,

Science-Metrix

Bibliometrics Indicators for the Science and Engineering Indicators 2024

Technical Documentation

October 2023

Science-Metrix

Bibliometrics Indicators for the Science and Engineering Indicators 2024

Technical Documentation

October 1st, 2023

Submitted to:
SRI International

Authors
Alexandre Bédard-Vallée
Guillaume Roberge

By:

Science-Metrix
1.438.945.8000 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com

Contents

Tables

Figures

1 Introduction

Science-Metrix, now part of Elsevier, has been commissioned by SRI International, on behalf of the National Science Foundation (NSF), to develop measures and indicators of research and patent activity using bibliometrics and patent data for inclusion in the Science and Engineering Indicators (SEI) 2024. This technical document details the various steps taken to implement the databases, clean and standardize the data, and produce statistics. This documentation is accompanied by a collection of external files that are necessary complements to perform these tasks. The list of accompanying external files is as follows:

External File 1: Databricks scripts
External File 2: Scopus discontinued title list
External File 3: DOAJ canceled title list
External File 4: Article eid to TOD and subfields
External File 5: Enriched Scopus country mapping
External File 6: Scopus U.S. addresses to U.S. states
External File 7: Scopus U.S. addresses to sectors
External File 8: Impact indicators NSF production

These external files are also introduced in the relevant sections of this documentation.

2 Bibliometric methods

Bibliometrics is, in brief, the statistical analysis of scientific publications, such as books or journal articles. Bibliometrics comprises a set of methods used to derive new insights from existing databases of scientific publications and patents. In this study, the bibliometric indicators are not computed on the original and complete text of the publications, but rather on the bibliographic information of a very comprehensive set of scientific articles published in peer-reviewed journals and indexed in the Scopus database. As Figure 1 exemplifies, the information used to compute the indicators is mostly derived from the bibliographic information contained in the first page of the document and in the list of references.

Only two databases offer extensive coverage of the international scientific literature and index the bibliographic information required to perform robust and extensive bibliometric analyses—both of which are aspects necessary for performing advanced bibliometric analyses on scientific activity. These databases are the Web of Science (WoS), which is produced by Clarivate Analytics and currently covers about 21,500 peer-reviewed journals, and Scopus, which is produced by Elsevier and covers about 39,000 peer-reviewed journals. Both these counts do not include the number of conference proceedings indexed in these databases.

The bibliometric indicators for SEI 2024 have been produced by Science-Metrix using an in-house implementation of the Scopus database that has been carefully conditioned for the production of large-scale comparative bibliometric analyses. A similar approach using Scopus was also employed to produce indicators of all editions of the SEI since 2016.

For this project, the indicators were computed on science and engineering scientific publications; this includes publications on the natural sciences, the applied sciences, the medical sciences, and the social sciences, but excludes the arts and humanities. Only peer-reviewed documents have been retained (i.e., articles, reviews, and conference papers). The peer-review process ensures that the research is of good quality and constitutes an original contribution to scientific knowledge. In the context of bibliometrics, these documents are collectively referred to as *papers*.

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biological Conservation 118 (2004) 583–592

BIOLOGICAL CONSERVATION

www.elsevier.com/locate/biocon

Comparison of Coleoptera assemblages from a recently burned and unburned black spruce forests of northeastern North America

Michel Saint-Germain [a,*], Pierre Drapeau [a], Christian Hébert [b]

[a] *Groupe de recherche en écologie forestière interuniversitaire, Département des sciences biologiques, Université du Québec à Montréal, CP 8888, succ., Centre-ville, Montréal, Que., Canada H3C 3P8*

[b] *Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Center, 1055 rue du PEPS, CP 3800, Sainte-Foy, Que., Canada G1V 4C7*

Received 2 April 2003; received in revised form 15 September 2003; accepted 14 October 2003

Abstract

Several insect groups have adapted to fire cycles in boreal forests, and can efficiently use new habitats created by fire. Our study aimed at producing a first characterization of post-fire Coleoptera assemblages of black spruce forests of eastern North America. For two years, we sampled Coleoptera using flight-interception traps in burned stands of contrasting age and structure in a 5097-ha wildfire and in neighbouring unburned mature stands. More than 40 species were exclusively captured in burned stands. Time elapsed since fire and proximity of unburned forests were the most significant parameters affecting Coleoptera assemblages. Stand age and structure had limited effects on assemblage structure; the Scolytid *Polygraphus rufipennis* Kirby was the only common species to clearly favor older stands. Fire-associated Coleoptera assemblages found in our study area were clearly distinct from those found in similar unburned stands; we should thus be conservative in our management approach concerning recently burned stands.

Keywords: Boreal forest; Forest fires; Habitat selection; Fire-associated Coleoptera; Salvage logging

References

Anhlund, H., Lindhe, A., 1992. Endangered wood-living insects in coniferous forests– some thoughts from studies of forest-fire sites, ourcrops and clearing in the province of Sörmland, Sweden. Enromologisk Tidskrift 113, 13–23 (in Swedish).

Bergeron, Y., Archambault, S., 1993. Decreasing frequency of forest fires in the southern boreal zone of Quebec and its relation to global warming since the end of the "Little Ice Age". The Holocene 3, 255–259.

- counts of papers by publication year (trends)
- delineation of scientific fields/subfields or research topics
- counts of papers by researcher (author)
- counts of papers by institution, sector, province, region and country
- citations counts, i.e. number of times paper appears in references of other papers to measure scientific impact

Figure 1 Bibliographic information for the computation of bibliometric indicators
Source: Prepared by Science-Metrix

2.1 Database implementation

Scopus data were acquired by Science-Metrix through the internal Elsevier network via its Databricks platform. Scopus data are stored on Elsevier-controlled Amazon servers in their original XML format in a distributed file format that is updated daily with new articles and modifications to existing records. The daily updates enable Science-Metrix to extract the most up-to-date and complete information available. Each article in Scopus is stored as its own XML and transformed using the Scala programming language into a set of intermediary tables, again hosted on Elsevier servers and accessed using the Databricks platform. The end result was three tables that contained the required information about each paper. For previous editions of the SEI, data were then exported to a Microsoft SQL Server database. However, since SEI 2022, all further processing, from initial data ingestion to the output of the final Excel files, takes place directly on the Databricks platform. This simplified the workflow for this edition, as it removed a number of data transfer steps from the production pipeline and ensured no compatibility issues (e.g., in data encoding) would arise.

Table I Link between XML items and columns in the article table

Column	Data type	XPATH
eid	bigint	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid[@idtype='SGR']
index_date	string	/xocs:doc/xocs:meta/xocs:indexeddate
orig_load_date	string	/xocs:doc/xocs:meta/xocs:orig-load-date
sort_date	string	/xocs:doc/xocs:meta/xocs:datesort
year	int	/xocs:doc/xocs:meta/xocs:sort-year
month	string	*Derived from sort_date*
day	string	*Derived from sort_date*
doi	string	/xocs:doc/xocs:meta/xocs:doi
doc_type	string	/xocs:doc/xocs:meta/cto:doctype
source_title	string	/xocs:doc/xocs:item/item/bibrecord/head/source/sourcetitle
source_abbr	string	/xocs:doc/xocs:item/item/bibrecord/head/source/sourcetitle-abbrev
source_id	bigint	/xocs:doc/xocs:item/item/bibrecord/head/source/@srcid
issn_ani	string	/xocs:doc/xocs:meta/xocs:issn
issn	string	/xocs:doc/xocs:item/item/bibrecord/head/source/issn[@type ='print']
issn2	string	/xocs:doc/xocs:item/item/bibrecord/head/source/issn[@type ='electronic']
issn3	string	/xocs:doc/xocs:item/item/bibrecord/head/source/issn[@type ='other']
subject	string	/xocs:doc/xocs:meta/xocs:subjareas/
source_type	string	/xocs:doc/xocs:item/item/bibrecord/head/source/@type
title	string	/xocs:doc/xocs:item/item/bibrecord/head/citation-title/titletext
title_original	string	/xocs:doc/xocs:item/item/bibrecord/head/citation-title/titletext/@original
title_lang	string	/xocs:doc/xocs:item/item/bibrecord/head/citation-title/titletext/@language
total_ref	string	/xocs:doc/xocs:item/item/bibrecord/tail/bibliography/@refcount
volume	string	/xocs:doc/xocs:meta/xocs:volume
issue	string	/xocs:doc/xocs:meta/xocs:issue
first_page	string	/xocs:doc/xocs:meta/xocs:firstpage
last_page	string	/xocs:doc/xocs:meta/xocs:lastpage
provider_timestamp	timestamp	/xocs:doc/xocs:meta/xocs:timestamp/@datetime
unique_auth_count	smallint	/xocs:doc/xocs:meta/cto:unique-auth-count
pmid	string	/xocs:doc/xocs:meta/xocs:pmid
source_filename	string	*Generated by the extraction code*

Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

Table II Link between XML items and columns in the author_address table

Column	Data type	XPATH
eid	bigint	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid[@idtype='SGR']
ordre_address	int	*Automatically incremented by the extraction script*
country_iso3	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/@country
country	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/country
city_group	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/city-group
afid	bigint	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/@afid
dptid	bigint	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/@dptid
ordre_author	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/@seq
auid	bigint	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/@auid
indexed_name	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:indexed-name
given_name	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:given-name
initials	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:initials
surname	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:surname
pref_indexed_name	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:indexed-name
pref_given_name	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:given-name
pref_author_initials	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:initials
pref_surname	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:surname
organization	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/organization OR /xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/ce:source-text
ordre_affil	int	*Automatically generated by the extraction code*
address_part	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/address-part
city	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/city
postal_code	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/postal-code
state	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/state
degree	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:degrees
email	string	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:e-address
is_collaboration	string	*Automatically generated by the extraction code*

Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

Table III Link between XML items and columns in the reference table

Column	Data type	XPATH
eid	bigint	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid attr=SGR
eid_ref	bigint	/xocs:doc/xocs:meta/cto:ref-id

Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

External File 1: Databricks scripts[1]

[1] These scripts have been converted to a format that is easily readable by humans, should someone wish to recode this without having access to the Databricks platform. However, they can still be reimported to Databricks easily without loss of content by using Databricks' import functionnality and uploading the full .zip file provided.

2.1.1 Completeness of the database

There is a time lag between when a document is published and when it is indexed in Scopus. Because of this, the documents published in any given year are not completely indexed in Scopus on 31 December of that year. In order to produce meaningful and robust indicators, we aimed to have at least 95% of the latest year's documents indexed in the database when we calculated the indicators. One of the challenges in determining the completeness of a database is to determine what is considered 100% of a publishing year's documents. As noted, new documents are indexed constantly. Most of these have been recently published, but each weekly update file from Elsevier includes some journals and documents published in previous years. Therefore, the total number of documents for any publishing year is always increasing, and the numerator in the calculation of the completeness is a moving target.

In order to be able to measure completeness, it was assumed that the completeness of a publication year is achieved within 30 months of the first day of a publication year. For example, publication year 2016 was assumed to be complete by 1 July 2018. This is somewhat arbitrary but is a reasonable compromise between a short time frame for analysis and a high completion rate.[2]

Based on Elsevier's modeling, a completion rate of at least 95% could be expected within 18 months of the first day of a publication year in the past edition. In our case, this meant a 95% completion rate on 1 July 2023 at the latest for the publication year 2022. However, numerous efforts have been made since 2019 to further accelerate the indexation of documents, resulting in the completion rate reaching 95% much earlier in the year, as soon as January, reducing the delay to about 12 months (i.e., 1 January instead of 1 July). For the estimation of completeness of a publication year on a specific date, the number of publications counted on that specific date is calculated against the number of publications counted or estimated after 30 months.

$$Completeness\ percentage = \frac{N_i}{N_j}$$

Where:

N_i is the number of publications after i months, in this case 12 months,

and N_j is the number of publications after j months, in this case 30 months.

As it is currently impossible to calculate the "30 months after" estimation for 2022, a prediction model was used to determine the expected number of publications for these years. The model used is simple. The compound annual growth rate for 2019 to 2021, the last three completed years, was first calculated using a simple regression model. This growth rate was then used for every year, starting in 2019.

Details about the complete set of filters applied to the Scopus database are available in sections 2.1.2 and 2.1.3. However, because some of these tests cannot be performed before the whole production version

[2] Although it is impossible to measure completeness, the rate at which new documents are added after 30 months after publication is low, and therefore the count of publications measured at that time is a fairly stable number against which one can benchmark completion.

of the Scopus database is built, a stripped-down version of the data filtering was used to have a quickly computable and easily reproducible model that did not require the full preparation of the complete database, given that the model was to be evaluated several times by Science-Metrix. The only filters used were document type (Article, Review, and Conference proceedings) and removal of the low-quality journals identified by Elsevier and DOAJ.[3]

As the continuous updating of the database by Elsevier includes substantial backward correction of already included papers, the full prediction model was recalculated at the beginning of each week. The results for the year 2022 as measured on the first weeks of every month between January and April 2023 is presented in Table IV.

Table IV Monthly follow-up of the completion rate for the year 2022

Month	Observed (papers)	Predicted (papers)	Completion (%)
January	3,407,099	3,725,856	91.4%
February	3,534,633	3,726,647	94.8%
March	3,568,763	3,725,214	95.8%
April	3,584,644	3,719,586	96.4%

Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

This year compared to previous SEI production years, estimated completion rates were lower than expected at a given moment. Although the number of publications indexed for 2022 in April roughly aligns with the faster growth observed in the database since 2017, it now appears that 2021 might have been a comparatively high-growth year, which led to an overestimation of the number of predicted papers. It is possible that the COVID-19 pandemic has perturbed growth in the 2019–2021 range. Secondary completion analyses were made to take into account the observed filling rate of the database, which showed that a more accurate completion estimation would have been of 98.8% in April, and that the situation was similar for most of the top 50 countries in terms of output. This evidence led to the April 1st snapshot being used for statistics computation for SEI 2024.

2.1.2 Filtering non-peer-reviewed documents

The Scopus database is composed mainly of original peer-reviewed documents, but it also includes other material produced by the scientific community. In order to identify the peer-reviewed documents, information on the type of media (source type) and the document type is used. The types of media included in Scopus are categorized into eight categories:

1. Journal
2. Conference Proceeding

[3] The numbers reported here only serve an administrative purpose and should not be compared to the values found in the SEI report; however, differences are in the end minimal given the limited reach of the more complex filters not applied in the model.

3. Book
4. Book Series
5. Major Reference Work
6. Trade Publication
7. Report
8. Preprint Repository

These include documents that are categorized into 18 categories:

1. Article
2. Article in Press
3. Preprint Article
4. Conference Paper
5. Conference Review
6. Review
7. Data paper
8. Letter
9. Book
10. Book Chapter
11. Editorial
12. Note
13. Short Survey
14. Report
15. Abstract Report
16. Business Article
17. Erratum
18. Retracted Document

For this project, the goal was to keep only documents that were peer reviewed and that presented new scientific results. The classification of documents by source type and document type in Scopus cannot be used directly to precisely identify all peer-reviewed papers in the database. An empirical approach has been developed by Science-Metrix to filter documents based on the source types and document types, and to maximize the recall of peer-reviewed papers while trying to minimize the inclusion of non-peer-reviewed documents. The approach is based on Elsevier's documentation and statistics on the number of references and citations per document for each combination of source type and document type. Science-Metrix also filters out documents that have a value of "0" for the "refcount" field, which indicates that the paper did not refer to any other works: a strong indication that it is not original, peer-reviewed research. This filter is applied before subsequent steps of data standardization.

Table V details the combinations that have been kept for the bibliometric analyses.

Table V Combinations of source types and document types used for the production of bibliometric indicators

Source Type	Document Type
Book Series	Article, Conference Paper, Review, Short Survey
Conference Proceeding	Article, Review, Conference Paper
Journal	Article, Conference Paper, Review, Short Survey

Source: Prepared by Science-Metrix

2.1.3 Filtering low-quality papers

The classical publication business model was based on printed journals competing for limited library shelf space. Since the 1990s, scholarly publishing has been transitioning from print-based to digital publishing. Many of these digital publications are available through open access (as evidenced by this edition's data, approximately 34% of all publications published between 2013 and 2022 were made available via journal-based open access[4]. Open access scholarly literature is free of charge to read and often carries less restrictive copyright and licensing barriers than traditionally published works. "Pay-to-publish" journals contain low-quality non-peer-reviewed articles, representing an abuse of the open access model.[5] Researchers may or may not be the victim of a dubious publisher, since early-career and naïve researchers may erroneously submit papers to such journals and low-quality researchers may seek to increase their publication numbers without the hard work that comes with performing original research.

Researchers have attempted to create lists of good and bad journals. However, it is challenging to create a transparent system because there are high-quality journals with few subscribers and new journals without a proven track record.

Science-Metrix applied a systematic approach to remove low-quality journals, rather than a journal-by-journal approach. The approach is to reject from this database two sets of journals identified as low quality. The first set is the list of journals removed by Elsevier from the Scopus database. Scopus undertakes periodic reviews of the quality of the journals already accepted into the database and those applying for entry into the database. The discontinuation of a journal in Scopus means that no new papers from this journal enter the database; however, the Scopus database retains the already indexed papers from canceled journals. To create a comparable time series, Science-Metrix removed all the previously indexed papers of the Elsevier canceled journals—this does not reflect upon the journal quality prior to cancelation but rather on the need to create a consistent time series.

The second list of excluded journals comes from the Directory of Open Access Journals (DOAJ),[6] which is a community-curated online directory of open access scientific journals. DOAJ has a set of inclusion criteria to ensure the directory only includes peer-reviewed, open access journals and diffuses the list of

[4] Journal-based open access excludes publications that are made open-access by uploading them to a repository (ArXiv, medRxiv, Figshare, etc.). The share presented excludes publications with unknown access types.

[5] Memon, A. R. (2019). Revisiting the term predatory open access publishing. *Journal of Korean Medical Science, 34*(13), e99. doi:10.3346/jkms.2019.34.e99

[6] https://doaj.org/

journals that have been excluded over the years, with a brief description of the reason for exclusion. Science-Metrix constructed a list of excluded journals based on a subset of these reasons (ISSN not registered, invalid, or not in the ISSN database; suspected editorial misconduct by publisher; no editorial board; use of fake impact factor). Journals that were excluded in the DOAJ only because they are no longer open access were retained in the database for analysis.

The two lists of journals excluded based on the Scopus and DOAJ criteria that were used for SEI 2024 are included as external files.

External File 2: Scopus discontinued title list

External File 3: DOAJ canceled title list

2.2 Data standardization

2.2.1 Linking TOD classification to the database

Since SEI 2020, the Taxonomy of Disciplines (TOD) classification has been used to classify science when calculating the indicators. This classification comprises 14 different fields of science, which are mutually exclusive. The approach used to classify articles from Scopus into those fields has also changed accordingly. Now, the categorization is done by first assigning the 176 subfields in Science-Metrix's own classification[7] to the different TOD fields. Papers that were part of a given subfield are automatically included in the corresponding TOD field. For example, papers from the subfield "Dentistry" are all assigned to the "Health Sciences" TOD field. This works well as the 176 subfields are much more granular than the TOD fields, which enables their easy classification in the larger categories. Some problematic subfields exist, however—namely, "Energy", "General Arts, Humanities & Social Sciences", and "General Science & Technology", the latter of which is used for generalist journals such as *Science* or *Nature*, which cannot be categorized in the TOD fields as a whole because of the great variety of articles they contain.

To address this, a classification method at the paper level was developed and used on the papers from those categories. This method is based on a machine learning algorithm that was trained on a subset of papers already categorized in the other 173 subfields. Thorough testing ensured that the algorithm produced results that were equivalent to those that could be obtained manually by Science-Metrix analysts. The algorithm was used to reclassify the papers from the three generalist subfields into the more specific ones that mapped to only one possible TOD field. This resulted in a classification in which every paper is assigned to only one TOD field, which can then be used to produce statistics for each area of science. Details about the reclassification methods are presented next.

[7] This classification is done at the journal level. Papers published in those journals automatically get assigned to the same category.

2.2.2 Paper-level reclassification of general subfields

Originally, peer-reviewed papers were all classified at the journal level. However, some journals are interdisciplinary and thus could not be classified. Often, the highest-impact publications within disciplines are published in these interdisciplinary journals. To properly compare the citation scores of publications, they should be normalized by discipline as each discipline has its own citation patterns. Thus, papers published in interdisciplinary journals had to be reclassified so that they could be normalized against documents from the same subfields of science. We conducted research to find the best approach to reclassify publications, and the most suitable model we were able to identify was based on artificial intelligence. As described below, its performance with the classification was eventually as good as or even slightly better than that of human experts.

Model description

The classifier used here was a neural network; it can be described as a character-based convolutional deep neural network. To clarify, a neural network is a type of machine learning algorithm. The neural network can be designed to receive words, letters, or other tokens and features. Here, *character-based* means that the model used to reclassify the papers uses letters as inputs. From those letters, the model learns words and discovers features. *Convolutional* refers to the architecture of the model. It is a well-studied and very performant architecture.[8] *Deep* means that there are several layers to the neural network architecture. This type of supervised machine learning approach has recently been found to be extremely performant to find patterns in noisy data, providing the network can be trained on a lot of data.[9]

Virtually any type of information relating to a publication can be provided to a deep neural network. The model performed best when it was given the following information: the author affiliations, the names of journals referenced in bibliography, the titles of the references, the publication's abstract, the publication's keywords, the publication's title, and the classification of a publication's reference. Each of these pieces of information were given a slot of a fixed length. For example, author affiliations were placed first in the input vector with 450 characters. Any text longer than its allocated slot was truncated and, if the slot was not completely filled, the text was padded with zeroes at the beginning of the text. Table VI presents a list of the length of each feature.

[8] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved from https://doi.org/10.1038/nature14539

[9] Zhang, X., & LeCun, Y. (2015). Text Understanding from Scratch, 1–9. Retrieved from http://arxiv.org/abs/1502.01710

Table VI Feature fixed length

Feature	Length
Author Affiliation	450
References' journals	1000
References' titles	500
Publication abstract	1750
Publications keywords	150
Publication title	175
References' subfields	70

Source: Prepared by Science-Metrix

Each character was embedded into a one-hot encoded vector of length 244. *One-hot encoding* is defined as a vector filled with zeroes and a one at the position assigned to the character. Table VII presents an example of character embedding for the word "cab".

Table VII Illustration of character embedding

	c	a	b
a	0	1	0
b	0	0	1
c	1	0	0

Source: Prepared by Science-Metrix

For the encoding, the 26 letters, the 10 arabic numerals, several punctuation signs (e.g., "-„;.!?:'_/\|@#$%ˆ&*˜+=<>()[]{}\) and the space character each occupied one position in the vector. Any character that was not in this list was encoded as an asterisk (i.e., *). The subfields were the only features that were not fed to the model as raw text. They were instead encoded by assigning one position to each subfield. Therefore, the first 68 slots of the vector were assigned to characters and 176 slots were added to the vector, one for each subfield.

The deep neural network was nine layers deep (Table VIII). The first six layers were one-dimensional convolutions, and the three remaining were dense. Rectified linear units were used as the activation function between each layer, except after the last one, in which case a softmax was used instead. The kernels had a width of seven for the first two convolutions and three for the others. The model was trained with a stochastic gradient descent as the optimizer and categorical cross-entropy as the loss function. The gradient descent had a learning rate of 0.02, with a Nesterov momentum of 0.9 and a decay of 0.0001. The learning rate was updated every 400,000 publications. The model was trained on batches of 64 publications at a time. For the training set, all Scopus publications, up to 9 January 2019, were

included as potential candidates (i.e., about 40 million publications). However, after having been trained on approximately 25 million publications, the model did not show signs of additional improvement.

Table VIII Deep neural network architecture

Layer architecture	Number of features	Kernel size	Activation	Pooling	Dropout
conv1D	500	7	Rectified linear unit	3	
conv1D	500	7	Rectified linear unit	3	
conv1D	500	3	Rectified linear unit		
conv1D	500	3	Rectified linear unit		
conv1D	500	3	Rectified linear unit		
conv1D	500	3	Rectified linear unit	3	
dense	1500		Rectified linear unit		0.5
dense	1500		Rectified linear unit		0.5
dense	1500		softmax		

Source: Prepared by Science-Metrix

Model evaluation

There is no absolute truth on the correct subfield in which a scientific publication should fall. Thus, we asked six analysts to classify the same set of 100 randomly selected scientific publications. Then, we used that corpus as a gold standard to evaluate our deep neural network. The analysts were asked to classify the publications as best as they could using whichever information they wanted. Most analysts used search engines to acquire additional information. Analysts were permitted to assign more than one subfield to a publication, when it was ambiguous, but they were asked to rank the chosen subfields in order of suitability.

We used six indicators to evaluate the deep neural network. We calculated three indicators at two levels of aggregation (i.e., subfield and TOD). The indicators were (1) the average agreement between an analyst's first choice and the deep neural network's first choice, (2) the percentage of time that the deep neural network's first choice was within one of any analyst's first choices, and (3) the percentage of time that the deep neural network's first choice was within any of the analyst's suggested subfields.

Model's performance

At the level of subfields, analysts agreed with one another on average 41% of the time, whereas they agreed with the deep neural network on average 42.3% of the time (indicator 1). The classifier thereby seems to be of good quality and maybe even slightly better than humans. The probability that the

classifier's prediction fell within one of the analysts' first choices was 81% (indicator 2). The probability that the classifier's prediction fell within any of the analysts' choices was 92%, indicating that subfields selected by the classifier were deemed relevant by at least one analyst for 92% of all cases.

At the higher level of the TOD fields, analysts agreed with one another on average 68% of the time, whereas they agreed with the deep neural network on average 70% of the time (indicator 1). For TODs, the classifier seems to be, once again, as good as, and maybe better than, a human. The probability that the classifier's prediction fell within one of the analysts' first choices was 93% (indicator 2), and the probability that the classifier's prediction fell within any of the analysts' choices was 99% (indicator 3).

External File 4: Article eid to TOD and subfields

2.2.3 Data standardization: country, country groups, regions

The attribution of a country to each author address listed on a paper is a complex task that faces several challenges, including the following:

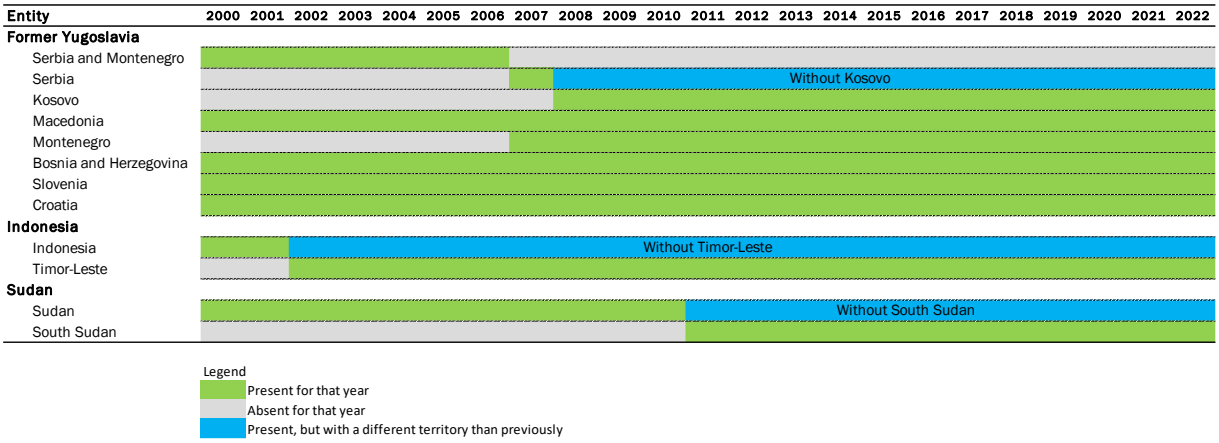
- There have been geopolitical changes in the world map over the time period covered by the database.
- Some parts of the world are disputed, and there is official and unofficial disagreement about some areas. For example, some authors claim to be publishing from particular countries years after those countries have been replaced by new entities.
- Scopus does not resolve these discrepancies and generally uses the country as provided by the author of the paper.
- The general process undergone by the metadata (first provided to the publisher by the author, then transferred by the publisher to Elsevier, which ultimately indexes the metadata in Scopus) entails various automatic and manual steps that may lead to errors—notably confusion between countries that sound alike (e.g., Australia and Austria) or that have similar ISO 3166-1 alpha-3 country codes (e.g., SLV for El Salvador and SVN for Slovenia).
- Older entries in Scopus often do not have information about the author's country.

In order to mitigate these challenges, a two-step process was developed:

1. Definition of a world map by year for the scope of this project.
2. Attribution of each institutional address to one of the countries available in this year.

For step 1, the list of countries is defined in SEI Appendix Table 5a-1. There were border changes in three regions of the world during the time period of interest in this report. These changes are summed in Table IX.

Table IX Geographic entities that changed over time



Note: Green: Present; Grey: Absent; Blue: Present, but with a territory removed.
Source: Prepared by Science-Metrix

Step 2 consisted in the use of several in-house heuristic algorithms based on external geographical sources to attribute each author address to a country. In addition to the name of the country as provided by Scopus, Science-Metrix developed an algorithm that used other information in the address field (city, affiliation, department, and author's unique ID) to identify the country. This algorithm clusters articles on other discriminating fields and looks for a suitable match based on both a frequency (at least 10 occurrences) and a ratio (at least 90%) threshold. For example, if an address lacks the country, but lists a city (e.g., Chicago) that in more than 90% of cases is associated with a given country (e.g., United States), then this country will be associated with this address.

Inconsistency and coding errors on the country were resolved using information about the city. For example, if an address was provided as the city of Bethlehem, but the country was attributed to Israel, then Science-Metrix overwrote the country to the West Bank. However, if the city was Bethlehem and the country was the United States, then the country was kept, as there are several cities and localities named Bethlehem in the United States.

An address attributed to a country that was subject to change on the world map as defined in step 1 was attributed to the entity encompassing the given city for the year of publication. For example, an address in the city of Juba in South Sudan would be attributed to South Sudan if published in 2015, but to Sudan if published in 2005.

A precision analysis demonstrated a precision level of above 99% for the assignment of countries to addresses. However, because a few countries account for a large share of the total output in the database, this high precision could still lead to lower precision levels for smaller countries. To alleviate the risk of important assignment errors for smaller countries, a stratified sampling approach was performed at the country level. A random sample of addresses was extracted for each country individually, and analysts manually validated the country assigned to these addresses, ensuring high precision (i.e., at least 95%) for each country.

External File 5: Enriched Scopus country mapping

Note on the impact of the United Kingdom's exit from the European Union

The withdrawal of the United Kingdom from the European Union (EU) on 31 January 2020, commonly referred to as Brexit, has had an impact on the data delivered for the 2022 and 2024 editions of the SEI. This is because the list of countries and regions used in previous editions included entries for the EU and for the rest of Europe under the name of "Other Europe". For data purposes, the United Kingdom effectively became a member of both regions (EU from 1973 to 2019, and Other Europe from 2020 onwards). Having a country with a large output volume switch regions during the period of analysis complicates matters as it makes the analysis of time trends for a given region very complex.

The solution we reached for this edition was to add new stable regions that had data presented for the full period of analysis, making the analysis of time trends for regions more reliable as none of these regions were affected by Brexit. The regions follow this hierarchy:

- Europe
    - EU-27 and UK
        - UK
        - EU-27
    - Other Europe
    - Other Europe, 2020
        - UK

The "Europe" region remains unchanged relative to past editions of the SEI, as it is a region defined geographically rather than politically. The "EU-28" group is composed of the 28 EU Member States prior to Brexit. In this group, the United Kingdom's publications are *always* included, including for 2020. The "EU-27" group is composed of the 27 Member States as of 2020. The United Kingdom's publications are *never* counted, even for the period 1996–2019.

The "Other Europe" region is composed of all European countries that were not members of the EU-28. It does not include the United Kingdom for all years. Meanwhile, the "Other Europe – 2020" region is composed of all European countries that were not members of the EU-27. It *includes* the United Kingdom for all years.

A significant difference with tables from the previous SEI editions is that there are now regions ("Other Europe" and "Other Europe-2020" as well as "EU-28" and "EU-27") that have overlapping country lists. Furthermore, the United Kingdom appears in both an EU and a non-EU region. However, the papers from all European countries are still only counted once at the level of Europe and the world. To further emphasize this, papers from all countries are still only shown once in the full data tables.

If it is needed to reconstitute a unified "EU" time series that reflects the change that occurred in 2020, it would always be valid to do so by combining the EU-28 and EU-27 time series appropriately. However, note that impact indicators are not affected by Brexit yet, as the time period for impact indicators ends

in 2018. If the current standards stay the same, Brexit will start having effects on the impact indicators of the EU and Other Europe regions starting from SEI 2024 (which would present impact indicators up to 2021) onwards.

2.2.4 Data standardization: U.S. states

Address information about the state or other subdivision of the author's listed institution was not available in the Scopus database until recently. The database contains a specific, standardized field that assigns a state or province to author addresses on papers; however, this field is not always populated, so it cannot be used to assign states to all U.S. addresses. Furthermore, the data in this field, although correct most of the time, are not perfect, and other information must be relied upon to accomplish the matching of U.S. addresses to states. Information about the city, the zip code, and the state as they appear on papers are also all contained in a separate single field named "city". Although the city is most often present in this field, the zip code and the state are not systematically recorded and are presented in an inconsistent format. However, for U.S. papers, most addresses somewhat fit the following convention: *city name, state abbreviation, zip code.*

Science-Metrix uses an algorithm to identify the state in U.S. addresses:

- A regular expression extracts the longest word that does not contain any digits from the city field. This word is a candidate for the city name.
- A regular expression extracts all five-digit numbers from the city field. These are potential zip codes.
- A regular expression (SQL script) extracts all two-capital-letter words that exist in the list of common U.S. state name abbreviations.
- The zip codes and city names are checked against a U.S. zip code/city database (e.g., https://www.unitedstateszipcodes.org/) to produce up to two candidate states per address.
- Full state names are searched for in addresses to assign these addresses to related states.
- Using decisions made in earlier steps, the distribution of the output of scientific papers across states for each distinct city name is computed; cities with at least five assignations and at least 90% of them pointing to a single state are assigned to the corresponding state.
- A city dictionary is also built from a U.S. geocoding database. City names that are always associated with the same state in this database get assigned to this state.
- Again, using decisions made in previous steps, the distribution of output across institutions, as defined by the Scopus "AFID" identifier, is computed for each institution; addresses linked to institutions with at least five assignations and at least 95% of all of them pointing to a single state are assigned to the corresponding state.
- Following the previous steps and including the Scopus state field, each address now has up to seven candidate states. All cases where a state gets a higher number of assignations than all other possibilities get definitively assigned to that state.
- Cases where there is a tie in the highest number of assignations get assigned by giving priority to the most reliable assignations.

- Ambiguous addresses are fixed manually in decreasing order of frequency.
- Extensive manual coding was performed on the remaining addresses with unknown states.

A general characterization of the precision on a sample of 100 U.S. addresses demonstrates that the global precision in the state assignment process stands at about 99%. This may appear quite high, but because a few states dominate in terms of scientific output in the U.S., there could still be important systematic errors for some states that would result in incorrect analysis if these cases were not detected. To avoid such situations, a stratified sampling approach was performed at the state level. A random sample for each state was manually validated to ensure that each state was properly assigned its rightful scientific output. Corrections were applied to the automated process when errors were detected, yielding precision levels of above 95% for each state individually. Completing this stratified process also ensured that no state was missing a significant portion of its output due to that output being wrongfully assigned to another state.

At the end of the process, the state remained unknown for 3.4% of U.S. addresses.[10] For most of these addresses, there was no information available that enabled coding at the state level.

External File 6: Scopus U.S. addresses to U.S. states

2.2.5 Data coding: U.S. sectors

All U.S. addresses were coded into one of the following sectors: *Academic, Federal Government, State/Local Government, Private Nonprofit, FFRDC,* and *Industry*. The Academic sector was also further divided between *Academic Private*, *Academic Public,* and *Academic Undefined.*

The coding was based on the organization provided in the addresses of authors using the following method:

- A sector conversion table provided by Elsevier was used to make a preliminary decision regarding the author's sector. This table provides a match between a unique organization ID (AFID) for each address and a sector (note that this is not the sector as used in this study, but one based on a specific sector ontology used by Elsevier). There are many inaccuracies in the attribution of AFID to organizations in Scopus, several errors also occur in the coding of AFID to sectors, and many lower-frequency addresses are not classified.
- All the highest frequencies (approx. the first 500 organizations) were verified manually. These 500 organizations accounted for 68% of the U.S. addresses in the database, so a large proportion of the coding was manually validated at this step.

The remaining untested matched sectors and the remaining unknown sectors were validated and/or coded following various approaches that can be synthetized as follows:

[10] A paper may contain more than one U.S. address. In a fictive example, with 10 papers having 10 U.S. addresses each, there are 100 U.S. addresses in total. If the state cannot be determined for 4 of these addresses, then the state remains unknown for 4% of the U.S. addresses.

Table X Coding papers by sector

Elsevier	Final Sector	Note
Academic	Private Academic Public Academic	• Manual validation of automatic coding in "Academic" • Manual coding (e.g., searches for univ*, polytech*) • Use NSF HERD file and then IPEDS and then Carnegie to code between Private/Public Academic • Manual verification of automatic coding between Private/Public (e.g., institution's website and Wikipedia) • Some automatic coding of remaining Academic using keywords (e.g., mainly looking for "state" in the name)
Government	Federal Government State/Local Government	• Manual coding of Federal vs. State & Local, with the help of some filters (e.g., national, federal, U.S., army, navy for the Federal, and state, regional and state/city names for the State/Local)
Other	Private nonprofit	• Manual validation of automatic coding (Elsevier's conversion table) • Use several lists of nonprofit organizations for automatic coding
Corporate	Industry	• Manual validation of automatic coding (Elsevier's conversion table) • Additional coding based on a list of company names • Additional coding with the help of some filters (e.g., Inc., Corp., Ltd.)
Medical	Private Academic Public Academic Federal Government State / Local Government Private nonprofit	• Use Medicare to split between sectors (Industry, Federal, State/Local Gov., Private nonprofit) • Extensive manual validation to identify hospitals that are affiliated with an academic institution, and coding in Private or Public Academic • Additional manual validation and coding of hospitals
[Not used]	FFRDC (Federally Funded Research and Development Center)	SQL queries and manual coding of FFRDCs

Source: Prepared by Science-Metrix

At the end of the process, 93.4% of all U.S. addresses were assigned a sector. The precision of the assignment process reached 98.5% globally. In addition, random samples for each sector were manually validated to ensure that high precision levels were observed across all categories and not only for categories dominating in terms of output (i.e., Academic). For most sectors, precision stood between 96% and 99%. Only two sectors presented lower precision levels: Private Nonprofit (93%) and Academic Undefined (90%).

External File 7: Scopus U.S. addresses to sectors

2.2.6 Open-access types

A new addition in SEI 2024 is the inclusion of statistics about open-access publications. In the last years, in a phenomenon facilitated by the switch from physical publishing to online publishing, open access has grown significantly. In 2003, only about 10% of all research articles were published in open access journals; by 2022, that number had increased to 42%.

Open access, as its name implies, is a publishing model that makes research information available to readers at no cost. This is in contrast to the traditional subscription model, in which readers have access to scholarly information by paying a subscription (usually via their institution). There are many different types of open access, all sharing the same goal of making research more accessible to everyone. For SEI 2024, we defined the following types (or "colors") of access:

- **Gold open access** (also known as "fully open access") is the most common type of open access. In gold open access, the final version of an article is made freely and permanently accessible to everyone, immediately after publication, under an open license. This is done by the publisher, who may charge an article processing charge (APC) to cover the costs of publishing. Cases where an APC is not charged are referred to as "platinum" or "diamond" open access, but given this is still a rare occurrence, no distinction is made from gold open access in this study.
- **Other journal-based open access.** For the purpose of this study, bronze and hybrid open access were merged into a single category. They are described as follows:
    - **Hybrid open access** is a hybrid of gold open access and subscription-based publishing. In hybrid open access, some articles in a journal are made open access, while others remain subscription-based. Authors of articles that are made open access in a hybrid journal may be charged an APC.
    - **Bronze open access** is a type of open access where the final version of an article is made freely and permanently available to everyone, but the article is not licensed under an open license, which means that the article may not be freely reused or redistributed.
- **Green open access** is a self-archiving model of open access. In green open access, the author of an article deposits a version of the article (usually the pre-print or post-print) in an open access repository, such as a university repository or a subject-based repository. The article is then made available to the public, free of charge.

Scopus' open access information is sourced from the Unpaywall database, which allows for a single publication to be coded to multiple open access types. For example, a paper made available both in the arXiv repository and in the journal PLOS One would be coded both to green and gold open access. In an effort to make statistics as meaningful as possible, in such cases, only the first encountered color in the order of the above list will be used for statistics computation. Therefore, in the example case above, the paper would be counted as gold open access only, and all green open access papers were *not* published in an open access journal. Statistics on green open access also do not count articles that are as of yet unpublished (i.e., preprints) as preprints were not retained for use in SEI 2024 (§2.1.2).

2.2.7 AI publications dataset

In SEI 2024, a bibliometric analysis of publications related to artificial intelligence (AI) focused on collaboration patterns is included. This analysis first required that a dataset of these publications was made available. To ensure uniformity with the rest of the bibliometric statistics provided as part of SEI 2024, these publications were identified from the same custom Scopus database used for the other analyses. The previously described paper inclusion criteria (Section 2.1) were therefore still applied to this analysis.

In an effort to make the identification of publications related to AI as reproducible and neutral as possible while also ensuring the comprehensiveness and quality of the resulting dataset, publications were identified by using the All Science Journal Classification (ASJC) scheme.Indeed, all journals and conference proceedings indexed in Scopus are first classified into one or multiple ASJC fields[11] to identify their thematic scope, and Artificial Intelligence has its own field within the ASJC[12]. The ASJC is a journal-level classification scheme defined by a team of experts in which every journal can be assigned one or many subject codes depending on its scope. These codes are self-reported by the journals' publishers when a journal is indexed into Scopus, ensuring by the very nature of the process that a wide variety of views are taken into account to define every subject category. Furthermore, ASJC is a comprehensive classification system: every journal or conference proceeding indexed into Scopus is classified, which ensures that coverage is good. This classification work is conducted when a journal is first indexed into Scopus and is periodically reviewed to ensure that journals that shift, widen, or restrict their thematic scope over time remain accurately classified.

In general, journals and conference proceedings classified in the ASJC Artificial Intelligence field are focused on a subset of AI subfields of research. Their scope can include both theoretical and applied research, as long as an aspect of AI is the main focus. For example, there are journals included in the AI ASJC field focusing on various subfields such as deep learning, knowledge-based computer vision, neural networks, or the integration of AI-driven tools in industrial processes. It should be noted that generally, papers making passing mentions of AI (e.g., an article mentioning the use of an AI-enabled method in a publication otherwise unrelated to AI) will generally be published in non-AI centered journals and will therefore not be included in the current dataset – the journals included here are those where the papers published have Artificial Intelligence research as a main focus.

Given publications are assigned the codes of their source, every peer-reviewed publication considered in this study is also classified under the ASJC. For example, all publications from the journal *IEEE Transactions on Neural Networks and Learning Systems* are assigned to the four ASJC subjects "Computer

[11] For a list of all ASJC Subject Areas and Codes, please refer to https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/.

[12] "Artificial Intelligence" corresponds to ASJC code 1702. It is possible to examine the current list of journals classified within a field by going to https://www.scopus.com/sources and selecting a subject area at the top of the page. A static version of the current Scopus source list can also be downloaded from https://www.elsevier.com/?a=91122.

Networks and Communications", "Computer Science Applications", "Software" and "Artificial Intelligence".

The fact that a single journal can have multiple codes is both an advantage and a drawback when the purpose is to identify relevant literature. It is an advantage as it prevents the classification of interdisciplinary journals into a single one of the relevant fields, therefore increasing the coverage of every field. However, it is also a drawback in that a journal with multiple independent foci is likely to select all of the relevant fields, which is true of the journal but not necessarily of all individual publications, which then risks reducing precision at the paper level. However, a manual investigation of the most publishing journals in the Artificial Intelligence ASJC field has shown that this issue is not as prevalent in the field of AI as in others, with most of the double-coded sources being also focused on related fields such as robotics, automation or computer vision — the risk of including altogether unrelated documents is therefore lower.

Of course, using a journal-based classification scheme to identify papers also has the drawback that it would fail to capture relevant documents published in more general journals (e.g., an AI publication in a general computer science journal).Prior to this study, two analyses were conducted: one used the ASJC-defined publication set, and the other used a publication set produced by the publication-level model used by Elsevier in the production of its AI report[13]. Although the exact sets of publications identified by both approaches differed, the results obtained at the world, national and institutional levels were similar, especially as it came to collaboration. At the institutional level, although the ordering is not exactly the same, 80 of the institutions within the top 100 most producing are the same. At the national level, the same is true for 49 out of the top 50 countries. The collaboration patterns and indicators were also found to lead to the same high-level findings.

In light of the above findings, the decision was then taken to go forward with the approach that could be the most easily reproduced independently in the future, and the ASJC-driven method could be exactly reproduced by anyone with access to Scopus data. To contextualize the findings correctly, care should however be taken to note that the publications used to drive the analysis are those that were published in journals for which Artificial Intelligence was in scope.

2.2.8 U.S. federal agencies mentioned in funding acknowledgements

A new addition in SEI 2024 was the inclusion of data on U.S. papers acknowledging support from U.S. federal agencies. For the purposes of these analyses a "U.S. paper" was defined as being any publication otherwise included in the SEI's database that listed at least one author as being affiliated with at least one U.S. institution.

This subset of papers was then analyzed for the presence of funding information, most of which can be found in the funding acknowledgements section. This section of a peer-reviewed paper is a brief

[13] ArtificiaI Intelligence: How knowledge is created, transferred, and used. Available online : https://www.elsevier.com/research-intelligence/resource-library/ai-report

statement at the end of a paper that acknowledges the support, financial or otherwise, that was received for the research that was conducted. The information that is typically included in the funding acknowledgements section is generally expected to include information such as:

- The name of the funding agency or organization that provided the support.
- The grant number or other identifier that was assigned to the grant.
- A brief description of the type of support that was provided.

Scopus already has text-mining algorithms in place, which are aimed at extracting relevant information from the funding acknowledgements sections of publications, which Science-Metrix knew to be especially good for large-scale funders, and in the U.S. context. As expected, coding precision of a random sample of U.S. papers having received federal funding was remarkably good, with 99% of the coded information being manually confirmed as correct. The sample included a variety of cases, ranging from complete mentions of the funder's name, a grant number, and a description of the support received down to a single mention of the funder using only the acronym, all of which were correctly handled. It would be normal to expect that precision is lower when only the acronyms are used compared with the use of the full name and a grant number, especially for institutes that have the same acronym as another. However, this precision benchmarking exercise has also shown that these very partial occurrences are rare compared to complete mentions, which reduces the risk of misattribution.

This can be helped by the fact that Scopus also uses reliable funding information sources past the funding acknowledgements section, also sourcing funding information directly from funders through the existing reporting platforms of federal agencies. In some instances, even publications with no funding acknowledgements text indexed still have funding information attached because of this.

After preliminary checks, very little room for improvement was identified by Science-Metrix for tagging federally-funded publications. However, some improvements were made to the classification of funders into federal and state level, and to the hierarchization of parent and children entities.

Indeed, funding data is hierarchical, which means that funders can be organized into parent-child relationships (e.g., the Directorate of Engineering is considered to be a "children" of NSF). In Scopus data, funding attribution is always done to the most precise level possible, and the attribution of funded publications to the parent organizations is not done automatically. To address this and to correctly report counts for parent institutions, Science-Metrix built and used an improved version of the hierarchical information available as part of Scopus funding data to roll-up all funded publications for all relevant parents. This way, a publication that acknowledged support from the Directorate of Engineering would also be counted for NSF, and this extends to any length of parenting chain. However, in a similar way as what is the case for regional data, if two or more children of a funder are acknowledged by the same publication, this publication will still only be counted once at the level of any parents these entities have in common.

A final important element that must be kept in mind when analyzing the funding data provided as part of SEI 2024 is that Scopus' coverage of funding information has increased with time: 27.2% of all U.S. publications from 2003 have funding information indexed in Scopus while in 2022, this share rises to

67.7%. Therefore, any annual data showing raw volumes over a long period will be affected by this changing coverage rate. However, this improvement in coverage rates has slowed down in recent years, as from 2018 onwards, it only grew 6%. It then stands to reason that more recent time trends (in the last five years or so) are less likely to be strongly affected by coverage rates. More detailed information on coverage rate was provided as part of the report.

2.3 Production database

Two databases were developed for this project: a basic Scopus database containing all the "original" data from Scopus, with minimal filtering and data transformation, and a production version of the database. The first database contains three tables, one for basic bibliographic information about each article**,** one presenting the information on authors and their addresses, and one presenting the references listed in each article. The production database is leaner as it contains only the necessary information to produce basic bibliometric indicators and is limited to relevant articles and journals. Essentially, the production database was obtained using the following filters:

- Peer-reviewed documents presenting new scientific results were first selected and retained (see Section 2.1.1).
- Only documents for which it was possible to identify the country of at least one author were retained (see Section 2.2.3).
- Only documents classified into one of the 14 TOD fields of research were retained (see Section 2.2.1).
- Documents that were identified as being published in a low-quality journal were removed (see Section 2.1.3).

Table XI presents the number of papers remaining after each step of filtering. Slightly less than 80% of the documents were kept for the analysis, and this was fairly consistent for all years.

Table XI Number of documents after each step of filtering performed by Science-Metrix

Year	All Documents		Peer-reviewed		Country is available		S&E Only		Low-quality removed	
	Papers	%	Papers	%	Papers	%	Papers	%	Papers	%
2003	1,587,005	100%	1,383,280	87%	1,287,973	81%	1,245,402	78%	1,236,106	78%
2004	1,701,217	100%	1,490,456	88%	1,401,298	82%	1,356,199	80%	1,345,967	79%
2005	1,937,149	100%	1,658,538	86%	1,571,903	81%	1,521,595	79%	1,508,837	78%
2006	2,038,354	100%	1,757,195	86%	1,673,828	82%	1,618,695	79%	1,599,532	78%
2007	2,142,260	100%	1,834,152	86%	1,749,013	82%	1,689,663	79%	1,668,894	78%
2008	2,243,305	100%	1,933,714	86%	1,852,589	83%	1,787,769	80%	1,762,986	79%
2009	2,363,493	100%	2,048,314	87%	1,972,475	83%	1,897,011	80%	1,862,709	79%
2010	2,487,546	100%	2,151,119	86%	2,077,707	84%	1,996,330	80%	1,950,178	78%
2011	2,650,257	100%	2,293,334	87%	2,223,540	84%	2,134,536	81%	2,045,351	77%
2012	2,787,630	100%	2,372,155	85%	2,306,820	83%	2,217,526	80%	2,105,157	76%
2013	2,918,579	100%	2,456,908	84%	2,394,284	82%	2,301,428	79%	2,168,061	74%
2014	2,957,371	100%	2,553,590	86%	2,490,447	84%	2,391,584	81%	2,245,240	76%
2015	2,973,724	100%	2,553,879	86%	2,495,876	84%	2,391,958	80%	2,302,230	77%
2016	3,089,547	100%	2,645,233	86%	2,595,816	84%	2,483,753	80%	2,386,418	77%
2017	3,347,922	100%	2,718,394	81%	2,674,374	80%	2,554,466	76%	2,467,960	74%
2018	3,518,423	100%	2,876,544	82%	2,835,235	81%	2,710,074	77%	2,613,116	74%
2019	3,739,769	100%	3,097,568	83%	3,056,065	82%	2,919,455	78%	2,811,110	75%
2020	3,957,841	100%	3,257,995	82%	3,213,014	81%	3,064,254	77%	2,955,562	75%
2021	4,224,732	100%	3,501,971	83%	3,456,303	82%	3,302,824	78%	3,240,279	77%
2022	4,330,312	100%	3,588,344	83%	3,550,187	82%	3,392,000	78%	3,344,037	77%
Total	**56,996,436**	**100%**	**48,172,683**	**85%**	**46,878,747**	**82%**	**44,976,522**	**79%**	**43,619,730**	**77%**

Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

2.3.1 Computation of the citations

The basic Scopus database contains the original printed reference string for every paper, and it also contains this information in a ready-to-use relational list of article identifiers. The schema for this "reference" table is presented in Figure 2. This set of references is smaller than in the original data as it only contains information about references to articles that are also indexed in Scopus.

Once the Scopus database is loaded, a query can be run to pre-compute variables at the article level, based on references. These variables are necessary for computing the bibliometric indicators for the SEI and are presented in Section 2.4.5.

External File 8: Impact indicators NSF production

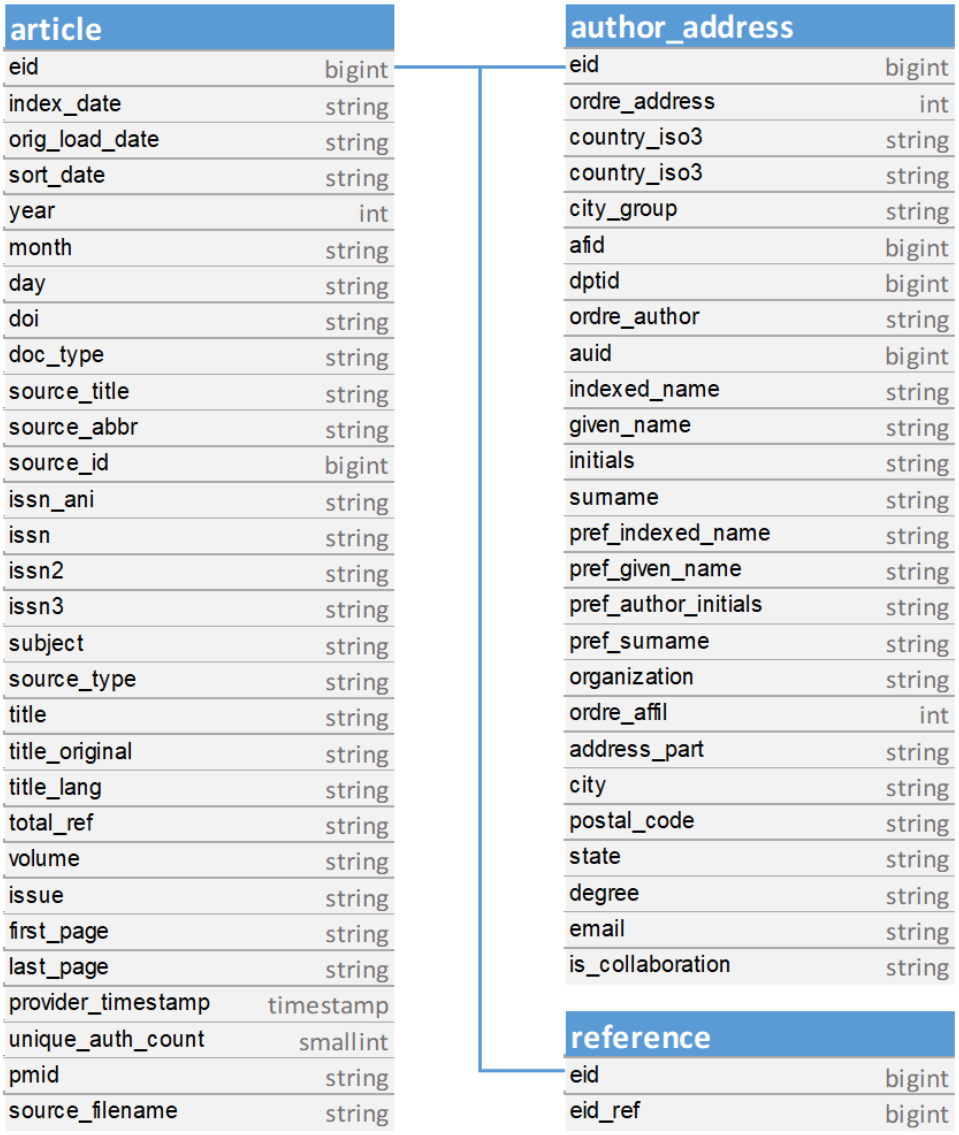


Figure 2 Basic Scopus database schema
Source: Prepared by Science-Metrix

2.3.2 Production database structure

As mentioned, Science-Metrix also computed a production version of the database, which is leaner than the basic Scopus database as it contains only the necessary information to produce bibliometric indicators. The filters described at the beginning of this section were applied to the total database to create the production database, and its structure is as follows.

The table "article" contains all the information at article level that supports the production of bibliometric indicators, including the ID, year of publication, elements of classification (TOD, subfield, and reclassified subfield) and various variables/indicators that were pre-computed. The table "country" presents the standardized country for each address of articles listed in the table "article", based on the

work presented in Section 2.2.3. The table "US_State" contains the standardized state for each U.S. address in the "country" table (country_nsf = "United States"), based on the work described in Section 2.2.4. Finally, the table "US_Sector" contains the results of the coding by sector of U.S. organizations (see Section 2.2.5).

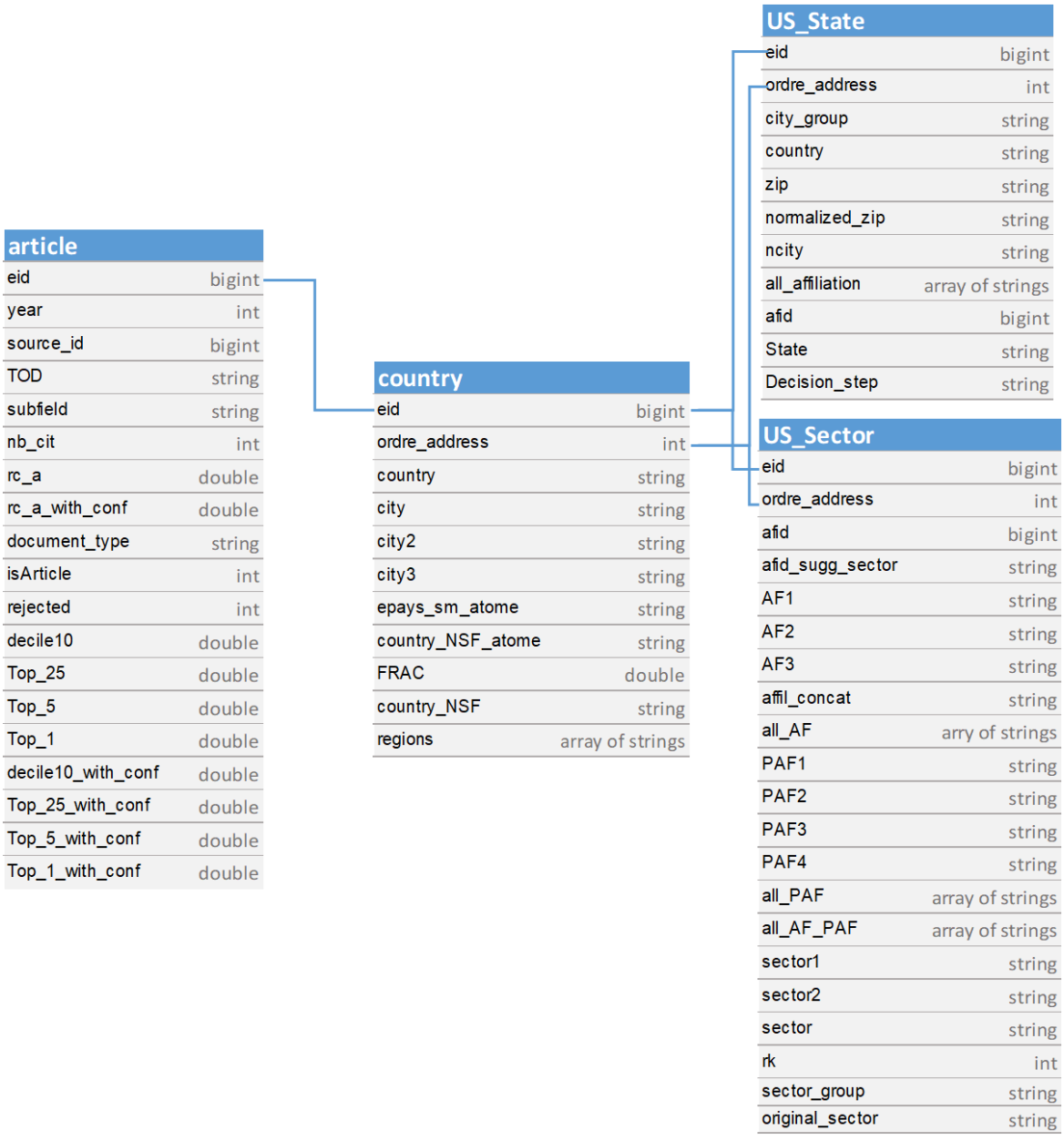


Figure 3 Production database schema
Source: Prepared by Science-Metrix

2.4 Indicators

This section presents the bibliometric indicators computed as part of this study.

2.4.1 Number of publications

The traditional, widespread publication count is one means of measuring and comparing the production of various aggregates (e.g., organizations, regions, and countries). It can also be used to evaluate output in individual disciplines, such as philosophy and economics, and to track trends in research fields, collaborative research, and many other aspects of research output. Several other indicators can also be derived from these simple counts. Full and fractional counting are the two main ways of counting the number of papers.

Full counting

In the full counting method, also called the whole counting method, each paper is counted once for each entity listed in the address field. For example, if a paper was authored by two researchers from the University of Oslo, one from the University College London (UCL) and one from the University of Washington, the paper would be counted once for the University of Oslo, once for UCL, and once for the University of Washington. It would also be counted once for Norway, once for the United Kingdom and once for the United States. When it comes to aggregating groups of institutions (e.g., research consortia) or countries (e.g., the European Union), double counting is avoided. This means that if authors from Croatia and France co-published a paper, this paper would be credited only once when counting papers for the European Union, even though each country had been credited with one publication count.

In cases where a single author has a double affiliation from multiple countries that no other author shares, that paper will still fully count toward the totals of both countries in whole counting; it is essentially the same as if there were multiple authors, each with a single affiliation. In fractional counting, these situations are treated differently.

Fractional counting

Fractional counting is used to ensure that a single paper is not counted several times in calculating totals. This approach avoids summing totals across entities (e.g., researcher, institution, region, country) that add up to more than the total number of papers, as is the case with full counting. Ideally, each author on a paper would be attributed a fraction of the paper that corresponds to his or her level of participation in the study. Since no reliable means exists for calculating the relative effort of authors on a paper, each author is granted the same fraction of the paper.

Using fractional counting on the example presented for full counting (two authors from the University of Oslo, one from UCL, and one from the University of Washington), half of the paper can be attributed to Norway and one-quarter each to the United Kingdom and the United States. Using the same approach for institutions, half of the paper would be counted for the University of Oslo and one-quarter each would be attributed to UCL and the University of Washington. For this study, fractions were calculated at the level of researchers.

As a final example, consider the more complex case presented by the fictitious paper described in Table XII, for which one author has signed using multiple affiliations from different countries.

Table XII Example of the article fractioning procedure when authors have multiple affiliations

Author	Affiliation	Country	Fraction
1	Harvard University	United States	1/5
2	Harvard University	United States	1/5
3	Princeton University	United States	1/5
4	University of Toronto	Canada	1/10
	Sapienza Università di Roma	Italy	1/10
5	Oxford University	United Kingdom	1/5

Note: Each author number represents a distinct author. The fourth author has a double affiliation, and both affiliations are in different countries. The fraction column shows what fraction of the paper is attributed to this specific author—affiliation combination. Fractions are computed such that each individual author gets an equal share of the article.
Source: Prepared by Science-Metrix

As presented, the article has five distinct authors. Each author is therefore attributed one-fifth of the article. The fourth author, however, has two distinct affiliations. This means that the fraction attributed to the fourth author is further split up across these two affiliations, such that each of these affiliations is attributed a tenth of the article. When aggregating at country level, the United States would get three-fifths (60%) of the paper, whereas the United Kingdom gets one-fifth (20%), Canada gets a tenth (10%) and Italy also gets a tenth, which sums to 1. In full counting, every country would instead be attributed the full credit from the article.

2.4.2 Collaboration

In the context of bibliometrics, scientific collaboration is measured by co-publications. A co-publication is defined as a publication that was co-authored by at least two authors. When a publication involves only authors from one country, it is defined as a national collaboration. When at least two different countries are identified among the addresses of authors on the publication, it is defined as an international collaboration. A publication can involve national and international partnerships simultaneously if more than two countries are involved and at least one of the countries is represented by more than one author on the publication. In some tables, the statistics have been presented for different types of co-authorship:

- **With multiple institutions:** Articles with two or more institutional addresses.
- **With domestic institutions only:** Articles with one or more institutional addresses all within a single country/economy.
- **With international institutions:** Articles with institutional addresses from more than one country/economy.

In a perfect scenario where metadata regarding all addresses in Scopus are fully available, the sum of publications falling under categories "With domestic institutions only" and "With international

institutions" should add up to the total number of publications in the database, as both categories are complementary. However, because a small fraction of addresses falls under the "unassigned" category[14] due to missing country affiliations, the sum of both categories is instead slightly lower than the total number of publications. This is because it is still possible to classify a paper under the "With international institutions" category even if there are unassigned country affiliations in its address field as there only needs to be two known distinct countries to classify a paper under this category (e.g., one Canadian address, one U.S. address and one unknown address). The same does not apply for the "With domestic institutions only" category. Publications with an address from a single country accompanied by unassigned addresses cannot be assigned to the "With domestic institutions only" category because the unassigned addresses may be from a different country, which would place these publications in the "With international institutions" category instead. Therefore, the collaboration status of these publications is unknown, and the total count of publications is the sum of the "With domestic institutions only" category, the "With international institutions" category and this third category of unknown collaboration status (data are not presented for this third category).

As a concrete example of this, 1,938,121 publications were published at the world level for year 2010, of which 1,576,003 fell under the "With domestic institutions only" category and 342,607 were classified under the "With international institutions" category. Summing both categories reveals a difference of 19,511 missing publications compared to the world total: these are the papers of unknown collaboration status.

2.4.3 Collaboration rates

Collaboration rates presented in the SEI are ratios of counts, using the numbers of co-authored publications as numerators and the total counts as denominators. These ratios can then be compared across countries to assess differences in collaboration patterns. Note that while cross-country comparisons are relevant, comparing data from different geographical aggregation categories (e.g., countries with the world level, countries with regions) should not be done because of the multilateral nature of co-publications (i.e., multi-country publications are counted only once at the world level, but multiple times across countries as they get counted once per country).

2.4.4 Index of collaboration

The index of collaboration, also named probabilistic affinity index (PAI) in the literature, provides an indication of the preference of two countries to collaborate. It compares the number of papers co-authored between the two countries with the number of co-authored articles that would have resulted from a uniform distribution of partnering countries. The index is based on full counts of papers and is calculated as follows:

[14] These are displayed as the last entries, labeled "Unassigned", in most tables presenting country and regional data.

$$C_{xy} = \frac{n_{xy}N}{n_x n_y}$$

where

C_{xy}	Index of collaboration between country x and country y
n_{xy}	Number of papers co-authored by country x and country y
n_x	Total number of internationally collaborative papers involving country x
n_y	Total number of internationally collaborative papers involving country y
N	Total number of internationally collaborative papers in the database

This index is symmetrical, such that $C_{xy} = C_{yx}$. Two countries that share an index of collaboration higher than 1.0 are collaborating more than would be expected to happen by chance, assuming a uniform distribution of collaboration links.

Essentially, the index of collaboration takes the share of a country's internationally collaborative papers cowritten with another specific country $\left(\frac{n_{xy}}{n_x}\right)$ and divides it by the share of the world's collaborative papers in which this other country was involved $\left(\frac{n_y}{N}\right)$. If the first share is higher than the second, then the index of collaboration is higher than 1 and both countries collaborate more than was expected.

2.4.5 Scientific impact analysis—citations

An important part of scientific excellence is gaining recognition from colleagues for one's scientific accomplishments. Although this recognition can be expressed in many ways, references to scientific publications are often considered to be explicit acknowledgments of an intellectual contribution. As a result, it is considered that the more a scientific article or publication is cited, the greater its impact on the scientific community, and the more likely it is to be a work of great quality. This is the basic assumption that underlines the various indicators grouped in this section (i.e., citation counts and the various ways to normalize them).

Before going into detail about specific indicators, it is important to highlight several issues related to the act of citing itself. One source of contention concerns what exactly is being measured through citation analysis. References are the practice of acknowledging previous work that has been important in the production of the referencing article. However, motivations for citing can be unclear, and not all of them are linked to the quality of the work in the cited article. This can undermine the idea that papers are cited because they are of high quality and make an important contribution to science. Critics have thus questioned the validity of citations as measures of research visibility, impact, or scientific quality,[15,16] but these measures remain widely used because few alternatives exist that would be more objective and cost-effective. When the law of large numbers is maintained and studies are correctly designed, the

[15] Tijssen, R. J. W., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381–397.

[16] Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics, 50*(3), 455–482.

idiosyncratic uses of citations are largely mitigated, and citations can therefore be used with a high level of confidence.

Citation count

The number of citations received by a scientific article or publication is considered a measure of the impact of that contribution on the scientific community: the higher the number of citations, the greater the scientific impact. The number of citations can be aggregated to establish citation counts for an individual scientist, a research group, a department, an institution, or a country.

A number of problems can be associated with absolute citation counts, notably since citation practices differ between subfields of science, such as physical chemistry and colloidal chemistry,[17,18] and citations accrue at different rates depending on the field or even the document type (e.g., article vs. conference paper). Citation counts are also affected by the period over which they are counted, and the importance of this factor has been characterized by a number of authors.[19,20,21]

Absolute citation counts are a very imprecise way to benchmark scientific performance, as some of the above critiques demonstrate.

Relative citation scores

A high-quality paper in a field where fewer citations are given could receive fewer citations than an average-quality paper in a field with heavy citing practices. It would not be rigorous to compare these papers on absolute terms. A number of indicators have been developed to take these field specificities into account. They are called relative citation measures and are based on relative citation (RC) scores, as detailed below.

One way to increase the fitness of citation counts using RC scores is to calculate them relative to the size of the publication pool analyzed, or better, to the citation performance expected for the scientific field or subfield. In the first instance, the number of citations accrued by an individual scientist, an institution, or a country for a specific set of articles is divided by the number of articles in that set. The assumption here is that the number of citations received by the individual, institution, or country is closely linked to the number of articles published. To further increase the fitness of the citation analysis, the results of this citation-per-publication ratio can be compared to an expected citation rate, which is the citation-per-publication ratio of all articles in the journal or the subfield where the research unit publishes. This additional sophistication is based on the assumption that practices in different scientific subfields have

[17] Braun, T. (2003). The reliability of total citation rankings. *Journal of Chemical Information and Computer Sciences, 43*(1), 45–46.

[18] Frandsen, T. F. (2005). Journal interaction. A bibliometric analysis of economics journals. *Journal of Documentation, 61*(3), 385–401.

[19] Frandsen, T. F., & Rousseau R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology, 56*(1), 58–62.

[20] Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. J. F. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy, 14*(3), 131–149.

[21] Van Raan, A. J. F. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Technikfolgenabschätzung, 12*(1), 20–29.

an impact on the citations normally received in that field, and that comparison of the unmodified citation-to-publication ratio between different fields is not rigorous.

The RC score computed by Science-Metrix is a normalization of the relative scientific impact of papers produced by a given entity (e.g., a country, an institution) that takes into consideration the fact that citation behavior varies between fields and years of publication. For a paper in a given subfield (based on the classification of journals described previously in this section) and publication year, the citation count is then divided by the average count of all papers in the relevant subfield (e.g., astronomy & astrophysics) and publication year to obtain an RC. When the RC is above 1, a paper scores better than the average paper; when it is below 1, it is not cited as often as the average paper. The relative citation score of a given paper i (RC_i) is calculated as follows:

$$\mathrm{RC}_i = \begin{cases} \frac{c_{s,y,i}}{\bar{c}_{s,y}} & \text{if } n_{s,y} \geq 30 \text{ and } y \leq Y-2 \\ \text{undefined} & \text{otherwise} \end{cases}$$

Where

RC_i	Relative citation score for paper i
$c_{s,y,i}$	Number of citations received by paper i, which belongs to subfield s, and was published in publication year y
$\bar{c}_{s,y}$	Average number of citations received by papers belonging to subfield s and published in publication year y
$n_{s,y}$	Total number of publications belonging to subfield s and published in publication year y
Y	The last year in the database with complete coverage

In SEI 2024, RC scores are not computed for conference articles except where explicitly indicated. The scores are used to compute the average of relative citations (ARC) and the highly cited articles (HCA) indicators. These are detailed below.

Average of relative citations

The ARC is the average of the RC scores of all the articles published by a given entity. The full-counting ARC for a given entity is computed as follows:

$$\mathrm{ARC}_S = \begin{cases} \frac{\sum_{i \in S} \mathrm{RC}_i}{n_{\mathrm{RC},S}} & \text{if } n_{\mathrm{RC}} \geq 30 \\ \text{undefined} & \text{otherwise} \end{cases}$$

Where

ARC_S	Average of relative citations for the set of publications S
RC_i	Relative citation score for paper i
$n_{\mathrm{RC},S}$	Number of publications in set S with a RC score

The ARC is normalized to 1, meaning that an ARC above 1 indicates that the entity's articles have higher-than-average impact, an ARC below 1 means that the entity's articles have lower-than-average impact, and an ARC near 1 means that the publications have near-average impact.

Because RC scores are known to be skewed in their distribution—with a small number of papers receiving a large share of the total citations—the ARC offers a useful snapshot of overall performance but can hide important underlying nuance. For this reason, Science-Metrix proposes to complement the ARC with the HCA described below.

Highly cited articles and citation percentiles

To compute the proportion of articles of an entity that are in the top x% most cited articles, the top x% most cited articles at the world level must first be determined. To account for the variations in citation behavior between the disciplines and over time, the top x% for the whole database is composed of the top x% for each discipline for each given year, in accordance with the RC scores presented above. However, as the highly cited articles are identified on a per-field, per-year basis, the same indicator could be computed using the raw citation counts instead, as long as the articles that are not attributed an RC score are also given an HCA score. These are the articles where either the combination of subfield and year of publication creates a group containing fewer than 30 publications, or the articles have been published after 2018 (2020 being the last complete year, –2 years to allow for a long enough citation window). Furthermore, conference proceedings are excluded from this analysis in SEI2022, except where explicitly specified. As a result, the average of citation counts for a subfield and year effectively plays no role in the computation of this indicator.

Because some publications are tied based on their citation score, to include all publications in the database that have a citation score equal to or greater than the x% threshold would often lead to the inclusion of slightly more than x% of the database. To ensure that the proportion of publications in the x% most cited publications in the database is exactly equal to x% of the database, publications tied at the threshold citation score are each given a fraction of the number of remaining places within the top x%.

For example, if a database contains 100 publications, then the top 10% should contain exactly 10 publications. Ranked in descending order of their citation score, if the 9th, 10th, 11th, and 12th publications all have the same score, they are each given a quarter of the remaining two places in the top 10% ($2\times\frac{1}{4}=\frac{1}{2}$ article of the top 10% each). This situation, using fictitious numbers, is summarized visually in Table XIII.

Table XIII Example of the fractioning procedure used to compute the HCA10% scores at article level

Article number	Number of citations	Rank	HCA10%	Number of articles in top 10%	Comment
1	100	1	1	1	
2	97	2	1	2	
3	81	3	1	3	
4	79	4	1	4	These articles are tied, but both can fit within the top 10 articles, so both get a score of 1
5	79	4	1	5	
6	65	6	1	6	
7	64	7	1	7	
8	63	8	1	8	
9	60	9	1/2	8.5	These articles are tied, but there are 4 of them and only 2 places left to fill in the top 10. They share the final 2 places, each getting a score of 2/4 = 1/2
10	60	9	1/2	9	
11	60	9	1/2	9.5	
12	60	9	1/2	10	
13	55	12	0	10	All other articles starting with this one are not in the top 10%.

Note: These data are based on a fictive case of 100 articles, such that the top 10% should contain 10 articles. Each row represents a single article. The ranks are attributed based on the number of citations. The "number of articles in top 10%" column counts the effective number of articles (based on HCA scores) already included in the top 10%, current row included.
Source: Prepared by Science-Metrix

In addition, in some cases the number of places in the top 10% most cited publications is not an integer (e.g., if there are 11 publications in the database, there should be $\frac{11}{10}$ publications in the top 10%). For example, if there are no ties in the citation score of papers at the threshold, the paper with the highest score is given a count of 1 and the second paper is given a count of $\frac{1}{10}$. In cases where there are also ties at the threshold, there is dual fractioning. For example, in the previous case, if three papers were tied in second place behind the first paper, they are each given a weight of $\frac{1}{30}$ (i.e., $\frac{1}{10}\times\frac{1}{3}$). Likewise, if the top two papers are tied, they are each given a count of $\frac{1}{2}$ (i.e., $1 \times \frac{1}{2}$).

Following this process, the proportion of papers of a given entity that are in the world's top x% most cited papers can be computed. An entity with x% of its papers in the top x% most cited papers is considered to be on a par with the world level. Both full and fractional counting of publications can be used. In fractional counting, there could thus be a triple fractionation of scores (i.e., in the case where there is a tie on the citation score at the threshold, and the x% is not an integer, and the paper is co-authored).

2.4.6 Relative citation index

The relative citation index (RCI) is used to characterize citation affinity between countries, accounting for the scientific output of the citing and cited countries. It is similar to the index of collaboration (Section 2.4.4), but it is computed from citation data instead of collaboration data. Mathematically, it is defined as:

$$\mathrm{RCI}_{xy} = \frac{c_{xy}N}{c_x n_y}$$

where

RCI_{xy}	Relative citation index between country x and country y
c_{xy}	Fractional number of citations by country x's papers to country y's papers
c_x	Total fractional number of citations made by country x's papers
n_y	Fractional number of papers authored by country y
N	Total number of papers in the database

For this indicator, data are presented according to the publication year of the cited papers, and citations are counted using a fixed citation window of three years (i.e., citations from a paper published in publication year up until publication year +2 years). For example, the RCI for 2018 is computed using cited papers published in 2018, and only citations coming from papers published in 2018, 2019, or 2020 are counted. Any citation received after this threshold is not counted. Contrary to the index of collaboration, the RCI is not symmetrical, such that, generally, $\mathrm{RCI}_{xy} \neq \mathrm{RCI}_{yx}$.

Normalizing citation counts by a country's publication output, as is done in the RCI, is essential for correct interpretation of the data. The expected share of citations that one country receives from another depends on the number of articles that the cited country produces. For instance, if the United States had authored about 22% of all 2018 papers, it would be assumed that, all things being equal, U.S. publications should account for about 22% of each country's citations to 2018 papers for the pool of articles covered during this period. The countries with a higher than 22% proportion of their outgoing citations would be showing a preference for citing the United States, and those below this level would be citing the country less frequently than expected. Dividing the share of a country's references to U.S. articles by the expected share given the size of output of the United States in 2018 results in a relative citation index. For instance, if 25% of China's outgoing citations to publications published in 2018 are to U.S. publications, and the United States published 22% of all articles released in 2018, China's RCI toward the United States would stand at 25%/22%=1.14.

To correctly account for all citations between country pairs, all citations must be fractioned across both citing and cited articles using double fractioning counts—that is, by fractioning citing papers and their corresponding cited papers at the same time at the author level. Table XIV presents this fractioning procedure for two hypothetical articles. The citing article was written by two authors from two distinct countries (such that there are two affiliations), and the cited article was written by four authors from four distinct countries, with one author having a double affiliation (such that there are five affiliations). This results in 10 distinct citation links, as shown in the table. When all citation fractions are summed, the expected result of one citation going from the citing article to the cited article is found.

Table XIV Citation counts between country pairs for a pair of citing–cited articles

Citation link	Citing article				Cited article				Citation fraction
	Author id	Affiliation id	Country	Article fraction	Author id	Affiliation id	Country	Article fraction	
1	1	1	United States	1/2	1	1	United States	1/4	1/8
2	1	1	United States	1/2	2	2	United States	1/4	1/8
3	1	1	United States	1/2	3	3	France	1/4	1/8
4	1	1	United States	1/2	4	4	United Kingdom	1/8	1/16
5	1	1	United States	1/2	4	5	Australia	1/8	1/16
6	2	2	Canada	1/2	1	1	United States	1/4	1/8
7	2	2	Canada	1/2	2	2	United States	1/4	1/8
8	2	2	Canada	1/2	3	3	France	1/4	1/8
9	2	2	Canada	1/2	4	4	United Kingdom	1/8	1/16
10	2	2	Canada	1/2	4	5	Australia	1/8	1/16

Note: These data are based on a fictive case of a citing–cited pair of articles. Each row represents one citation link from one affiliation to the next. Fractions are determined by attributing equal weight to each author and then refractioning for each of the authors' affiliations if required.
Source: Prepared by Science-Metrix

From this, it is possible to compute the fractional citation counts at the country level. Table XV shows what would be obtained from the data presented in Table XIV.

Table XV Aggregated citation counts between country pairs for a pair of citing–cited articles

Cited country	Citing country	Fraction of citation
United States	United States	1/4
	Canada	1/4
	Total	**1/2**
France	United States	1/8
	Canada	1/8
	Total	**1/4**
United Kingdom	United States	1/16
	Canada	1/16
	Total	**1/8**
Australia	United States	1/16
	Canada	1/16
	Total	**1/8**

Note: The calculation refers to the case presented at Table XIV. These data are based on a fictive case of a citing–cited pair of articles.
Source: Prepared by Science-Metrix

Note that both citing and cited papers must be peer-reviewed documents to be included in this analysis. Fractioning both citing and cited papers and looking at all the possible combinations in the database results in billions of pairs. In the end, the number of citations made from one country to another is simply the sum of the fractioned scores associated with each pair, with the sum across all possible pairs adding up to the total number of citations made at the world level.

The share of citations of a country at the end of the process is simply the sum of citations received from all countries divided by the total number of citations at the world level. In the cases presented in Table XV, the United States' share of the citations stands at 50%, that of France at 25%, and those of the United Kingdom and Australia at 12.5% each, which adds up to 100%.

2.4.7 Network indicators

Network indicators were used in SEI 2024 to analyze collaboration related to artificial intelligence research. In this situation, network analysis is a method of analyzing scientific collaboration by looking at the patterns of co-authorship among researchers between pairs of entities. In the context of SEI 2024, "entities" could refer to either the authors' institutional affiliations or to the countries of such affiliations.

In a simple co-publication network, each node represents an entity, and each edge represents the number of publications that both of these entities have co-written, using whole counting. In this approach, the exact number of authors from a given entity on any given publication has no impact: an article co-authored by 4 researchers from an U.S. institution and 2 researchers from a Canadian institution is still counted as a single collaboration between both countries.

At the national level, a variant of the co-publication network normalized for size effects was also prepared. This network has the same basic structure, but instead of using the raw number of co-publications between countries as the weight of edges, the index of collaboration (also called the probabilistic affinity index in literature) was used. The index of collaboration essentially compares the number of papers co-authored by an entity pair with the number of co-authored articles that would have resulted from a random selection of partnering countries, considering the production size of countries. The index is based on full counts of co-authored papers and is calculated as follows:

$$\mathrm{IC}_{xy} = \frac{C_{xy}}{C_w}\left(\frac{C_x}{C_w}\frac{C_y}{C_w}\right)^{-1} = \frac{C_{xy}C_w}{C_xC_y}$$

where

IC_{xy}	Index of collaboration for the xy country pair
C_{xy}	Number of papers co-authored by the xy country pair
C_x	Total number of internationally co-authored papers by country x
C_y	Total number of internationally co-authored papers by country y
C_w	Total number of internationally co-authored papers in the dataset

An index of collaboration of 1 signifies that countries from the xy dyad have co-authored exactly as many papers as they were expected to, given their respective proportions of internationally co-written papers. An index of collaboration higher (lower) than 1 signifies that countries have collaborated more (less) than expected, meaning they have (a lack of) collaboration affinity. This index is defined in a symmetric manner, meaning that $IC_{xy} = IC_{yx}$, or for example that the U.S. has the same affinity with Canada than Canada has with the U.S.

For both edge weights above, multiple indicators, were presented. They are each described in detail in this section. All of the following indicators were all computed using the open source iGraph software package.[22]

Degree centrality

The degree of a node is the number of edges connected to the node. In the context of an international collaboration network, this corresponds to the number of other countries with which the country has collaborated. In this case, the maximum value of this indicator is the number of nodes in the network minus 1, as the node that has its degree computed cannot have collaborated with itself.

Strength

The node strength is the sum of the weights of edges connected to the node. For international collaboration, a single paper can generate multiple collaboration links. For example, if one author from the United States co-wrote an article with two authors from France and one author from Canada, this article generates three collaboration links: U.S.–France, U.S.–Canada, and Canada–France, each with a weight of 1, regardless of the number of authors.

Betweenness centrality

Betweenness centrality measures how often a given node in a network lies along the shortest paths between two other nodes that are not directly connected to one another. For example, this indicator would highlight entities that play an important "brokering" role, acting as a connecting link between entities that do not co-publish with one another directly. Nodes with a high betweenness centrality score are the bridges that connect relatively isolated islands of research communities within the overall topography. These entities play an important role in the interconnection of subgroups within the network as a whole.

Closeness centrality

Closeness centrality assesses the degrees of separation between one node and other nodes within a network. That is, it assesses the length of the chains that connect a given node to the rest of its community. Whereas, for example, betweenness centrality highlights entities that play an interconnecting role for their community, closeness centrality measures the level of access that a given entity has to its surrounding community. It highlights those who can tap into a large section of a network without passing through many degrees of separation, or through distant and mediated connections. When calculating closeness centrality, a node directly connected to every other node in the network would score 1, the highest possible closeness centrality score.

Weighted eigenvector centrality

Weighted eigenvector centrality is a measure of the level of integration of a node in a collaboration network. The level of integration of nodes within a collaboration network is reflected by the number of

[22] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, *Complex Systems*, 1695. https://igraph.org.

nodes to which they are connected and the quality of their collaborations (that is, the strength of the ties measured by the number of co-authored publications and the importance of the nodes to which they are connected in the network). The mathematical definition of eigenvector centrality is such that the centrality score of a node in a network is proportional to the sum of the centrality scores of all nodes connected to it. Thus, this indicator offers a good appreciation of both the number and quality of an entity's collaborations because connections to high-scoring nodes contribute more to the score of that entity than equal connections to low-scoring nodes. A node scoring high for this indicator operates closer to the core of the network than a low-scoring node. High-scoring nodes are central and highly important to the network's structure. Eigenvector centrality provides a good appreciation of the integration of individual entities within a network—that is, the higher the score, the more integrated the entity. The weighted version of the indicator accounts for the size of the tie between nodes. Centrality scores are typically normalized between 1 (most central node) and 0 (least central node).

Weighted PageRank

PageRank, made famous by its use by the Google search engine, is a variant of eigenvector centrality. It can be thought of as the result of a random walk, meaning that the PageRank score of a given node corresponds to the probability that someone starting on a random node of the network and randomly following edges will end the walk on a particular node. The weighted version of the algorithm makes stronger links more likely to be followed than weaker links. PageRanks are shown as a percentage to clearly indicate the share of random walks where the end point was the given node. All scores sum to 100%. In undirected networks, weighted PageRank yields results very similar to node strength.

2.4.8 Network visualization

Network visualization is a way of representing a network of nodes and links in a visual way. It is a powerful tool for understanding the structure of networks, and it can be invaluable in identifying patterns and trends that would be difficult to see otherwise. However, care must be taken to interpret the visuals only as helpful representations of the data: many graph layout methods and techniques exist, and each will generally produce a different result. This is due to the fact that a network's node doesn't have a single "correct" position in 2D space, and by extension edges don't have a single correct "length".

For SEI 2024, all network visualizations were done by using force-driven positioning. The purpose of such algorithms is to position the nodes of a graph in 2D space by assigning forces to each node in a network. These forces can be attractive (e.g., between two nodes connected together by an edge) or repulsive (e.g., between nodes not connected by an edge), and they can be based on a variety of factors such as the weight of edges between nodes, the number of connections between nodes, the total number of nodes on the graph, etc. The nodes are then moved according to these forces, and the process is repeated until a stable configuration is reached.

The algorithms used here (either Fruchterman-Reingold or GraphOPT as they are implemented by iGraph) will generally provide stable results given stable input data. However, it is possible that small changes are introduced through time if the package's implementation of the algorithms changes. The graphs' defining features should however generally be the same: if a set of nodes forms a clear community

in one visualization, then this community will be visible in all visualizations of the graph that do not vary the inputs and are using the same algorithm. However, the graph could be rotated, mirrored, or some nodes might have slightly different distances between them.