Science-Metrix

Bibliometrics Indicators for the Science and Engineering Indicators 2020

Technical Documentation

December 2019

Science-Metrix

Bibliometrics Indicators for the Science and Engineering Indicators 2020

Technical Documentation

November 20, 2019

Submitted to:
SRI International

Authors
Grégoire Côté
Guillaume Roberge
Matt Durning
Alexandre Bédard-Vallée
Maxime Rivest

Project Leader
Grégoire Côté

By:

Science-Metrix
1.514.495.6505 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com

Contents

Tables

Figures

1 Introduction

Science-Metrix 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) 2020. 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: Postgresql scripts
External File 2: Scopus canceled title list
External File 3: DOAJ canceled title list
External File 4: Article id to TOD and subfields
External File 5: Scopus country
External File 6: Scopus US addresses to U.S. states
External File 7: Scopus U.S. sectors
External File 8: Impact 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 13,000 peer-reviewed journals, and Scopus, which is produced by Elsevier and covers about 23,000 peer-reviewed journals.

The bibliometric indicators for SEI 2020 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 for SEI 2016 and SEI 2018.

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 (mostly 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*.

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

For the SEI 2018 edition, Science-Metrix received a hard drive containing the Scopus data. For this edition, Scopus data were instead 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. These tables were then exported as tab-separated flat text files. The resulting text files were UTF8 encoded and could therefore have contained non-Latin characters, which was not desirable. Furthermore, the next analysis step was performed using a Microsoft SQLServer database, which does not easily support UTF8 encoding. Text files were therefore transliterated in the CP1252 encoding using the iconv software under Linux.

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

Column	Data type	XPATH
id	Varchar(50)	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid attr=SGR
index_date	varchar(50)	/xocs:doc/xocs:meta/xocs:indexeddate
orig_load_date	varchar(50)	/xocs:doc/xocs:meta/xocs:orig-load-date
sort_date	varchar(50)	/xocs:doc/xocs:meta/xocs:datesort
year	varchar(8)	/xocs:doc/xocs:meta/xocs:sort-year
month	varchar(4)	derived from sort_date
day	varchar(4)	derived from sort_date
doi	varchar(120)	/xocs:doc/xocs:meta/xocs:doi
doc_type	varchar(10)	/xocs:doc/xocs:meta/cto:doctype
source_title	varchar(500)	/xocs:doc/xocs:item/item/bibrecord/head/source/sourcetitle
source_abbr	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/source/sourcetitle-abbrev
source_id	varchar(20)	/xocs:doc/xocs:item/item/bibrecord/head/source attr=SRCID
issn	varchar(50)	/xocs:doc/xocs:item/item/bibrecord/head/source/issn
issn2	varchar(50)	/xocs:doc/xocs:item/item/bibrecord/head/source/issn
subject	varchar(400)	/xocs:doc/xocs:meta/xocs:subjareas/xocs:subjarea
source_type	varchar(2)	/xocs:doc/xocs:item/item/bibrecord/head/source attr=TYPE
title	varchar(1000)	/xocs:doc/xocs:item/item/bibrecord/head/citation-title/titletext
title_lang	varchar(20)	/xocs:doc/xocs:item/item/bibrecord/head/citation-title attr=@language
total_ref	Int	/xocs:doc/xocs:item/item/bibrecord/tail/bibliography attr=refcount
volume	varchar(200)	/xocs:doc/xocs:meta/xocs:volume
issue	varchar(200)	/xocs:doc/xocs:meta/xocs:issue
first_page	varchar(200)	/xocs:doc/xocs:meta/xocs:firstpage
last_page	varchar(200)	/xocs:doc/xocs:meta/xocs:lastpage
id_permanent	bigint	Automatically generated by extraction script
Provider_timestamp	varchar(1000)	/xocs:doc/xocs:meta/xocs:timestamp
unique_auth_count	int	/xocs:doc/xocs:meta/cto: unique-auth-count
pmid	bigint	/xocs:doc/xocs:meta/xocs:pmid
filename	varchar(200)	Automatically generated by extraction script

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

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

Column	Data type	XPATH
id	Varchar(22)	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid attr=SGR
Id_permanent	bigint	Created by PostgreSQL to dissociate the various versions of the same papers
Provider_timestamp	varchar(1000)	/xocs:doc/xocs:meta/xocs:timestamp
ordre_address	int	Automatically incremented by extraction script
Country_ISO3	varchar(6)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation attr=COUNTRY
country	varchar(250)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/country
City_group	varchar(300)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/city-group
afid	varchar(18)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation attr=AFID
dptid	varchar(18)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation attr=DPTID
ordre_author	int	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author attr=seq
auid	varchar(22)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author attr=AUID
indexed_name	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:indexed-name
given_name	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:given-name
author_initials	varchar(20)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:initials
surname	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:surname
pref_indexed_name	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:indexed-name
pref_given_name	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:given-name
pref_author_initials	varchar(20)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:initials
pref_surname	varchar(400)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/preferred-name/ce:surname
ordre_affil	int	Automatically incremented by extraction script
affiliation	varchar(800)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/ce:text OR /xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/organization/
address_part	varchar(300)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/address-part
city	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/city
postal_code	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/postal-code
state	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/affiliation/state
degree	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:degree
email	varchar(120)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/ce:e-address attr=EMAIL
orcid	varchar(200)	/xocs:doc/xocs:item/item/bibrecord/head/author-group/author/orchid

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

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

Column	Data type	XPATH
id	Varchar(50)	/xocs:doc/xocs:item/item/bibrecord/item-info/itemidlist/itemid attr=SGR
id_ref	varchar(22)	/xocs:doc/xocs:meta/cto:ref-id

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

External File 1: Postgresql scripts

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 December 31 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.[1]

Based on Elsevier's modeling, a completion rate of at least 95% is expected within 18 months of the first day of a publication year. In our case, this meant a 95% completion rate on 1 July 2019 at the latest for the publication year 2018. 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 18 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 2018, 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 (CAGR) for 2015 to 2017, the last three completed years, was first calculated using a simple regression model. This growth rate was then used for every year, starting in 2015.

Details about the complete set of filters applied to the Scopus database are available at sections 2.1.2 and 2.1.3, but because some of these tests cannot be performed before the whole Scopus database is built on the Science-Metrix server, a stripped-down version of the data filtering was used to have a quickly computable and easily reproducible model which did not require the full preparation of the complete database given that the model was to be evaluated several times by Science-Metrix.[2] The only filters used

[1] 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.

[2] Elsevier performed similar tests to these presented in this report for the SEI 2018 after adopting Science-Metrix filtering approach, so our assumption is that the same tests were performed this time again for the SEI 2020 prior to Science-Metrix validating that completion had indeed reached 95% as we were not informed of any change to the method.

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 month. The result for the year 2018 as measured between April and July 2019 is presented in Table IV.

Table IV Monthly follow-up of the completion rate for the year 2018

Month	Observed (papers)	Predicted (papers)	Completion (%)
April	2,656,717	2,784,781	95.4%
May	2,686,708	2,790,230	96.3%
June	2,706,610	2,795,110	96.8%
July	2,707,198	2,797,511	96.8%

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

The full prediction model calculated on 1 June 2019 is presented in Figure 2.

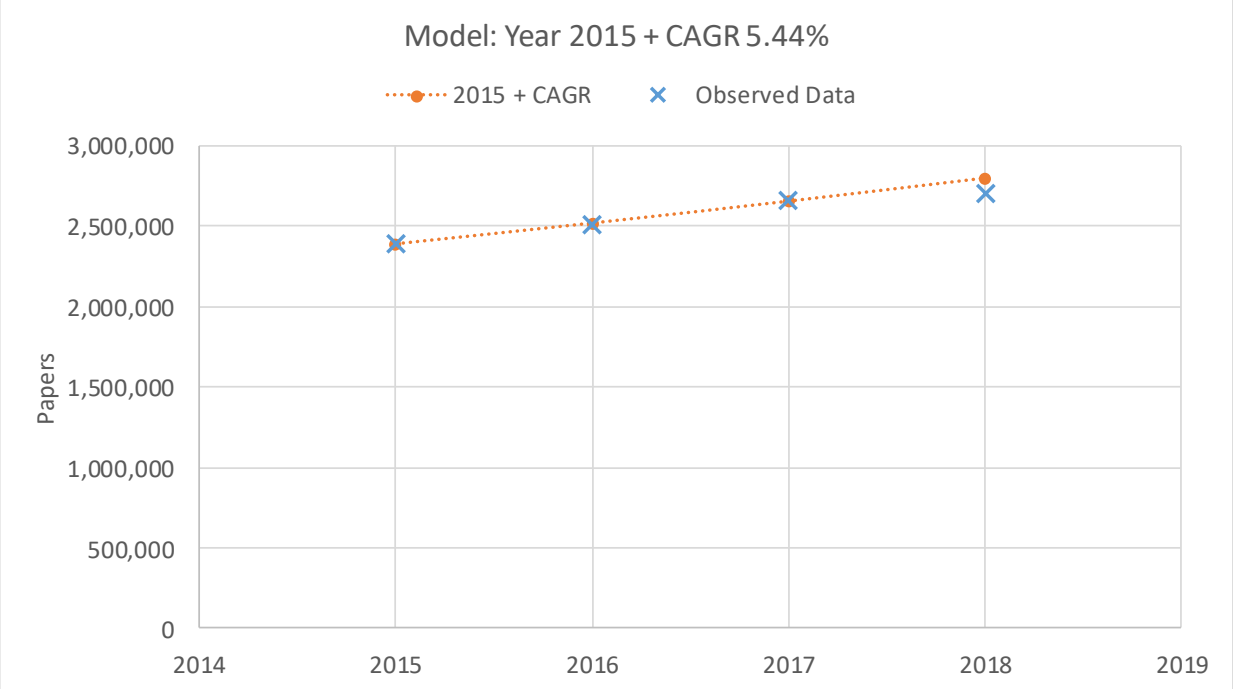


Figure 2 Observed data and evaluation of the completeness, July 2019
Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

[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.

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 six categories: *Journal, Conference Proceeding, Book Series, Trade Publication, Book* and *Report*. These include documents that are categorized into 15 categories: *Article, Conference Paper, Review, Letter, Book Chapter, Editorial, Note, Short Survey, In Press, Erratum, Book, Conference Review, Report, Abstract Report* and *Business Article*.

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 (58% of U.S. publications between 2006-15 are available via Open Access[4]).Open access scholarly literature is free of charge 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]

[4] NSB Indicators: Academic Research and Development. https://www.nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/outputs-of-s-e-research-publications

[5] Memon AR. Revisiting the Term Predatory Open Access Publishing. *J Korean Med Sci.* 2019;34(13):e99. Published 2019 Apr 8. doi:10.3346/jkms.2019.34.e99

Researchers may or may not be the victim of a dubious publisher since early-career and naïve researchers may erroneously submit while low quality researchers may seek to increase their publication numbers without the hard work of original research.

Researchers have attempted to create lists of good and bad journals but 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 cancelation of a journal 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 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 are included as external files.

External File 2: Scopus canceled title list

External File 3: DOAJ canceled title list

2.1.4 Missing addresses

A previous report[7] prepared by Science-Metrix for SRI on behalf of the NSF compared the SEI 2016 bibliometric indicators (prepared with Scopus) to those from SEI 2014 (prepared with the WoS). This report discussed at length the fact the Scopus database is missing institutional addresses of authors for papers published between 2000 and 2002. For the most part, these missing addresses are the result of

[6] https://doaj.org/

[7] Côté, G., Roberge, G., & Archambault, É. (2016). *Bibliometrics and patent indicators for the Science and Engineering Indicators 2016: Comparison of 2016 bibliometric indicators to 2014 indicators*. Montréal, QC: Science-Metrix. Retrieved from http://www.science-metrix.com/files/science-metrix/publications/science-metrix_comparison_of_2016_bibliometric_indicators_to_2014_indicators.pdf.

limitations within the data sources used by Elsevier to build the database. Commonly referred as a "hole" in the Scopus database, these missing institutional addresses may result in an underestimation of levels of output and collaboration rates for years affected by this issue. Figure 3 highlights the address issue for 2000, 2001 and 2002.

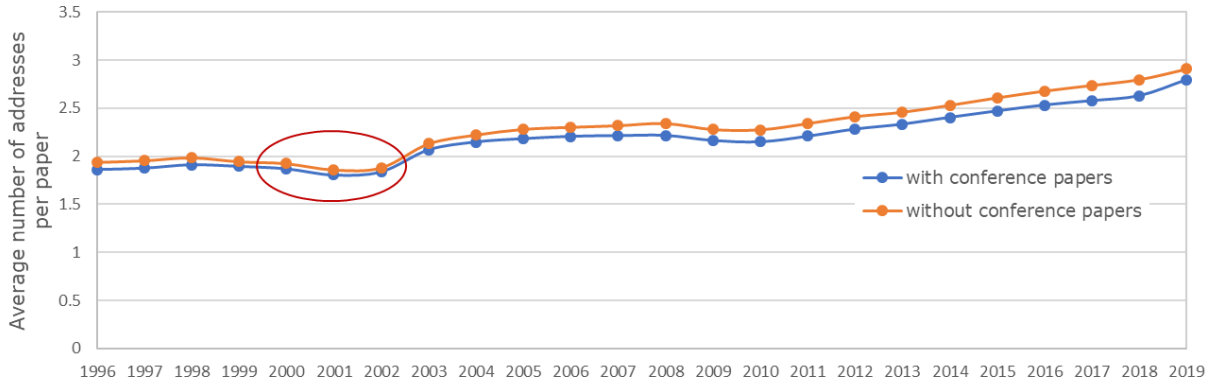


Figure 3 Average number of addresses on publications in Scopus, 1996–2019
Source: Prepared by Science-Metrix using the Scopus database (Elsevier)

During data production for the SEI 2016, it was decided that years affected by this artifact would be left in the SEI because data to be presented in the main report were mostly outside this hole. The impact of this decision was that some data presented in annex tables did indeed present the characteristic decrease in output associated with the hole. However, for the SEI 2018 the window for the data presented shifted, and the data to be presented in the main report would have been themselves directly impacted by the hole. As discussions regarding the hole became more prominent during data production, it was decided that all metrics directly impacted by it would be left out of the SEI, unless it was assessed that missing addresses did not have a notable effect on the data. For SEI 2020, the data points presented in the report are now mostly out of the range of year affected by this artifact, so the data are included in the report. Other metrics based on averages and related to citation counts were mostly unaffected by these missing addresses, so the SEI contains longer time frames for those metrics.

2.2 Data standardization

2.2.1 Linking TOD classification to the database

One change for SEI 2020 is that the WebCASPAR classification of science used for previous editions is to be replaced by the Taxonomy of Disciplines (TOD) classification. The decision to change classification was supported by three factors: (1) almost 50% of the content in Scopus was not covered by the WebCASPAR classification as only older journals were covered; (2) this classification was officially terminated, meaning nobody would try to update it; and (3) there was a need to align with the current classification at the NSF for ease of data alignment during analysis. This new 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[8] 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" all get assigned to the "Health Sciences" TOD field. This works well as the 176 subfields are much more restrictive 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" (this latter 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.

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.[9] *Deep* means that there are several layers to the neural network architecture. This

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

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

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.[10]

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.

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

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

For the encoding, the 26 letters, the 10 arabic numbers, several punctuation signs (e.g., "-„;.!?:'_/\|@#$%^&*~+=<>()[]{}\) 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 metrics to evaluate the deep neural network. We calculated three metrics at two levels of aggregation (i.e., subfield and TOD). The metrics 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 (metric 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% (metric 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 (metric 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% (metric 2), and the probability that the classifier's prediction fell within any of the analysts' choices was 99% (metric 3).

External File 4: Article id 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 inserts the metadata in its database) 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).

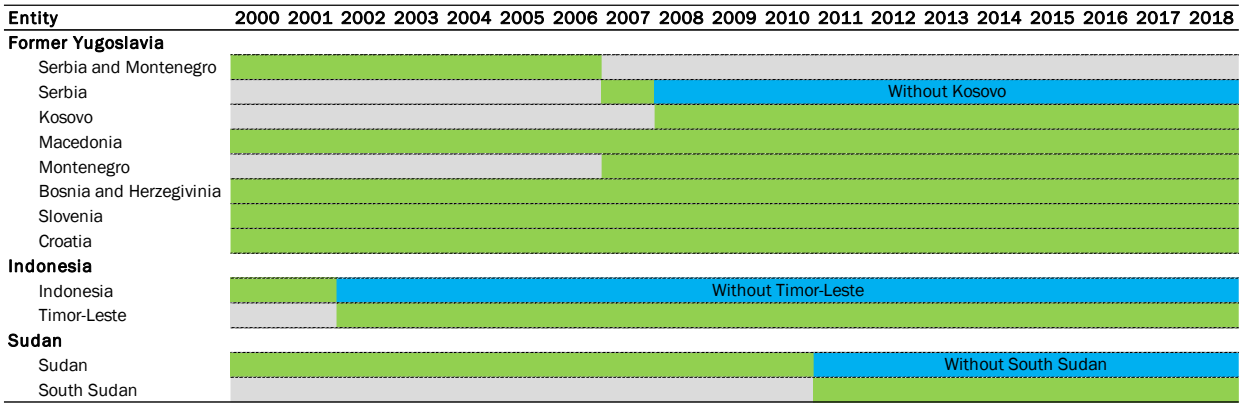
- Older entries in Scopus often do not have information about the author's country.

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 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 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: Scopus country

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 (Python 2.7 script) 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 (Python 2.7 script) extracts the first encounter of a five-digit number from the city field. This is assumed to be the zip code.
- A regular expression (SQL script) extracts the first encounter of a two-capital-letter word that exists 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 (SQL script) 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 4% of U.S. addresses.[11] For most of these addresses, there was no information available that enabled coding at the state level.

External File 6: Scopus US 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*Private Academic*, *Public Academic* 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 errors in the attribution of AFID to

[11] 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.

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:

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% of all U.S. addresses were assigned a sector. The precision of the assignment process reached 98% 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 97% and 99%. Only two sectors presented lower precision levels: Private Nonprofit (87%) and Academic Undefined (88%).

External File 7: Scopus U.S. sectors

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. About 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	%
1996	1 163 537	100%	1 001 670	86%	959 233	82%	939 862	81%	939 288	81%
1997	1 192 245	100%	1 062 422	89%	1 005 871	84%	986 555	83%	985 898	83%
1998	1 195 424	100%	1 065 118	89%	1 007 451	84%	987 212	83%	986 598	83%
1999	1 203 917	100%	1 073 115	89%	1 016 797	84%	996 638	83%	995 889	83%
2000	1 279 411	100%	1 121 438	88%	1 072 236	84%	1 050 369	82%	1 049 636	82%
2001	1 385 745	100%	1 162 815	84%	1 106 130	80%	1 083 029	78%	1 082 118	78%
2002	1 452 460	100%	1 224 119	84%	1 157 644	80%	1 132 651	78%	1 131 633	78%
2003	1 528 235	100%	1 294 933	85%	1 220 443	80%	1 194 174	78%	1 192 446	78%
2004	1 646 729	100%	1 398 251	85%	1 339 953	81%	1 310 732	80%	1 309 104	79%
2005	1 878 778	100%	1 575 256	84%	1 520 001	81%	1 484 258	79%	1 481 643	79%
2006	1 978 734	100%	1 659 499	84%	1 612 358	81%	1 572 424	79%	1 567 422	79%
2007	2 100 465	100%	1 758 171	84%	1 710 183	81%	1 668 296	79%	1 661 634	79%
2008	2 198 887	100%	1 859 148	85%	1 809 038	82%	1 762 651	80%	1 752 431	80%
2009	2 308 269	100%	1 973 903	86%	1 927 553	84%	1 875 264	81%	1 861 148	81%
2010	2 444 914	100%	2 090 149	85%	2 041 917	84%	1 983 086	81%	1 958 948	80%
2011	2 602 588	100%	2 251 243	87%	2 195 994	84%	2 132 926	82%	2 070 735	80%
2012	2 720 421	100%	2 349 050	86%	2 282 935	84%	2 218 388	82%	2 137 315	79%
2013	2 816 031	100%	2 437 136	87%	2 372 542	84%	2 305 985	82%	2 217 046	79%
2014	2 919 614	100%	2 535 123	87%	2 468 099	85%	2 398 510	82%	2 300 684	79%
2015	2 845 730	100%	2 468 001	87%	2 405 847	85%	2 334 683	82%	2 305 909	81%
2016	2 830 950	100%	2 445 909	86%	2 403 857	85%	2 326 470	82%	2 295 608	81%
Total	**41 693 084**	**100%**	**35 806 469**	**86%**	**34 636 082**	**83%**	**33 744 163**	**81%**	**33 283 133**	**80%**

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 4. 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 NSF production

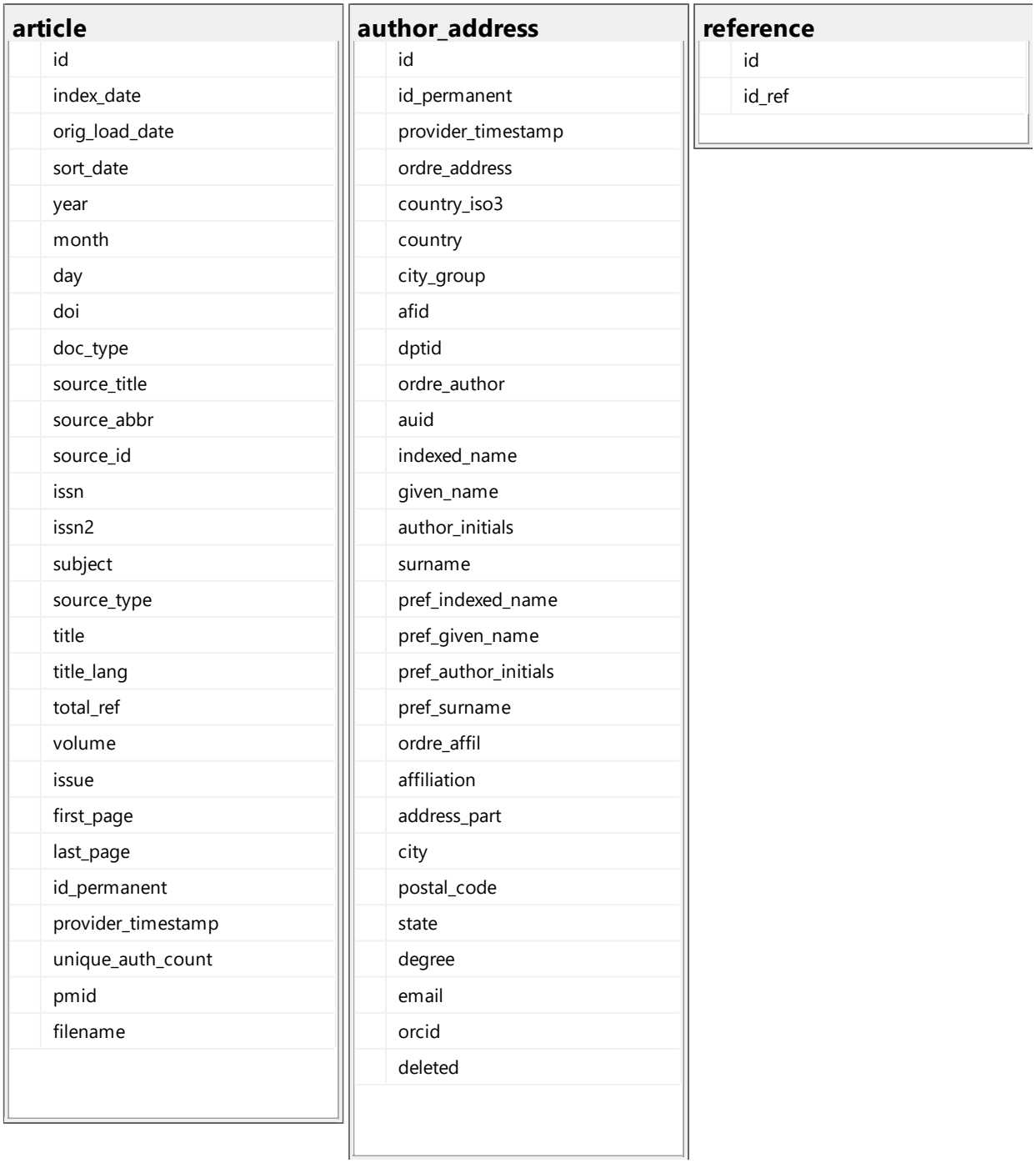


Figure 4 Basic Scopus database schema
Source: Prepared by Science-Metrix

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 basic 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).

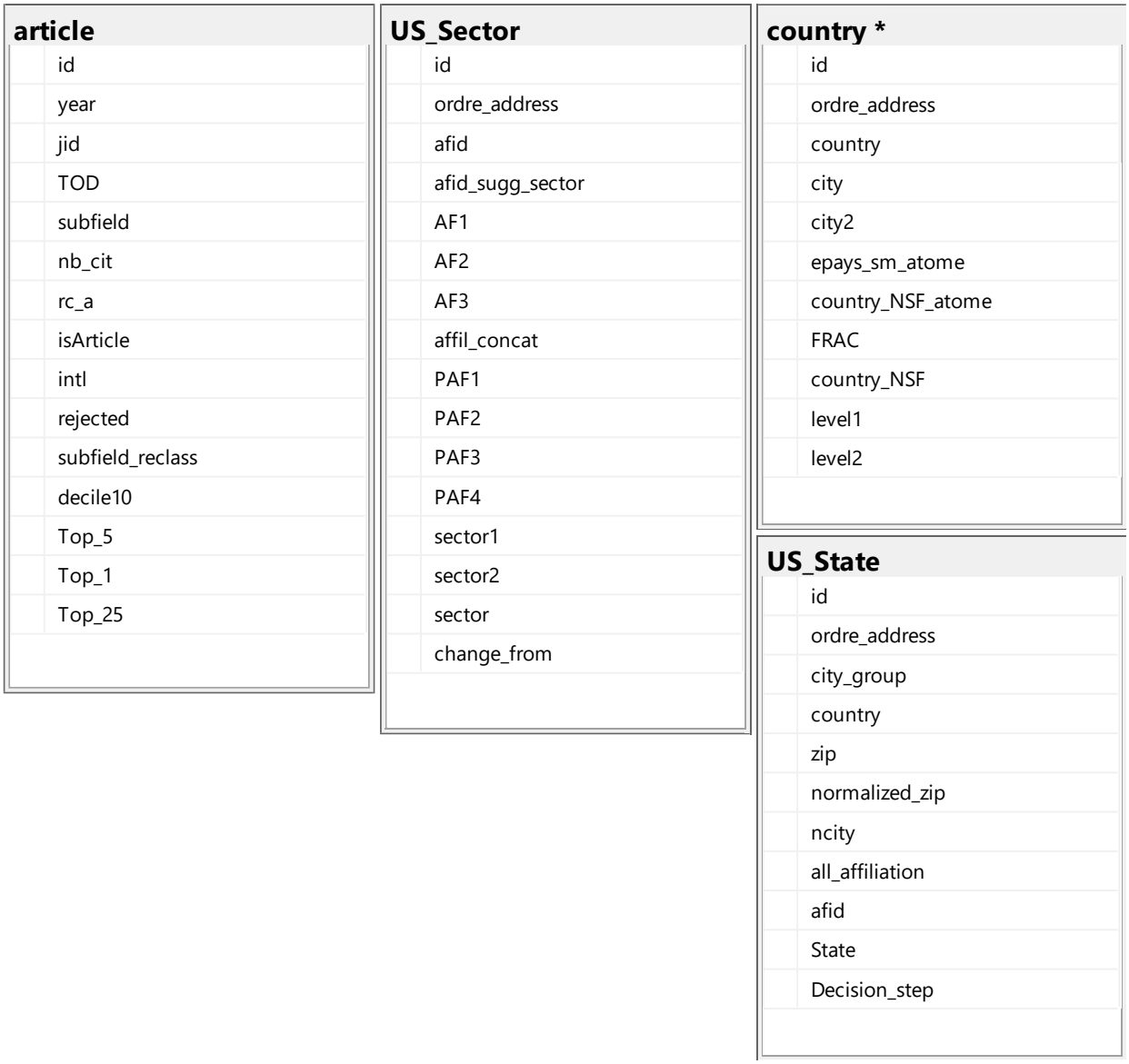


Figure 5 Production database schema
Source: Prepared by Science-Metrix

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 principal ways of counting the number of papers.

Full counting

In the full 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.

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. when the fractions are calculated at the level of researchers. 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.

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[12] because of 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 applies for the "With domestic only institutions" category. Indeed, in this latter case, on publications where there are only unassigned addresses accompanying addresses from a single country, it is not possible to assign these publications to the "With domestic only institutions" category as it is not possible to know for sure that the remaining unassigned addresses are not from another distinct country, which would force these addresses in the "With international institutions" category instead. Therefore, the collaboration status of these publications is ambiguous, and the total count of publications is the sum of the two categories presented above, plus this third category of ambiguous collaboration status (date are presented for this third category).

Using numbers to better exemplify this, there were 1,071,952 publications published at the world level for year 2000, of which 912,227 fell under the "With domestic only institutions" category and 145,362 were instead classified under the "With international institutions" category. When summing both categories, there is a difference of 14,363 missing publications compared to the world total, which accounts for all the papers that fit the description presented earlier, that is, papers with only known addresses from a single country, but with remaining unknown addresses, making it impossible to classify the papers under either "with domestic only institutions" or "with international institutions".

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 multi-lateral 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).

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

2.4.4 Index of collaboration

The index of collaboration (IC) 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 random selection of partnering countries. The index is based on full counts of papers and is calculated as follows:

$$IC_{xy} = \left(\frac{C_{xy}}{C_x}\right) \Big/ \left(\frac{C_y}{C_w}\right)$$

where

IC_{xy}	Index of collaboration between country x and country y
C_{xy}	Number of papers co-authored between country x and country y
C_x	Total number of international co-authorship by country x
C_y	Total number of international co-authorship by country y
C_w	Total number of international co-authorship in the database

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 different 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 here under "citation analysis" (i.e., citation counts and the various ways to normalize them).

Before going into detail about specific indicators, it is important to highlight a number of 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,[13,14] but these measures remain widely used as 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 idiosyncratic uses of citations are largely mitigated, and citations can therefore be used with a high level of confidence.

[13] 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.

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

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,[15,16] 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.[17,18,19]

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

Highly cited publications and citation percentiles

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 the relative citation scores.

One way to increase the fitness of citation counts using these relative citation 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 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 relative citation score (RC) 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

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

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

[17] 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.

[18] 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.

[19] 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.

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:

$$RC_i = \frac{C_{i_{sp}}}{\sum_{i=1}^{N_{sp}} C}$$

Where

C	All *c*itations to a paper
$C_{i_{sp}}$	All *c*itations to paper *i*, which belongs in subfield *s*, and publication year *p*, made by articles, reviews and conference papers
N_{sp}	Total number of papers from subfield *s* published in publication year p

In SEI 2020, only the scores on highly cited publications are based on these relative citation scores. To compute the proportion of papers of an entity that are in the top x% most cited papers, the top x% most cited papers at the world level must first be determined. To account for the variations in citation behavior between the disciplines, document types 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. 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 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% (0.5 publications of the top 10% each). 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 1.1 publications in the top 10%). In this case, there is a dual fractionation if there are ties at the threshold. 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 0.1. If three papers are tied in second place behind the first paper, they are each given a weight of 0.033 (i.e., 0.1*1/3); if the top two papers are tied, they are each given a count of 0.55 (i.e., 1.1/2); and so on.

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 would be 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 (i.e., a tie on the citation score, the x% is not an integer and the paper is co-authored).

2.4.6 Specialization index

The specialization index is a metric using fractional counts of publications to normalize an entity's share of the world output in a given topic with this entity's share of the world output in science overall, providing a normalized indicator highlighting whether an entity is publishing more or less than expected in a given topic relative to its share of all scientific output at the world level.[20] Equipped with this indicator, it is now much easier to make an assessment regarding focus in research activity. Furthermore, it also becomes much easier to extend these comparisons to include other countries, since the normalization makes it possible to account for the size of the field at both the world level and for each national level.

Because of the way the indicator is constructed, a country cannot be specialized in all areas of research, which reflects one of the realities of research: resources are limited and investments and focus in some areas will necessarily result in less focus in other areas. Therefore, the SI is a zero-sum game indicator, with specialization in some areas necessarily resulting in a lack of specialization elsewhere.

[20]More details about the computation of the specialization index and its multiple uses are available in a memo prepared by Science-Metrix dedicate to this indicator (http://www.science-metrix.com/?q=en/publications/reports/memo-on-specialization).