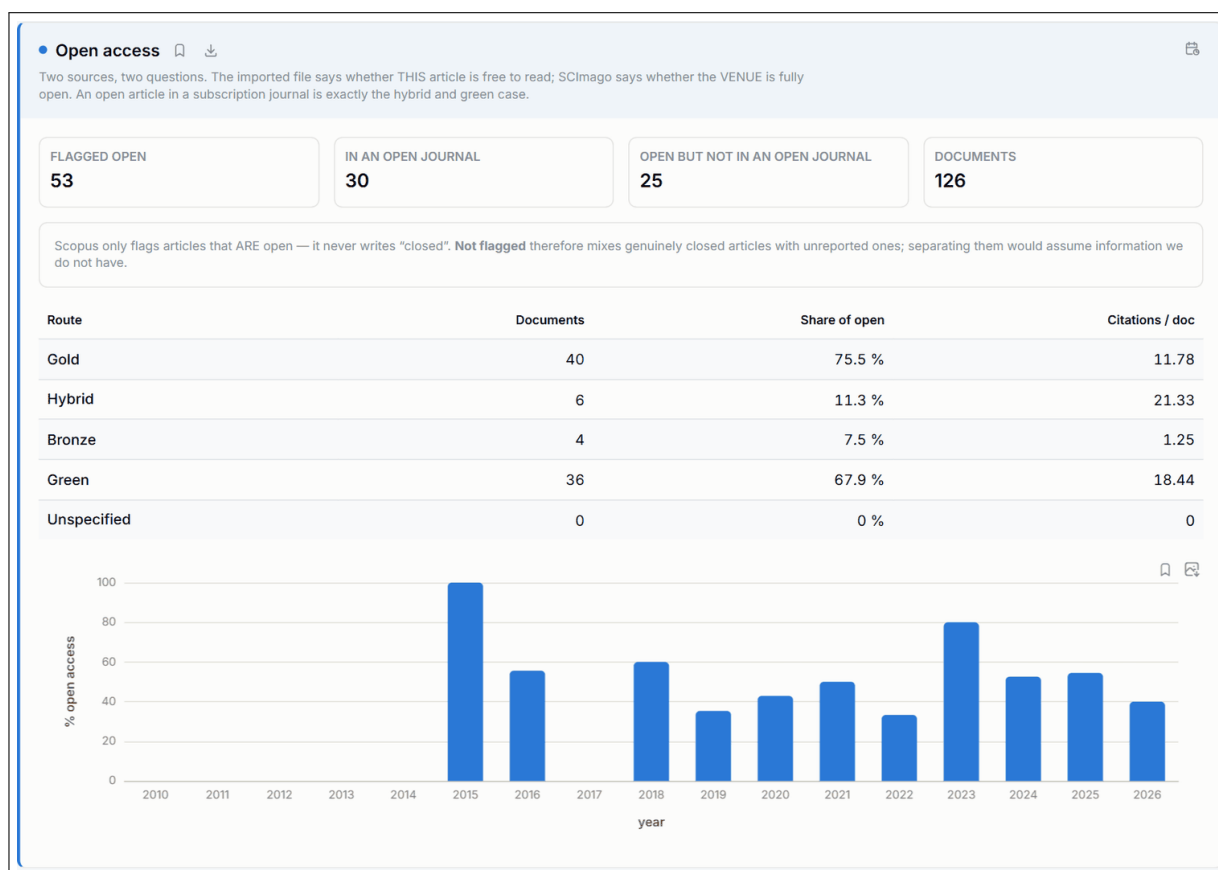
How it is computed

With p_k the share of documents in area k (K areas):

$$H = - \sum_k p_k \ln p_k \quad (\text{Shannon}), \quad D = 1 - \sum_k p_k^2 \quad (\text{Simpson}), \quad J = \frac{H}{\ln K} \quad (\text{evenness})$$

Here $K = 18$, $H = 2.26$, $D = 0.841$, $J = 0.782$; the first area, Computer Science, carries 46 % of the documents. Coverage: 75.4 % of the documents (those whose journal is in SCImago).

6.5 Open access

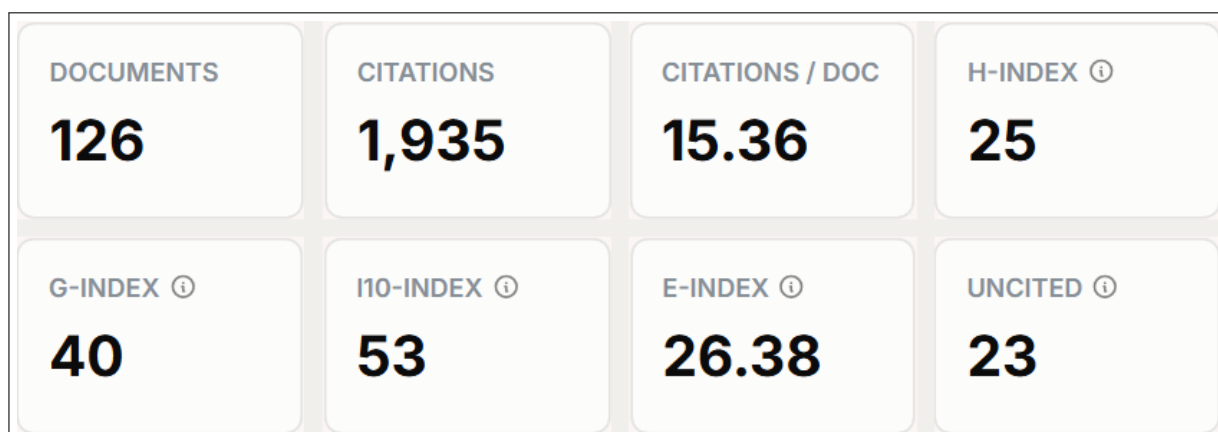


Open access: status, route (gold, hybrid, bronze, green) and annual evolution.

How it is computed

Share of the documents flagged as open access by Scopus (*Open Access* column), route by route, and citations per document of each route. A “not flagged” document is not necessarily closed: Scopus simply does not say.

7 Authors



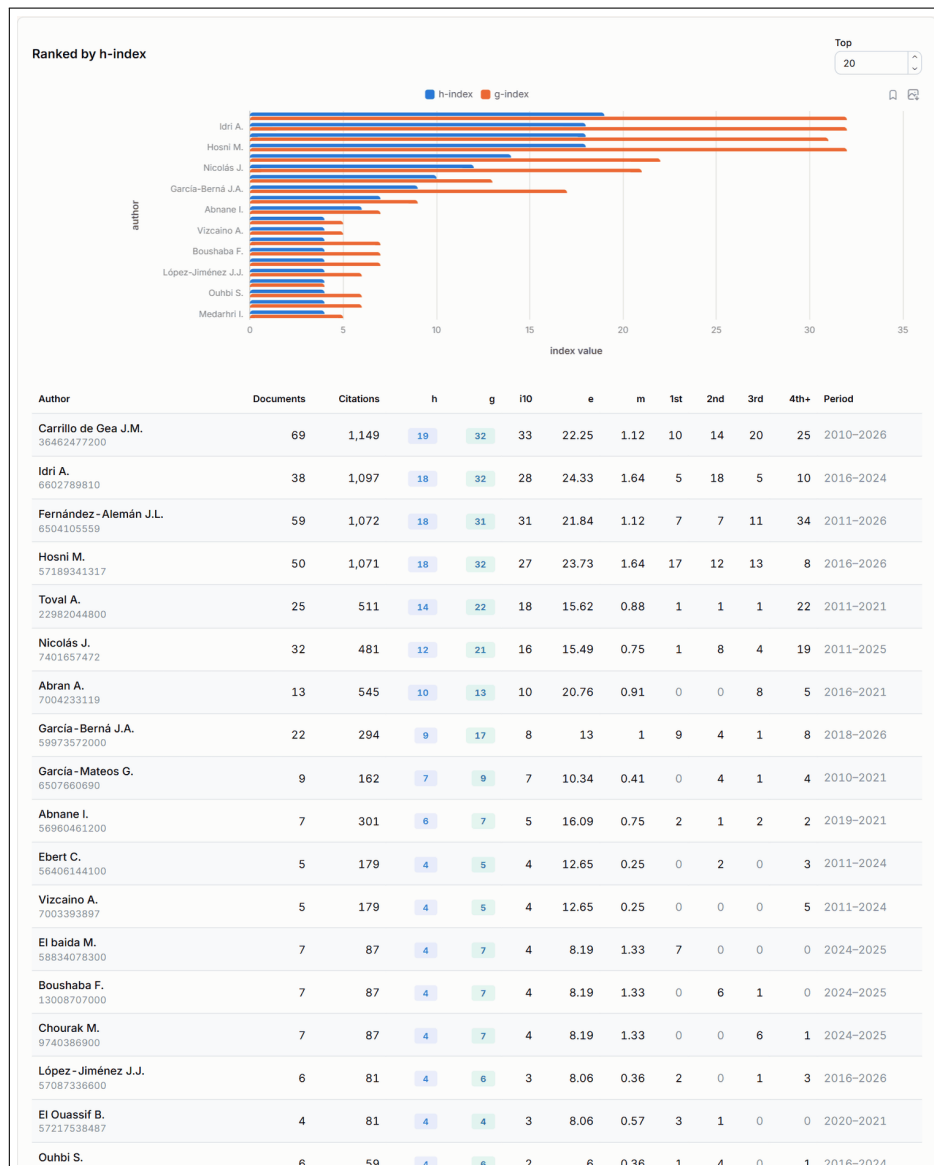
The impact indices of the whole corpus.

How it is computed

With the sorted citations $C_{(1)} \geq C_{(2)} \geq \dots$:

$$\begin{aligned}
 h &= \max\{h : C_{(h)} \geq h\} & (\text{Hirsch, 2005}): h &= 25 \\
 g &= \max\{g : \sum_{i=1}^g C_{(i)} \geq g^2\} & (\text{Egghe, 2006}): g &= 40 \\
 i_{10} &= \#\{d : C_d \geq 10\} & &= 53 \\
 e &= \sqrt{\sum_{i=1}^h C_{(i)} - h^2} & (\text{Zhang, 2009}): e &= 26.38 \\
 m &= \frac{h}{y_{\text{last}} - y_{\text{first}} + 1} & (\text{per author or institution}) &
 \end{aligned}$$

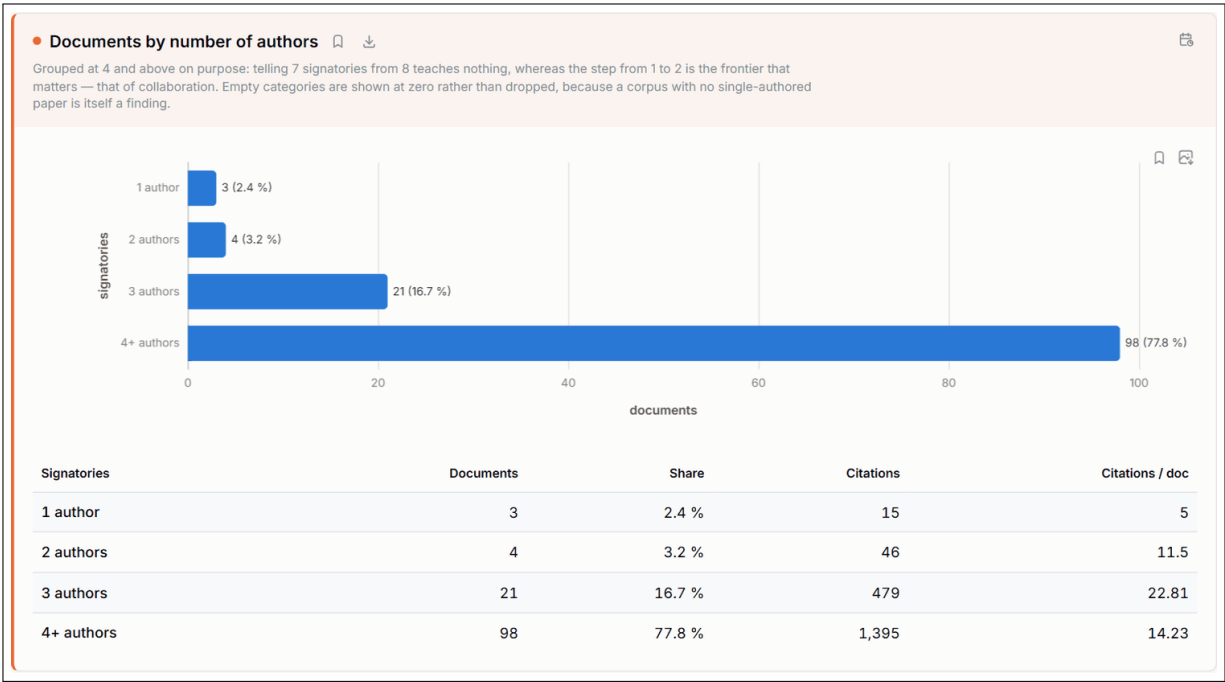
The e measures what h ignores: the citations of the core beyond h^2 . Two authors with the same h are told apart by their e . “Uncited”: documents with 0 citations (23).



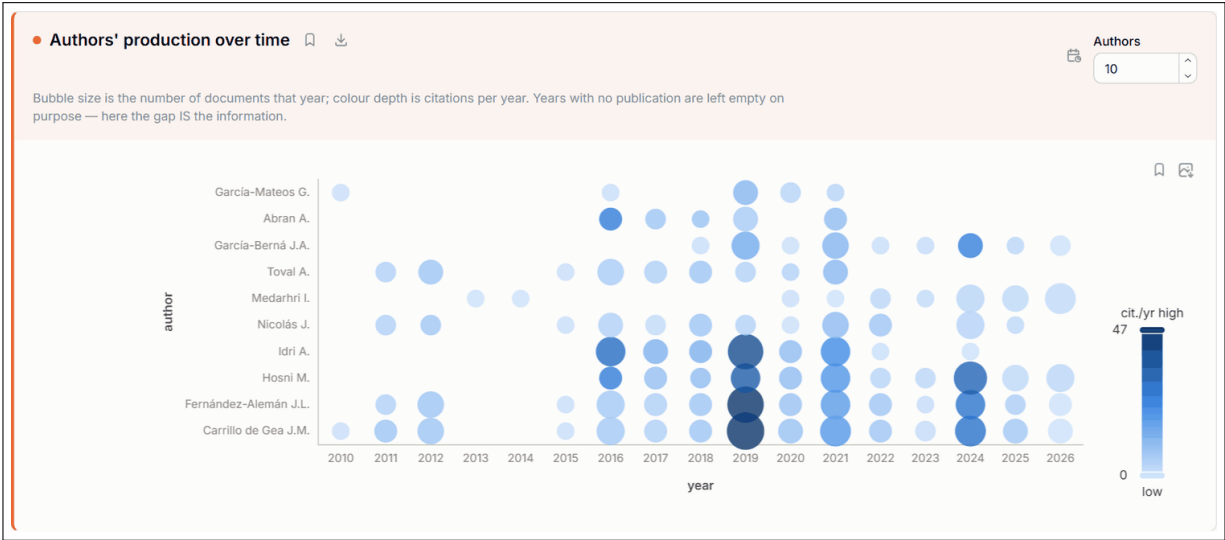
Authors ranked by h-index (excerpt): h and g per author, then the full table (documents, citations, m , authorship rank).

How to read it

Carrillo de Gea J.M.: 69 documents, $h = 19$, $g = 32$, $m = 1.12$; Idri A.: 38 documents, $h = 18$, $m = 1.64$. Idri's m is higher: the same impact, over a shorter career in the corpus.

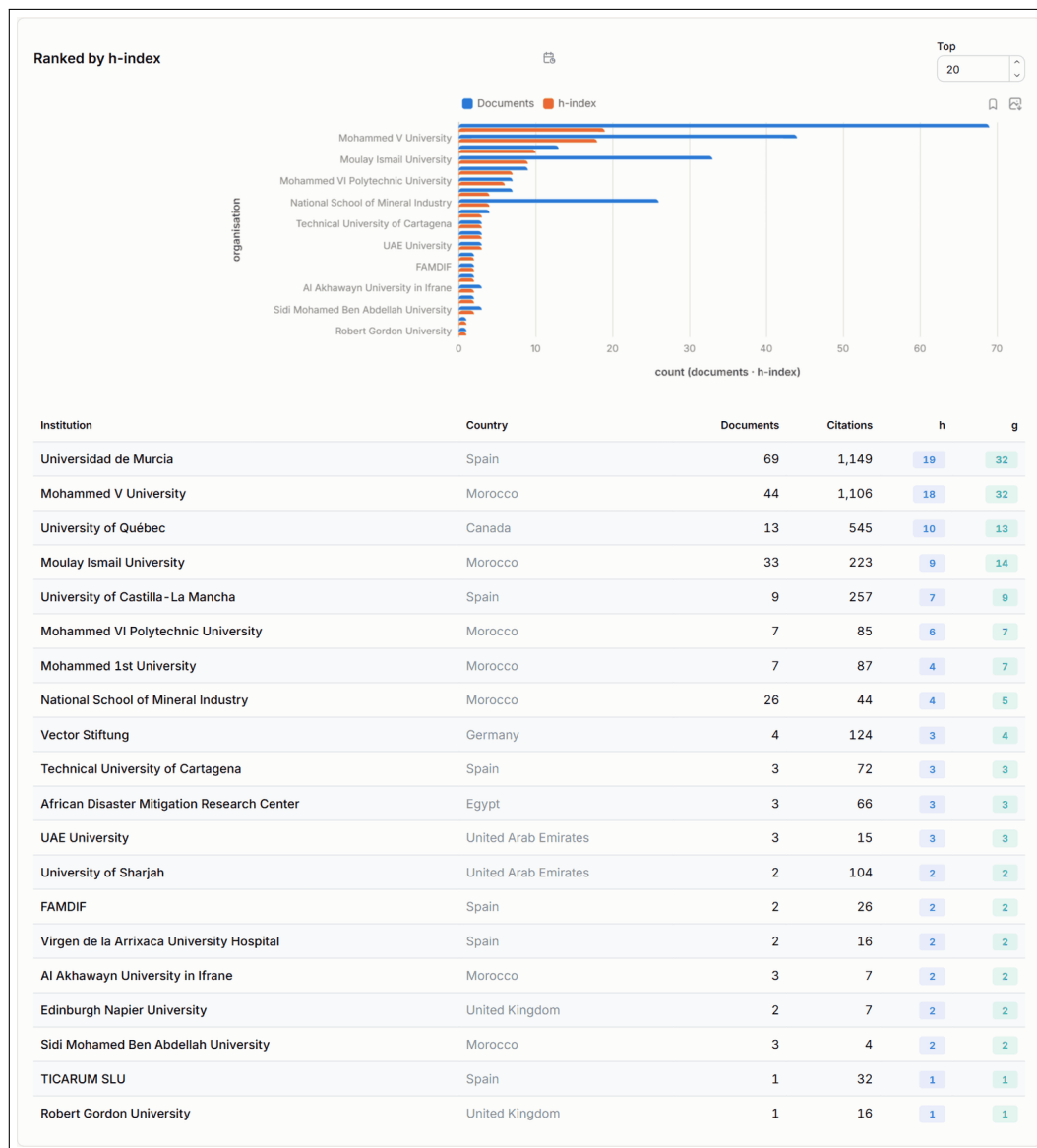


Documents by number of authors, and citations per document of each group.

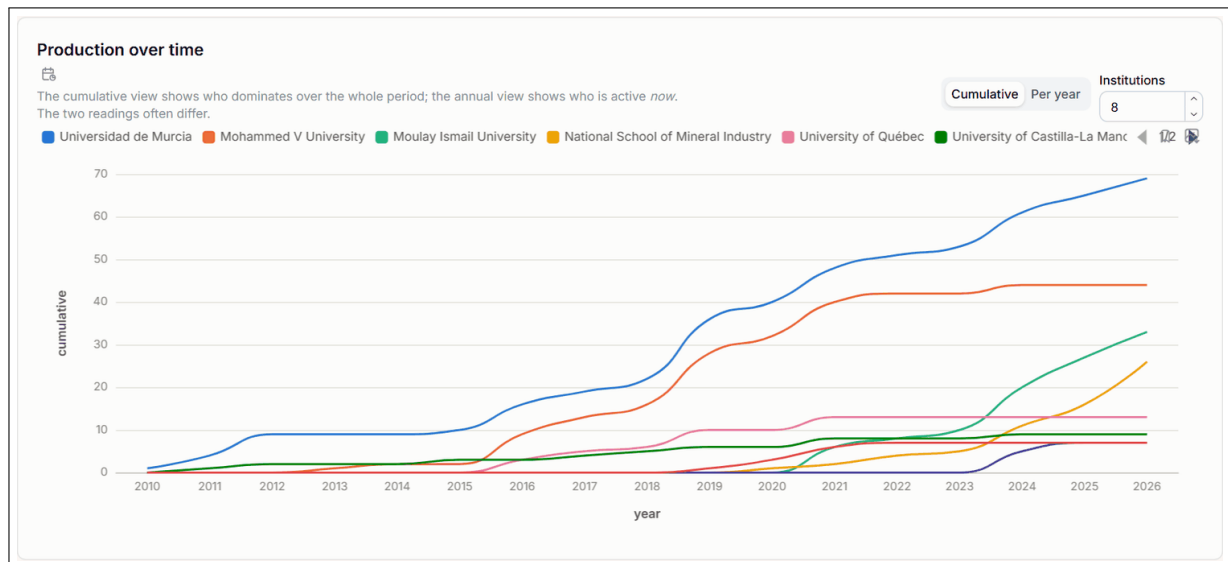


Author production over time: one disc per author and per year, larger as the author published more.

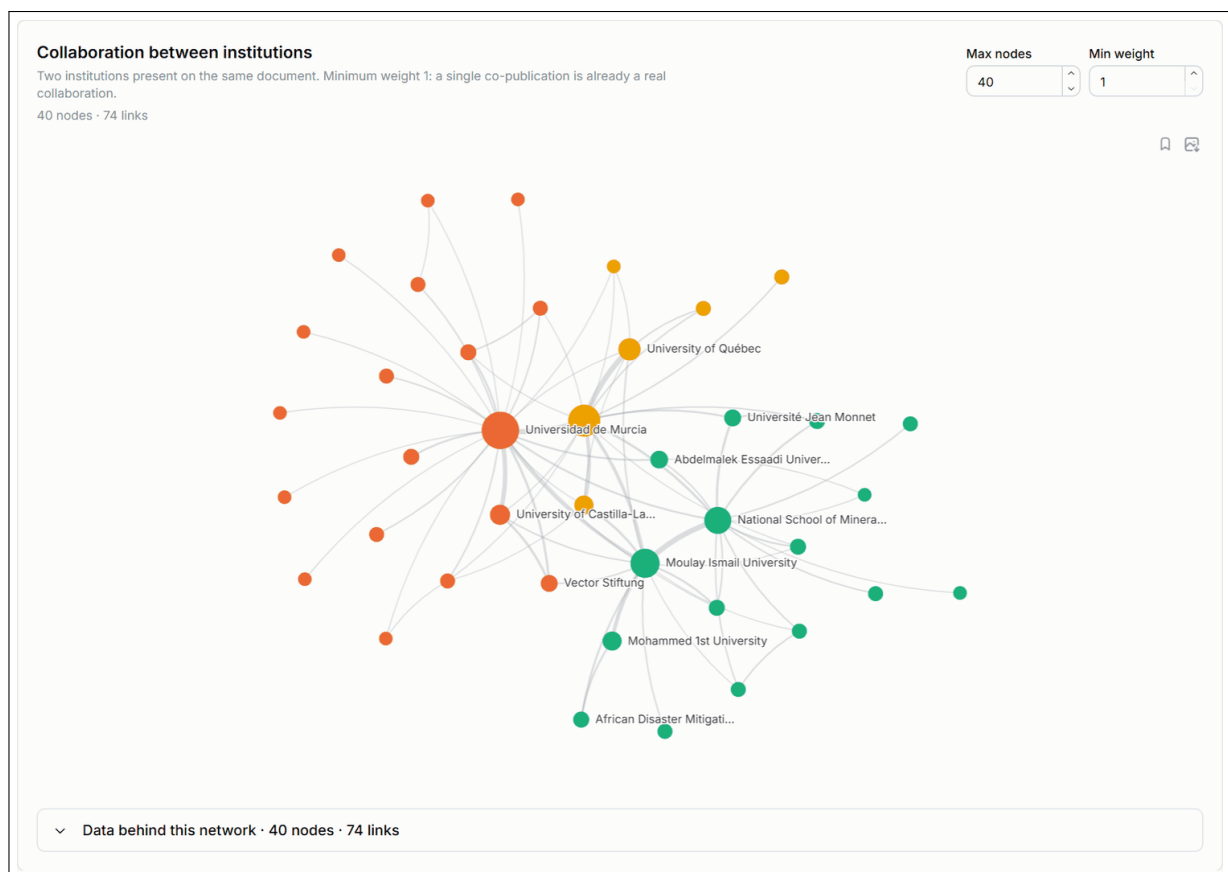
8 Institutions



Institutions ranked by h-index (excerpt). Same formulas as for authors, on the documents of each parent organisation.



Institution production over time (annual or cumulative).



Collaboration network between institutions: a link = co-signed documents.

How it is computed

Weight of the link between two institutions = number of documents signed by both. The rules shared by all networks are in section 22.

Institution → units hierarchy

As established by the cleaning step. The **share** tells whether a unit carries most of its institution's output or is only one component of it. It is computed over **all** the documents of the institution, including those with no identified unit.

Institution	Unit	Docs	Citations	Share
Universidad de Murcia	Faculty of Computer Science	54	1,028	78.3 %
↳	Faculty of Medicine	3	51	4.3 %
↳	Facultad de Educación	2	31	2.9 %
↳	Faculty of Informatics	2	55	2.9 %
↳	School of Medicine	1	12	1.4 %
Mohammed V University	ENSIAS	38	1,097	86.4 %
↳	Mohammadia Engineering School	5	9	11.4 %
↳	Faculty of Sciences	1	0	2.3 %
↳	High School of Technology - Sale	1	0	2.3 %
Moulay Ismail University	ENSAM	33	223	100 %
University of Québec	ETS	13	545	100 %
University of Castilla - La Mancha	Institute of Information Technologies and Systems	5	151	55.6 %
↳	Escuela Superior de Informática	2	56	22.2 %
↳	Faculty of Computer Science	2	50	22.2 %
Mohammed 1st University	National School of Applied Sciences of Oujda	7	87	100 %
↳	Faculty of Sciences Oujda - FSO	4	38	57.1 %
Abdelmalek Essaadi University	ENSATE	2	0	40 %
↳	Ecole Normale Supérieure de Tétouan (ENS Tétouan)	1	0	20 %
↳	Polydisciplinary Faculty of Larache	1	1	20 %
Université Jean Monnet	Institut Camille Jordan	4	6	100 %
Al Akhawayn University in Ifrane	School of Science and Engineering	3	7	100 %
Sidi Mohamed Ben Abdellah University	FST	3	4	100 %
↳	ENSA	1	2	33.3 %
Cadi Ayyad University	ENSA - S	2	0	100 %
University of Coventry	Faculty of Engineering and Computing	1	3	50 %
Virgen de la Arrixaca University Hospital	Institute of Biomedical Research of Murcia	2	16	100 %
Abdelmalek Essaadi University Tetuan	National School of Applied Science of Tetuan	1	0	100 %
Univ. Bordeaux	ESTIA - Institute of Technology	1	0	100 %
University of Calgary	Werklund School of Education	1	0	100 %
University of Sfax	Faculty of Economics and Management of Sfax	1	14	100 %
War Studies University	Faculty of Management and Command	1	0	100 %
↳	National Security Faculty	1	0	100 %

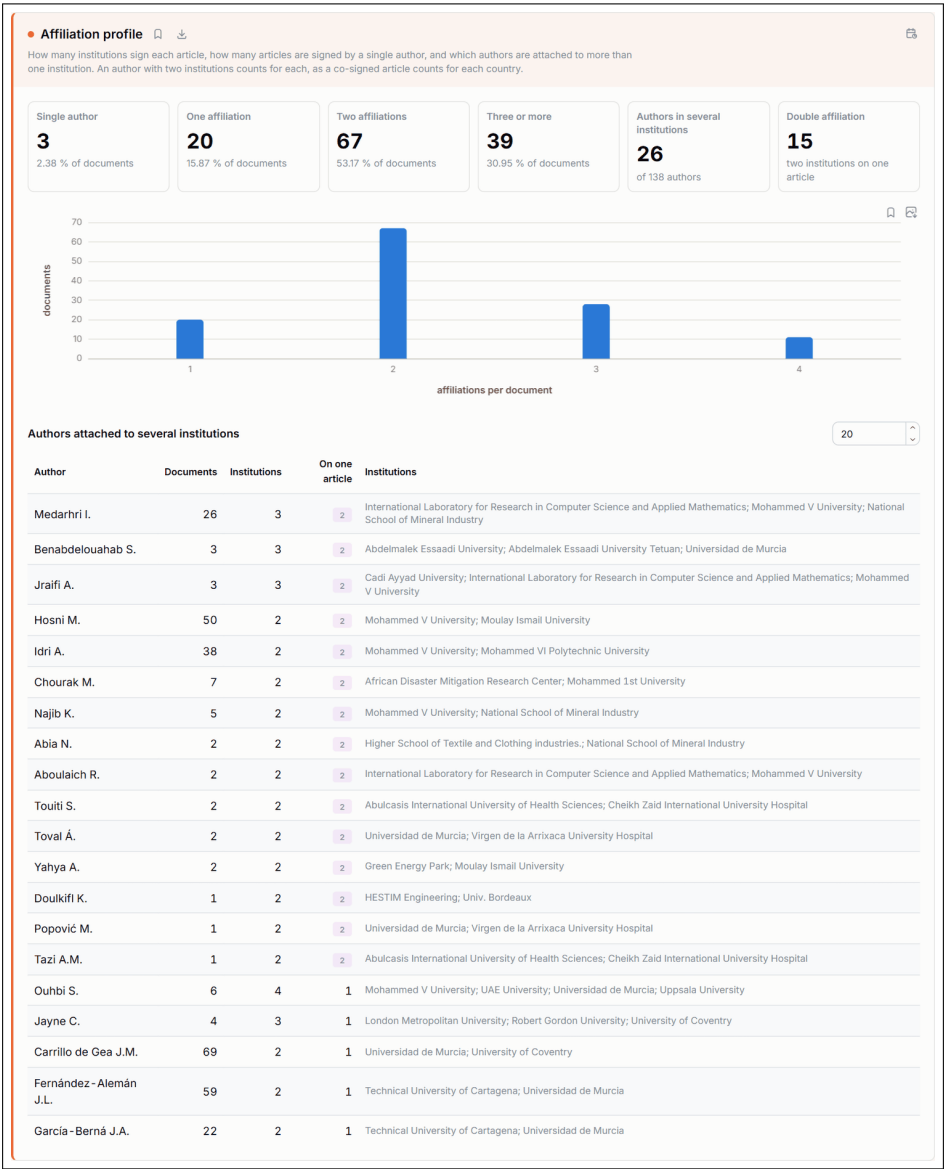
Hierarchy: each organisation and its sub-units (faculties, schools, laboratories), as the cleaning structured them.

Breakdown by country

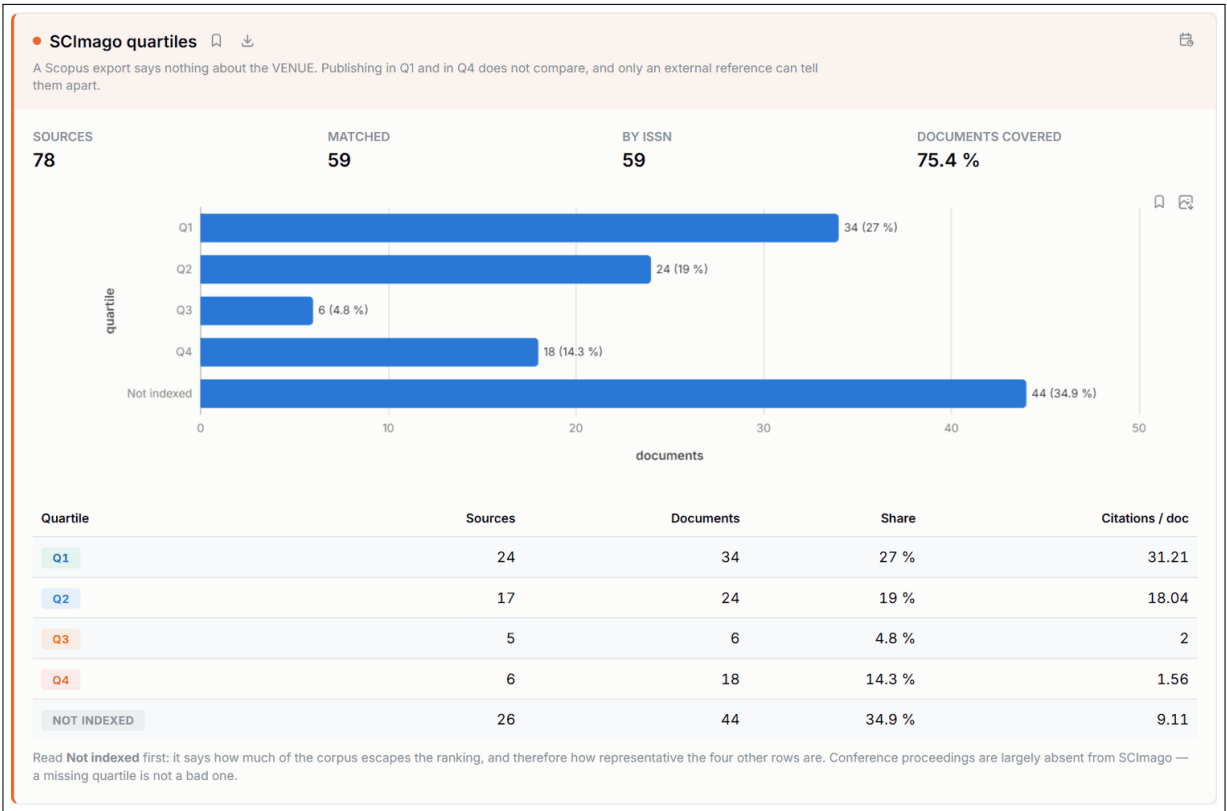
How many institutions each country mobilizes, and which one dominates.

Country	Institutions	Documents	Leading
Morocco	18	85	Mohammed V University
Spain	9	69	Universidad de Murcia
Canada	3	17	University of Québec
United Kingdom	4	6	Edinburgh Napier University
France	3	6	Université Jean Monnet
United Arab Emirates	2	5	UAE University
Germany	1	4	Vector Stiftung
Egypt	1	3	African Disaster Mitigation Research Center
Latvia	1	2	Transport and Telecommunication Institute
Poland	2	1	Kazimierz Wielki University
Lithuania	1	1	European Humanities University
Mexico	1	1	Autonomous University of Guerrero
Sweden	1	1	Uppsala University
Tunisia	1	1	University of Sfax
United States	1	1	Louisiana State University

Breakdown by country: number of institutions and documents, and the leading institution in each country.



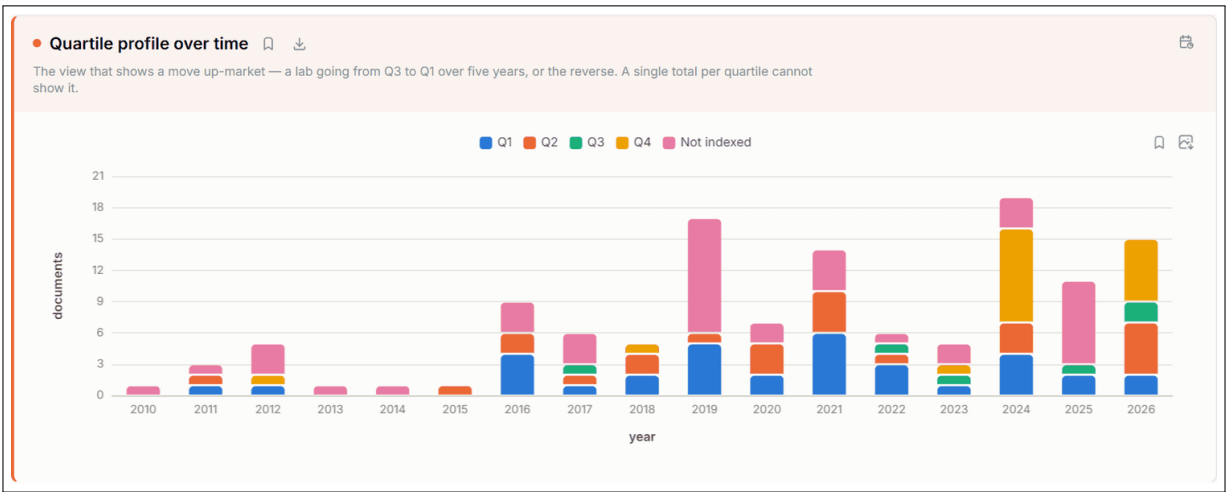
9 Sources



SCImago quartiles: documents by the quartile of their journal.

How it is computed

Each journal is matched with the SCImago index through its **ISSN** (never by title similarity). Its quartile is SCImago’s *SJR Best Quartile*. Share of a quartile = documents in journals of that quartile / *N*. A journal absent from SCImago is “Not indexed”, never “Q4” by default. Here 58 of the 78 sources are recognised, i.e. 75.4% of the documents.



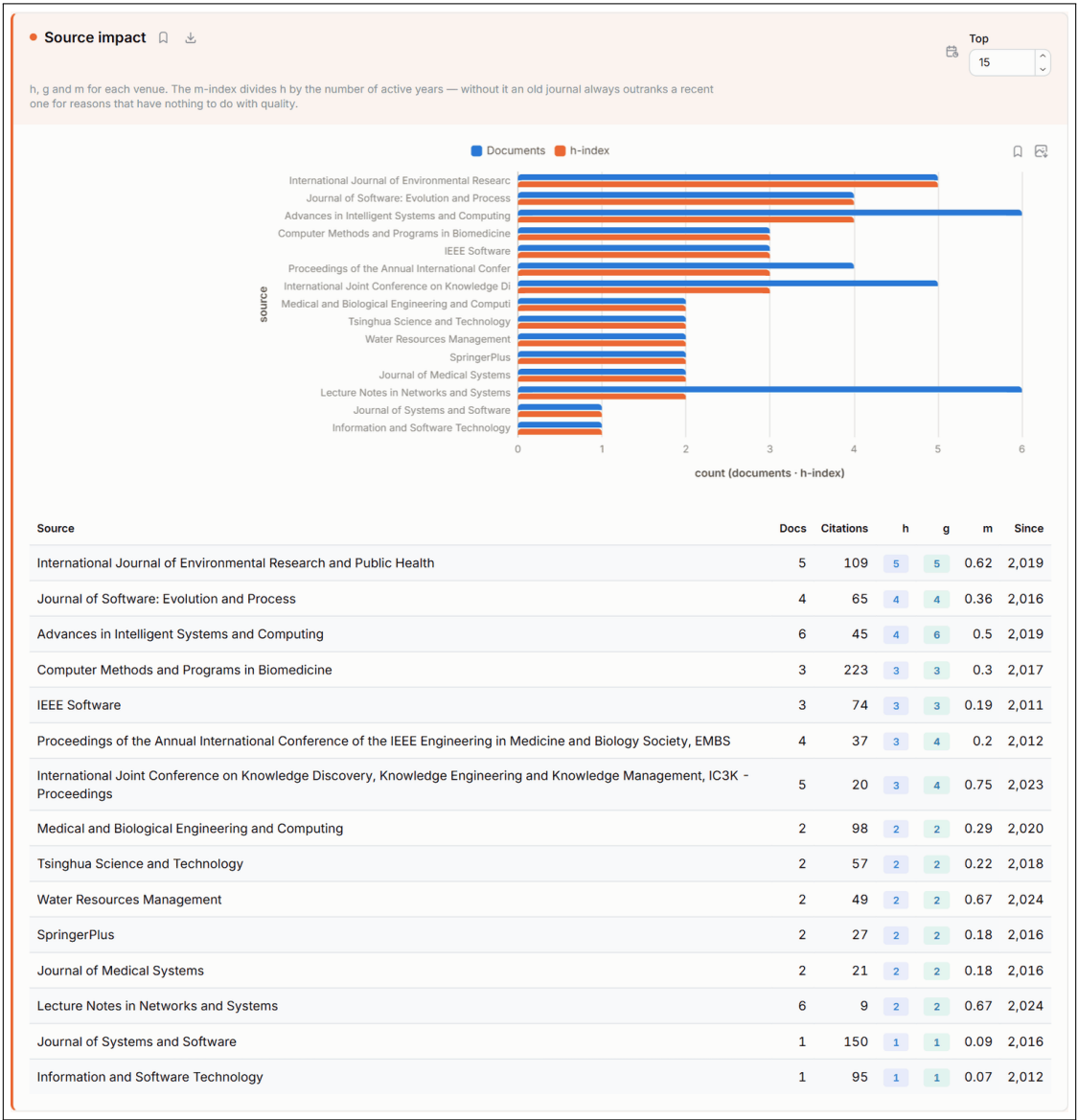
Quartile profile over time.

● **SCImago details**  

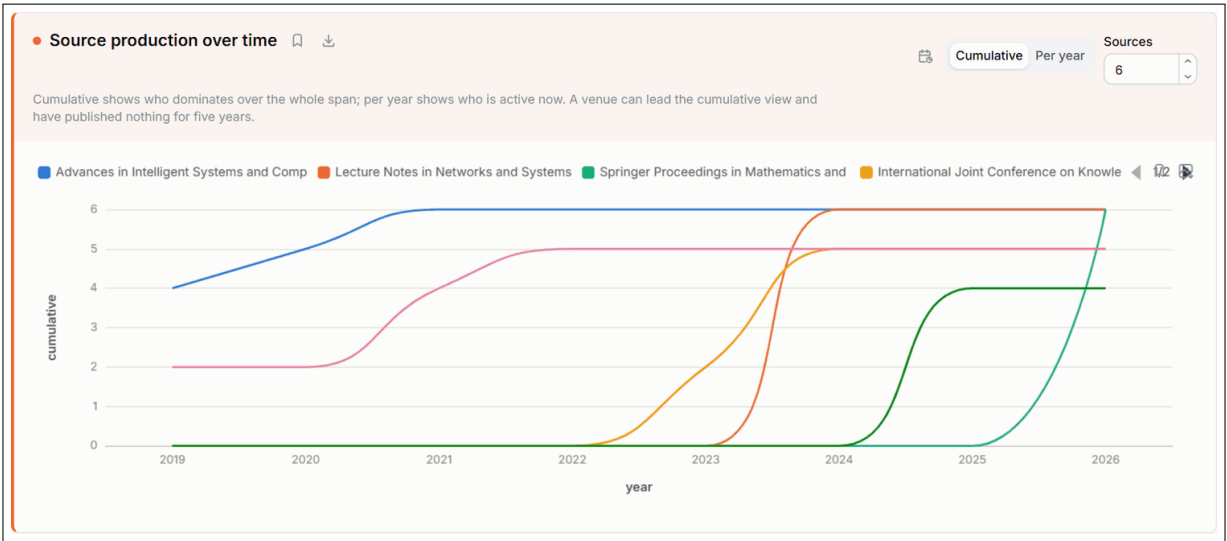
Everything the reference knows about each venue. An unmatched source keeps EMPTY cells — writing a default quartile would be an invention.

Source	Docs	Quartile	SJR	h	Cit/doc 2y	Type	Publisher	Country	OA	Matched
Advances in Intelligent Systems and Computing	6	nan	—	—	—	-	-	-	-	-
Lecture Notes in Networks and Systems Computer Networks and Communications (Q4); Control and...	6	Q4	0.165	57	0.56	book series	Springer International...	Switzerland	-	issn
Springer Proceedings in Mathematics and Statistics Mathematics (miscellaneous) (Q4)	6	Q4	0.152	35	0.3	book series	Springer	United States	-	issn
International Journal of Environmental Research and Public Health Public Health, Environmental and Occupational Health (Q1); Healt...	5	Q1	0.927	256	4.17	journal	Multidisciplinary Digita...	Switzerland	c	issn
International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, ... Management of Technology and Innovation; Software; Strategy a...	5	nan	0.169	7	0.56	conference and proceedings	Science and Technolo...	Portugal	-	issn
Journal of Software: Evolution and Process Software (Q2)	4	Q2	0.473	67	2.56	journal	John Wiley and Sons...	United Kingdom	-	issn
Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, ...	4	nan	—	—	—	-	-	-	-	-
International Conference for Artificial Intelligence: Applications, Innovation and Ethics, AIZE 2025	4	nan	—	—	—	-	-	-	-	-
Computer Methods and Programs in Biomedicine Computer Science Applications (Q1); Health Informatics (Q1);...	3	Q1	1.233	165	6.66	journal	Elsevier Ireland Ltd	Ireland	-	issn
IEEE Software Software (Q2)	3	Q2	0.627	134	3.27	journal	IEEE Computer Society	United States	-	issn
Mathematical Modeling and Computing Computational Mathematics (Q4); Computational Theory and...	3	Q4	0.227	14	1.12	journal	Lviv Polytechnic...	United Kingdom	c	issn
E3S Web of Conferences Earth and Planetary Sciences (miscellaneous); Energy...	3	nan	0.195	52	0.73	conference and proceedings	EDP Sciences	France	c	issn
Medical and Biological Engineering and Computing Biomedical Engineering (Q2); Computer Science Applications (Q2)	2	Q2	0.672	120	4.14	journal	Springer Science and...	Germany	-	issn
Tsinghua Science and Technology Multidisciplinary (Q1)	2	Q1	1.155	69	7.06	journal	Tsinghua University	China	c	issn
Water Resources Management Civil and Structural Engineering (Q1); Water Science and...	2	Q1	1.207	144	6.49	journal	Springer Science and...	Netherlands	-	issn
SpringerPlus	2	nan	—	—	—	-	-	-	-	-
Journal of Medical Systems Health Informatics (Q1); Health Information Management (Q1);...	2	Q1	1.299	127	8.34	journal	Springer	United States	-	issn
Journal of Universal Computer Science Computer Science (miscellaneous) (Q3); Theoretical Computer...	2	Q3	0.308	61	1.37	journal	Technische Universita...	Austria	c	issn
Numerical Algorithms Applied Mathematics (Q1)	2	Q1	0.823	85	2.61	journal	Springer Netherlands	Netherlands	-	issn
SoftwareX Computer Science Applications (Q2); Software (Q2)	2	Q2	0.499	59	2.23	journal	Elsevier B.V.	Netherlands	c	issn
Journal of Systems and Software Hardware and Architecture (Q1); Information Systems (Q1);...	1	Q1	0.95	141	5.88	journal	Elsevier Inc.	United States	-	issn
Applied Soft Computing Journal Software (Q1)	1	Q1	1.456	223	9.14	journal	Elsevier B.V.	Netherlands	-	issn
Information and Software Technology Computer Science Applications (Q1); Information Systems (Q1);...	1	Q1	1.054	137	6.24	journal	Elsevier B.V.	Netherlands	-	issn
Soft Computing	-	nan	—	—	—	-	-	-	-	-

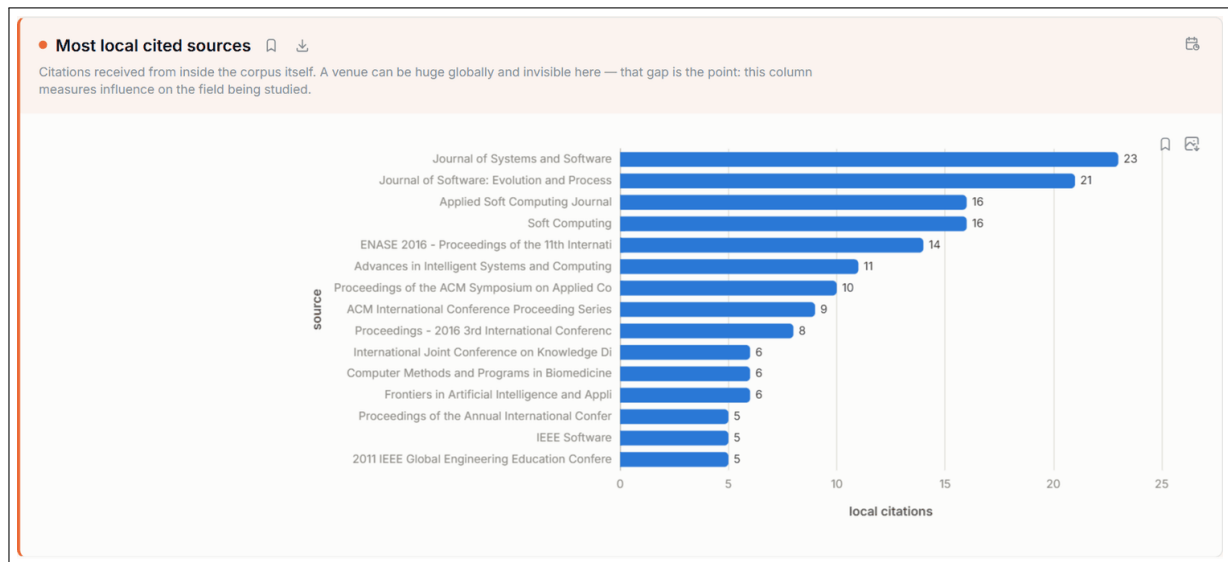
The SCImago details of each journal (excerpt): SJR, quartile, journal h-index, publisher, area.



Source impact: *h* and *g*-index computed on the corpus documents published in each journal.

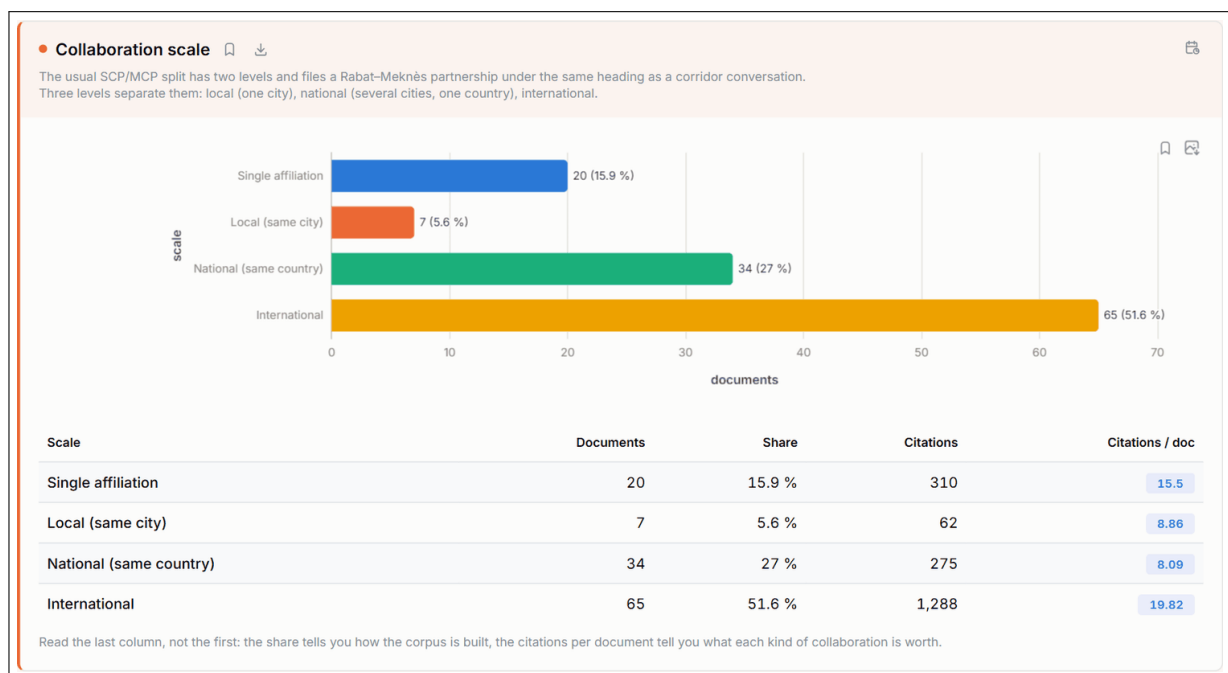


Source production over time (cumulative).



Most locally cited sources: $\sum LC_d$ over the documents of each journal.

10 Cities

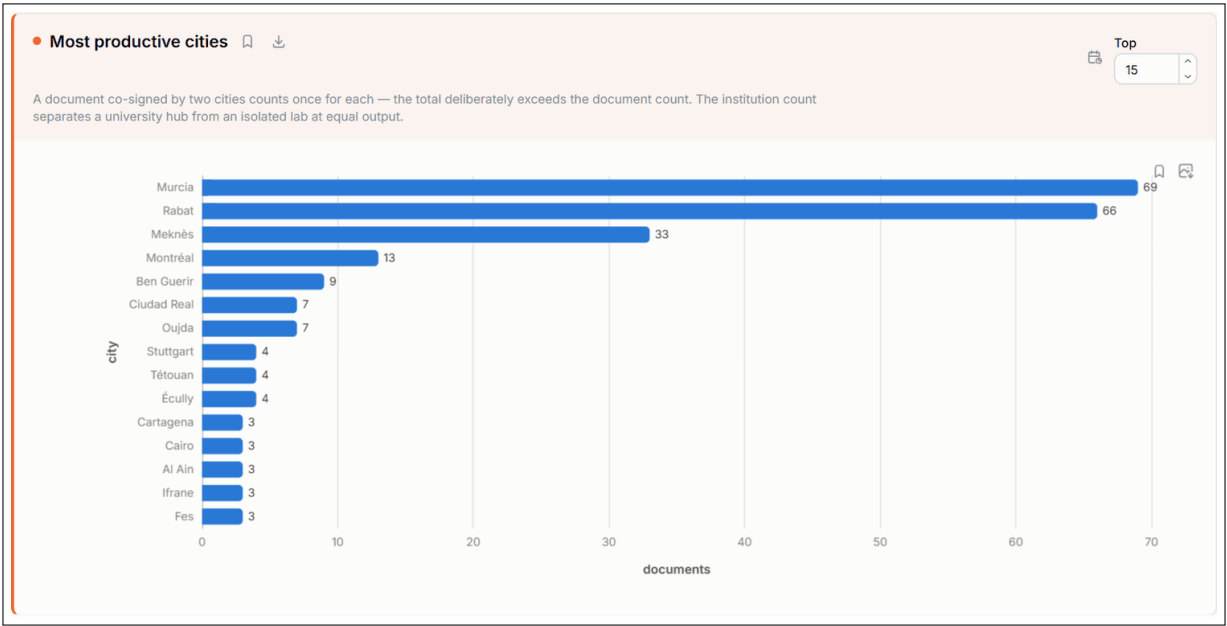


Collaboration scale: each document classified by the scope of its affiliations.

How it is computed

- **Single affiliation:** a single identified affiliation (this is not collaboration);
- **Local:** several affiliations, a single city;
- **National:** several cities, a single country;
- **International:** at least two countries.

Here: 15.9 % / 5.6 % / 27.0 % / 51.6 %. International documents are the most cited (19.8 citations per document, against 8.1 for national ones).



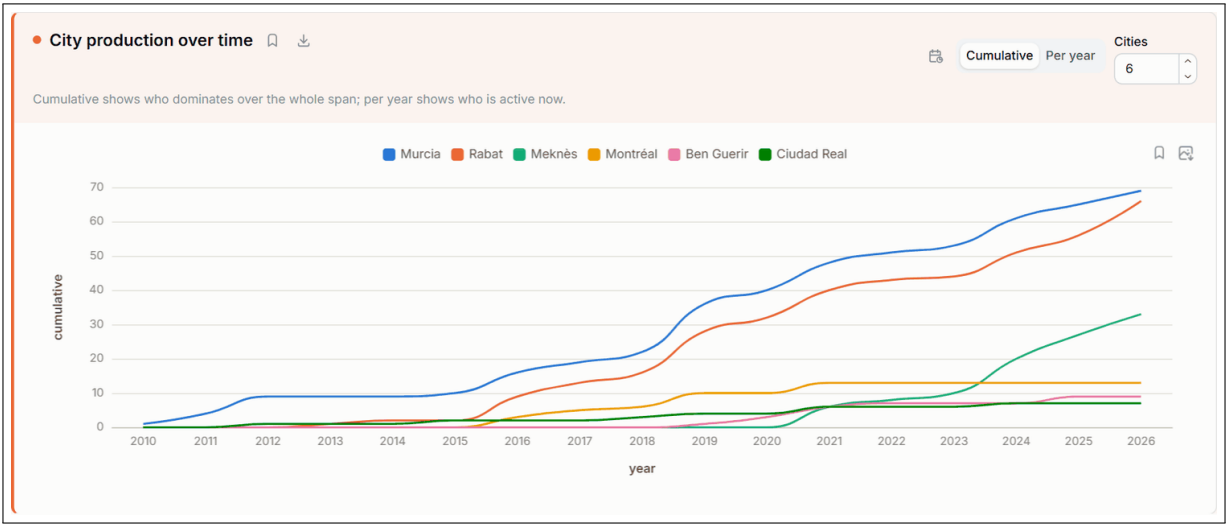
Most productive cities (full counting).

● **City impact**  

h, g and m per city. Compare h with m: a young hub can have a modest h and the best m of the corpus, because m divides by the number of active years.

City	Country	Documents	Citations	h	g	m	Since
Murcia	Spain	69	1,149	19	32	1.12	2,010
Rabat	Morocco	66	1,144	18	33	1.29	2,013
Montréal	Canada	13	545	10	13	0.91	2,016
Meknès	Morocco	33	223	9	14	1.5	2,021
Ciudad Real	Spain	7	201	6	7	0.4	2,012
Ben Guerir	Morocco	9	85	6	9	0.75	2,019
Oujda	Morocco	7	87	4	7	1.33	2,024
Stuttgart	Germany	4	124	3	4	0.2	2,012
Cartagena	Spain	3	72	3	3	0.38	2,019
Cairo	Egypt	3	66	3	3	1	2,024
Al Ain	United Arab Emirates	3	15	3	3	0.43	2,020
Sharjah	United Arab Emirates	2	104	2	2	0.2	2,017
Ifrane	Morocco	3	7	2	2	0.67	2,024
Edinburgh	United Kingdom	2	7	2	2	0.67	2,024
Fes	Morocco	3	4	2	2	1	2,025
Alarcón	Spain	2	56	1	2	0.06	2,011
Aberdeen	United Kingdom	1	16	1	1	0.09	2,016
Sfax	Tunisia	1	14	1	1	0.2	2,022
Riga	Latvia	2	10	1	2	0.12	2,019
Vilnius	Lithuania	1	10	1	1	0.12	2,019

City impact: documents, citations, h and g-index.



City production over time.



Collaboration between cities: a national link (same country) or an international one.

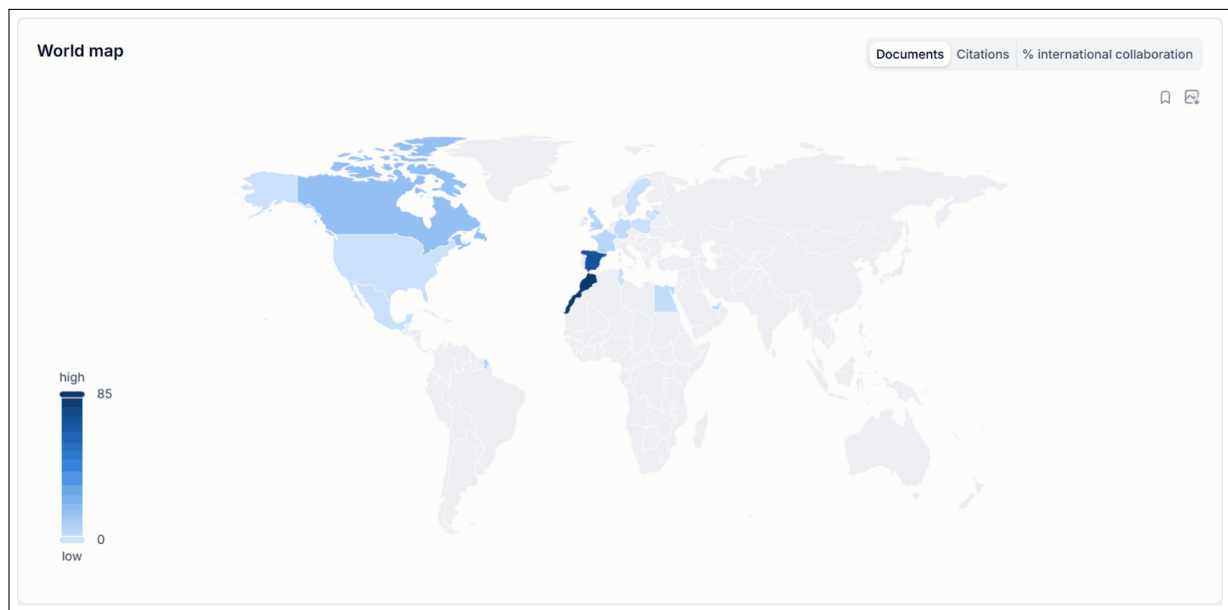
● Country → city → institutions  

The share tells whether a city carries most of its country's output or is only one component of it. It is computed over ALL the country's cities, before any truncation.

Country	City	Institutions	Documents	Share of country	Institutions (affiliations)
Canada	Montréal	1	13	76.5 %	University of Québec (13)
Canada	Moncton	1	3	17.6 %	Université de Moncton (3)
Canada	Calgary	1	1	5.9 %	University of Calgary (1)
Egypt	Cairo	1	3	100 %	African Disaster Mitigation Research Center (3)
France	Écully	1	4	80 %	Université Jean Monnet (4)
France	Bidart	1	1	20 %	Univ. Bordeaux (1)
Germany	Stuttgart	1	4	100 %	Vector Stiftung (4)
Latvia	Riga	1	2	100 %	Transport and Telecommunication Institute (2)
Lithuania	Vilnius	1	1	100 %	European Humanities University (1)
Mexico	Chilpancingo	1	1	100 %	Autonomous University of Guerrero (1)
Morocco	Rabat	5	66	77.6 %	Mohammed V University (47) National School of Mineral Industry (26) Abulcasla International University of Health Sciences (2) Cheikh Zaid International University Hospital (2) Hassan II Institute of Agronomy and Veterinary Medicine (2)
Morocco	Meknès	1	33	38.8 %	Moulay Ismail University (34)
Morocco	Ben Guerir	2	9	10.6 %	Mohammed VI Polytechnic University (7) Green Energy Park (2)
Morocco	Oujda	1	7	8.2 %	Mohammed 1st University (18)
Morocco	Tétouan	2	4	4.7 %	Abdelmalek Essaadi University (3) Abdelmalek Essaadi University Tetuan (1)
Morocco	Casablanca	2	3	3.5 %	Higher School of Textile and Clothing Industries. (2) HESTM Engineering (1)
Morocco	Fes	1	3	3.5 %	Sidi Mohamed Ben Abdellah University (4)
Morocco	Ifrane	1	3	3.5 %	Al Akhawayn University in Ifrane (3)
Morocco	Safi	1	2	2.4 %	Cadi Ayyad University (2)
Morocco	El Jadida	1	1	1.2 %	University Chouaib Doukkali (2)
Morocco	Larache	1	1	1.2 %	Abdelmalek Essaadi University (1)
Morocco	Martil	1	1	1.2 %	Abdelmalek Essaadi University (1)
Morocco	Sale	1	1	1.2 %	Mohammed V University (1)
Poland	Bydgoszcz	1	1	100 %	Kazimierz Wielki University (1)
Poland	Warsaw	1	1	100 %	War Studies University (2)
Spain	Murcia	7	69	100 %	Universidad de Murcia (78) FAMDIF (2) Virgen de la Arrixaca University Hospital (2) Comunidad Autónoma de la Región de Murcia (1) Quirónsalud Hospital (1) TICARUM SLU (1) Universidad Politécnica de Cartagena (1)
Spain	Ciudad Real	1	7	10.1 %	University of Castilla-La Mancha (7)
Spain	Cartagena	1	3	4.3 %	Technical University of Cartagena (5)
Spain	Almería	1	1	1.2 %	University of Castilla-La Mancha (1)

Country → city → institutions (excerpt).

11 Countries



World map: documents, citations or share of international collaboration per country. Countries too small for the base map (Singapore, Bahrain...) are placed as a dot at their capital.

How it is computed

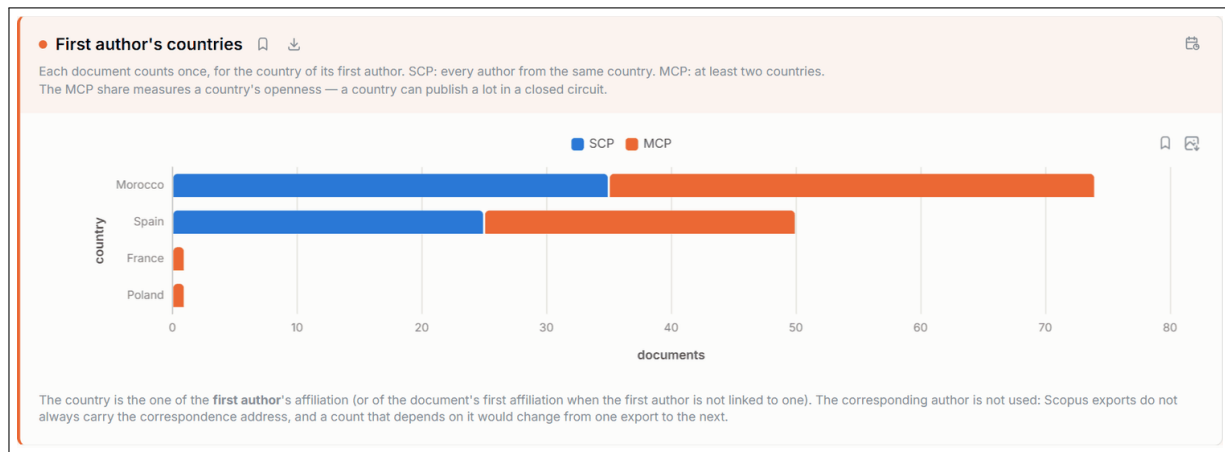
For each country: documents (full counting), citations, and

$$\text{international share} = \frac{MCA}{\text{documents}} \times 100$$

where *MCA* (*multiple country articles*) counts its documents co-signed with at least one other country, and *SCA* = documents – *MCA*.



Collaboration map: one arc per pair of countries, thicker as they co-sign more documents.



First author's countries: SCP / MCP per country of the first author.

How it is computed

Each document is assigned to the country of the affiliation of its **first author** (failing that, of its first affiliation). SCP (*single country publication*): all signatories from the same country; MCP: at least two countries. $MCP \text{ rate} = MCP / (SCP + MCP) \times 100$: the international openness of the country.

Caution

This is not the country of the *corresponding author*. Scopus exports do not always carry the correspondence address: a count that depended on it would change from one export to another. The first author, on the other hand, is always known. In an article, write “country of the first author”.

• **Country impact**  

h, g and m per country. The m-index divides h by the number of active years — without it a country present since 2016 always outranks one that started in 2022, for reasons that have nothing to do with quality.

Country	Documents	Citations	h	g	m	Since
Morocco	85	1,273	19	34	1.36	2,013
Spain	69	1,149	19	32	1.12	2,010
Canada	17	546	10	17	0.91	2,016
United Arab Emirates	5	119	4	5	0.4	2,017
Germany	4	124	3	4	0.2	2,012
Egypt	3	66	3	3	1	2,024
United Kingdom	6	31	3	5	0.2	2,012
Tunisia	1	14	1	1	0.2	2,022
Latvia	2	10	1	2	0.12	2,019
Lithuania	1	10	1	1	0.12	2,019
Mexico	1	10	1	1	0.09	2,016
France	6	7	1	2	0.08	2,014
United States	1	4	1	1	0.12	2,019
Sweden	1	3	1	1	0.33	2,024
Poland	1	0	0	0	0	2,026

Country impact: documents, citations, h, g and m-index.

Breakdown by country					
Country	Documents	Citations	Domestic	International	% intl.
Morocco	85	1,273	35	50	58.8 %
Spain	69	1,149	25	44	63.8 %
Canada	17	546	0	17	100 %
France	6	7	0	6	100 %
United Kingdom	6	31	0	6	100 %
United Arab Emirates	5	119	0	5	100 %
Germany	4	124	0	4	100 %
Egypt	3	66	0	3	100 %
Latvia	2	10	0	2	100 %
Lithuania	1	10	0	1	100 %
Mexico	1	10	0	1	100 %
Poland	1	0	0	1	100 %
Sweden	1	3	0	1	100 %
Tunisia	1	14	0	1	100 %
United States	1	4	0	1	100 %

Domestic: all authors from the same country. International: at least two signatory countries.

Breakdown by country: national and international documents.

12 Local impact

How it is computed

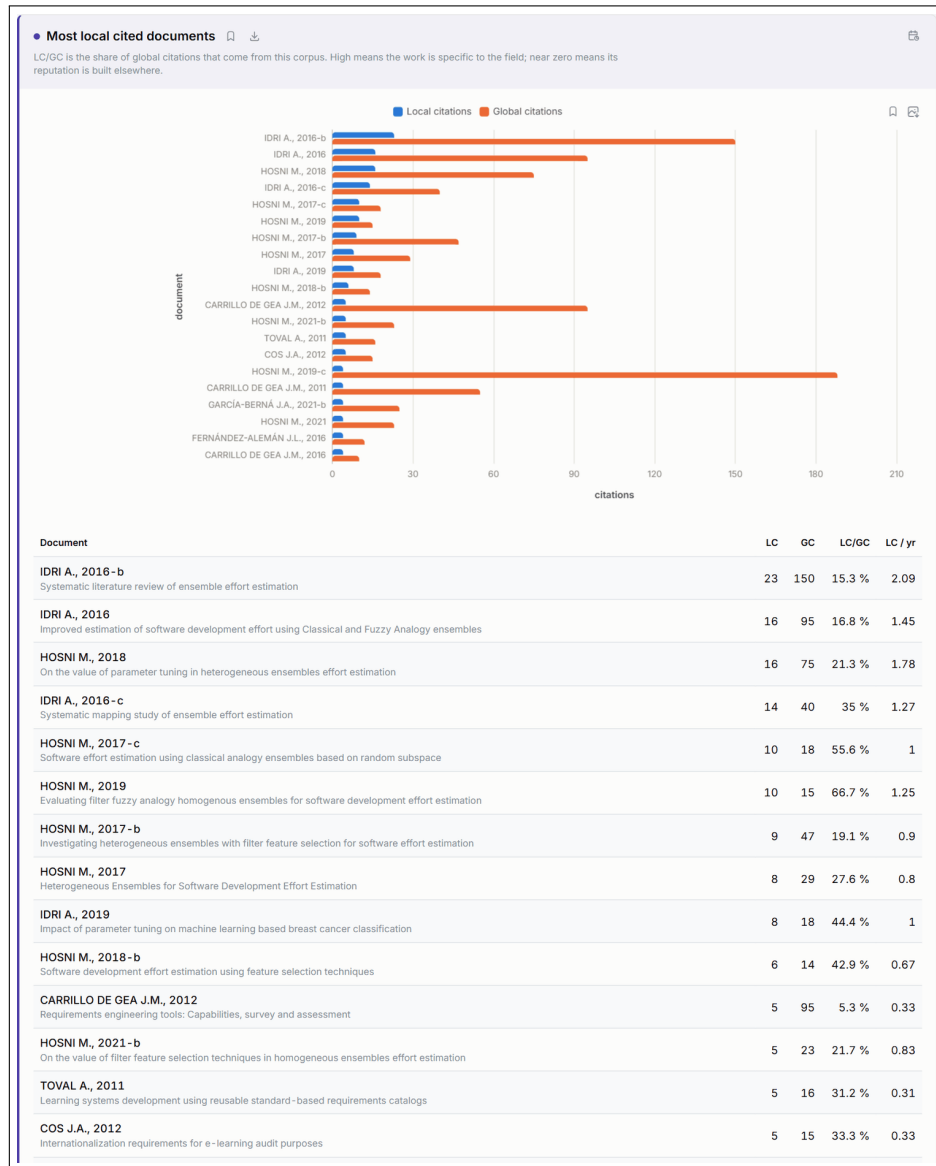
A **local citation** is a citing $\rightarrow$ cited link *between two documents of the corpus*. A reference of the citing document is matched with a document:

1. by **DOI** when both have one;
2. otherwise by **normalised title** (lowercase, without accents or punctuation), provided it has at least 25 characters: a short title (“Introduction”) would create false links.

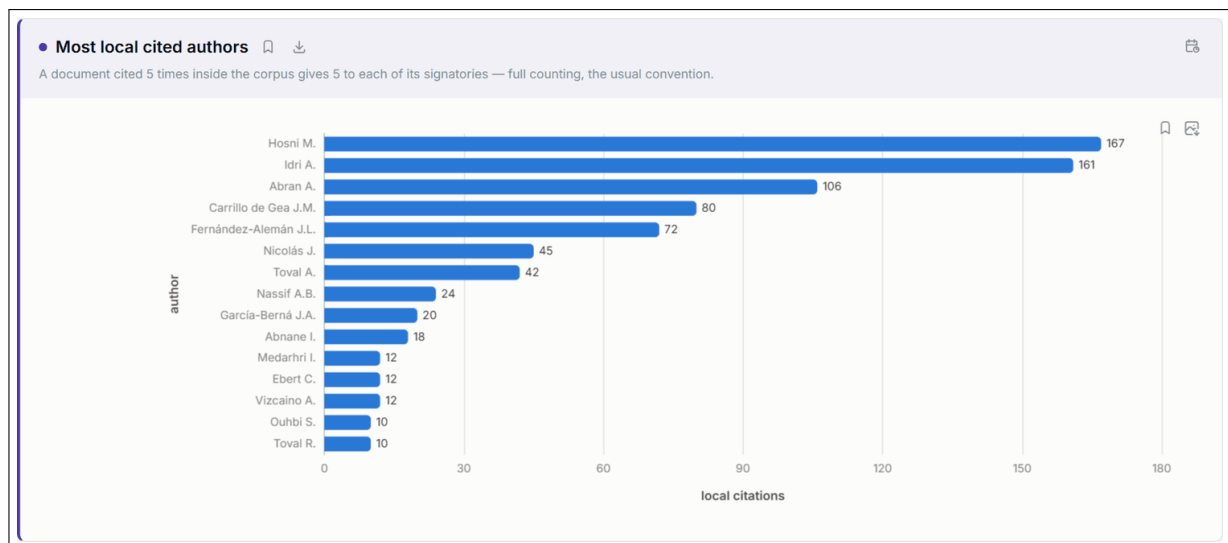
A pair counts once; a document does not cite itself. $LC_d = \#\{\text{documents of the corpus that cite } d\}$.

How to read it

C_d measures influence in the world; LC_d , influence *in the field studied*. An article with 800 global citations and 0 local ones matters elsewhere, not here.



Most locally cited documents (excerpt), with their global citations.

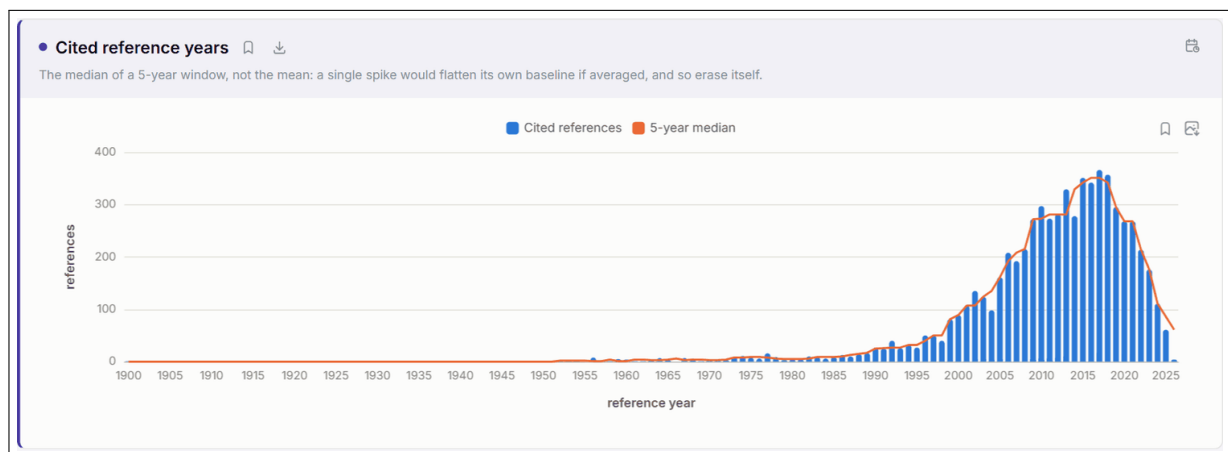


Most locally cited authors: $\sum LC_d$ over the documents of the author.

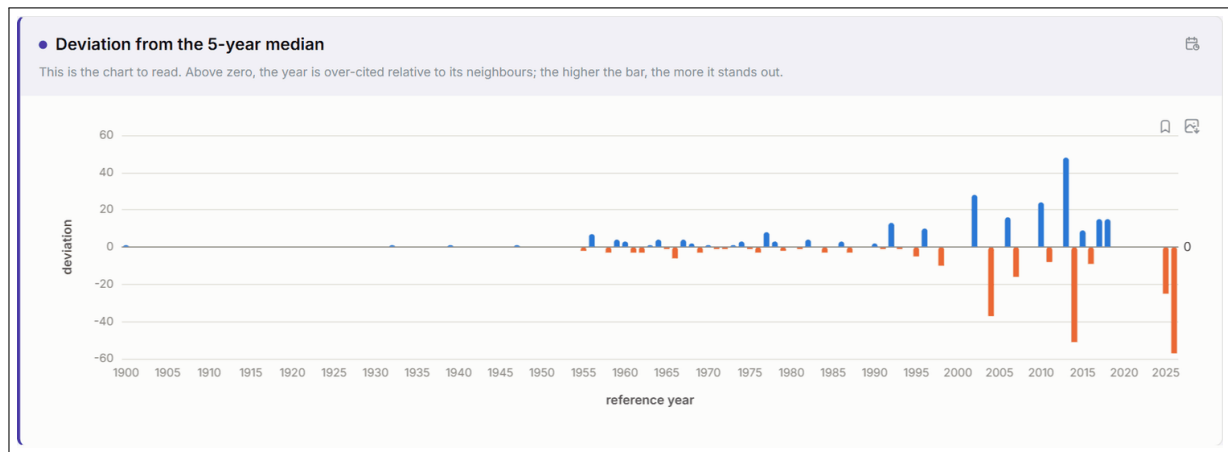


The historiograph: the most locally cited documents, placed by year, linked by their direct citations.

13 Reference spectroscopy



Publication years of the cited references, and their 5-year rolling median.



Deviation from the median: the peaks signal years of foundational works.

How it is computed

With $R(t)$ the number of cited references published in year t (Marx et al., 2014):

$$\text{med}_5(t) = \text{median}(R(t-2), \dots, R(t+2)), \quad D(t) = R(t) - \text{med}_5(t)$$

A year is a **peak** if $D(t) > 0$ and $D(t) \geq \text{median}\{D(s) : D(s) > 0\}$. The median, not the mean: a single peak would inflate its own reference.

• **Peak years**

For each peak, the reference most frequently cited that year — the work to name when commenting the chart.

Year	References	Median	Deviation	Most cited reference of that year
2013	329	281	+48	Mahmoud O. Elish, Tarek Helmy, Muhammad Imtiaz Hussain — Empirical Study of Homogeneous and Heterogeneous Ense (9×)
2002	135	107	+28	Ambrosio Toval, Joaquín Nicolás, Begoña Morós, Fernando de Cuadra — Requirements Reuse for Improving Informati (8×)
2010	297	273	+24	José Luis Fernández-Alemán, Dominic Palmer-Brown, Chrisina Jayne — Effects of Response-Driven Feedback in Comp (6×)
2006	208	192	+16	Arjun Chandra, Xin Yao — Ensemble Learning Using Multi-Objective Evolutionary Algorithms (12×)
2017	366	351	+15	Mohamed Hosni, Ali Idri, Alain Abran, Ali Bou Nassif — On the value of parameter tuning in heterogeneous ensem (11×)
2018	357	342	+15	Mohamed Hosni, Ali Idri, Alain Abran — Evaluating filter fuzzy analogy homogenous ensembles for software devel (8×)
1992	40	27	+13	Jacob Cohen — A power primer (9×)
1996	50	40	+10	Leo Breiman — Bagging Predictors (12×)
2015	351	342	+9	Mohammad Azzeh, Ali Bou Nassif, Leandro L. Minku — An empirical evaluation of ensemble adjustment methods for (10×)
1977	16	8	+8	John Richard Landis, Gary G. Koch — The Measurement of Observer Agreement for Categorical Data (6×)
1956	8	1	+7	M. H. Quenouille — ESTIMATION (5×)
1959	5	1	+4	E. T. Hall — The Silent Language (2×)

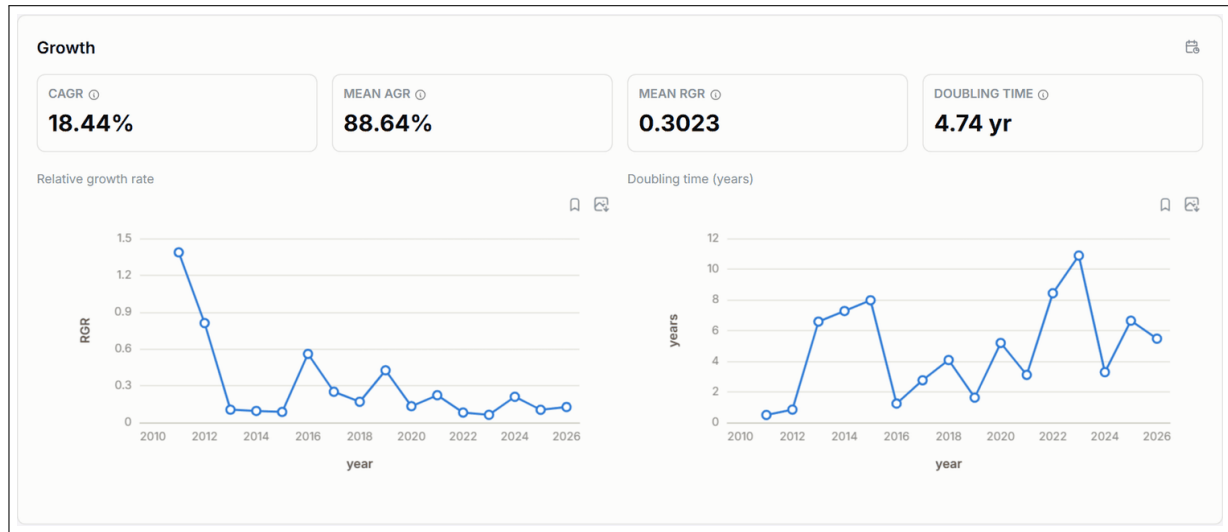
The peak years, with the most cited reference of each.

How to read it

First peak: 2013 ($R = 329$, median 281, deviation +48), then 2002 (+28) and 2010 (+24).

14 Indicators

14.1 Growth



CAGR, mean AGR, mean RGR and mean doubling time, with the annual evolution.

How it is computed

With $W_t = \sum_{s \leq t} n_s$ the *cumulative* production:

$$CAGR = \left[\left(\frac{n_{\text{end}}}{n_{\text{start}}} \right)^{1/k} - 1 \right] \times 100$$

k = number of annual intervals (18.44%)

$$AGR_t = \frac{n_t - n_{t-1}}{n_{t-1}} \times 100$$

mean = 88.64%

$$RGR_t = \ln W_t - \ln W_{t-1}$$

(Mahapatra, 1985), mean = 0.302

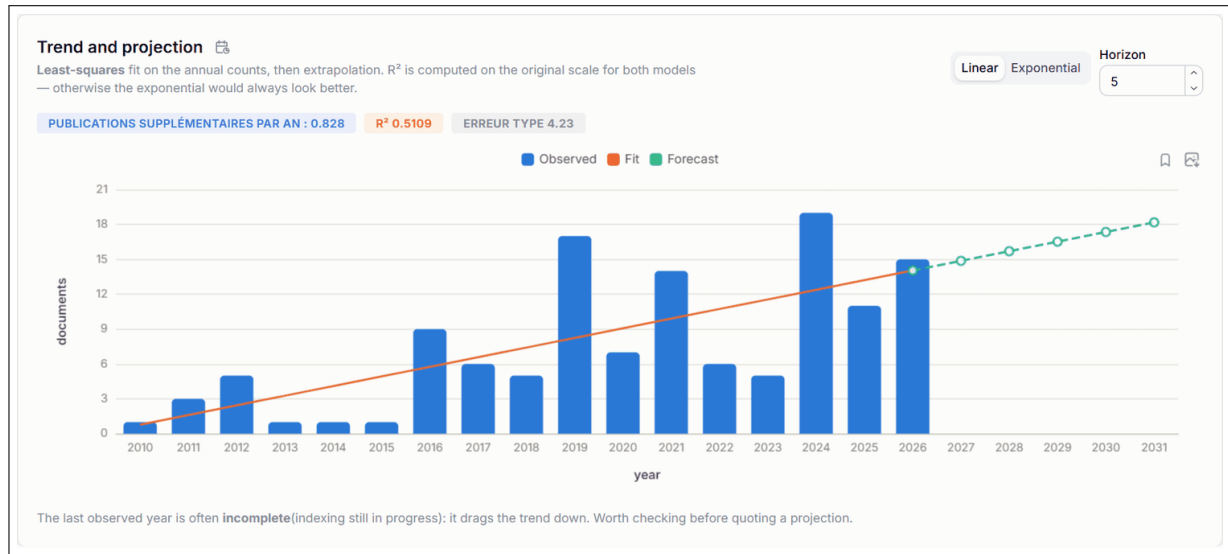
$$DT_t = \frac{\ln 2}{RGR_t}$$

mean = 4.74 years

Caution

The mean AGR is very sensitive to small counts: a year going from 1 to 5 documents counts as +400%. On a small corpus, quote the CAGR.

14.2 Trend and projection



Fitted trend and 5-year projection.

How it is computed

Linear model $\hat{n}_t = a + bt$, or exponential $\ln \hat{n}_t = a + bt$ (i.e. $\hat{n}_t = e^a e^{bt}$), fitted by least squares, t counted from the first year. Quality:

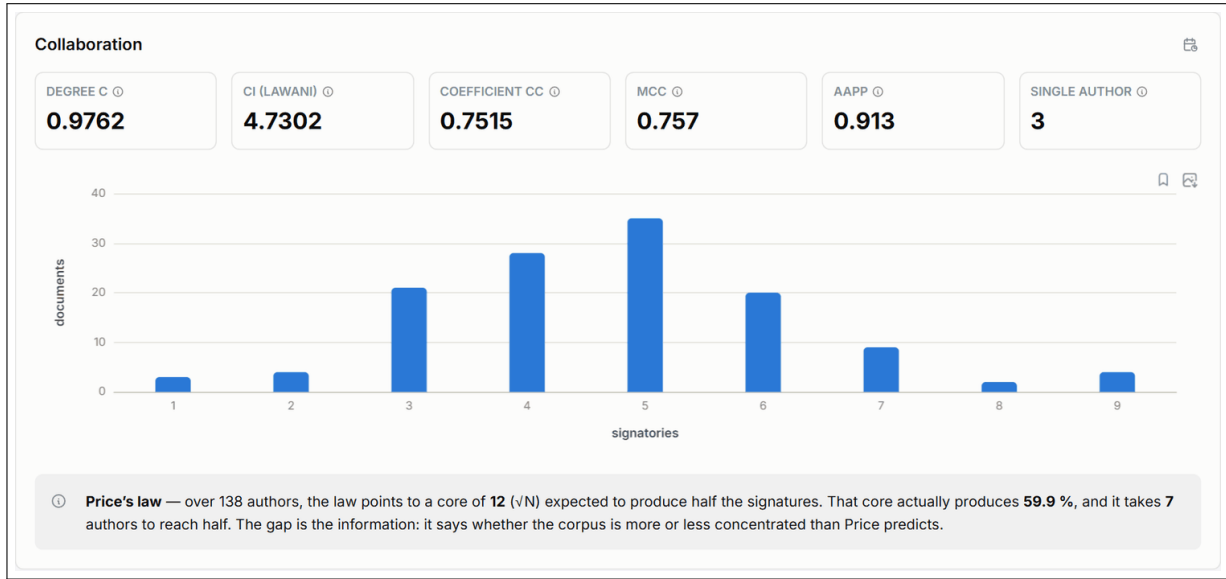
$$R^2 = 1 - \frac{\sum_t (n_t - \hat{n}_t)^2}{\sum_t (n_t - \bar{n})^2}, \quad RMSE = \sqrt{\frac{\sum_t (n_t - \hat{n}_t)^2}{T - 2}}$$

R^2 is computed on the count scale for both models, otherwise their R^2 would not be comparable. Here: $b = 0.83$ more document per year, $R^2 = 0.51$.

Caution

The last year (2026) is incomplete: indexing is in progress. It pulls the trend down.

14.3 Collaboration



Collaboration indicators, documents by number of authors, and Price's law.

How it is computed

With f_j the number of documents with j authors, N_m the co-signed documents and A the distinct authors:

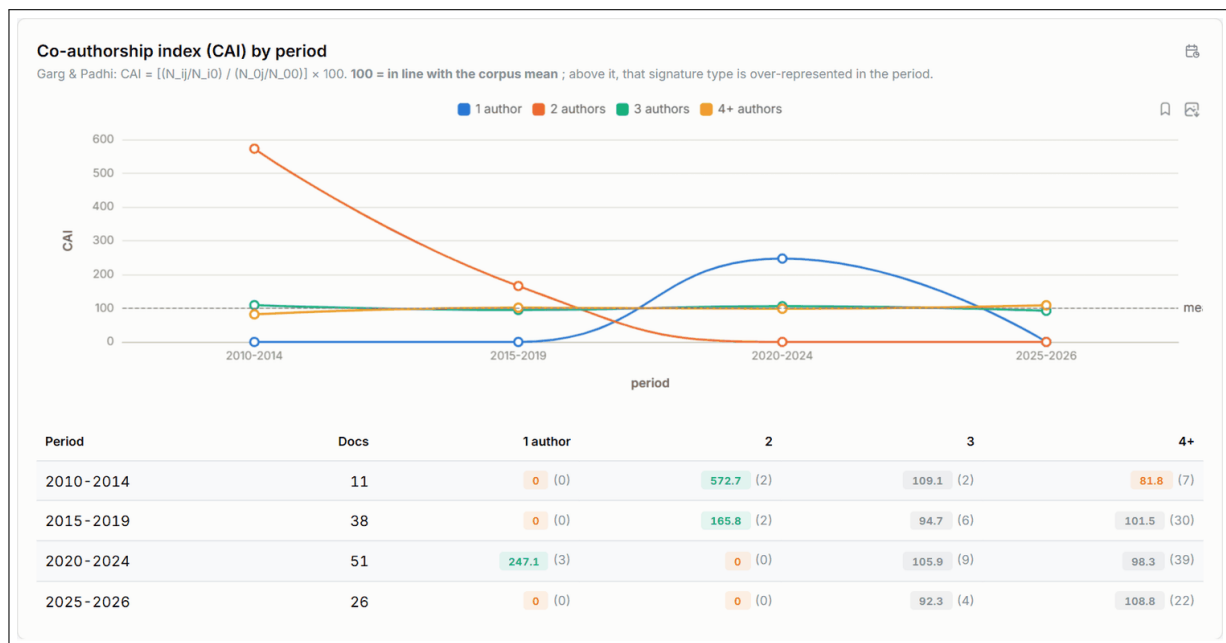
$DC = \frac{N_m}{N}$	degree of collaboration (Subramanyam, 1983) = 0.976
$CI = \frac{\sum_j j f_j}{N}$	collaboration index (Lawani, 1980) = 4.730
$CC = 1 - \frac{\sum_j f_j/j}{N}$	coefficient (Ajiferuke et al., 1988) = 0.752
$MCC = \frac{A}{A-1} CC$	modified coefficient (Savanur & Srikanth, 2010) = 0.757
$AAPP = \frac{N}{A}$	average productivity per author = 0.913

Price's law: $\sqrt{A}$ authors should produce half of the authorships. Expected core = $\text{round}(\sqrt{138}) = 12$ authors; they actually produce 59.9% of the 596 authorships, and 7 authors are enough to reach half: the corpus is more concentrated than the law.

Caution

Only the **CI (Lawani)** is the collaboration index: the mean number of signatories over *all* documents (4.73), equal to “Authors / document” in Overview. “Authors / co-authored document” (4.82) leaves out the single-author documents: it is not the same indicator, and the interface no longer gives it that name. In an article, quote Lawani's CI.

14.4 Co-authorship index (CAI)



CAI per 5-year period and per authorship type (1, 2, 3, 4 or more authors).

How it is computed

For period i and authorship type j (Garg & Padhi, 2001):

$$CAI_{ij} = \frac{N_{ij}/N_{i0}}{N_{0j}/N_{00}} \times 100$$

N_{ij} : documents of type j in period i ; N_{i0} : documents of the period; N_{0j} : documents of type j over all periods; N_{00} : total. $CAI = 100$: in line with the mean; above, the type is over-represented. Here, two-author documents are strongly over-represented in 2010–2014 ($CAI = 572.7$) and then disappear.

14.5 Sample size (Cochran)

Sample size (Cochran)

How many documents must be read for the conclusions to hold for the whole corpus? $n_0 = Z^2 \cdot p \cdot q / e^2$, then the finite-population correction $n = n_0 / (1 + (n_0 - 1)/N)$.

Confidence %: 95 Margin %: 5

POPULATION: **126** Z: **1.96** REQUIRED SAMPLE: **96** SAMPLING RATE: **76.2 %**

How many documents to check by hand to validate the whole corpus.

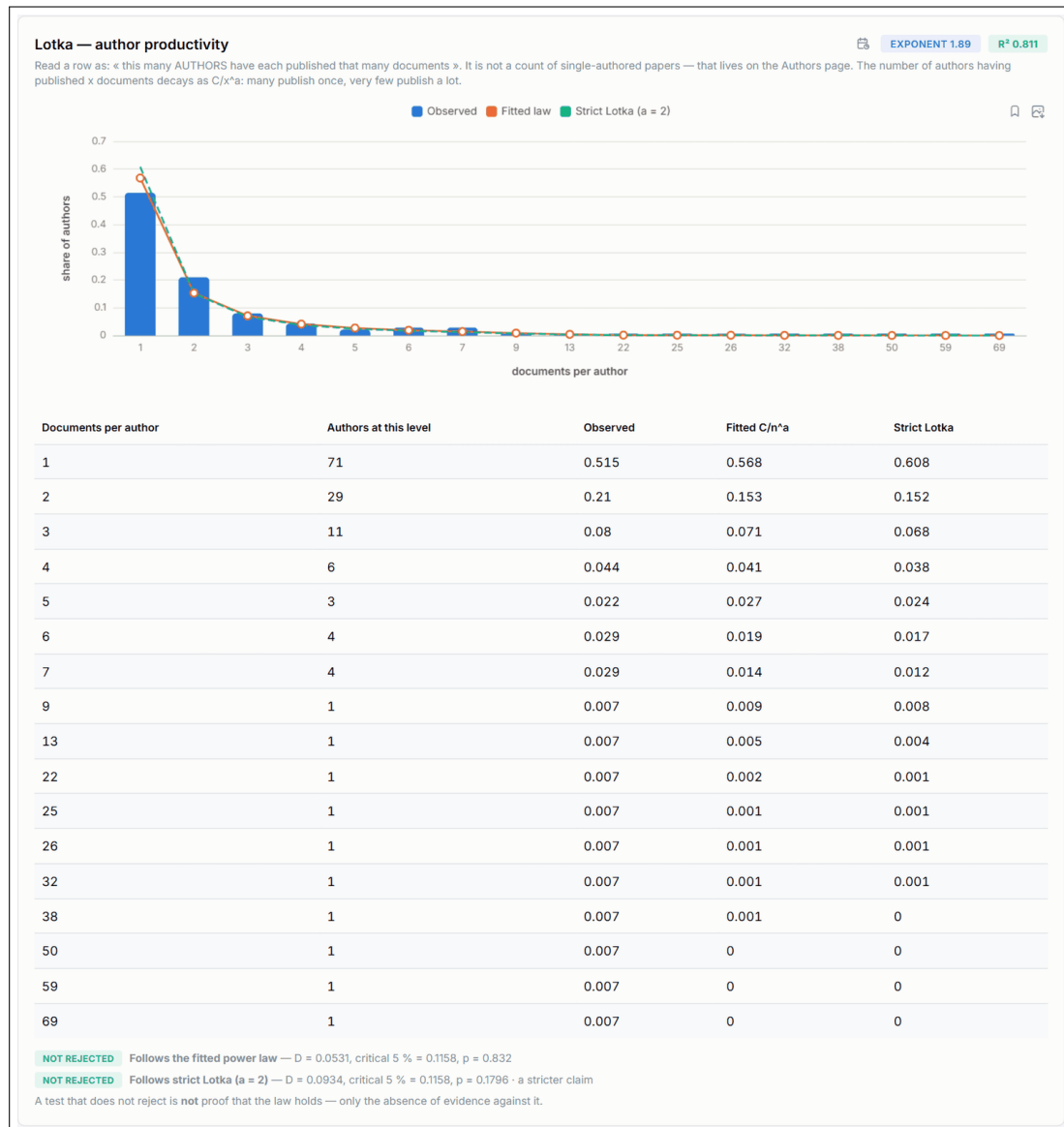
How it is computed

$$n_0 = \frac{z^2 p(1-p)}{e^2}, \quad n = \frac{n_0}{1 + (n_0 - 1)/N}$$

With a 95 % confidence ($z = 1.96$), a margin $e = 5\%$ and $p = 0.5$ (the most cautious case): $n_0 = 385$, and for $N = 126$, $n = 96$ documents, i.e. 76.2 % of the corpus.

15 Laws: Lotka, Bradford, Zipf

15.1 Lotka: author productivity



Share of authors who wrote x documents: observed, fitted law, strict Lotka law; table and Kolmogorov-Smirnov test.

How it is computed

The share of authors who wrote x documents follows a power law $f(x) = C x^{-a}$.

- **Exponent** by maximum likelihood (discrete zeta law): a is the solution of

$$\frac{\sum_{x \geq 1} \ln x x^{-a}}{\zeta(a)} = \frac{1}{A} \sum_{i=1}^A \ln x_i, \quad \zeta(a) = \sum_{x \geq 1} x^{-a}$$

solved by bisection; then $C = 1/\zeta(a)$. Here $a = 1.89$.

- **Strict Lotka**: $a = 2$, $C = 1/\zeta(2) = 6/\pi^2 \approx 0.608$.
- **Kolmogorov-Smirnov test** on the cumulative distribution functions:

$$D = \max_x |F_{\text{obs}}(x) - F_{\text{th}}(x)|, \quad D_{\text{crit}} = \frac{1.36}{\sqrt{A}} \text{ (5 \%)}$$

with A the number of *authors*, not of productivity levels. $p = 2 \sum_{k \geq 1} (-1)^{k-1} e^{-2k^2 AD^2}$.

Here: $D = 0.053 < D_{\text{crit}} = 0.116$, $p = 0.83$: the corpus follows the fitted law; and strict Lotka too ($D = 0.093$, $p = 0.18$).

Caution

The least-squares fit on log-log, very widespread, gives an exponent of 0.96 here: the tail (a few very productive authors, one per level) dominates the regression. Maximum likelihood weighs each *author*, not each level (Clauset, Shalizi & Newman, 2009).

15.2 Bradford: the dispersion of sources



The sources ranked by productivity, the Bradford zones, and the core (zone 1).

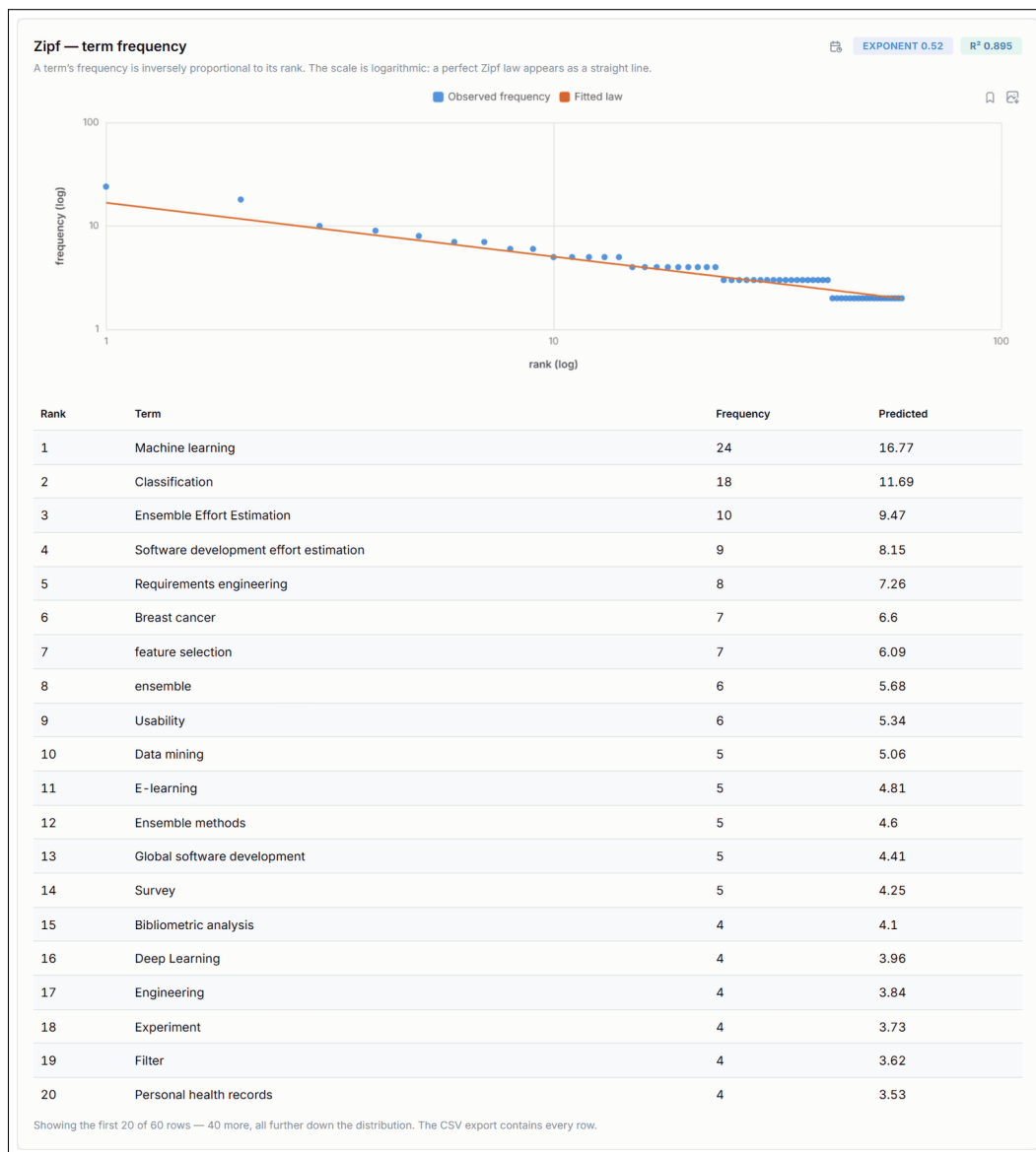
How it is computed

The sources are sorted by decreasing number of documents, then split into 3 zones each holding about $N/3$ documents (a zone changes as soon as the cumulative count reaches $z \times N/3$). With s_z the number of sources of zone z :

$$k = \frac{1}{2} \left(\frac{s_2}{s_1} + \frac{s_3}{s_2} \right)$$

Here: 9 sources for 43 documents, then 27 for 84, then 42 for 126; $k = 2.28$. A roughly constant multiplier from one zone to the next signals that the law holds. Zone 1 is the core of the field.

15.3 Zipf: keyword frequency



Keyword frequency as a function of rank, observed and fitted (log-log scale).

How it is computed

Frequency of a keyword = number of documents that use it (case-insensitive). Fit $f(r) = C r^{-s}$ by least squares on $\log_{10} f$ against $\log_{10} r$:

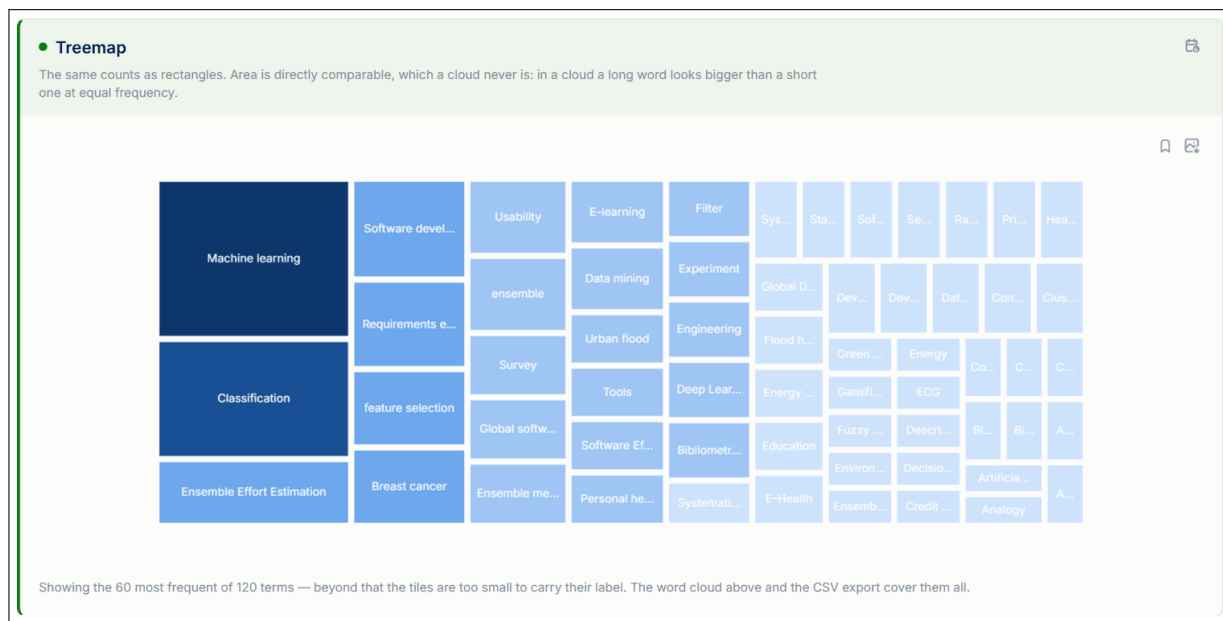
$$\log_{10} f = \log_{10} C - s \log_{10} r, \quad R^2 = 1 - \frac{\sum (\log f - \widehat{\log f})^2}{\sum (\log f - \overline{\log f})^2}$$

Here $s = 0.52$, $C = 16.8$, $R^2 = 0.895$.

Word cloud

Size follows the number of documents the term appears in. Colour follows the same value — it carries no separate meaning, and a random palette would suggest one that is not there.

Word cloud: the size follows the number of documents that use the keyword.



Treemap: the area of each rectangle is proportional to the number of documents.



Keyword dynamics: cumulative occurrences year after year.

How it is computed

$\text{occ}(t, w) = \#\{d : y_d = t, w \in \text{keywords}(d)\}$; the cumulative curve is $\sum_{s \leq t} \text{occ}(s, w)$. Every year of the interval is present, at zero if needed, otherwise the cumulative count would be wrong.

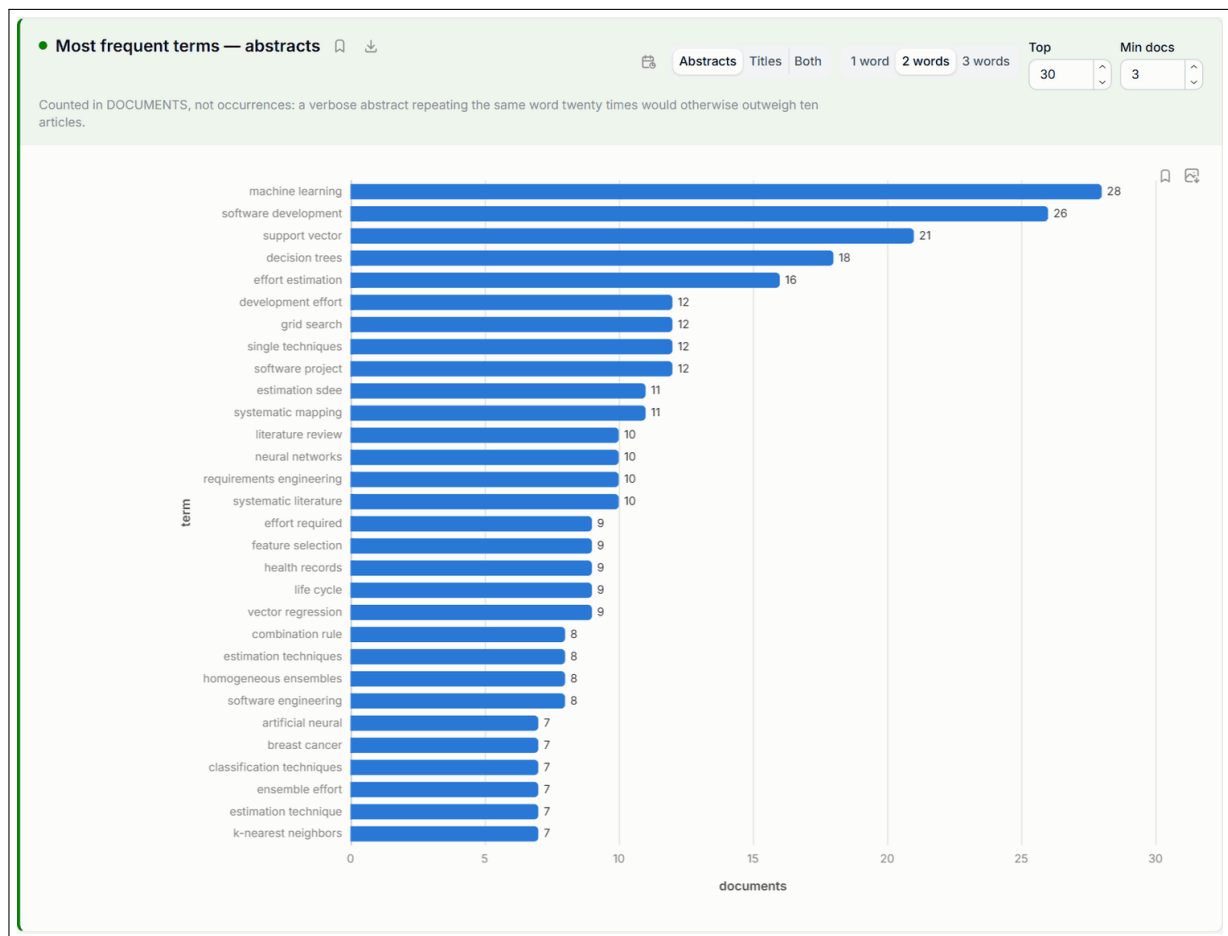


Co-word network: two keywords are linked when they appear in the same documents.

How it is computed

Link weight $c_{ij} = \#\{d : i \in d \text{ and } j \in d\}$, on the $N_{\max}$ most frequent keywords and above a minimum weight. The normalisations and the communities are described in section 22.

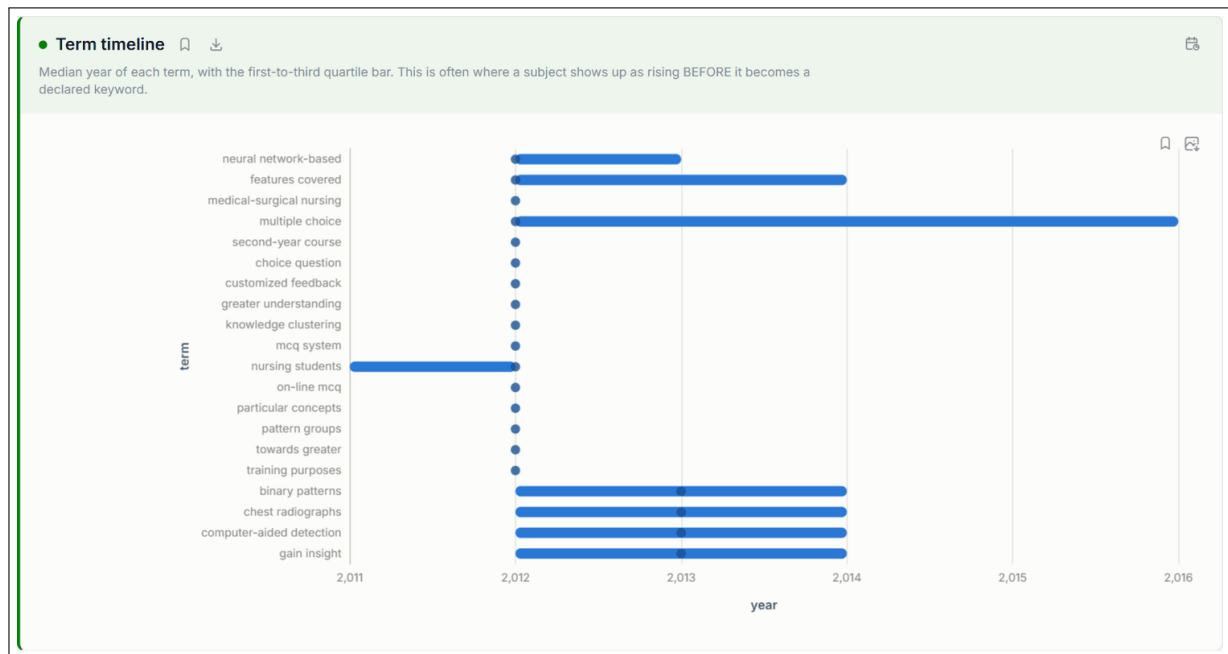
17 Text mining: what the abstracts say



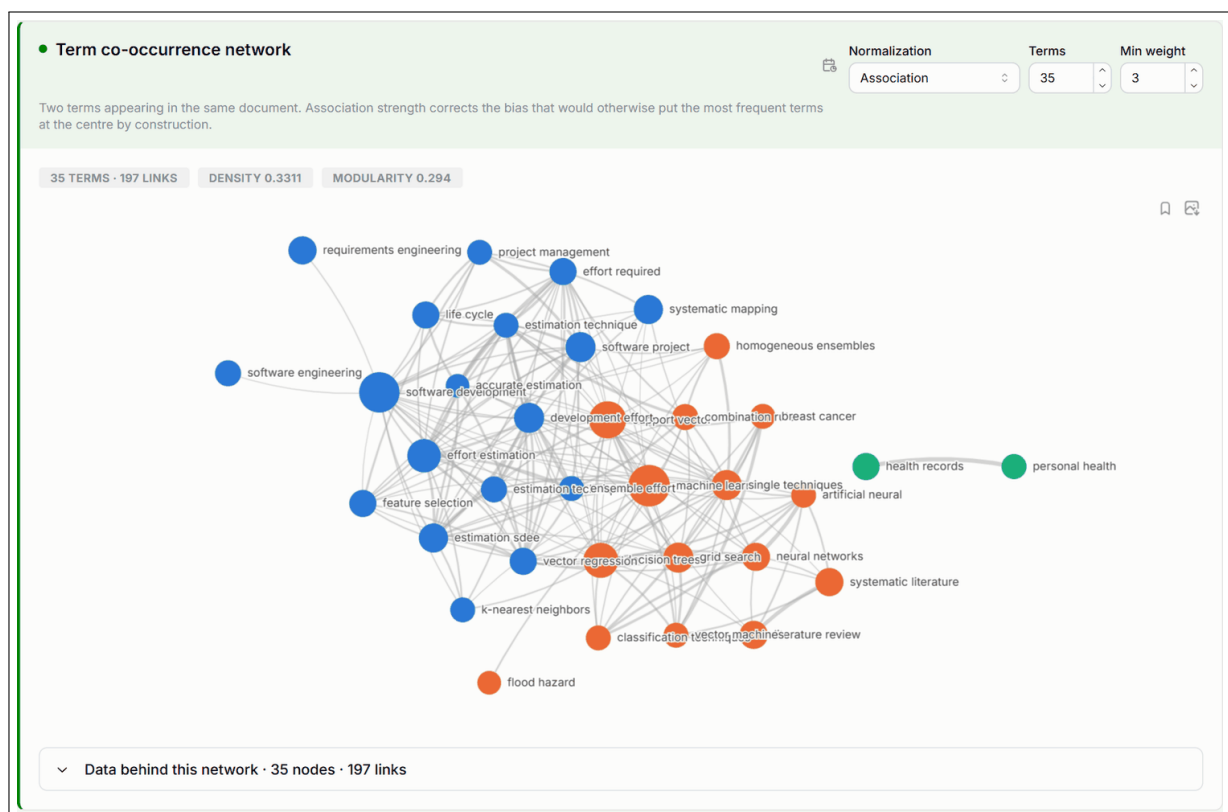
Most frequent terms of the abstracts (or titles), in 1, 2 or 3 words.

How it is computed

Titles or abstracts are split into n -grams ($n = 1, 2, 3$). Stopwords are removed, including the vocabulary of academic writing (“paper”, “results”, “proposed”); an n -gram is only kept if it neither starts nor ends with a stopword. Frequency of a term = number of documents that contain it (a document counts once).



Term timeline: for each term, the first quartile, the median and the third quartile of the years in which it appears.



Co-occurrence network of the terms in the same abstracts.

18 Thematic map: Callon's strategic map



The themes (groups of keywords) placed by their centrality and density, in four quadrants.

How it is computed

On the co-word network, each link is weighted by the **equivalence index** $e_{ij} = c_{ij}^2 / (c_i c_j)$, between 0 and 1 (c_i : occurrences of i). The groups are formed by greedy modularity. For a group of w terms (Callon et al., 1991):

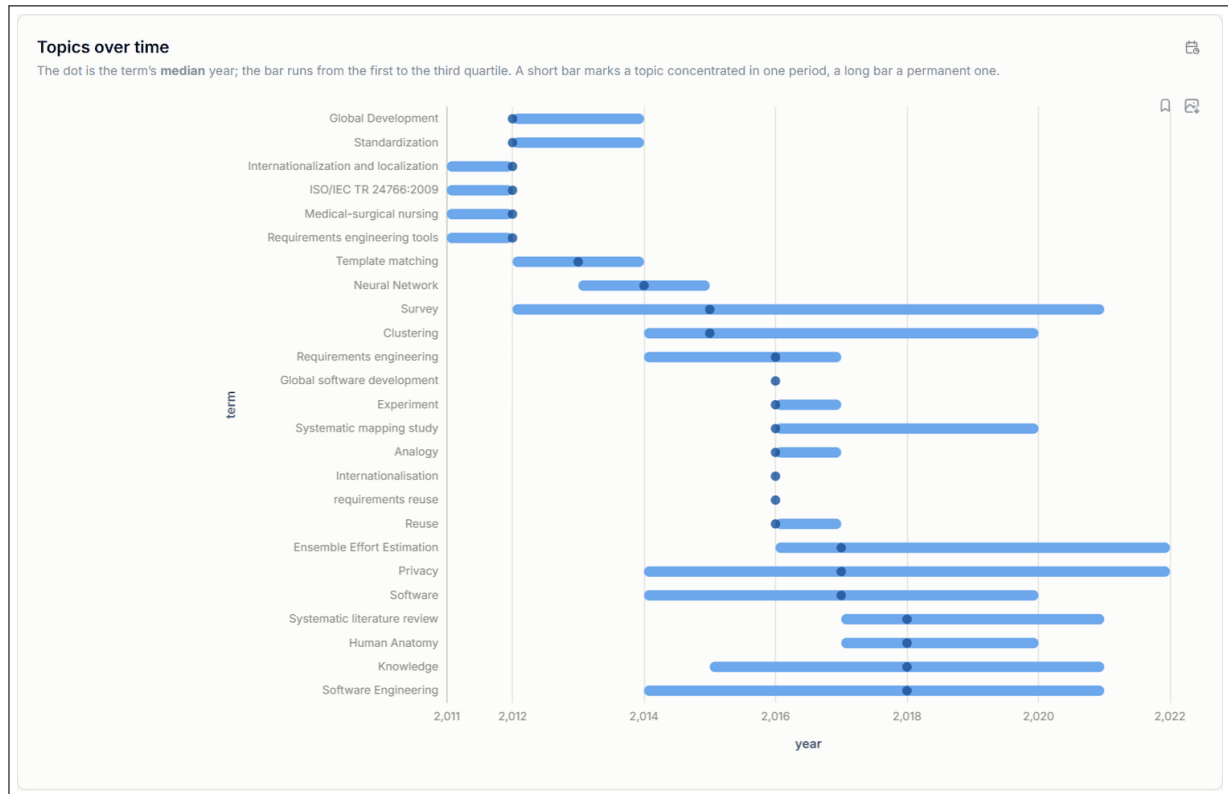
$$\text{centrality} = 10 \times \sum_{i \in G, j \notin G} e_{ij}, \quad \text{density} = 100 \times \frac{\sum_{i, j \in G} e_{ij}}{w}$$

Both are measured on *all* co-occurrences between the retained terms. The quadrants are cut at the medians (here centrality 1.53, density 50):

	Low centrality	High centrality
High density	niches (developed, isolated)	motor themes
Low density	emerging or declining	basic themes

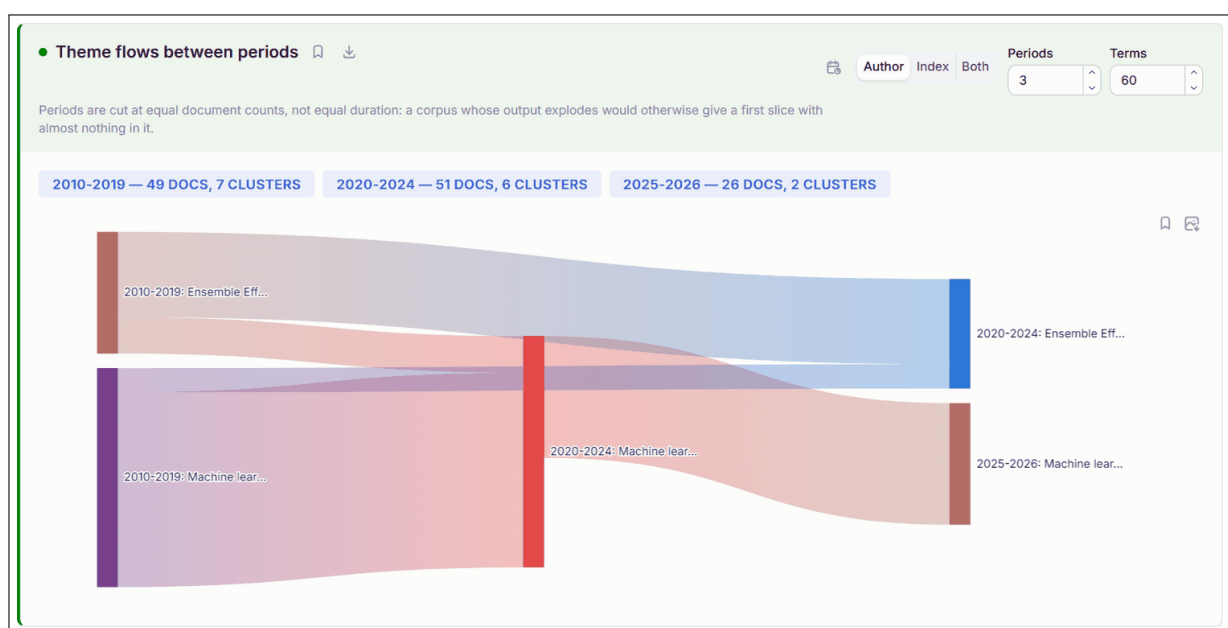
How to read it

“Machine learning”: centrality 15.3, density 32.0: a basic theme, linked to everything but little structured in itself. “Requirements engineering”: 14.6 and 66.2: a motor theme. The bottom-left quadrant does not say whether a theme is being born or dying: thematic evolution decides.



Topics over time: first quartile, median and third quartile of the years of each keyword.

19 Thematic evolution



The themes of each period, and their flows from one period to the next.

How it is computed

The corpus is cut into periods; a thematic map is computed for each. The link between a theme A of one period and a theme B of the next is the **weighted inclusion index**:

$$I(A, B) = \frac{\sum_{t \in A \cap B} \min(\text{occ}_A(t), \text{occ}_B(t))}{\min(\sum_{t \in A} \text{occ}_A(t), \sum_{t \in B} \text{occ}_B(t))}$$

It equals 1 when a small theme is entirely absorbed by a large one.

20 Conceptual structure



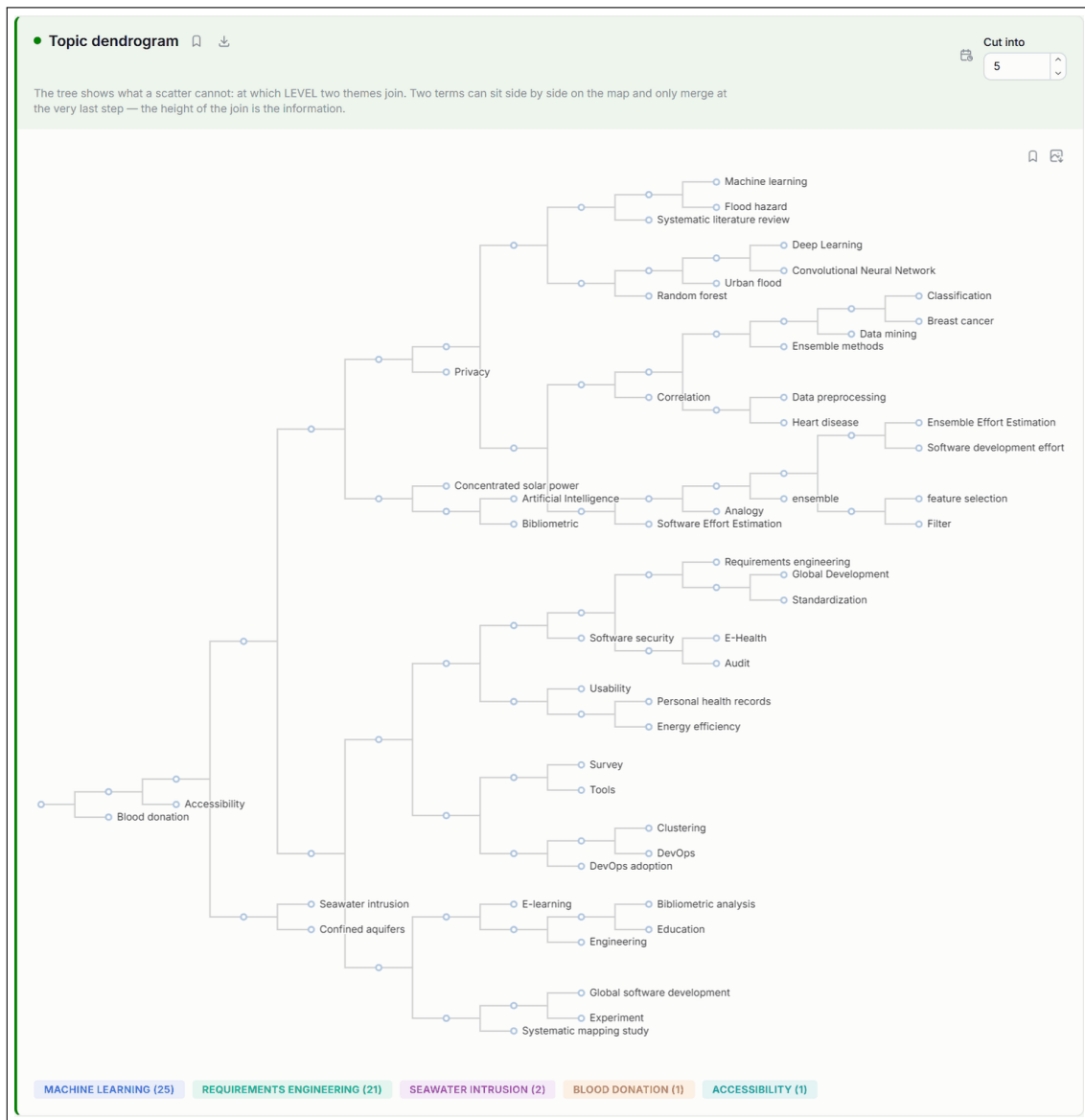
Factorial map of the keywords: two close terms appear in the same documents.

How it is computed

Correspondence analysis (CA). On the binary documents $\times$ terms table X : $P = X / \sum X$, row masses $r = P\mathbf{1}$ and column masses $c = P^\top \mathbf{1}$. Matrix of standardised residuals, then singular value decomposition:

$$S = D_r^{-1/2} (P - r c^\top) D_c^{-1/2} = U \Sigma V^\top, \quad F = D_c^{-1/2} V \Sigma$$

F gives the coordinates of the terms; axis k carries the share of inertia $\sigma_k^2 / \sum_l \sigma_l^2$ shown on the map. **MDS** (optional): classical Torgerson scaling on the dissimilarity $1 - \text{Salton}$ (see below). The groups are formed by k -means, k being chosen by the silhouette coefficient.



Dendrogram: at which level two themes join.

How it is computed

Agglomerative hierarchical clustering, average linkage (UPGMA), on the dissimilarity

$$\delta_{ij} = 1 - \frac{c_{ij}}{\sqrt{c_i c_j}}$$

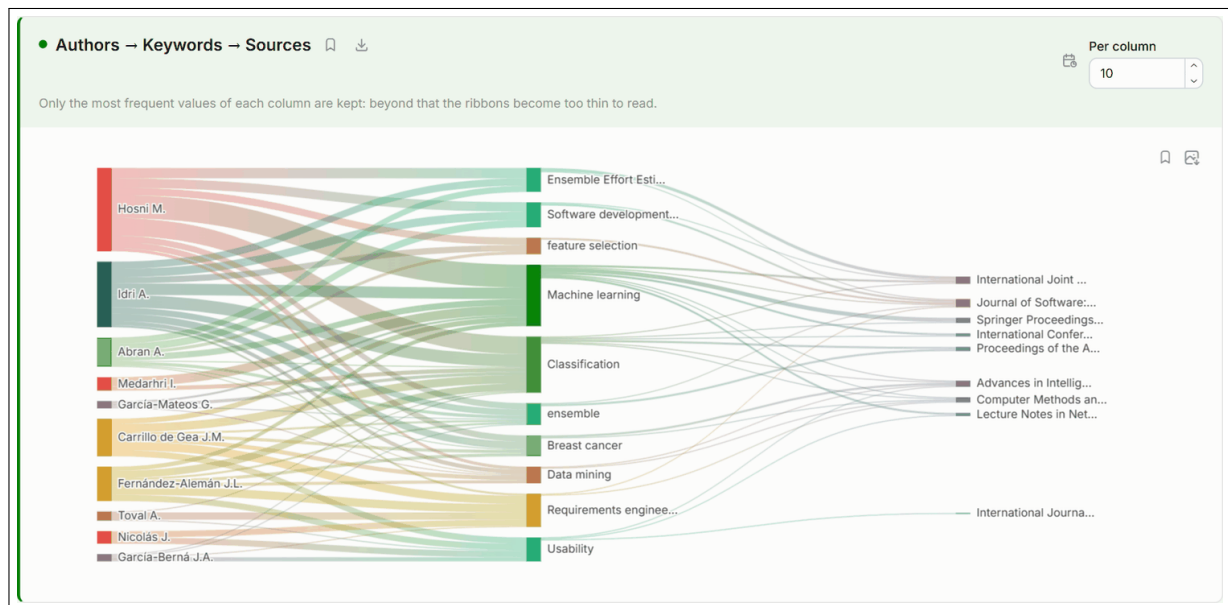
(1– Salton index). The tree is cut into at most 6 groups.

● **Clusters**

Cluster	Terms	Size	Occurrences
MACHINE LEARNI...	Machine learning, Classification, Ensemble Effort Estimation, Software development effort estimation, Requirements engineering, Breast cancer, feature selection, ensemble	46	219
SEAWATER INTR...	Seawater intrusion, Confined aquifers	2	5
ACCESSIBILITY	Accessibility	1	2
BLOOD DONATION	Blood donation	1	2

The groups, with their terms and their occurrences.

21 Three fields

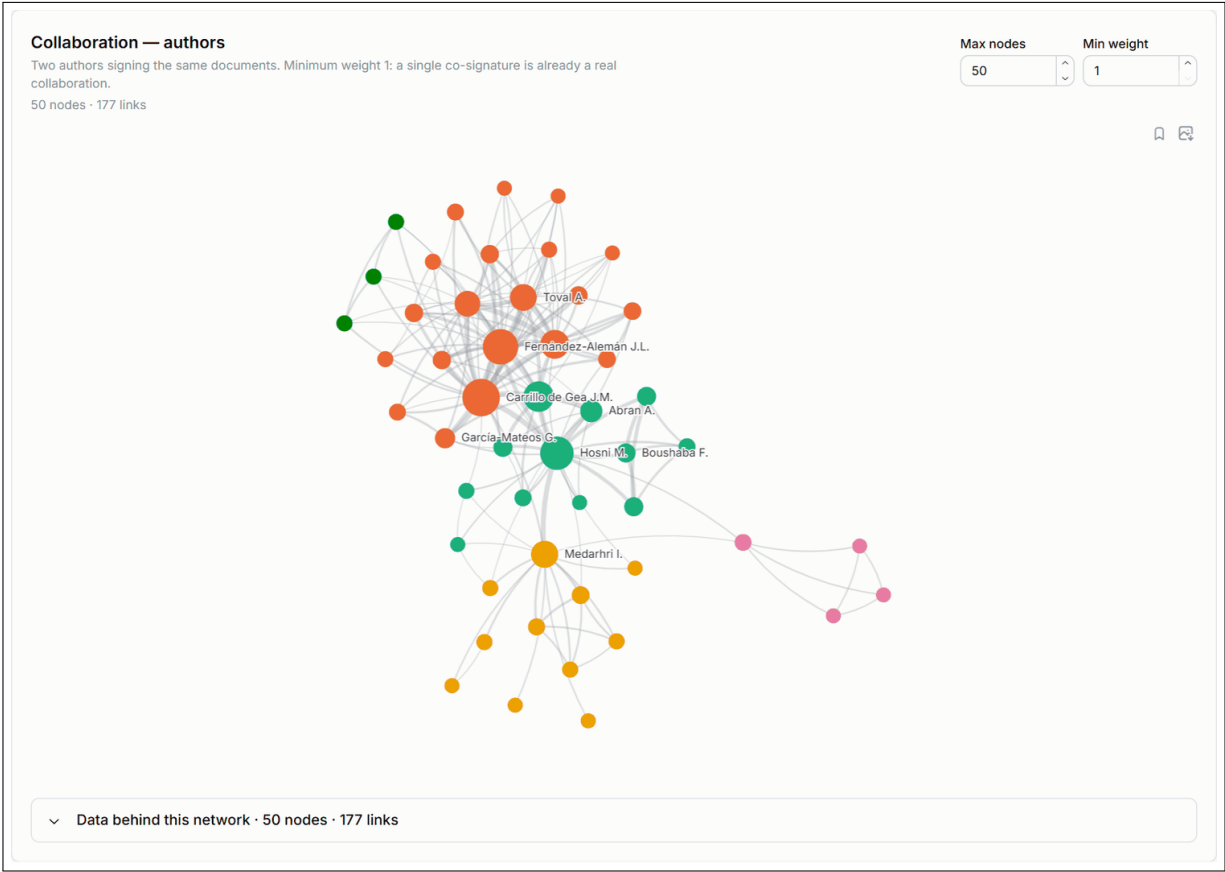


Authors → keywords → sources: who works on what, and publishes where.

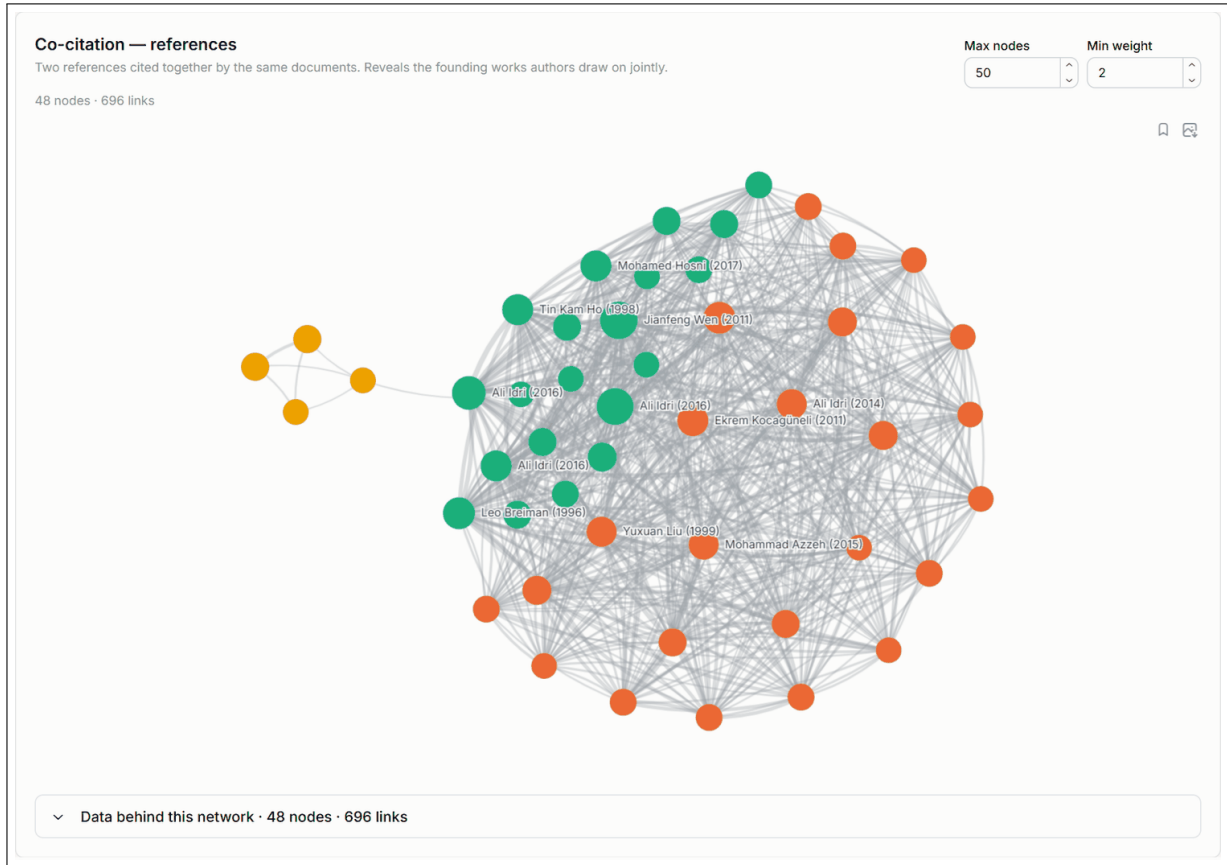
How it is computed

In each column, the n most frequent values. Width of a flow between two values of neighbouring columns = number of documents that carry both. A node is identified by its column and its name: a keyword and a journal with the same name stay two nodes.

22 Networks: the common principle



Collaboration network between authors.



Co-citation network of the references.

How it is computed

Construction. Two entities are linked when they appear in the same document; weight c_{ij} = number of shared documents. Depending on the network, the entity is:

- an **author** (co-authorship), an **institution**, a **country**, a **city**;
- a **keyword** (co-words);
- a **cited reference** (co-citation: two references cited by the same document); a **cited author** (author co-citation);
- a **document** (bibliographic coupling: weight = number of shared references).

The $N_{\max}$ most frequent entities are kept, and the links with weight $\geq$ a minimum.

Normalisations (s_i : total occurrences of i):

$$\text{association} = \frac{c_{ij}}{s_i s_j}, \quad \text{Jaccard} = \frac{c_{ij}}{s_i + s_j - c_{ij}}, \quad \text{Salton} = \frac{c_{ij}}{\sqrt{s_i s_j}},$$

$$\text{inclusion} = \frac{c_{ij}}{\min(s_i, s_j)}, \quad \text{equivalence} = \frac{c_{ij}^2}{s_i s_j}$$

Communities: Louvain algorithm (fixed seed), which maximises the modularity

$$Q = \frac{1}{2m} \sum_{i,j} \left(w_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j)$$

(m : total weight, k_i : weighted degree, g_i : group of i). The colour of a node is its group; its size, its occurrences.

Layout: multidimensional scaling on the shortest paths, refined by a repulsion force, with a fixed seed: the same network is always drawn the same way. Groups without links between them are arranged in rows to the format of the map, and each disc is pushed away from its neighbours according to its drawn size: none covers another, even at 250 nodes. Weak links fade when the network is dense.

How to read it

Under each network, “Data behind this network” gives the nodes and the links as tables, exportable: a network is not read by eye alone.

23 Network lab: what the eye does not see

Network

Co-authorship

Link normalization

Association strength

Max nodes

50

Min weight

1

Graph

Density

Cluster resolution

1

Colour by mean year (overlay)

☐

Two authors signing the same documents. - c / (s_i + s_j) — VOSviewer's default. Corrects the bias that favours frequent entities. - Resolution above 1 gives more, smaller clusters; there is no correct value — it is a choice of granularity.

The settings: network type, normalisation, number of nodes, minimum weight, community resolution.



The chosen network, here co-authorship, normalised by association strength.

Centrality ranking  

Betweenness first: it is the measure that finds bridges, and the one that most often disagrees with plain degree.

Node	Occ.	Degree	Betweenness	Closeness	PageRank	Cluster coef.	Community
Hosni M.	50	0.4694	0.5519	0.008	0.02132	0.0057	4
Medarhri I.	26	0.3061	0.3138	0.0062	0.03249	0.0122	2
Belhaj Z.	2	0.0816	0.2942	0.0067	0.02741	0.0558	6
Zine R.	3	0.0612	0.2772	0.0066	0.0163	0.0603	6
Abnane I.	7	0.1429	0.2151	0.0076	0.01396	0.0244	3
García - Berná J.A.	22	0.4082	0.2083	0.0068	0.02872	0.014	1
García - Mateos G.	9	0.1429	0.1769	0.0072	0.01689	0.0258	3
Mayordomo - Martínez D.	4	0.102	0.1701	0.0071	0.01589	0.0389	3
Zerhouni J.	4	0.102	0.1173	0.0052	0.01615	0.1908	5
Idri A.	38	0.3265	0.1097	0.0073	0.01952	0.0106	3
Nicolás J.	32	0.3265	0.085	0.0066	0.02622	0.0189	1
Toval Á.	2	0.1224	0.0621	0.0068	0.01508	0.0332	1
Ebert C.	5	0.1224	0.0561	0.0068	0.02159	0.0441	1
Vizcaino A.	5	0.1224	0.0561	0.0068	0.02159	0.0441	1
Toval A.	25	0.3265	0.0408	0.0064	0.02661	0.0187	1
López - Jiménez J.J.	6	0.2041	0.0306	0.0058	0.02558	0.0232	1
Sobrino - Duque R.	3	0.102	0.0272	0.0059	0.01586	0.0452	1
Pérez - Sánchez J.	6	0.1429	0.0196	0.0064	0.01535	0.0298	1
Mejía - Granda C.M.	3	0.0816	0.0162	0.0059	0.01084	0.0418	1
Misnevs B.	2	0.1224	0.0145	0.0062	0.02151	0.059	1

Ranking of the nodes by centrality.

How it is computed

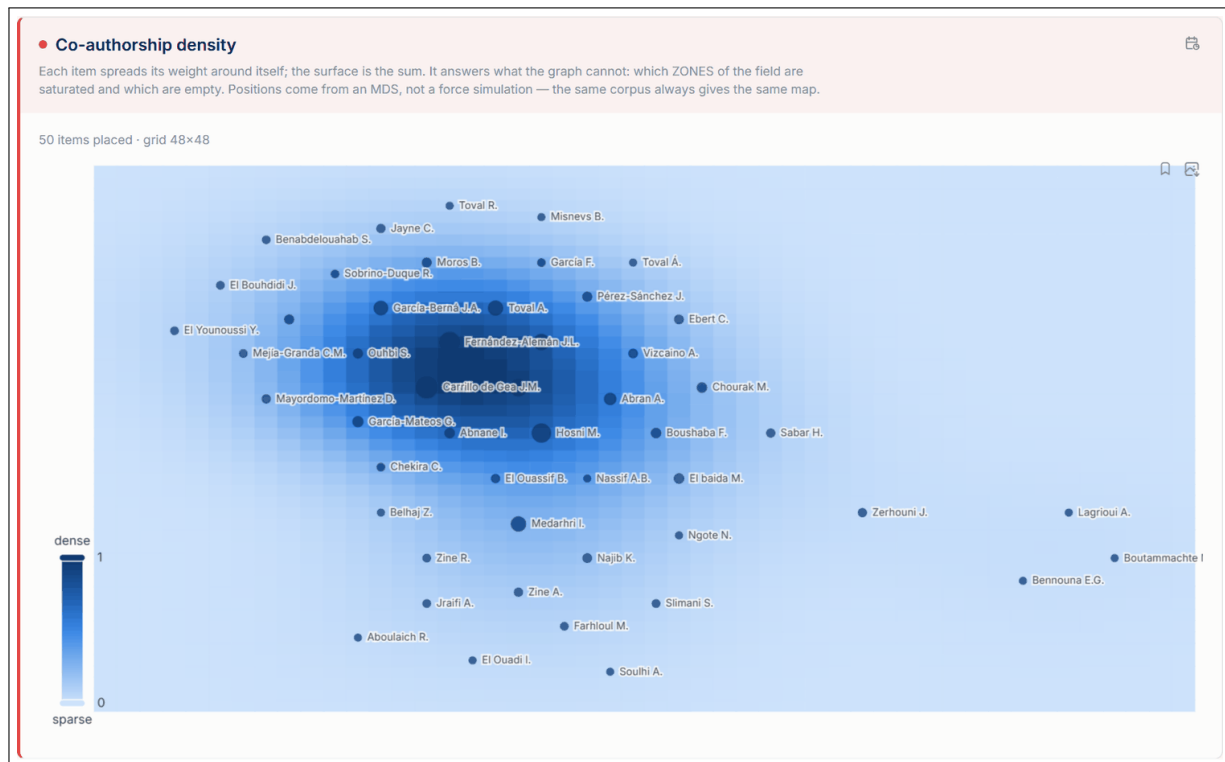
For a node i (network with n nodes; distance of a link $= 1/w_{ij}$):

- **Degree:** $\frac{k_i}{n-1}$, with k_i the number of neighbours of i .
- **Betweenness:** $\sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$, the share of the shortest paths between s and t that go through i (normalised).
- **Closeness:** $\frac{n-1}{\sum_j d(i,j)}$, with d the length of the shortest path.
- **PageRank:** $\pi = \frac{1-\alpha}{n} \mathbf{1} + \alpha W^\top \pi$, $\alpha = 0.85$, W the weight matrix normalised by row.
- **Eigenvector:** x such that $Ax = \lambda_{\max} x$.
- **Clustering:** the (weighted) share of the pairs of neighbours of i that are themselves neighbours.

For the whole network: density $= 2E/(n(n-1))$, transitivity, number of components; mean path length and diameter on the largest component.

How to read it

Betweenness finds the *bridges*: a node with few links but the only link between two groups. It is the measure that most often contradicts plain degree.



“Density” view: where the network is dense, and with what.

How it is computed

Each node i , placed at p_i and with weight w_i (its occurrences), spreads a Gaussian kernel:

$$\rho(x) = \sum_i w_i \exp\left(-\frac{\|x - p_i\|^2}{2h^2}\right), \quad h = \text{median}_i\left(\min_{j \neq i} \|p_i - p_j\|\right)$$

then ρ is brought back to $[0, 1]$ by dividing by its maximum. The width h follows the typical gap between neighbours: a kernel too narrow gives a map of dots, too wide a uniform blot.

24 Citation network

Aggregate by

Max nodes

Min citations

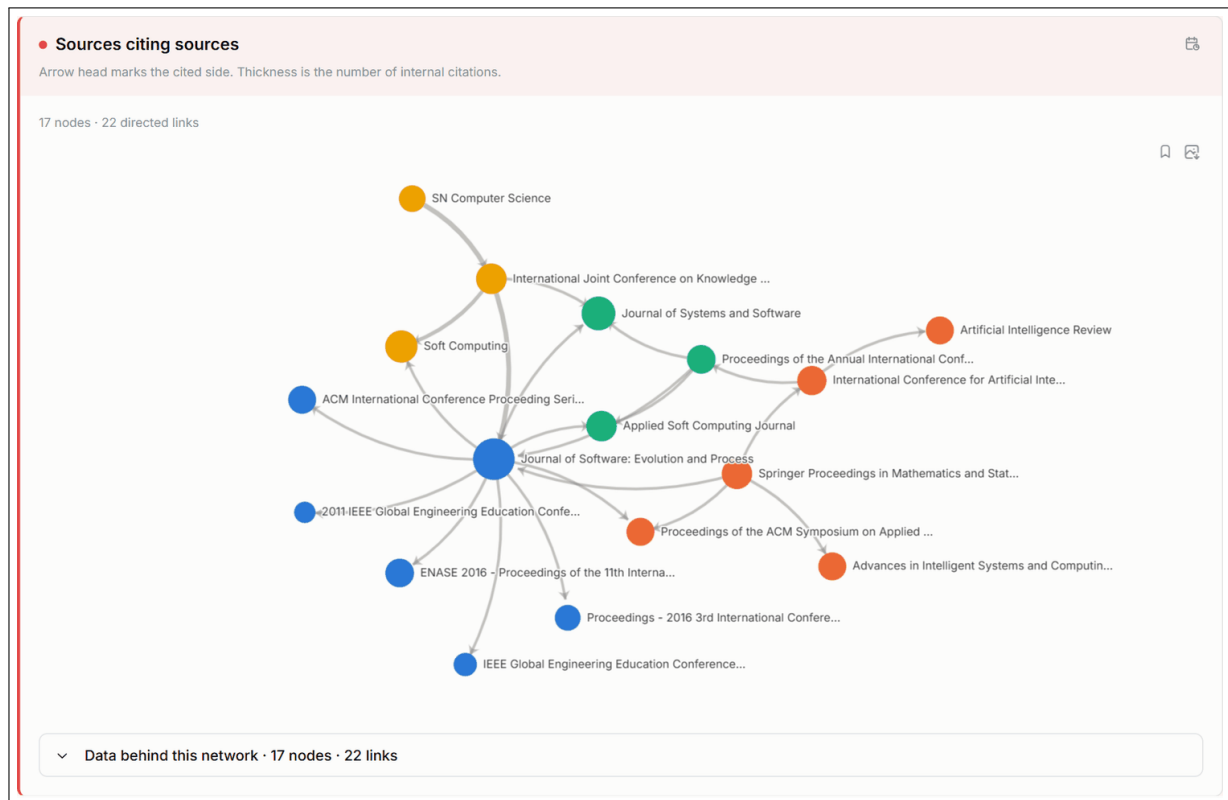
Sources

30

2

Which venues feed which.

The aggregation unit: sources, authors, countries, institutions or documents.



Who cites whom: directed network of the direct citations between sources of the corpus.

How it is computed

From the citing $\rightarrow$ cited pairs (section 12), aggregated at the chosen unit: weight of $A \rightarrow B$ = number of citations from documents of A to documents of B . An entity citing itself is set aside.

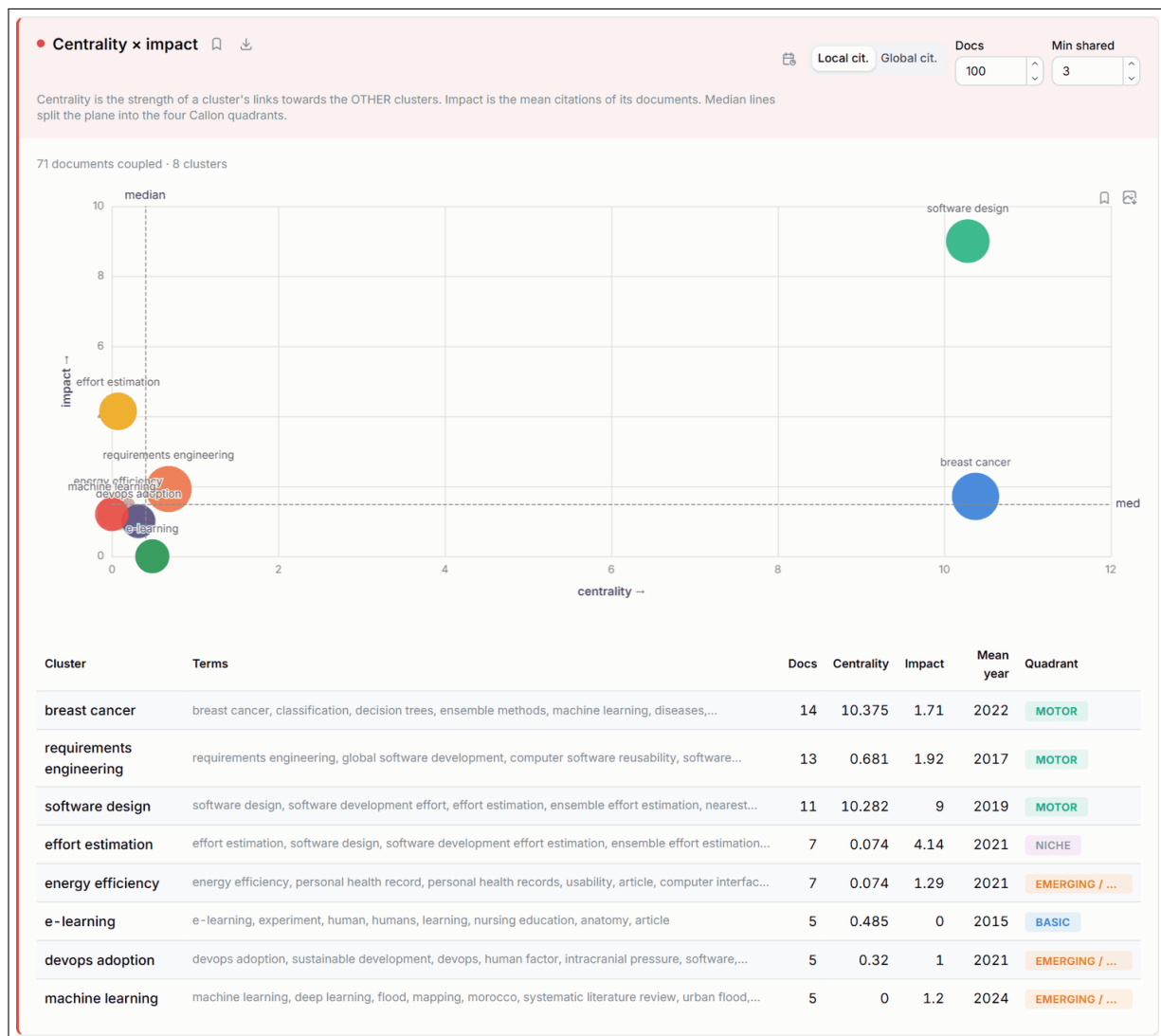
● Citation balance Received minus emitted. Positive means the entity supplies more ideas to the corpus than it takes from it — the clearest reading this network offers.			
Entity	Received	Emitted	Balance
Journal of Systems and Software	6	0	+6
Soft Computing	5	0	+5
Applied Soft Computing Journal	4	0	+4
Proceedings of the ACM Symposium on Applied Computing	4	0	+4
ENASE 2016 - Proceedings of the 11th International Conference on Evaluation of Novel Software Approaches to Software Engineering	2	0	+2
ACM International Conference Proceeding Series	2	0	+2
Advances in Intelligent Systems and Computing	2	0	+2
Artificial Intelligence Review	2	0	+2
Proceedings - 2016 3rd International Conference on Soft Computing and Machine Intelligence, ISCMi 2016	2	0	+2
IEEE Global Engineering Education Conference, EDUCON	2	0	+2
2011 IEEE Global Engineering Education Conference, EDUCON 2011	2	0	+2
International Conference for Artificial Intelligence: Applications, Innovation and Ethics, AI2E 2025	2	4	-2
Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS	2	6	-4
SN Computer Science	0	4	-4
International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings	4	9	-5
Springer Proceedings in Mathematics and Statistics	0	8	-8
Journal of Software: Evolution and Process	8	18	-10

Citation balance: received minus given.

How it is computed

$\text{balance}(A) = \sum_B w_{B \rightarrow A} - \sum_B w_{A \rightarrow B}$. Positive: the entity is a source of ideas for the corpus; negative: it consumes more than it provides.

25 Coupling clusters



Groups of coupled documents, placed by their centrality and impact.

How it is computed

The bibliographic coupling network (two documents linked by their shared references) is normalised by **association strength** $c_{ij}/(s_i s_j)$, otherwise an article with 300 references would couple with everyone. The groups are formed by Louvain. For each group G :

$$\text{centrality} = 100 \times \sum_{i \in G, j \notin G} w_{ij}, \quad \text{impact} = \frac{1}{|G|} \sum_{d \in G} LC_d$$

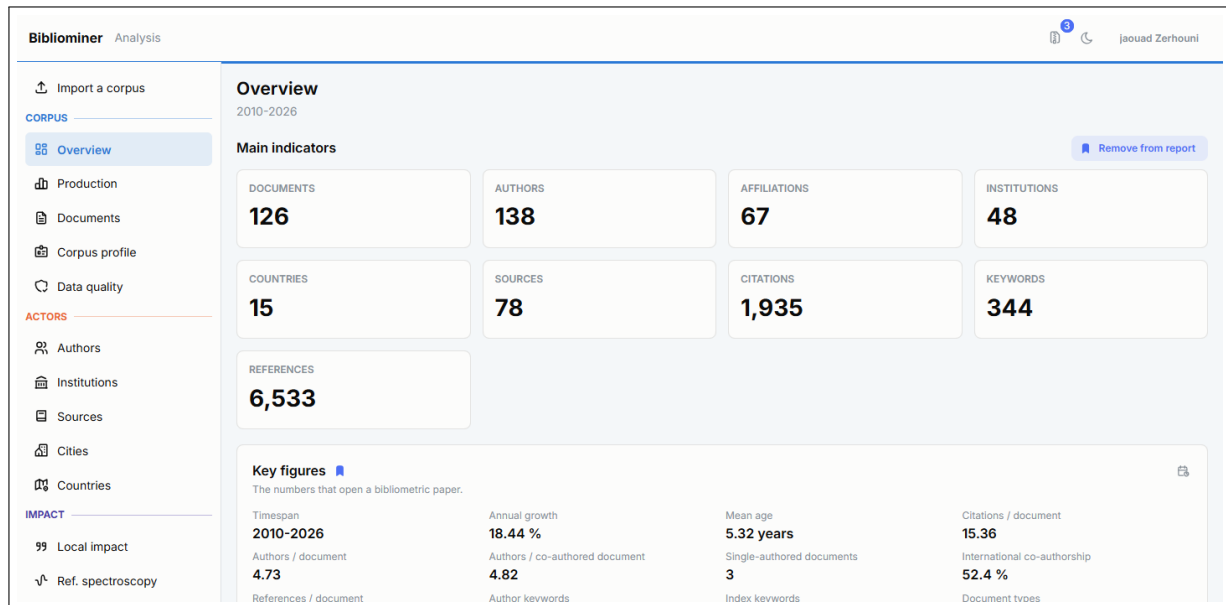
Impact is local by default (the citations received within the corpus). The quadrants are cut at the medians, like Callon's map.

How to read it

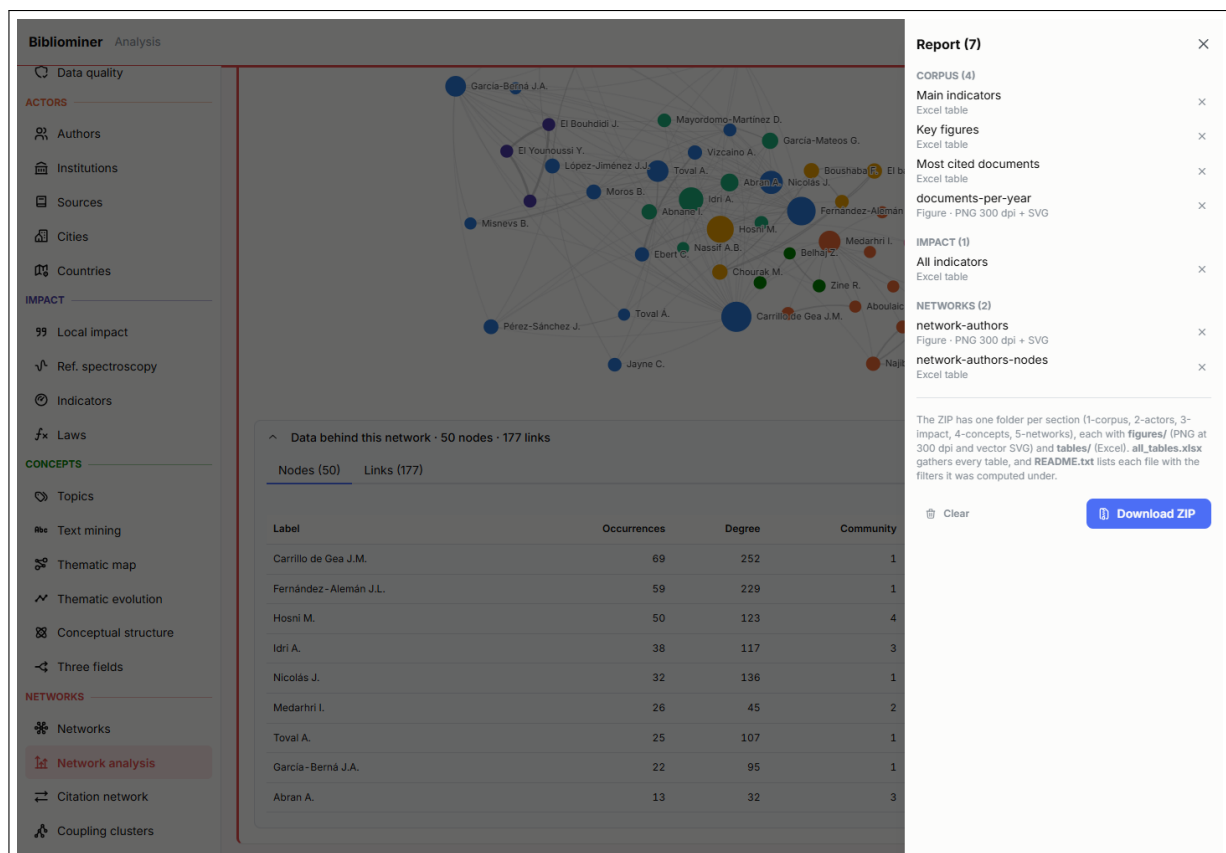
The thematic map groups *words*; this one groups *documents* that draw on the same sources. It finds the proximities that keywords miss, between communities that name the same thing differently.

26 The report: Add to report

Each figure, each table and each group of indicators carries a bookmark: [Add to report](#). Browse the pages, add what you keep, then the basket icon, at the top right, downloads everything in a single ZIP. On [Overview](#) and [Indicators](#), a button adds the indicators of the page at once ([Add all indicators to report](#)).



Overview: the main indicators are added to the report as a table (indicator, value, definition).



The basket: each entry states its nature (figure or Excel table) and the filters with which it was computed.

In the ZIP	Content
README.txt	each file, its title and the filters that produced it
all_tables.xlsx	all the tables, one sheet each
1-corpus/figures/	figures as 300 dpi PNG and as SVG (vector)
1-corpus/corpus_tables.xlsx	the tables of the section, one sheet each, after a contents sheet
2-actors/ ... 5-networks/	the other sections of the menu, likewise

How to read it

An entry is a *snapshot*: it keeps the filters of the moment it was added, including the period specific to a block. Two rankings of the authors, 2016–2019 and 2020–2025, are two distinct entries and the README says so. The basket survives a page reload; **Clear** empties it.

27 The same results in Python

Everything the interface shows comes from the `bibliominer-analysis` package, published on PyPI (<https://pypi.org/project/bibliominer-analysis/>), which can also be used on its own, in a notebook. It installs with a single command (Python 3.9 or later):

```
pip install bibliominer-analysis
```

It takes the cleaned file as input; a raw Scopus export is refused. The options of the screen (period, top N , unit of a network...) are arguments:

```
from bibliominer_analysis import Corpus, Report

corpus = Corpus.from_csv("merged_cleaned.csv")
recent = corpus.filter(years=(2023, 2025))      # the filter bar
recent.authors_impact(n=10)                     # a top 10
Corpus.figure_catalog()                        # the figures, page by page
recent.figure("top-authors", n=10, path="top_authors.svg")
recent.figure("collaboration-map", path="collaboration.png") # 300 dpi

report = Report(filters={"years": "2023-2025"})
report.add_corpus_figure(recent, "top-authors", n=10)
report.add_indicators("Impact", "Collaboration",
                      recent.collaboration_indicators())
report.save("report.zip")                      # the same ZIP as the interface
```

How to read it

The name of a figure is that of its file exported by the interface; its default options are those of the screen. Without options, the figure of the interface is obtained; with them, the same figure set differently. The README of the package gives, page by page, the Python call of each table and each figure.