Science-Metrix

Bibliometrics and Patent Indicators for the Science and Engineering Indicators 2016

Comparison of 2016 Bibliometric Indicators to 2014 Indicators

January 2016

Science-Metrix

Bibliometrics and Patent Indicators for the Science and Engineering Indicators 2016

Comparison of 2016 Bibliometric Indicators to 2014 Indicators

14 January 2016

Submitted to:
SRI International

Authors
Grégoire Côté
Guillaume Roberge
Éric Archambault

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

The Science & Engineering Indicators (SEI) provide a series of quantitative data on science and engineering activities in the United States (U.S.), along with contextual data on other countries. The SEI report and accompanying documents are prepared by the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES) on behalf of the National Science Board, which has the ultimate responsibility to report to the president and Congress about the state of science and engineering in the U.S. The SEI report comprises eight chapters that cover the following aspects of science and engineering:

- Chapter 1 – Elementary and Secondary Education
- Chapter 2 – Higher Education in Science and Engineering
- Chapter 3 – Science and Engineering Labor Force
- Chapter 4 – Research and Development: National Trends and International Linkages
- Chapter 5 – Academic Research and Development
- Chapter 6 – Industry, Technology, and the Global Marketplace
- Chapter 7 – Science and Technology: Public Attitudes and Understanding
- Chapter 8 – State Indicators

Indicators on scientific production have been included in the SEI since its first edition in 1972. At that time, bibliometrics was in its early stages of development, and the SEI was a pioneer in the use of bibliometric indicators outside of academia. A group led by Francis Narin at Computer Horizon Inc. (CHI) developed and computed the bibliometric indicators for the first edition of the SEI. This group and its successors eventually added indicators based on patents and continuously improved and expanded the indicators that were included in subsequent editions of the SEI, up until the most recent edition published in 2014.

For the 2016 edition, for the first time in more than 40 years, an entirely different team was selected to compute the bibliometric statistics for the SEI. More specifically, Science-Metrix was commissioned by SRI International and the NSF to compute the bibliometric and patent statistics for chapters 5, 6 and 8 for SEI 2016. In addition, whereas the bibliometric indicators have historically always been computed using data from the scientific publications indexed in the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI)—two citation indexes currently owned and distributed by Thomson Reuters as part of the Web of Science (WoS)—the SEI 2016 bibliometric are computed using Scopus, a similar database developed by Elsevier. Because of these changes, it is important to document carefully the changes between the 2014 and 2016 versions of the SEI, which is the goal of the present report.

For this study, an in-house version of Scopus was developed by Science-Metrix from the database sent by Elsevier. Several obstacles limit or prohibit the transformation of bibliographic data into bibliometric data when these data are accessed remotely. Firstly, remote databases are not editable, so it is impossible to standardize the data at the source (online databases are by and large "read-only"). Another factor is that remote access licenses limit the number of records that can be downloaded; consequently, it is not possible to standardize thousands of records (in order to, for example, properly code the countries of the authors). In many cases, downloading multiple notices, such as those from entire journals, is not permitted by the licenser. Standard licenses also prohibit the use of the database for the systematic retrieval and compilation of material, such as would be necessary to produce statistics.

1.1 Context and considerations regarding bibliographic databases

Until a decade ago, bibliographic databases distributed by Thomson Reuters—namely, the WoS and its predecessors—were the sole source of data available to conduct large-scale bibliometric analyses. The WoS began its journey in 1960 as the Science Citation Index (SCI), which was produced by the Institute for Scientific Information (ISI); the ISI was acquired by Thomson Scientific & Healthcare in 1992. Scopus, introduced by Reed Elsevier in 2004, quickly became a major competitor to Thomson Reuters' product.

The comparison of these two major databases has been an active area of study and discussion, and has yielded several scientific publications. However, most of these comparisons were conducted using the web versions of the WoS and Scopus and/or were performed using only a limited subset of the databases (e.g., single topic, single organization, or limited time frame). Frequently, these comparisons were aimed at informing the intended use of these databases (i.e., searching for scientific literature), but compared their use for bibliometric purposes less often. In most cases, the studies that compared the WoS and Scopus from a bibliometric perspective obtained similar results between the two databases across a variety of indicators, at least when comparing statistics computed on large numbers. For example, Archambault et al. showed a fairly high level of consistency between the two databases: at the country level, they observed a high degree of correlation between them ($R^2 > 0.99$), both in the number of publications and the number of citations.[1]

The previous editions of the SEI, including the most recent edition (2014), were compiled from the set of journals covered by the Science Citation Index (SCI) and Social Sciences Citation Index (SSCI), which were distributed by Thomson Reuters in CD format. These products were subsequently integrated into the WoS along with the Arts and Humanities Citation Index (AHCI), which can be accessed online and is widely subscribed to by university libraries. The SCI and SSCI CD formats are subsets of the online WoS that provide more recent bibliographic data for the most recent data year than is available from the more inclusive WoS or Scopus. Although the WoS in its entirety offers a more comprehensive coverage of scientific publications than the SCI and SSCI alone (Figure 1), a direct comparison of the WoS with Scopus is beyond the scope of this report. This report aims to illustrate and explain the differences between the indicators to be included in SEI 2016 and the indicators published in SEI 2014, which were based on data from SCI and SSCI alone.

[1] Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, *60*(7), 1320–1326. http://doi.wiley.com/10.1002/asi.21062

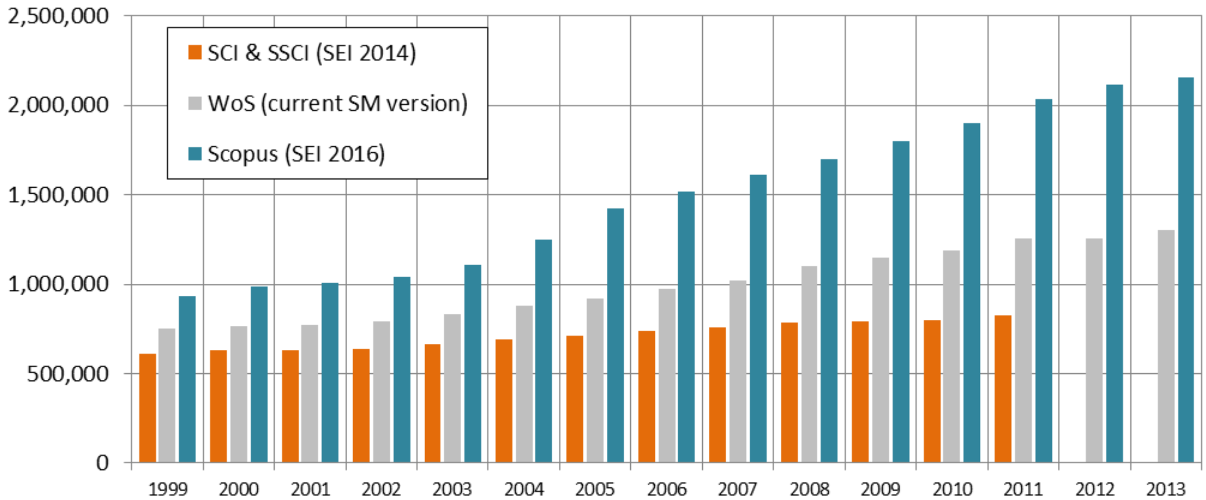

Figure 1 Number of papers from SEI 2016 in Scopus, SEI 2014 in the SCI & SSCI, and from the current Science-Metrix version of the WoS, 1999–2013

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI; current statistics from the WoS calculated by Science-Metrix

Science-Metrix has an in-house, licensed version of the WoS that can be used to support comparisons that cannot be derived from the data that are available online from SEI 2014. However, Science-Metrix did not have an in-house version of the SCI and SSCI, nor is Science-Metrix able to precisely replicate these two indices from its in-house version of the WoS, even if these two databases are a subset of the WoS. Thus, for some analyses, the comparison is performed between Scopus and the WoS, and the findings will be used to explain the differences in the statistics presented in SEI 2014 and SEI 2016. To avoid confusion, each table, figure and analysis presented in the report acknowledges the relevant source of data used in the comparison.

1.2 Organization of this report

Section 2.1 presents the differences in coverage between the SCI and SSCI, the WoS and Scopus, as well as differences in data quality and data cleaning procedures. Section 2.2 compares the results presented in SEI 2014 with those computed for SEI 2016, examining the impact of the database selection for country comparison, and highlighting what appear to be the most representative trends in the scientific output of the U.S., Europe and China. Section 2.4 presents a comparison of the output of U.S. sectors, while Section 2.5 examines the scientific production of U.S. states in SEI 2014 and in SEI 2016. Section 2.6 assesses the impact of the database change on the computation of co-authorship indicators, and Section 2.7 addresses the impact on indicators that are based on citations.

2 Scopus, SCI & SSCI and the WoS

2.1 General comparison

As discussed in Section 1.1, the bibliometric data produced for SEI 2014 were based on journals indexed in the SCI and SSCI, while those produced for SEI 2016 are based on Scopus, which has more extensive coverage—that is, it indexes more journals. For example, about 5,000 journals were indexed in the SCI and SSCI in 2011, compared to approximately 17,000 journals in Scopus (i.e., 3.4 times as many journals). But because the journals that are uniquely covered in Scopus publish fewer papers on average than those covered in the SCI and SSCI, the number of papers in Scopus in 2011 is only 2.5 times the number in the SCI and SSCI. Moreover, Scopus also indexed about 300,000 "other" publications in 2011: mainly peer-reviewed publications published in book series, or conference papers published via channels other than in journals and book series.

Figure 2 presents a comparison of the number of papers used for the bibliometric analyses in SEI 2014 and SEI 2016. In 1999, Scopus comprised about 1.5 times the number of peer-reviewed papers in the SCI and SSCI (153%), with the difference between the two databases increasing to 2.5 times the number (246%) in 2011. It is important to restate here that this is not a comparison of the online versions of Scopus and the WoS. Rather, this comparison reflects an important methodological choice made by the NSF to continue using solely the SCI and SSCI to compute the SEI over the years in the interest of obtaining a stable set of statistics, despite the availability of products from Thomson Reuters offering more coverage. Compared to the WoS as a whole, the 2014 SEI statistics omitted the AHCI, all the journals in the SCI Expanded that are not present in the SCI, and all the proceedings in the Conference Proceedings Citation Index (science as well as social sciences editions). These three databases are included in the WoS Core Collection. Some articles in book series were also possibly omitted. Figure 2 therefore compares a truncated portion of the offering by Thomson Reuters to the complete offering of Elsevier in Scopus. This comparison is solely useable to understand how the change of dataset affects the statistics for the subsequent editions of the SEI, and should not be used to compare the two commercial offerings.

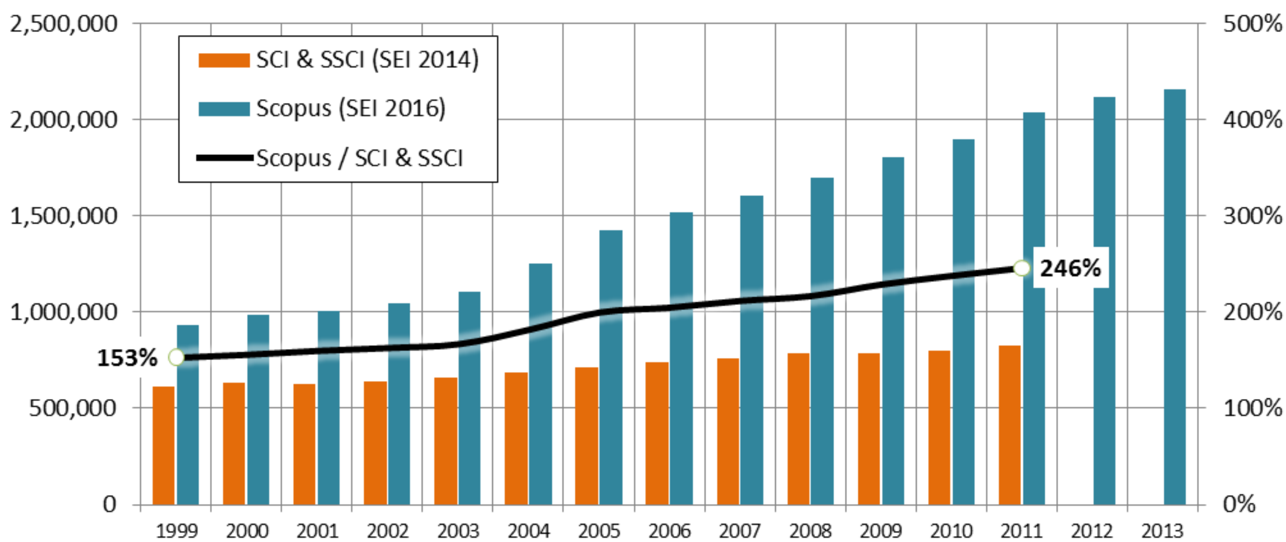

Figure 2 Number of papers from SEI 2016 in Scopus and SEI 2014 in the SCI & SSCI, 1999–2013

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

The difference in coverage between the two databases is often attributed to the divergence in strategies used to gather journals and publications in the WoS and Scopus.[2] Different indexing policies of databases can affect the bibliometric results, as shown by Bar-Ilan.[3] In an effort to control quality, the WoS includes journals using criteria based on the average citation rate of articles in the journals. This selection is periodically monitored: if a journal falls below a certain threshold it is dropped, whereas journals that rise above the threshold are included. On the other hand, Scopus favors a comprehensive strategy that results in a larger number of journals and publications being indexed compared to the WoS.

Although no database covers the entirety of the scientific publications published globally, and the exact number of peer-reviewed papers is not known, the sets of documents indexed in the SCI and SSCI, the WoS and Scopus each represent a very large, non-random sample of this population. In other words, these databases, although offering a large representation of the overall pool of papers published worldwide, do not provide an objective representation of the entire population of peer-reviewed journals. They both have a bias toward large, highly cited journals and a bias for articles written in English and published in English-language journals. To determine precisely the biases of a given database, the population of all peer-reviewed publications would first have to be characterized. However, gathering the information on the population in its entirety would be a colossal task, and Science-Metrix is not aware of any comprehensive dataset that has already been compiled and could be used for this purpose. Instead, Ulrich's Periodical Directory (known simply as Ulrich), which provides information about serials such as magazines, and trade and academic journals, is used as a benchmark by academics because it provides the most representative view of the population. Using the Ulrich database as a benchmark, it has been demonstrated that both the WoS and Scopus have underrepresented the arts, humanities and social sciences.[4] This study is not covering arts and humanities, thus this finding is only relevant for the social sciences in the current context.

The language and regional biases of bibliographic databases are therefore discussed below based on available data and/or research. This is followed by a discussion on the inclusion of "quality reviewed" articles and other document types in the databases, as well as a brief review of data quality issues and associated data cleaning considerations.

Language biases

Bibliometric indicators are quite representative of the language distribution of scientific publications for the natural sciences and engineering (NSE) fields, but are generally much less reliable in the social sciences and humanities (SSH) fields.[5,6] This is partly explained by the fact that phenomena studied in the NSE tend to have a more "universal" character than those examined in the SSH. For instance, molecules studied in chemistry behave the same way no matter in which country they are studied and chemists all around the world may be interested in knowing their specificities. This explains, to some

[2] Ball, R., & Tunger, D. (2006). Science indicators revisited — Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services and Use*, *26*(4), 293–301. Retrieved from http://epub.uni-regensburg.de/5244/1/ubr11317_ocr.pdf

[3] Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271. http://doi.org/10.1007/s11192-008-0216-y

[4] Mongeon, P., & Paul-Hus, A. (2014). The journal coverage of bibliometric databases: A comparison of Scopus and Web of Science. École de bibliothéconomie et des sciences de l'information, Université de Montréal. Retrieved from http://www.researchgate.net/profile/Adele_Paul-Hus/publication/268223674_The_journal_coverage_of_bibliometric_databases_A_comparison_of_Scopus_and_Web_of_Science/links/54664afb0cf2f5eb18016c0a.pdf

[5] Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342. Retrieved from http://www.akademiai.com/index/34W733263N36H661.pdf

[6] Moed, H. F., Luwel, M., & Nederhof, A. J. (2002). Towards research performance in the humanities. *Library Trends, 50*(3), 498–520. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/8415/librarytrendsv50i3m_opt.pdf?sequence=1

extent, why a common language—English—is used in a quasi-universal or at least very extensive manner in the NSE. Some topics examined in the SSH have a more local character, both in the nature of the topics under study and in the interest in these topics. For example, a study of how social workers are relied upon to remediate gang violence in Brooklyn may be most useful to scholars and practitioners in the U.S., and less so to those in Norway. Conversely, studies examining specific problems facing those dealing with violence in Oslo may be better conducted and shared in Norwegian.

To characterize the difference in linguistic coverage in the SSH between Scopus and the WoS (whole database comparison, online version), a sample of publications extracted from the research reports of Canadian researchers supported by the Social Sciences and Humanities Research Council was used to determine how both databases covered English- and French-language papers. Table I shows that although the WoS covered 43% of the papers published in English by Canadian researchers, it covered only 7% of those in French. Scopus covered 53% of the papers in English and 16% of those in French.

Table I Coverage of Scopus and WoS for a sample of Canadian SSH papers in English and French

	Scopus		WoS		Scopus & WoS		Sample
Paper's language	*Coverage*	*(n)*	*Coverage*	*(n)*	*Coverage*	*(n)*	
English (E)	53%	120	43%	97	58%	132	226
French (F)	16%	10	7%	4	20%	12	61
Coverage Canadian sample	**45%**	**130**	**35%**	**101**	**50%**	**145**	**289**
Times more coverage E/F	**3.2**		**6.5**		**3.0**		

This confirms that both databases have a bias towards English-language papers in the SSH. Overall, although there remains a large discrepancy in favor of papers written in English in Scopus, its coverage of French-language papers is more than twice that of the WoS. Consequently, Scopus could represent a better choice if one is looking for more extensive coverage of the SSH, especially for non-English-speaking countries.

Regional distribution of journals and publishers

Scientific publications are published in journals that vary in terms of geographical representation. Journals can focus on regional subjects (e.g., *Southwestern Naturalist*) or cover more universal topics (e.g., *Journal of Plasma Physics*). For historical reasons, even journals covering universal topics have at least some regional influence. Many journals were first published by universities, government agencies or regional associations, and therefore are better known by proximate researchers. Although this trend is decreasing over time, in part due to the globalization of science (e.g., increased mobility, electronic media, and the Internet), and in part due to the concentration of ownership of journals by large editors following takeovers, journals still demonstrate country biases. For example, for the 1999–2013 period, of the 21,000 or so journals in Scopus with at least 30 articles, nearly 12,000 had more than 50% of their articles signed by authors affiliated with a particular "dominant" country (as calculated by Science-Metrix). Therefore, the selection of journals that are indexed in a database can directly influence the regional bias of bibliometric indicators.

A simple analysis of the geographical coverage of publishers in the WoS and Scopus can explain, at least partially, the difference in coverage between China and the U.S. across the two databases. A comparison of the publishers' country in the WoS and Scopus for the 1999–2011 period shows a more balanced regional coverage in Scopus than in the WoS. In fact, 45% of the publications in the WoS were from publishers located in the U.S., while U.S. publishers account for 33% of the papers in Scopus. Conversely, Chinese publishers are more comprehensively covered in Scopus: only 1% of the papers in the WoS were from Chinese publishers, compared to 6% of the papers indexed in Scopus. This represents more than 700,000 additional Chinese papers in Scopus, which could partially explain the better coverage of Chinese papers in Scopus. Indeed, 97% of papers published in Chinese journals are authored by Chinese researchers.

When the distribution of journals by country of publisher in Scopus and the WoS is compared to the benchmark provided by the Ulrich database, the regional distribution in both Scopus and in the WoS appears to favor some countries and overlook others. These differences in coverage are not the same in Scopus and in the WoS. For example, the United Kingdom's health sciences journals are overrepresented by 46% in Scopus compared to 78% in the WoS.[7]

Peer-reviewed and quality reviewed articles

Only publications that are peer reviewed are deemed to be suitable for the compilation of bibliometric indicators. The peer-review process prior to publication ensures that the research is of good quality and constitutes an original contribution to scientific knowledge.

Both the WoS and Scopus primarily index scholarly and research-related documents, most of which are peer reviewed. There are exceptions to this; for instance, Scopus previously included *Business Week* and *The Economist* magazines. *Business Week* was only covered for a few years, whereas *The Economist* was indexed until March 2014. Neither of these magazines was included in the WoS database. On the other hand, magazines such as *Scientific American* (which clearly notes that "as a consumer magazine, *Scientific American* is not a peer-reviewed journal") are covered in both databases. Normally, articles from these publications should not affect bibliometric statistics as they rarely contain authors' addresses and as such would not be used in statistics for countries, institutions, and so forth. However, a number of articles from these magazines would still mistakenly be included in the analysis because of the small proportion of them that do have authors' addresses.

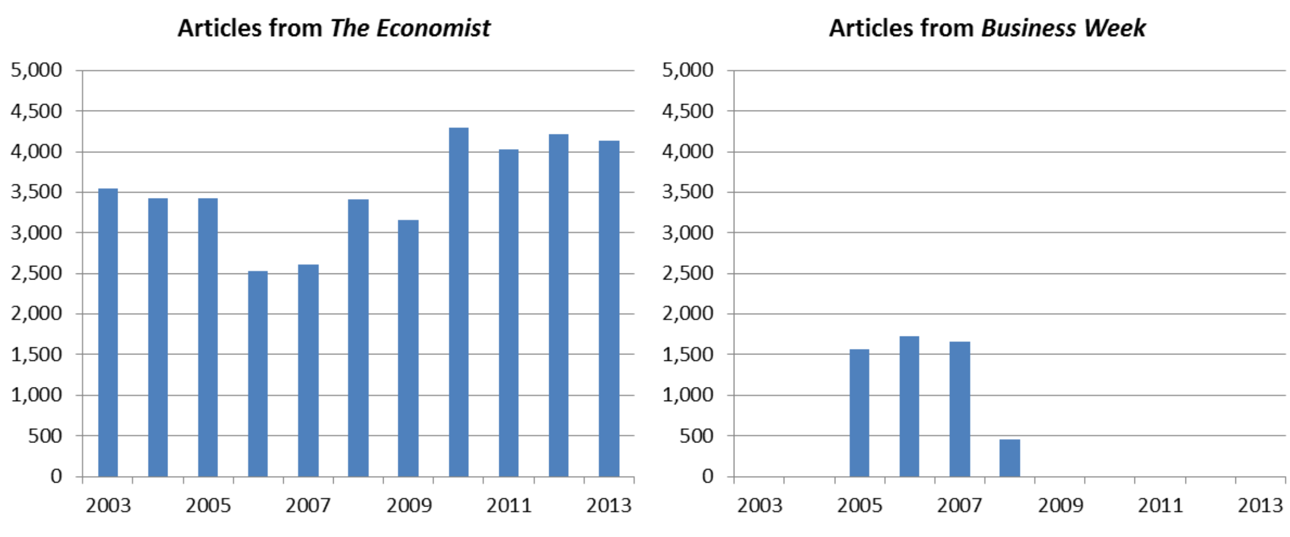

Figure 3 Examples of non-peer-reviewed articles in Scopus, 2003–2013
Source: Prepared by Science-Metrix using Scopus (Elsevier)

[7] Mongeon, P., & Paul-Hus, A. (2014). The journal coverage of bibliometric databases: A comparison of Scopus and Web of Science. École de bibliothéconomie et des sciences de l'information, Université de Montréal. Retrieved from http://www.researchgate.net/profile/Adele_Paul-Hus/publication/268223674_The_journal_coverage_of_bibliometric_databases_A_comparison_of_Scopus_and_Web_of_Science/links/54664afb0cf2f5eb18016c0a.pdf

Document types

The types of documents covered in bibliographic databases such as the WoS and Scopus are mostly articles, reviews and full-length conference proceedings. In the context of bibliometrics, these are collectively referred to as "papers."

The SCI and SSCI contain only documents published in periodicals (journals). For the production of bibliometric indicators for SEI 2014, and in most bibliometric studies based on the SCI, the SSCI or the WoS, only two document types were considered: articles and reviews.

Scopus comprises more types of media, organized into six categories: *Journal, Conference Proceeding, Book Series, Trade Publication, Book* and *Report*. Scopus therefore records information on the type of media (source type) in addition to the document type. Document type includes 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 use a similar filter to that used in the WoS (and the SCI and SSCI) and keep only documents that present original contributions to knowledge. However, an important difference between Scopus and the SCI and SSCI is that Scopus includes peer-reviewed conference papers. Even though this type of document respects the criteria of being peer reviewed and of presenting novel results, conference papers are not indexed in the SCI and SSCI. As explained below, these document types are retained when computing the indicators for SEI 2016.

The classification of document by source type and by document type in Scopus is imperfect and cannot be used directly to identify all relevant papers in the database. An empirical approach has been developed by Science-Metrix to filter documents based on the source types and document types, maximizing the recall of peer-reviewed papers while minimizing the inclusion of non-peer-reviewed documents. The approach is based on the Scopus documentation available on the Internet, as well as on statistics on the number of references and citations per document for each combination of source type and document type.

Table II presents the combinations of source type and document type that have been used for the bibliometric analyses in Scopus for SEI 2016.

Table II Combinations of source types and document types used for the production of bibliometric indicators for SEI 2016

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

Although it is difficult—or almost impossible—to apply the exact same filter for document types in Scopus as that used in the WoS, the resulting coverages generally converge. The only important source of variation between Scopus and the SCI and SSCI is the inclusion of conference papers in Scopus. This inclusion explains a large part of the difference between Scopus and the WoS. Figure 4 presents the number of papers that can be considered as original contributions to knowledge in the WoS, in Scopus, and in Scopus without the conference papers. As clearly shown in the figure, the number of papers in the WoS and in Scopus would be more broadly similar if conference papers were not considered in Scopus.

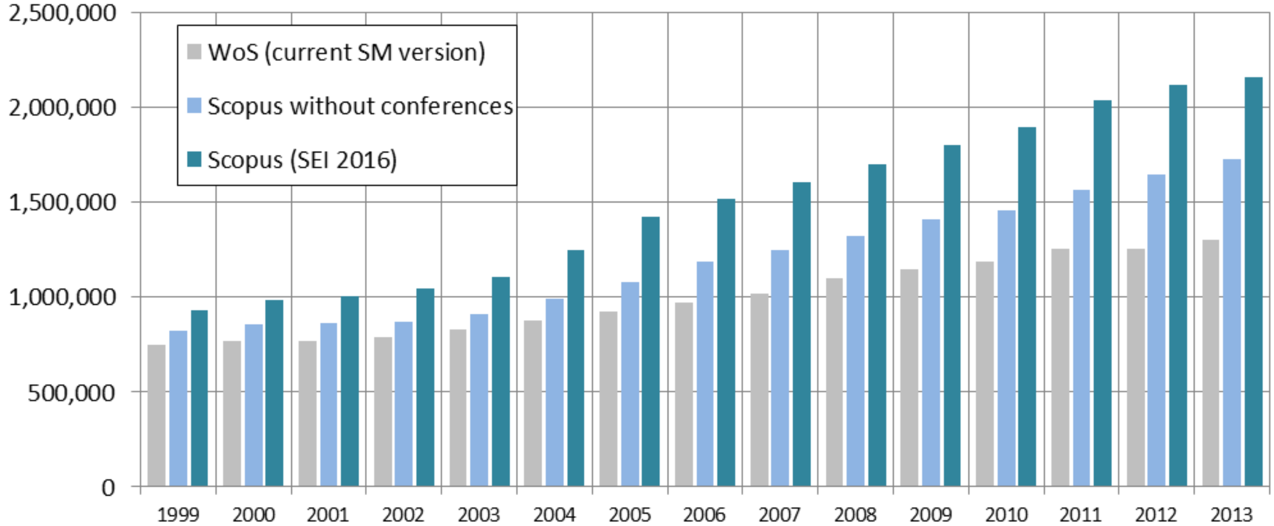

Figure 4 Number of papers in the WoS, Scopus, and Scopus without conference papers, 1999–2013

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

In the applied and physical sciences in particular, conference proceedings are very important for disseminating knowledge (Figure 5). Conference papers are particularly prevalent in computer sciences (72%), engineering (45%), physics (35%), and astronomy (20%). For these areas of research, the inclusion of conference papers in Scopus supports a more representative measure of scientific activity and output. However, conference papers are not as important in several other fields of research. For example, conference papers are scarce in medical sciences (4.7%), biological sciences (4.6%), psychology (4.1%), and the social sciences (2.8%).

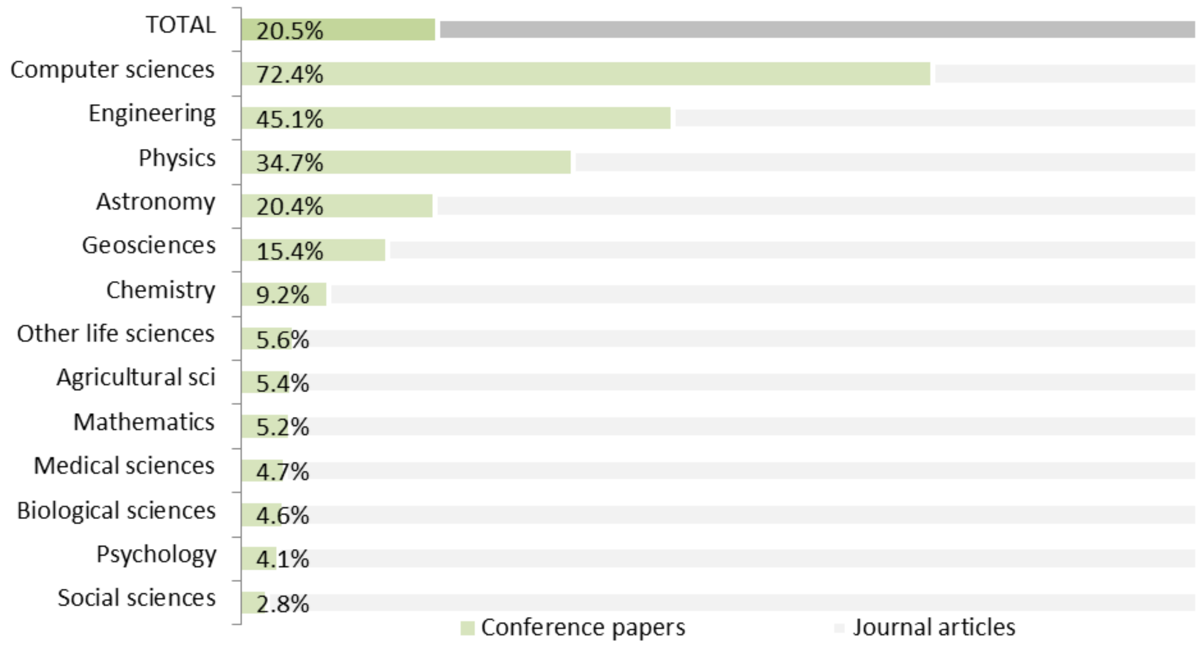

Figure 5 Share of conference papers and journal articles in Scopus, by Caspar field, 1999–2013

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

The effect of including conference papers in Scopus is a notable increase in the share of papers in computer sciences and in engineering. For example, while computer sciences only represented 1% of the dataset used for the SEI 2014 indicators, the field accounted for 8.5% in the SEI 2016 indicators, where conference papers were covered. A comparison of the shifts in distribution of 2011 papers by Caspar field from SEI 2014 to SEI 2016 as a result of the inclusion of conference papers in the latter is presented in Figure 6.

Thus, because of the coverage of conference papers in Scopus, countries (or other entities) that are specialized in computer sciences and/or engineering will perform relatively better in SEI 2016 for indicators based on publication counts. Variations between the two data sources in the inclusion of the various fields of science over time directly impact the measured output of countries, U.S. sectors and U.S. states. Figure 24 and Figure 25, in Annex I, present the trend in the share of the Caspar fields in SEI 2014 and SEI 2016.

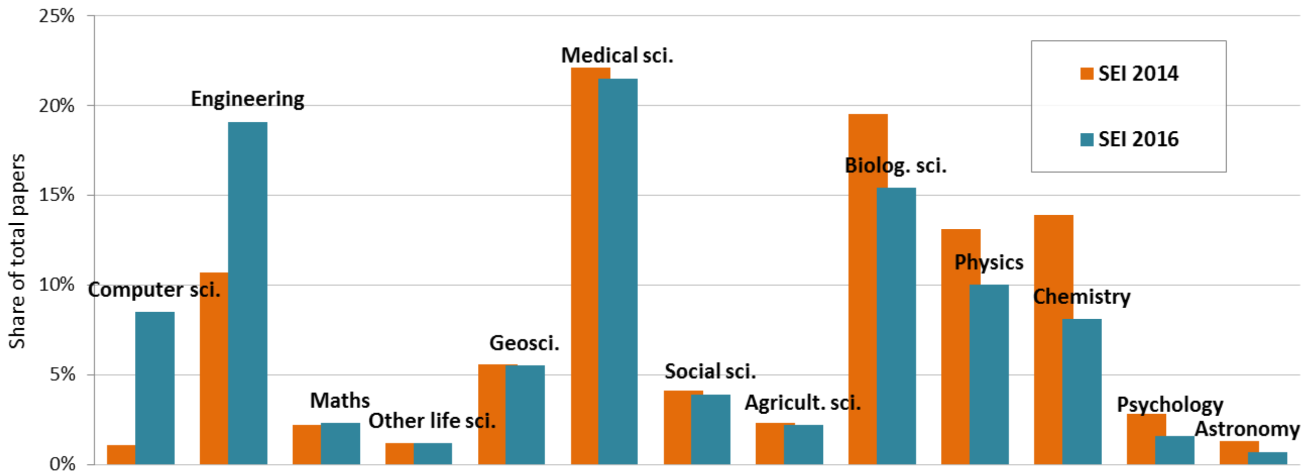

Figure 6 Distribution of 2011 papers by Caspar field, SEI 2014 and SEI 2016

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

Data quality

The quality of bibliometric indicators is highly dependent on the accuracy and completeness of data in a database: missing or incorrectly entered data can have significant ramifications on search results used to delineate the datasets for bibliometric analysis. Scopus, the WoS, the SCI and the SSCI have all been developed primarily to facilitate the retrieval of scientific publications, whereas analytics are increasingly becoming a key afterthought. The number of records in these databases is in the tens of millions, and their bibliographic information has been recorded through various means, including manual entry and optical character recognition (OCR). Importantly, the level of data cleaning and standardization that is optimal for a database used solely for bibliographic purposes does not match the level that is required for the production of bibliometric statistics. Although Eugene Garfield, the creator of the SCI, was mindful of the potential for the computation of statistics on science, the colossal amount of work and the limitation of computer space and power in the 1960s limited the standardization process, especially as verification of the best names/forms for institutions and addresses was certainly not a convenient undertaking in the absence of the Internet.

Despite a number of widely documented shortcomings, the data recorded in the SCI and SSCI is generally more curated than the data in Scopus. The SCI and SSCI are the result of a carefully planned data collection that started about 50 years ago and that has been progressively fine-tuned over the years. These databases focus on a limited set of highly cited journals, which are generally more amenable to economical indexation. Scopus was only launched in 2004, and covers substantially more journals than the SCI and SSCI, as well as book series and conference proceedings. It also includes more journals from small editors and from journals published in languages other than English. This is certainly a remarkable achievement, but it comes at the expense of data quality. Scopus is a combination of data that have been collected from several sources, and there are a number of telltale signs of this decision.

The following are examples of variations in data quality between the two data sources:

- While the country is provided for all addresses in the WoS, the country is missing for about 10% of addresses in Scopus.
- The state is provided for all U.S. addresses in the WoS, but is not readily available in Scopus.
- The city and postal code have been parsed in the WoS but not in Scopus.
- Journal names and ISSN have been thoroughly standardized in the WoS, but only partially in Scopus.
- Volume, issue and pages are more standardized in the WoS than in Scopus.

Figure 7 presents the trends in the number of addresses and the number of authors per paper in both Scopus and the WoS from 1999 to 2013. The chart on the left shows that the number of authors per paper is higher in the WoS and this divergence is increasing over time. This could probably be at least partially explained by a lower number of participants in the documents that are better covered in Scopus (e.g., computer journals, conference papers). However, overall, the curves are fairly smooth and similar for both databases.[8]

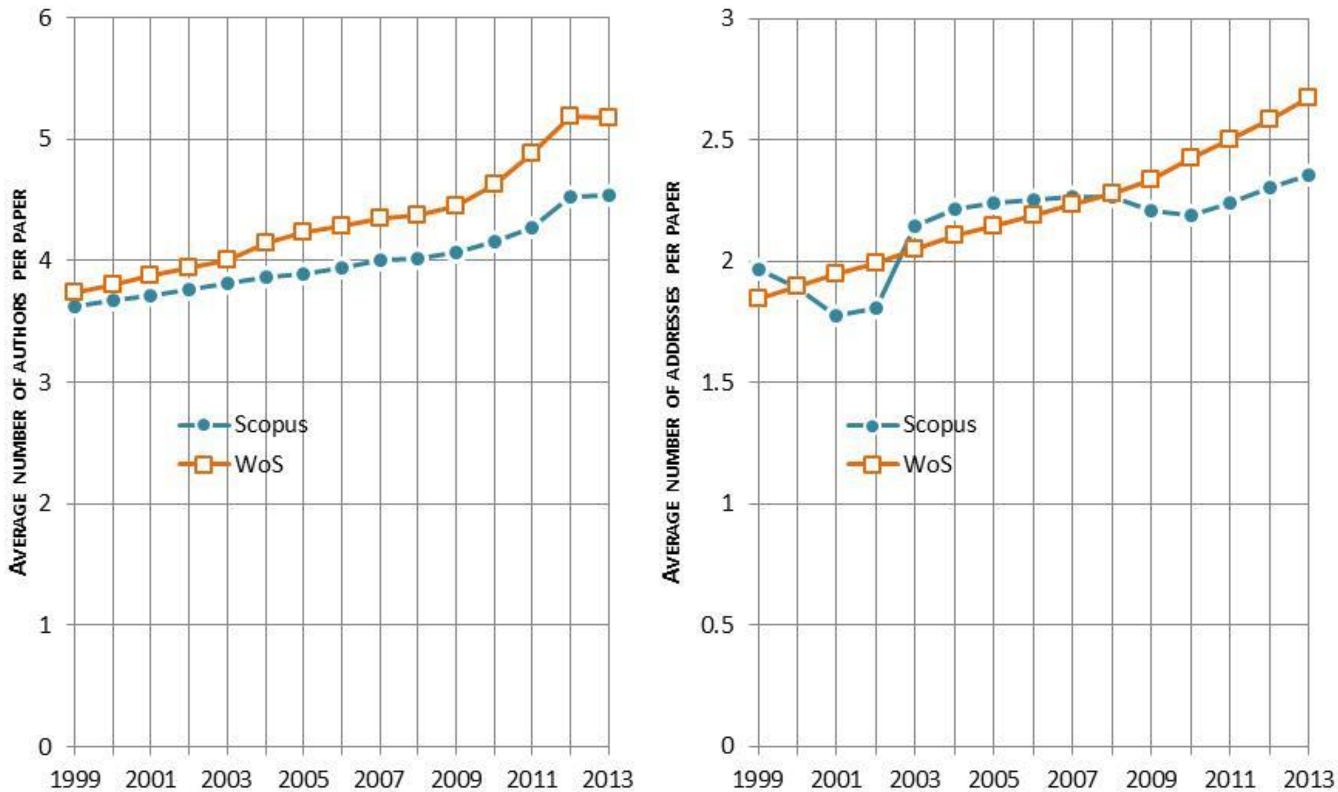

Figure 7 Number of authors and addresses per paper, WoS versus Scopus, 1999–2013

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

The curves presenting the number of addresses per paper in each database are more dissimilar (Figure 7, chart on the right). For Scopus, the curve reveals how the aggregation of various existing databases created marked irregularities compared to the extremely smooth curve observed for the WoS. Indeed, the averages presented for the WoS follow a steady increase over the period presented, which is to be expected given the general increase in co-authorship. In contrast, the curve for Scopus follows a more complex pattern. The "hole" prior to 2003 can be explained as follows: in order backfill the missing data for the years before 2004 (the year Scopus was officially commercialized), Elsevier included some of the data directly from Medline, a freely available biomedical bibliographic database, as well as from EMBASE and COMPENDEX. However, while Medline records the names of all authors, it only records the address of the first author. Therefore, several papers with authors from multiple addresses

[8] The abrupt decrease in 2013 compared to 2012 for both databases has not yet been investigated, but we feel it should not impact the conclusion about the differences and similarities observed between the two databases.

have only one address indexed in Scopus. Similarly, EMBASE and COMPENDEX captured many articles with only the first author and that author's affiliation, and captured many articles with all authors, but only with the affiliation of the first author. Elsevier is aware of this limitation and the years before 2000 have been reprocessed to increase the quality. However, for some years, particularly for 2000, 2001 and 2002, some information is still missing and this drags down the average number of addresses per paper. While we do not have explanations for the other irregularities in the curve, it has to be remembered that any analysis of time variations based on indicators produced with Scopus should take into account the variations in data indexation.

Regardless of the source of data, it is always necessary to perform data cleaning and standardization before the production of robust bibliometric statistics. Moreover, it is important to extensively characterize the dataset from which the indicators are computed, and to take these characteristics into account to inform both the computation and the interpretation of results.

Data cleaning and, more particularly, data standardization have to be performed according to guiding principles that cannot always be separated from the research questions examined. For instance, the aggregation of countries for data compilation and presentation must be determined at project outset. A frequent question in this regard is should statistics be compiled separately for Taiwan or should publications from Taiwan be counted with China? In some cases, a country does not exist in the database and has to be identified based on other information, such as the city or the name of the organization. This is the case for Kosovo, for example, which was mostly coded as Serbia in Scopus. Several steps of data cleaning, standardization and gap filling have been performed on Scopus data in order to have a dataset that is as comparable as possible to the dataset that was used for SEI 2014.

Because the complete methodology for the data cleaning and standardization of the SCI and SSCI in earlier editions of Science and Engineering Indicators was not available for this comparison, there will always be an element of subjectivity in the various steps of data cleaning that are performed on data from Scopus.

2.2 Country comparison

Scientific outputs of countries vary due to a number of factors. Some countries are more productive than others and contribute significantly more toward the global scientific output. This is a direct reflection of the effort these countries expend to foster scientific research. However, it is not uncommon to see the same country covered differently due to database-specific characteristics that stem from journal coverage policy, language or content bias, and overall quality of the database (as discussed in the Section 2.1). In most cases these differences tend to be marginal; nonetheless, it is important to consider these resulting country-specific variances when comparing statistics computed using two different databases.

In SEI 2016, countries have been grouped and presented in the same way as in the previous edition, with three exceptions:

- The Cook Islands has been added.
- Croatia is now listed in countries from the European Union (EU28).
- In this edition, statistics are computed for the West Bank and Gaza. In the previous editions, they were included in Israel.

In order to measure the effect of the change in database on the overall output of countries, a regression analysis was performed between the numbers of papers in both databases for the same publication year—2011. This year was chosen because it was the latest available for the two sources (Figure 8). Overall, countries' production is similar when measured in both databases ($R^2 = 0.97$). For the top two most active countries—namely, the U.S. and China—the U.S. has better coverage in the SCI and SSCI, while China is better covered in Scopus.

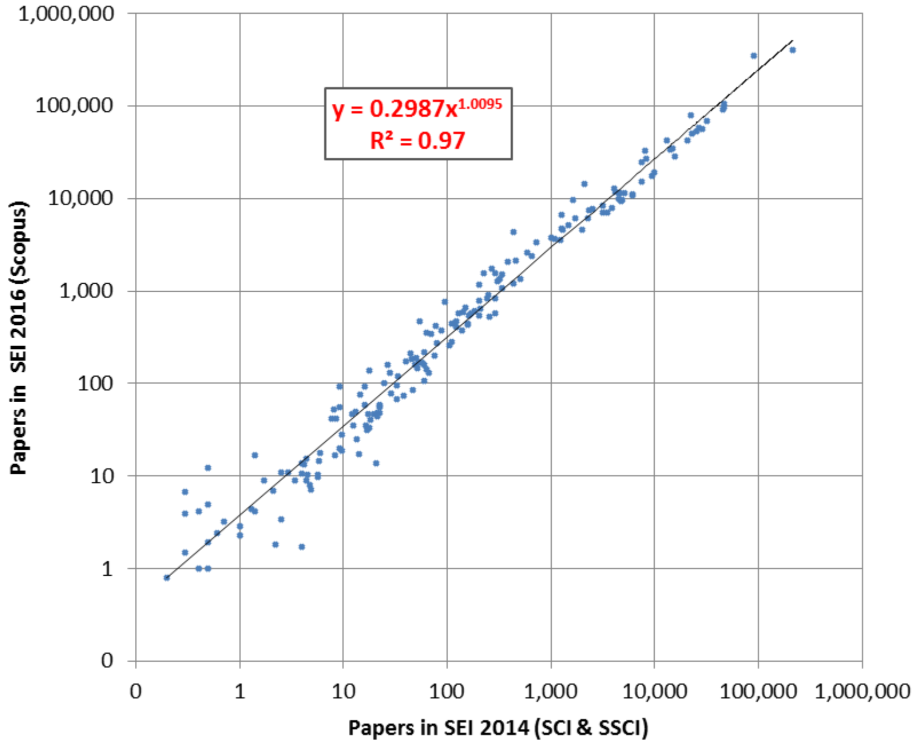

Figure 8 Correlation in papers published in 2011 by country, SEI 2014 (SCI & SSCI) versus SEI 2016 (Scopus)

Note: Fractional counting of papers.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

Although there are noticeable differences in the overall number of papers by country (Table III), what matters more is the relative ranking of these countries. The comparison of countries' output shows that the ranking of the top five countries (the U.S., China, Japan, Germany and the United Kingdom) does not change in 2011 using SEI 2016 data compared to when SEI 2014 data was used. However, the relative rankings change for most other countries. Most countries change position by one or more ranks in SEI 2016 compared to SEI 2014. For example, an increase of five ranks in SEI 2016 (Scopus) for India results in France, Canada, South Korea and Spain losing one or two ranks. Similar increases in rank are observed for Brazil, Turkey, Iran, Poland, Greece, Mexico, the Czech Republic and Malaysia. In general, the rank changes tend to be more common in the lower half of the ranking than the upper half, and in many cases are not relevant in a policy context. Nevertheless, the table does not show the trends over time, and sometimes a larger coverage in Scopus may lead to important changes in rank. For example, although China is behind the U.S. in 2011, a dynamic view of their scientific output would show that China is likely to overtake the U.S. in this regard and that China's output will continue to increase faster than that of the U.S. in the future. A comparison of trends in the U.S., China and Europe is examined in more detail in the following section (Section 2.3).

The inclusion of conference papers in SEI 2016 appears to be an important factor in the variation observed between the two editions (Table III; column "%conf"). The increased coverage of applied sciences in Scopus also appears to be an important factor contributing to the increase of some countries in SEI 2016 (column "%applied"). However, these two factors do not explain all the observed variations, and for some countries, other explanation are clearly needed (e.g., Taiwan, Turkey, Poland).

Table III Comparison of countries' scientific output in 2011, SEI 2014 (SCI & SSCI) versus SEI 2016 (Scopus)

Country / economy	SEI 2014	rk2014	SEI 2016	rk2016	SEI 2016/ SEI 2014	Δ rk	%conf	%applied
WORLD	**827,704.9**	N/A	**2,035,864.0**	N/A	2.46	N/A	21%	28%
United States	212,394.2	1	403,849.9	1	1.90		19%	20%
China	89,894.4	2	351,111.5	2	3.91		32%	45%
Japan	47,105.7	3	104,533.9	3	2.22		25%	26%
Germany	46,258.8	4	97,054.9	4	2.10		22%	21%
United Kingdom	46,035.4	5	92,458.2	5	2.01		16%	18%
France	31,685.5	6	69,575.7	7	2.20	-1	22%	23%
Canada	29,113.7	7	55,727.9	9	1.91	-2	18%	26%
Italy	26,503.4	8	58,102.8	8	2.19		21%	22%
South Korea	25,592.7	9	53,740.8	10	2.10	-1	24%	35%
Spain	22,910.3	10	50,374.5	11	2.20	-1	17%	24%
India	22,480.5	11	80,017.9	6	3.56	5	19%	27%
Australia	20,602.6	12	42,161.2	13	2.05	-1	17%	23%
Netherlands	15,508.3	13	28,758.9	17	1.85	-4	16%	19%
Taiwan	14,809.3	14	34,729.6	14	2.35		28%	39%
Russia	14,150.9	15	33,395.0	15	2.36		16%	20%
Brazil	13,148.1	16	42,894.8	12	3.26	4	13%	27%
Switzerland	10,018.6	17	19,308.4	20	1.93	-3	18%	19%
Sweden	9,472.9	18	17,747.4	21	1.87	-3	17%	21%
Turkey	8,328.4	19	26,923.3	18	3.23	1	10%	26%
Iran	8,175.5	20	32,723.0	16	4.00	4	16%	34%
Poland	7,564.2	21	24,646.5	19	3.26	2	17%	26%
Belgium	7,483.9	22	15,280.0	22	2.04		18%	21%
Israel	6,096.0	23	10,886.3	30	1.79	-7	16%	20%
Denmark	6,071.3	24	11,149.4	29	1.84	-5	16%	20%
Austria	5,102.5	25	11,417.8	26	2.24	-1	22%	24%
Finland	4,877.6	26	9,610.8	32	1.97	-6	20%	24%
Norway	4,777.1	27	9,316.8	34	1.95	-7	18%	24%
Portugal	4,621.1	28	11,391.2	28	2.47		26%	31%
Singapore	4,542.8	29	9,792.6	31	2.16	-2	28%	38%
Greece	4,534.1	30	11,395.1	27	2.51	3	20%	27%
Mexico	4,172.8	31	11,799.5	25	2.83	6	20%	28%
Czech Republic	4,126.5	32	12,831.5	24	3.11	8	25%	25%
Argentina	3,862.8	33	7,793.1	36	2.02	-3	10%	18%
New Zealand	3,471.8	34	7,030.0	40	2.02	-6	14%	21%
Ireland	3,186.4	35	7,067.1	39	2.22	-4	20%	24%
South Africa	3,124.6	36	8,291.3	35	2.65	1	12%	20%
Egypt	2,514.9	37	7,714.8	37	3.07		12%	27%
Thailand	2,304.0	38	7,494.7	38	3.25		22%	31%
Hungary	2,289.3	39	6,188.7	42	2.70	-3	17%	20%
Malaysia	2,092.2	40	14,507.1	23	6.93	17	30%	33%

Note: Fractional counting of papers. Only the top 40 countries in SEI 2014 are presented. SEI2014 = number of papers published in 2011 in SEI 2014; rk2014 = rank of country based on the number of papers published in 2011 in SEI 2014; SEI2016 = number of papers published in 2011 in SEI 2016; rk2016 = rank of country based on the number of papers published in 2011 in SEI 2016; Δ rk = rk2014-rk2016; %conf = proportion of the country's papers that are conference papers; %applied = proportion of the country's papers that are in the applied sciences (Information & Communication Technologies; Engineering; Enabling & Strategic Technologies; Built Environment & Design; Agriculture, Fisheries & Forestry).

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

2.3 The U.S., Europe and China

The change of database used to produce the SEI affects the relative position of the U.S. compared to other countries, particularly China, which is a strong contender to take over the pole position that the U.S. holds as the most active country in research. This section will examine in more detail the trends in U.S. scientific output compared to that of China and Europe, in order to show the differences in the statistics produced for both editions, and to discuss the source of those differences. Key questions in this regard include the following: what results are artifacts of the databases and of the indicators, and what appear to be real trends in the global and country-level scientific systems?

Figure 9 presents the annual scientific output (number of papers) in the SCI and SSCI for Europe, the U.S. and China for the 1999–2011 period. These statistics have been pulled directly from the online version of SEI 2014. Based on these statistics, and applying a forecast based on exponential trend lines, China is foreseen to overtake the U.S. sometime around 2016 or 2017 and Europe in 2018 or 2019.[9]

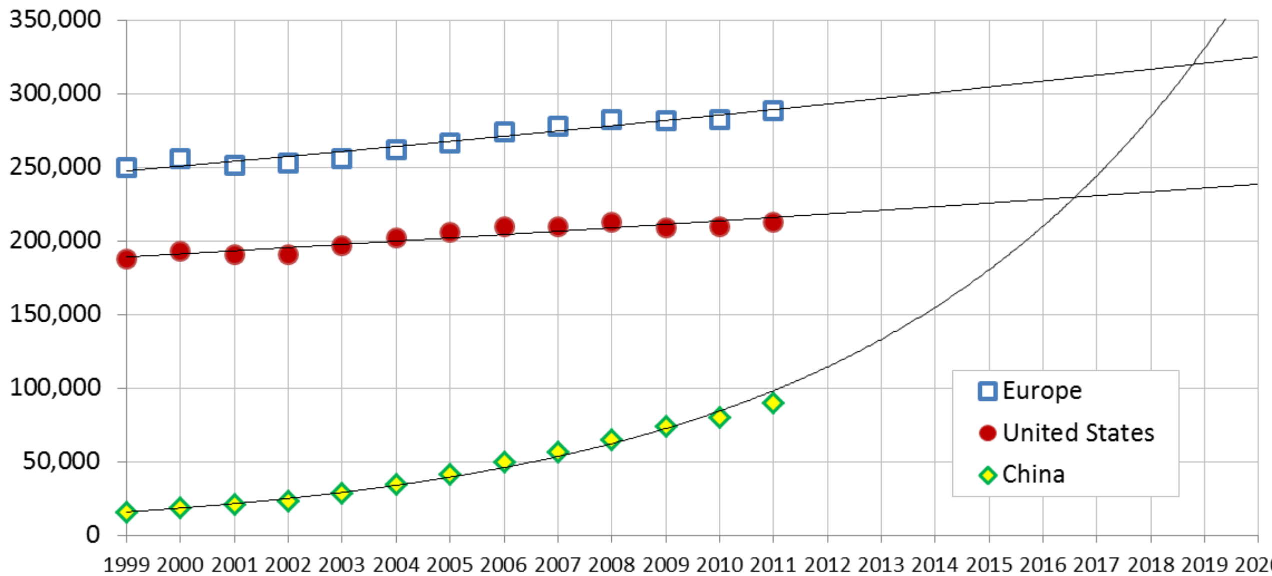

Figure 9 Trends in the number of papers from Europe, the U.S. and China in SEI 2014 (SCI & SSCI) and forecast until 2020

Note: Fractional counting of papers.
Source: Prepared by the Patent Board using SCI & SSCI

If one examines the data produced for SEI 2016, which are compiled from Scopus, it appears that China's annual output has already surpassed that of the U.S., having overtaken the U.S. in 2013 (Figure 10). Other authors have found similar results.[10,11] Moreover, the statistics produced with Scopus suggest that Europe's output is growing much faster than that of the U.S., and that it may publish twice as many papers per year as the U.S. by 2020.

---

[9] All trend lines in Figure 9 and 10 are exponential. For the United States and Europe the growth is almost constant, and therefore the trend lines appear linear. However, they are in fact exponential.

[10] Leydesdorff, L. (2012). World shares of publications of the USA, EU-27, and China compared and predicted using the new Web of Science interface versus Scopus. *El Profesional de La Informacion, 21*(1), 43–49. http://doi.org/10.3145/epi.2012.ene.06

[11] The Royal Society. (2011). *Knowledge, networks and nations: Global scientific collaboration in the 21st century* (No. 03/11 DES2096). London, UK: The Royal Society. Retrieved from http://www.interacademies.net/File.aspx?id=25069.

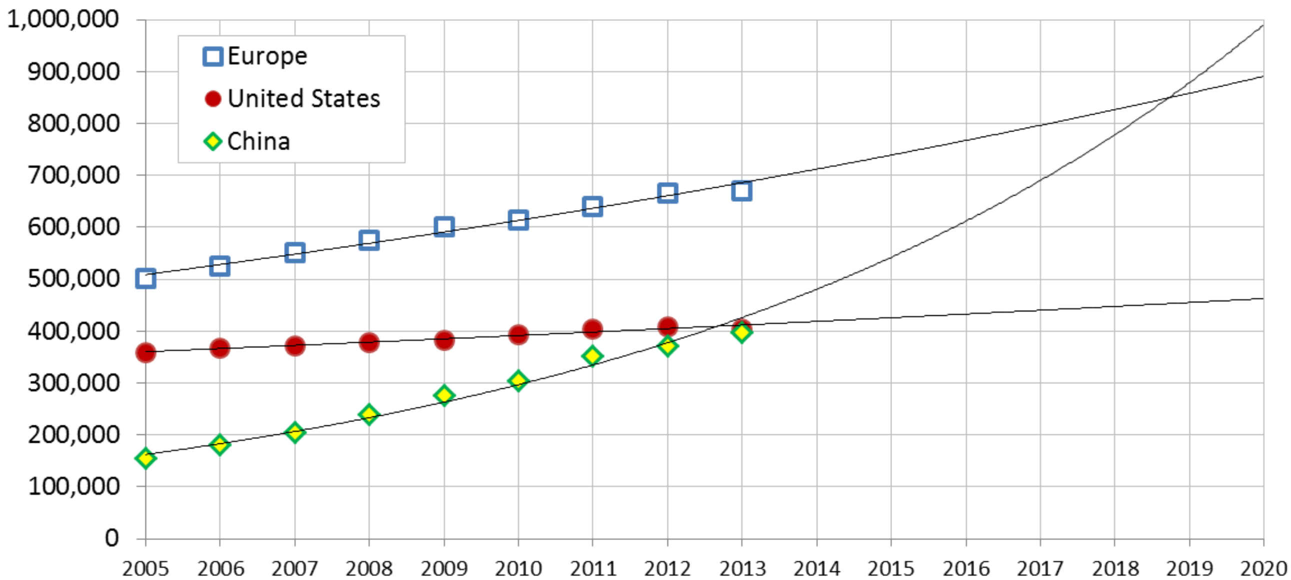

Figure 10 Trends in the number of papers from Europe, the U.S. and China in SEI 2016 (Scopus) and forecast until 2020

Note: Fractional counting of papers.
Source: Prepared by Science-Metrix using Scopus (Elsevier)

Although the predictions based on the two databases diverge somewhat, they both support the highly probable hypothesis that China will overtake the U.S. and Europe within four to five years. There are more papers indexed in Scopus, which explains the upward shift in output observed in Figure 10 compared to Figure 9: the U.S., Europe and China all have more papers in Scopus. As shown in Figure 2 the difference between the two databases is also increasing over time.

While an annual growth of about 2.5% over the previous 10 years was observed in SEI 2014, this growth was calculated as approximately 6% in SEI 2016 (Scopus). This faster growth in Scopus favors the U.S. and Europe more than it does China. The annual growth of the U.S. in recent years in SEI 2014 was about 1%, whereas it was 2% in SEI 2016. Similarly, Europe's output grew annually by 1.5% in SEI 2014, but by 4% in SEI 2016. For China, the growth as calculated in each of the two databases is roughly the same at 15%. Thus China's output overtakes that of the U.S. and Europe sooner in SEI 2016 than in SEI 2014 because it starts in a better position in SEI 2016: China has always had better coverage in Scopus.

As mentioned previously in this report, it is difficult to establish a clear picture of the linguistic and regional biases inherent in the bibliographic databases because the overall population of scientific papers is not precisely known; instead, Ulrich is used by academics as a benchmark as it provides the most representative view of that population. However, we know that Scopus has a more extensive coverage of non-English-language journals than the WoS. This would only slightly impact the measure of output for the U.S. as we can presume that American authors mostly published their results in English. But it certainly increases the representation of Europe, which includes several non-English-speaking countries. Similarly, one can reasonably suppose that Chinese researchers publish a large proportion of their papers in Chinese and that these are seldom indexed in the WoS and only marginally more frequently indexed in Scopus. Therefore, both in SEI 2014 and SEI 2016, China and Europe are relatively underrepresented compared to the U.S., and we can conclude that this underrepresentation was more pronounced in SEI 2014.

The representation of the journals' publishers in the databases can also impact comparisons between countries. Again, there is no exhaustive list of journal publishers or of the number of papers they publish. An underrepresentation of publishers in a given country will likely impact the measure of that country's scientific output. As mentioned above, for the 1999–2011 period, only 1% of the papers in the WoS are from Chinese publishers, compared to 6% in Scopus. But this is small in comparison to the number of papers authored by Chinese researchers: about 6% in the WoS and 13% in Scopus. In comparison, 45% of papers in the WoS are from U.S. journals, compared to 33% in Scopus. In any case, these proportions are high compared to the proportion of papers from U.S. authors in each database: 28% in the WoS and 24% in Scopus. European publishers are also slightly better covered in Scopus (49% of papers) than in the WoS (46% of papers).

Variations in the coverage of journals in the different fields of research can also affect indicators for each country differently. Hence a database with a better coverage of mathematics would favor a country that specializes in this field. Moreover, changes over time in the coverage of the various fields will likely impact the trends in the measure of countries' output over time. Figure 26, Figure 27, Figure 28 and Figure 29, in Annex I, present the evolution of the number of papers from the U.S., China and Europe in each Caspar field over time, for SEI 2014 and SEI 2016.

2.4 U.S. institutional sectors

This section compares the scientific output of U.S. institutional sectors, which are determined based on institution type: academic, private nonprofit, industry, federal or state/local government, or federally funded research and development centers (FFRDCs). Although statistics on publications produced by U.S. sectors have been computed in two different databases, and the coding of institutions by sector has been performed independently,[12] Figure 11 shows that the results for SEI 2014 and SEI 2016 are fairly similar.

The academic sector accounts for the majority of U.S. papers in both editions of the SEI, representing about 73% in 1999 and 76%–77% in more recent years. The increase in the academic share is somewhat greater in Scopus in the most recent years. The share of the government sectors (federal, FFRDCs and state/local) is similar in both editions. While the shares of the local/state governments (1%) and of the FFRDCs (3%) were fairly stable over time in both databases, the share of the federal government has decreased by about one percentage point during the period, from 7% to 6%. The share of industry is also rapidly decreasing in both databases, by approximately two percentage points. The share of industry is larger in Scopus, which can be explained by the importance of conference papers for industrial researchers (data not shown) and the preponderance of engineering and computer sciences in industrial research (Table IV), both being more extensively covered in Scopus.

The most important divergence for sector-level coverage is observed for the contribution of private nonprofit institutions (NFP). In SEI 2014, this sector's share was not only greater than in SEI 2016, but it slowly increased over the 1999–2011 period. In contrast, in SEI 2016, the NFP's share is lower and has markedly decreased over the same period. The concentration of NFP papers in medical sciences and biological sciences (Table IV), two fields with a smaller share in Scopus, easily explains the difference in the NFP sector's share in the two editions.

[12] Science-Metrix coded the institution for SEI 2016 based only on general definitions and without having access to the coding performed for previous editions.

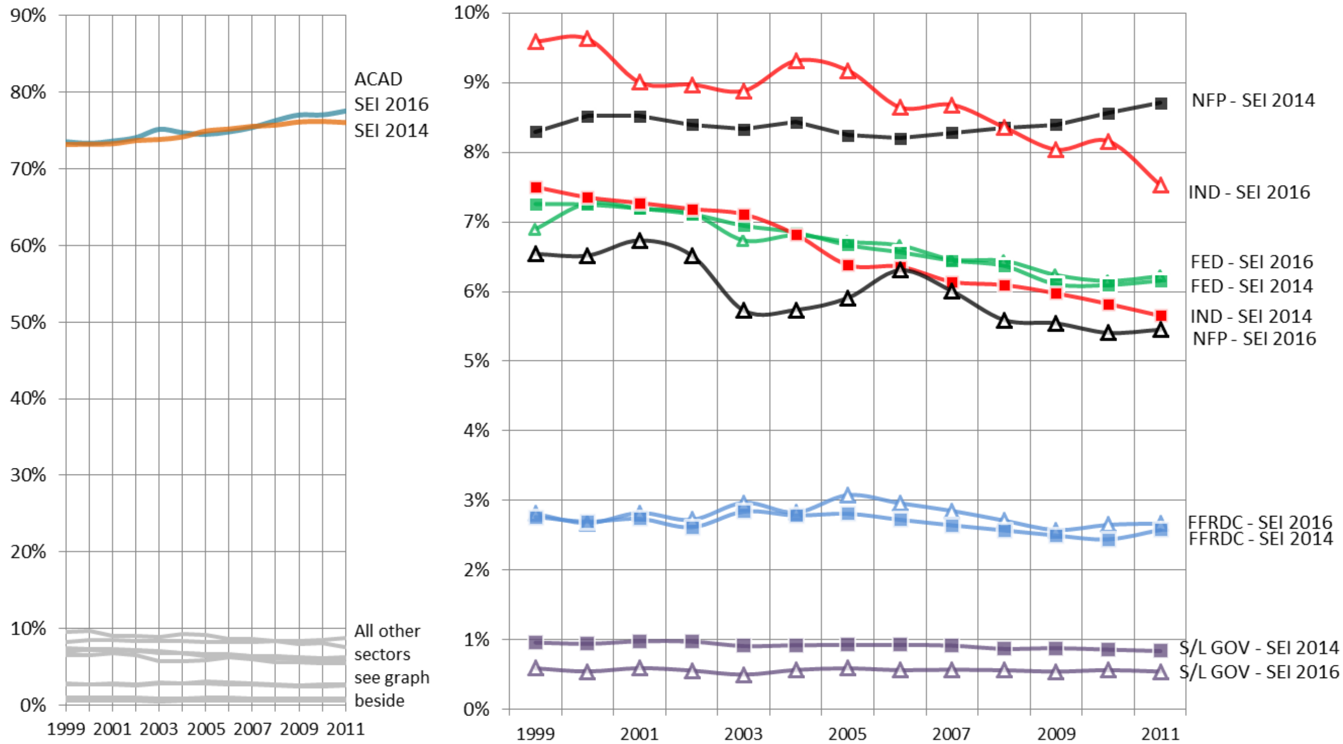

Figure 11 Share of U.S. publications by institutional sectors; a comparison between SEI 2014 (SCI & SSCI) and SEI 2016 (Scopus), 1999–2011

Note: Fractional counting of papers. ACAD = academic; NFP = private nonprofit; IND = industry; FED = federal government; FFRDC = federally funded research and development centers; S/L GOV = state/local government.

Source: SEI 2016 calculated by Science-Metrix from Scopus, SEI 2014 calculated by the Patent Board from SCI & SSCI

Table IV Distribution of U.S. scientific publications by field (Caspar) in Scopus, by sector, 1999–2011

Sector	Agricultural sci	Astronomy	Biological sciences	Chemistry	Computer sciences	Engineering	Geosciences	Mathematics	Medical sciences	Other life sciences	Physics	Psychology	Social sciences
US total	1%	1%	20%	6%	6%	12%	5%	2%	27%	2%	10%	3%	5%
Federal government	4%	2%	26%	4%	2%	13%	11%	1%	22%	1%	10%	1%	3%
Industry	1%	0%	14%	9%	10%	29%	6%	1%	11%	1%	16%	1%	1%
Academic	1%	1%	20%	6%	6%	10%	4%	3%	27%	2%	9%	4%	6%
FFRDCs	0%	3%	7%	12%	4%	23%	7%	1%	3%	0%	40%	0%	0%
Private nonprofit	0%	0%	24%	2%	1%	5%	2%	0%	55%	5%	1%	1%	3%
State/local government	1%	0%	30%	2%	1%	8%	20%	0%	29%	3%	1%	2%	3%

Note: Fractional counting of papers.
Source: Calculated by Science-Metrix from Scopus (Elsevier)

2.5 U.S. states

Adopting the Scopus database for SEI 2016 has not impacted the statistics at the state level as much as it influenced the comparison of countries. As a group, U.S. states (with the addition of Puerto Rico and the District of Columbia) are more homogeneous than the world's countries; in particular, differences in coverage by language and publishers' location affect the states more or less equally. The patterns of specialization in individual fields of research are also more similar across states than across countries, which means that the enhanced coverage of applied sciences in SEI 2016 is less discriminating at the state level. Figure 12 presents the correlation between the numbers of papers in SEI 2014 and in SEI 2016 for each state. The correlation is even higher than that observed at country level (R^2 = 0.996). Moreover, the ranking of states according to the number of papers is fairly stable (Table V). In fact, only a few minor changes can be observed, with the most notable being North Carolina's drop from 7th in SEI 2014 to 9th in SEI 2016.

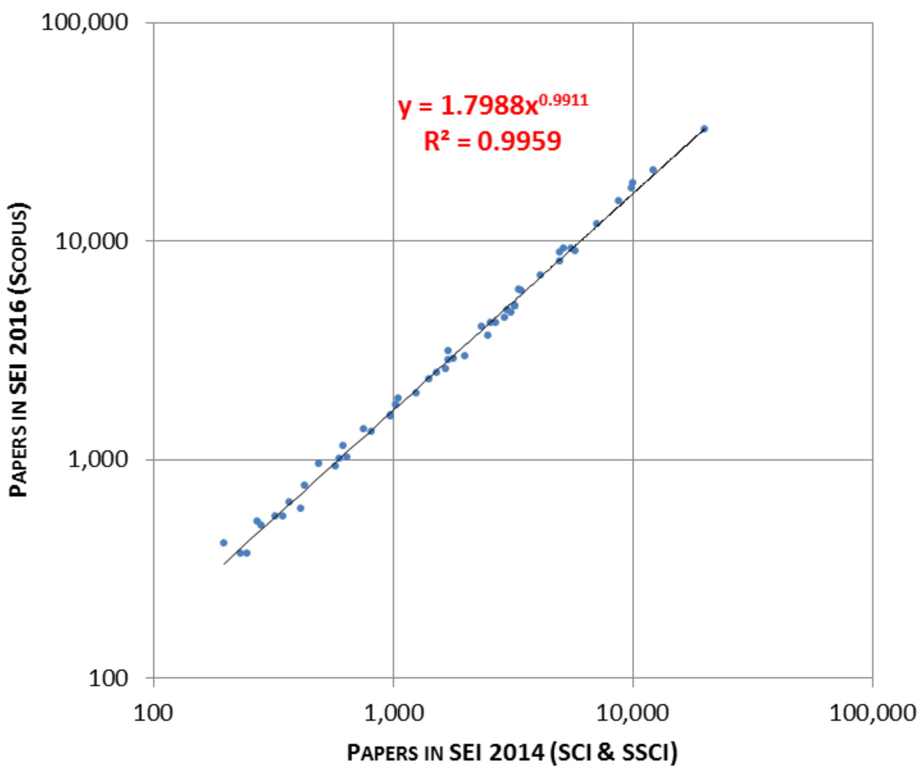

Figure 12 Correlation in papers published in 2011 by U.S. state, SEI 2014 (SCI & SSCI) versus SEI 2016 (Scopus)

Note: Fractional counting of papers.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

Table V Comparison of U.S. states' scientific output in 2011, SEI 2014 (SCI & SSCI) versus SEI 2016 (Scopus)

State	SEI 2016		SEI 2014		2016/ 2014	Δ rank	State	SEI 2016		SEI 2014		2016/ 2014	Δ rank
	n*	rk	n	rk				n*	rk	n	rk		
United States	**261,185**		**154,331**		**1.69**		**United States**	**261,185**		**154,331**		**1.69**	
California	32,394	1	19,781	1	1.64		Alabama	3,131	24	1,688	27	1.85	3
New York	20,966	2	12,075	2	1.74		Oregon	2,585	28	1,648	28	1.57	
Massachusetts	18,458	3	9,945	3	1.86		S. Carolina	2,483	29	1,509	29	1.65	
Texas	17,451	4	9,842	4	1.77		Kentucky	2,338	30	1,415	30	1.65	
Pennsylvania	15,330	5	8,723	5	1.76		Kansas	2,014	31	1,241	31	1.62	
Illinois	11,908	6	7,089	6	1.68		D.C.	1,905	32	1,044	32	1.82	
North Carolina	9,062	9	5,740	7	1.58	-2	Oklahoma	1,794	33	1,020	33	1.76	
Michigan	9,229	7	5,549	8	1.66	1	Nebraska	1,600	34	977	34	1.64	
Florida	9,187	8	5,152	9	1.78	1	Rhode Isl.	1,571	35	971	35	1.62	
Ohio	8,834	10	4,969	10	1.78		Mississippi	1,340	37	805	36	1.66	-1
Maryland	8,100	11	4,922	11	1.65		New Mexico	1,374	36	753	37	1.83	1
Georgia	6,995	12	4,084	12	1.71		N. Hampshire	1,025	39	638	38	1.61	-1
Indiana	5,894	14	3,415	13	1.73	-1	Arkansas	1,163	38	616	39	1.89	1
Virginia	5,974	13	3,332	14	1.79	1	Hawaii	1,009	40	595	40	1.70	
Washington	5,034	15	3,216	15	1.57		Delaware	926	42	572	41	1.62	-1
Missouri	5,032	16	3,202	16	1.57		Nevada	953	41	491	42	1.94	1
Wisconsin	4,713	18	3,106	17	1.52	-1	W. Virginia	759	43	425	43	1.79	
New Jersey	4,837	17	2,977	18	1.62	1	Vermont	594	45	410	44	1.45	-1
Connecticut	4,455	19	2,910	19	1.53		North Dakota	637	44	369	45	1.73	1
Colorado	4,207	21	2,673	20	1.57	-1	Montana	553	46	348	46	1.59	
Tennessee	4,214	20	2,555	21	1.65	1	Idaho	551	47	321	47	1.72	
Minnesota	3,680	23	2,483	22	1.48	-1	Maine	499	49	280	48	1.78	-1
Arizona	4,039	22	2,342	23	1.72	1	Puerto Rico	522	48	271	49	1.93	1
Iowa	2,962	25	1,980	24	1.50	-1	Alaska	374	51	245	50	1.52	-1
Louisiana	2,882	26	1,770	25	1.63	-1	Wyoming	373	52	230	51	1.62	-1
Utah	2,858	27	1,692	26	1.69	-1	South Dakota	416	50	196	52	2.12	2

Note: Fractional counting of papers. *Number of papers for the U.S. in SEI 2016 excludes fractions for unknown state. Includes Puerto Rico and the District of Columbia.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

2.6 Co-authorship

Indicators based on co-authorship are highly influenced by the quality of the information about the addresses of authors on each paper. When information is missing, erroneous or ambiguous, the computation of indicators on collaboration based on co-authorship is directly affected. Here is a hypothetical example of a paper with two authors in which the country of one author was incorrectly identified as the United Kingdom:

Author	Institution	Country
John Smith	University of Florida	United States
Helen Jackson	Yale University	United Kingdom

In this case, the paper would be erroneously counted as an international collaboration. Similarly, if only one address is listed for this paper (i.e., one address being omitted in the database), then the inter-institutional collaboration would not be counted.

It was shown previously in this report that the quality of data in Scopus is lower than in the WoS, including the important fact that some author addresses appear to be missing in Scopus. Recall that Figure 7 showed that the annual average of number of addresses per paper in Scopus follows a complex pattern during the 1999–2011 period, especially when compared to the WoS, for which the number of addresses per paper is increasing steadily and smoothly over the same period. A highly similar pattern is observed in Figure 13, where the evolution of inter-institutional and international collaboration in SEI 2016 (Scopus) and in SEI 2014 (SCI & SSCI) is examined. Consequently, it is probably safe to assume that the variation observed in co-authorship in Scopus is highly impacted by database coverage and data quality, and that it is more difficult to analyze the "real" trends in scientific collaboration using data from SEI 2016.

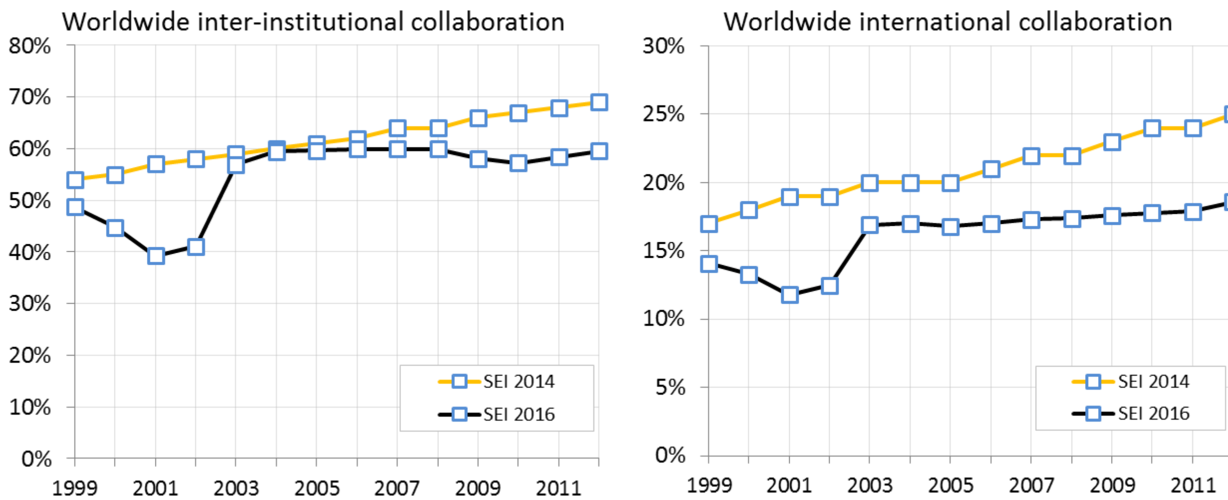

Figure 13 Trends in inter-institutional and international collaboration rates in SEI 2014 and SEI 2016, 1999–2011

Note: Inter-institutional collaboration rate = proportion of papers with authors from at least two addresses; international collaboration rate = proportion of papers with authors from at least two countries.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

When comparing the collaboration patterns of countries, these variations in the database over time should be taken into account. Figure 14 presents the trends in inter-institutional and international collaboration for China and the U.S., in both SEI 2014 and SEI 2016, for 1999 to 2011. The pattern with a "hole" between 1999 and 2003 that is observed for total inter-institutional and international collaborations in Scopus (Figure 13) is also observed for China and the U.S. individually. Here again, the collaboration rates are lower in SEI 2016 than as measured in SEI 2014. This could probably also be explained by missing addresses in Scopus. It can further be linked to a higher proportion of conference papers in Scopus, which are generally associated with fewer authors. Nevertheless, international collaboration is increasing rapidly for the U.S. in both SEI 2014 and SEI 2016, while it is fairly flat for China in both editions.

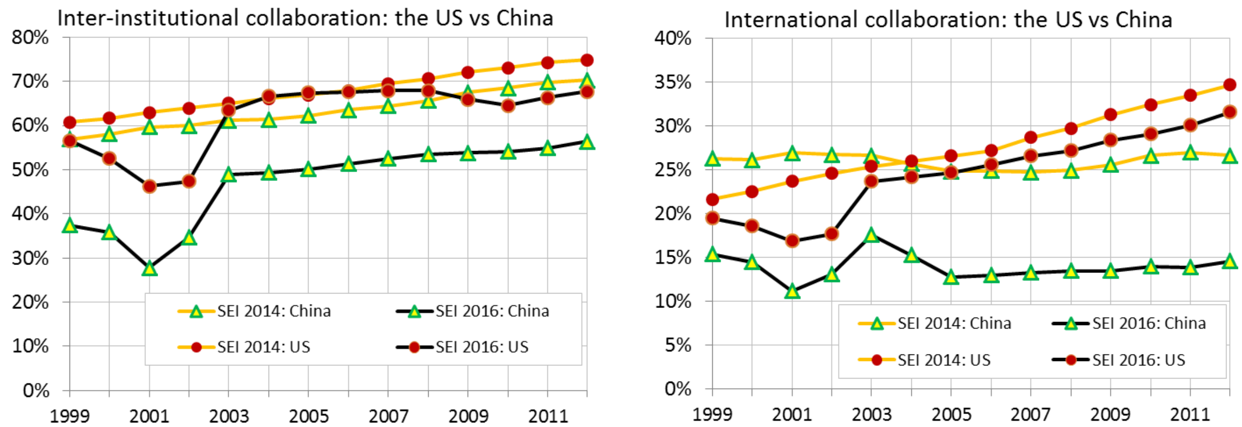

Figure 14 Trends in inter-institutional and international collaboration rates in SEI 2014 and SEI 2016, China and the U.S., 1999–2011

Note: Inter-institutional collaboration rates = proportion of papers from the given country with authors from at least two addresses; international collaboration rates = proportion of papers from the given country with authors from at least one foreign country.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

Statistics on international collaboration have also been computed at the Caspar field level in order to show the variations in collaboration patterns in various areas of research. These statistics are presented in Annex I (Figure 30).

2.7 Impact indicators

Comparing impact indicators stemming from two databases is a challenging exercise because of the many differences that can exist between the two sources of data. In particular, having different sets of documents indexed in the databases, and different matching algorithms to link references in publications to other indexed articles, could both potentially lead to notable differences when preparing impact indicators, which is why it is crucial to control for these factors.

To understand how the differences between the databases affect citation-based impact indicators, a first level of analysis was performed at the article level, on a precise set of articles. DOI information, which is a unique and standard identifier, was used to retrieve information for articles indexed in both databases. Using the set of documents that were found to be indexed in both the WoS and Scopus, a comparison of citations made to these documents was prepared (Figure 15). Overall, citations in both databases are remarkably similar, following a linear relationship with a coefficient of determination (R^2) of 0.96. The observed relationship indicates that citations in Scopus tend to be slightly higher than in the WoS. This could be expected given that Scopus indexes more documents than the WoS, resulting in a higher probability that documents citing a specific article are indexed in the databases, thus leading to a higher citation score. Nevertheless, this is not a systematic finding, as can be observed in Figure 15, with many documents obtaining slightly higher citations scores in the WoS than in Scopus.

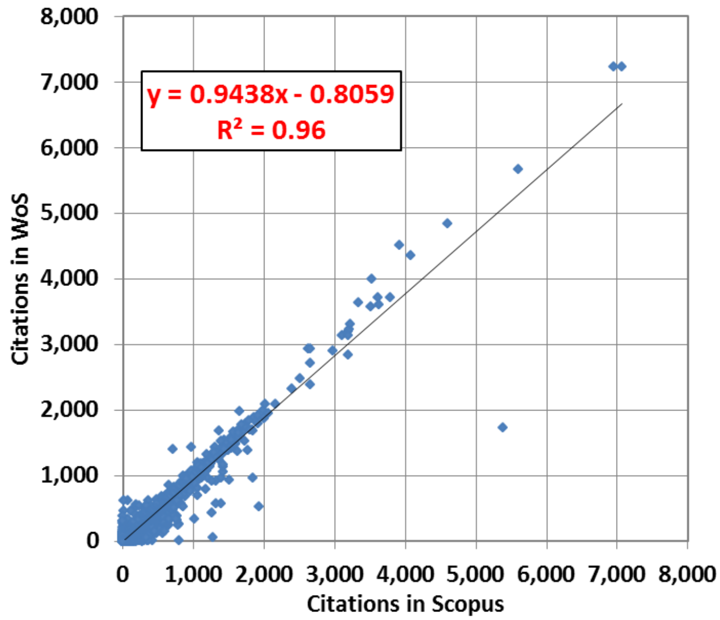

Figure 15 Citation scores in the WoS and Scopus for a set of documents indexed in both databases

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

While citations counts are of interest to assess the comparability of the two databases, the comparison of relative citation scores (Figure 16) is even more important as it addresses the issue of whether a document obtains similar levels of impact using both databases. Relative citation scores, also called impact scores, take into account the different citation patterns across fields and subfields of science, to provide a more meaningful indicator of scientific impact than citation scores alone. There is a strong correlation between impact scores from both databases ($R^2 = 0.81$), meaning that high-impact documents in the WoS also obtain high levels of impact in Scopus, while low-impact documents should obtain low scores in both databases. Consequently, using Scopus or the WoS to assess scientific impact should normally result in similar findings, although some variability may still occur.

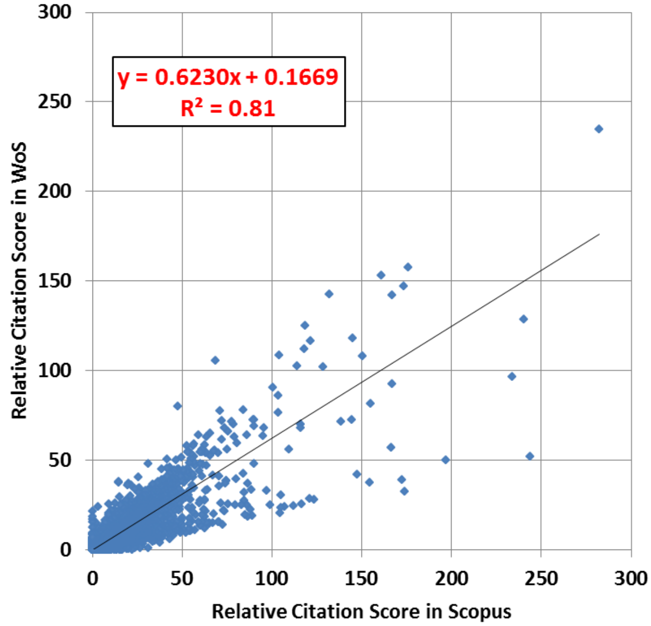

Figure 16 Relative citation scores in the WoS and Scopus for a set of documents indexed in both databases

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

Another observation made by using documents that are indexed in both databases is that relative citation scores in Scopus tend to be higher than those observed for their counterparts in the WoS, as expressed by the linear coefficient factor of 0.62 (with 1.00 representing a situation where impact scores are equal in both databases). To illustrate this, the number of documents that obtained each possible citation count was determined for each database (i.e., the frequency distribution). Then, in Figure 17, the proportion of documents associated with each number of citations was presented for Scopus (red) and the WoS (blue); only results for citation counts below 100 are displayed in this figure.

For documents with fewer than 100 citations, the WoS's share is higher for the first 10 categories (i.e., 0–9 citations), but for the remaining 91 categories, 90 obtain higher shares in Scopus. Furthermore, given the overall weight of the 0–9 citation categories in the total number of documents (40.5% in the WoS; 36.1% in Scopus), these higher shares in the WoS for lower citation counts have a direct impact on the relative citation scores, which probably explains an important part of the difference observed in Figure 16 (i.e., the trend of lower relative citations scores in the WoS). With WoS citation scores being more condensed at lower levels of citations, the relative citation scores will tend to stay lower on average, at least in comparison with what is observed in Scopus, where higher frequencies at higher levels of citations may yield higher relative citation scores. For comparison, documents cited at least 100 times represent 4.3% of all documents in the sample (which includes more than 545,000 articles) in Scopus, while these types of documents only represent 3.8% of the total for WoS. In the end, all these factors result in these sampled documents receiving about 10% more citations in Scopus than in the WoS (1.53 million in Scopus [27.9 cit./paper]; 1.4 million in the WoS [25.5 cit./paper]).

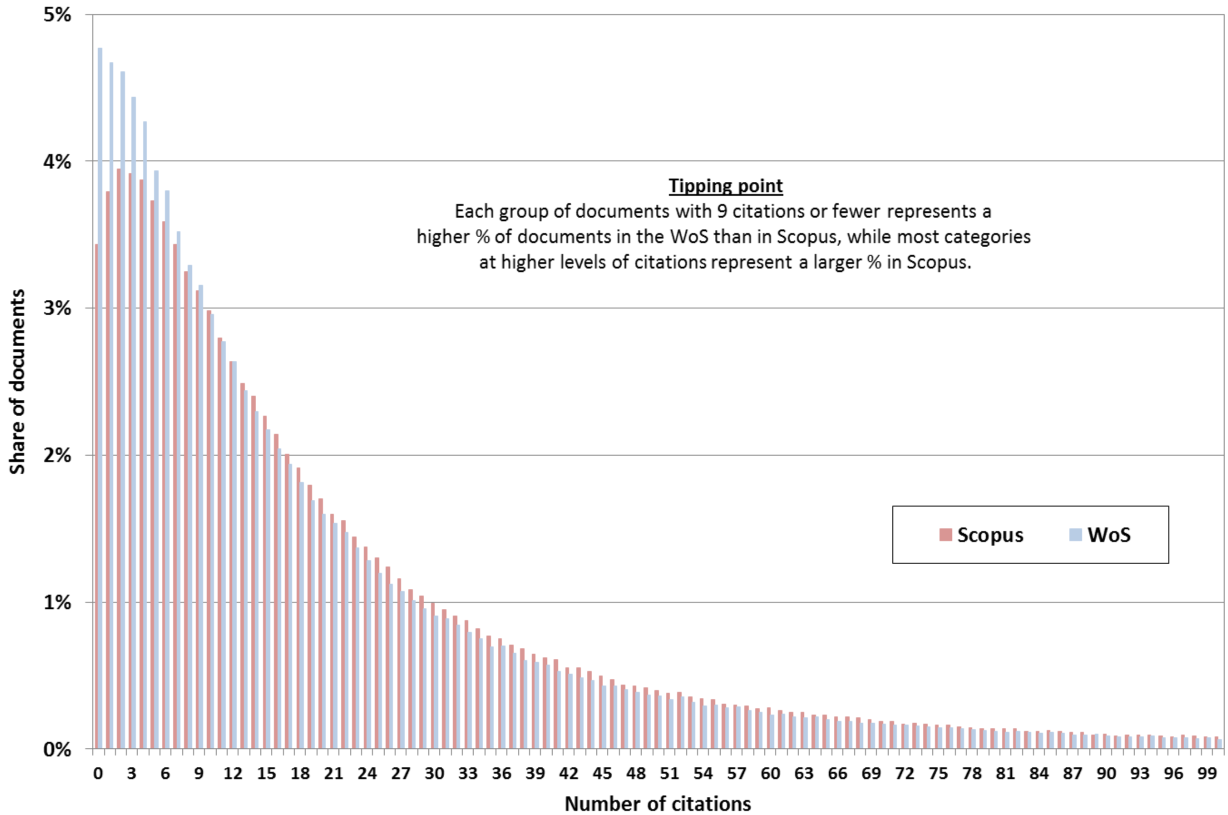

Figure 17 Citation scores in the WoS and Scopus for a set of documents indexed in both databases

Note: Only results for citation counts below 100 are displayed here. The sum across all citation counts for both databases (including those above 100 citations) adds up to 100%.

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

The differences in indexation of documents in the WoS and Scopus not only result in direct variations between SEI 2014 and SEI 2016 in terms of scientific output, as discussed in previous sections, but they also lead to complex disparities in impact measures computed for these two editions. Indeed, with citations to indexed documents being counted from references made by other indexed documents, the global sets of indexed documents in the databases are critical to the determination of impact indicators because citations coming from documents that are not indexed are not included in the computation of impact scores. Figure 18 presents trends in the WoS and Scopus databases regarding the average number of citations per article. Unsurprisingly, scores are slightly higher in the WoS than in Scopus, which is in line with indexation policies in the WoS that focus on quality (i.e., highly cited journals) over quantity, whereas Scopus instead indexes many more journals with a low number of citations. For both databases, trends over time are similar: the average number of citations starts to decrease after 2000 because more recent publications have not yet had the time to accumulate all potential citations.

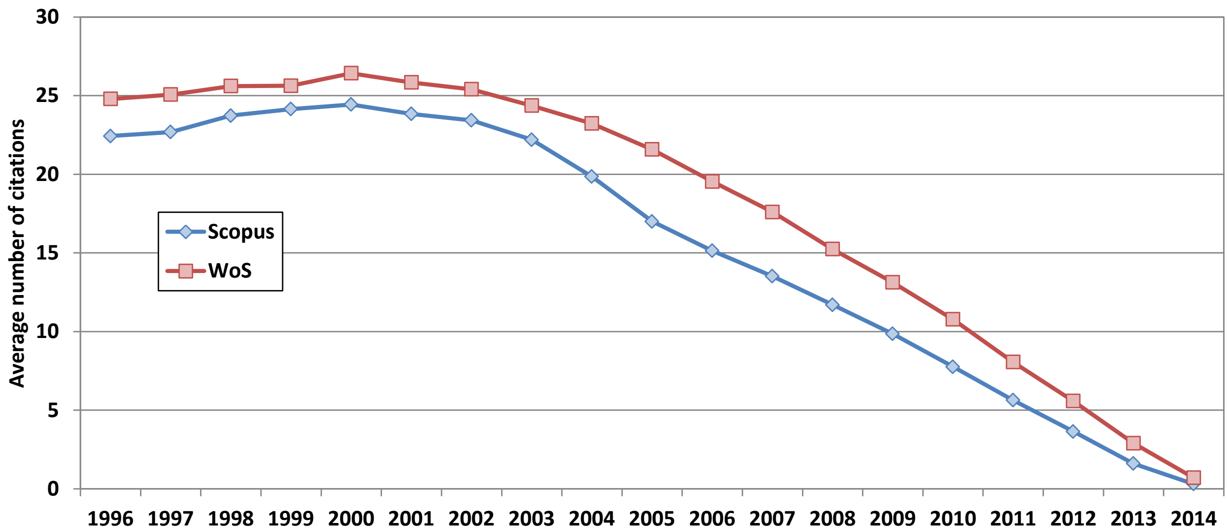

Figure 18 Average number of citations per article in the WoS and Scopus, 1996–2014

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

Another way to look at this data is to track the proportion of articles that are cited at least once in the databases. Again, this proportion is higher in the WoS than in Scopus (Figure 19), which confirms the stronger focus of the WoS on journals achieving a minimum level of citations. Indeed, for 1996 and up until 2010, about 85%–90% of articles in the WoS are cited at least once, as opposed to 69%–83% of articles in Scopus for the same period.

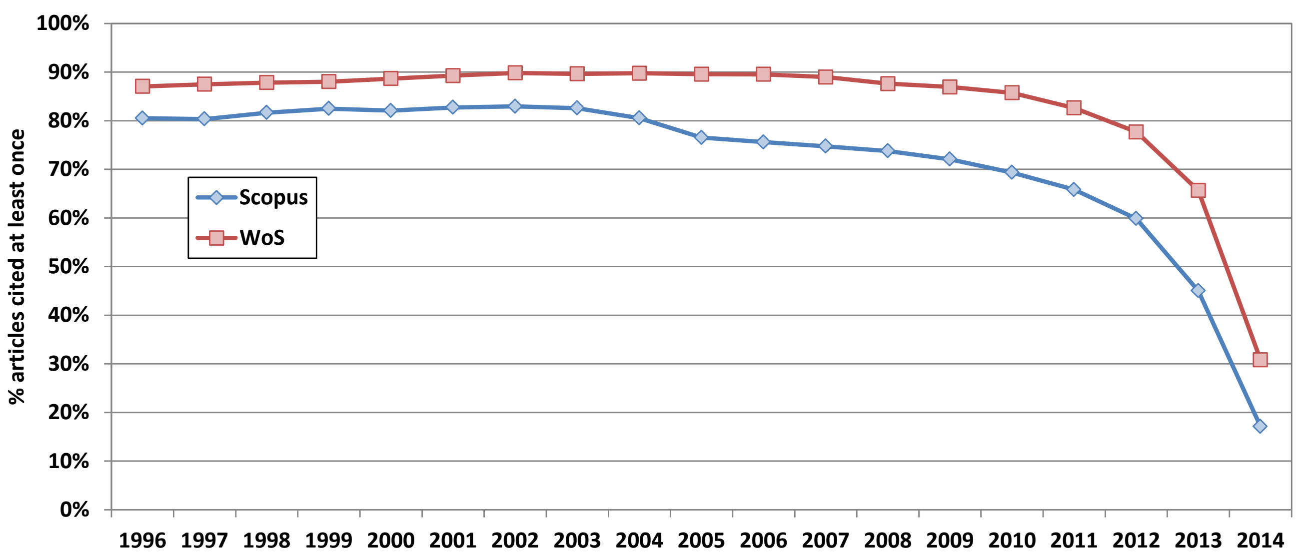

Figure 19 Proportion of articles cited at least once in the WoS and Scopus, 1996–2014

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

One notable difference between the WoS and Scopus is the coverage of papers from conference proceedings in Scopus, which are not covered in this version of the WoS. This has strong implications for the computation of impact scores because conference proceedings tend to be much less cited than journal articles. As a result, the average number of citations per document in Scopus is heavily affected. Indeed, as presented in Figure 20, conference papers published between 1996 and 2004 in Scopus received on average 7–8 citations, which is much lower than the average for articles, which stood at about 25 citations per document. In fact, if only articles are accounted for in both the WoS and Scopus, the figure clearly shows that the average number of citations is almost identical in both databases. As such, the impact of the inclusion of conference proceedings in Scopus has important implications, as it lowers the average number of citations per document.

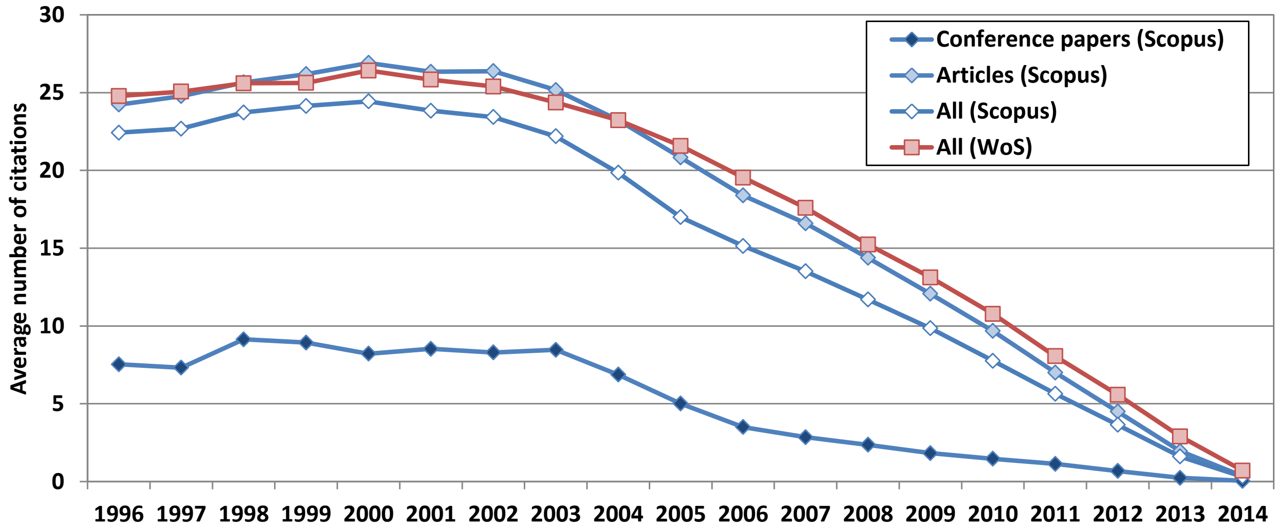

Figure 20 Average number of citations per document for articles and conference papers in Scopus and globally for the WoS, 1996–2014

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

Looking at potential regional effects, shares of citations received by countries are generally similar between the WoS and Scopus (Table VI). For 2011, the ranking of the top 10 countries in terms of citations is identical in both databases, and only a handful of other countries shift in the ranking, and most by one position. The only notable exception is Malaysia, which moves up by five positions with the change from the WoS to Scopus, passing and moving ahead of New Zealand, Mexico, South Africa, Argentina and Thailand. This striking variation is reflected in the ratio between the shares of citations in Scopus over the WoS (2011): at 1.39, it is by far the highest departure from equality within the top 40 countries. Indeed, India comes in second regarding this ratio with a score of 1.14. All 38 other countries score between 0.92 and 1.08 for this indicator, which is a strong indication of the stability of the citation share indicator across databases.

Table VI Country shares of world citations in Scopus and the WoS, 2000–2011

Country	Scopus														WoS		Scopus 2011 / WoS 2011	ΔRk
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Trend	Rk 2011	2011	Rk 2011		
United States	40.5%	40.4%	40.5%	38.2%	38.0%	36.8%	35.6%	34.2%	33.2%	32.0%	31.2%	30.1%		1	31.0%	1	0.97	-
China	1.6%	1.9%	2.3%	3.1%	3.8%	4.5%	5.4%	6.5%	7.5%	8.4%	9.3%	10.5%		2	11.4%	2	0.93	-
United Kingdom	8.2%	7.8%	8.0%	7.5%	7.3%	7.2%	7.0%	6.9%	6.7%	6.6%	6.4%	6.2%		3	6.3%	3	0.99	-
Germany	7.1%	7.2%	7.1%	6.7%	6.6%	6.7%	6.5%	6.6%	6.3%	6.2%	6.1%	6.0%		4	6.1%	4	0.98	-
Japan	7.2%	7.2%	6.9%	6.6%	6.5%	6.1%	5.8%	5.5%	5.3%	5.0%	4.6%	4.4%		5	4.5%	5	0.97	-
France	4.6%	4.5%	4.4%	4.2%	4.0%	4.1%	4.0%	4.0%	4.0%	3.9%	3.8%	3.7%		6	3.9%	6	0.97	-
Canada	3.5%	3.4%	3.4%	3.4%	3.5%	3.6%	3.6%	3.6%	3.6%	3.5%	3.5%	3.4%		7	3.4%	7	0.98	-
Italy	3.1%	3.2%	3.3%	3.2%	3.3%	3.4%	3.4%	3.5%	3.4%	3.4%	3.3%	3.3%		8	3.2%	8	1.03	-
Spain	1.9%	2.0%	2.0%	2.0%	2.1%	2.3%	2.4%	2.5%	2.5%	2.6%	2.6%	2.6%		9	2.8%	9	0.96	-
Australia	1.9%	2.0%	2.0%	2.0%	2.1%	2.1%	2.3%	2.3%	2.4%	2.4%	2.5%	2.6%		10	2.7%	10	0.97	-
India	0.8%	0.9%	1.0%	1.1%	1.2%	1.3%	1.5%	1.7%	1.8%	2.0%	2.2%	2.4%		11	2.1%	13	1.14	2
South Korea	1.0%	1.1%	1.2%	1.3%	1.5%	1.5%	1.6%	1.7%	1.9%	2.0%	2.2%	2.3%		12	2.3%	11	0.99	-1
Netherlands	2.2%	2.2%	2.2%	2.3%	2.2%	2.3%	2.2%	2.2%	2.2%	2.3%	2.3%	2.2%		13	2.2%	12	0.97	-1
Switzerland	1.7%	1.7%	1.6%	1.6%	1.7%	1.6%	1.6%	1.5%	1.5%	1.6%	1.5%	1.5%		14	1.6%	14	0.98	-
Brazil	0.6%	0.6%	0.7%	0.7%	0.8%	0.9%	1.0%	1.1%	1.2%	1.3%	1.3%	1.3%		15	1.2%	15	1.06	-
Sweden	1.6%	1.6%	1.6%	1.5%	1.4%	1.4%	1.4%	1.3%	1.2%	1.2%	1.2%	1.1%		16	1.2%	16	0.96	-
Iran	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%	0.4%	0.5%	0.6%	0.7%	0.9%	1.0%		17	0.9%	18	1.08	1
Belgium	0.9%	1.0%	0.9%	1.0%	1.0%	1.1%	1.0%	1.0%	1.0%	1.0%	1.0%	1.0%		18	1.0%	17	1.01	-1
Denmark	0.8%	0.8%	0.8%	0.9%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%		19	0.8%	19	0.97	-
Turkey	0.3%	0.3%	0.4%	0.5%	0.6%	0.6%	0.7%	0.8%	0.8%	0.8%	0.8%	0.8%		20	0.8%	20	0.97	-
Singapore	0.2%	0.3%	0.3%	0.4%	0.5%	0.5%	0.5%	0.5%	0.6%	0.6%	0.7%	0.7%		21	0.7%	21	1.00	-
Poland	0.5%	0.5%	0.5%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.7%		22	0.7%	23	1.01	1
Israel	0.9%	0.9%	0.9%	0.9%	0.9%	0.8%	0.8%	0.8%	0.7%	0.7%	0.7%	0.7%		23	0.7%	22	0.95	-1
Austria	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.6%	0.7%	0.6%	0.6%	0.6%	0.6%		24	0.6%	25	1.01	1
Russia	0.7%	0.7%	0.7%	0.7%	0.7%	0.6%	0.6%	0.5%	0.5%	0.5%	0.5%	0.6%		25	0.6%	24	0.92	-1
Finland	0.7%	0.8%	0.7%	0.7%	0.7%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.5%		26	0.6%	26	0.96	-
Greece	0.3%	0.3%	0.4%	0.4%	0.4%	0.5%	0.5%	0.6%	0.6%	0.6%	0.5%	0.5%		27	0.5%	29	1.08	2
Norway	0.4%	0.4%	0.4%	0.4%	0.4%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%	0.5%		28	0.5%	27	0.97	-1
Portugal	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.4%	0.4%	0.4%	0.5%	0.5%	0.5%		29	0.5%	28	0.98	-1
Ireland	0.2%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.4%	0.4%	0.4%	0.4%		30	0.4%	30	0.99	-
Malaysia	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%	0.1%	0.2%	0.3%	0.4%		31	0.3%	36	1.39	5
Czech Republic	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%	0.4%	0.4%	0.4%	0.4%	0.4%		32	0.4%	31	1.02	-1
New Zealand	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.4%	0.4%		33	0.4%	32	0.99	-1
Mexico	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%		34	0.3%	33	1.01	-1
South Africa	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%		35	0.3%	34	0.97	-1
Argentina	0.2%	0.2%	0.3%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%		36	0.3%	35	0.93	-1
Thailand	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%		37	0.2%	37	1.07	-
Egypt	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%	0.2%	0.2%		38	0.2%	38	1.07	-
Hungary	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%		39	0.2%	39	1.01	-
Saudi Arabia	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.1%	0.2%		40	0.2%	40	0.99	-

Note: Fractional counting of citations.
Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

In terms of the average of relative citations (ARC) scores, rankings of countries in the WoS and Scopus are mostly stable (Figure 21), where the R^2 of 0.91 indicates a clear correlation between ARC scores from both databases. As such, both databases should normally yield similar conclusions when looking at the scientific impact of countries. However, disparities still exist between the databases. Within the list of the top 40 countries with the highest level of output in 2011 (Table VII), most countries shift rank based on the ARC with the change from the WoS to Scopus, but these shifts are only by few ranks at most. A few notable exceptions include Israel (17th in the WoS, 9th in Scopus), China (26th in the WoS, 35th in Scopus), Turkey (39th in the WoS, 34th in Scopus), the Republic of Korea (29th in the WoS, 25th in Scopus) and Argentina (30th in the WoS, 26th in Scopus). Within the top seven countries, Singapore moves up three places from 5th in the WoS to 2nd in Scopus, passing ahead of Denmark (2nd in the WoS, 3rd in Scopus), the Netherlands (3rd in the WoS, 4th in Scopus) and Belgium (4th in the WoS, 5th in Scopus); Switzerland ranks 1st in both databases.

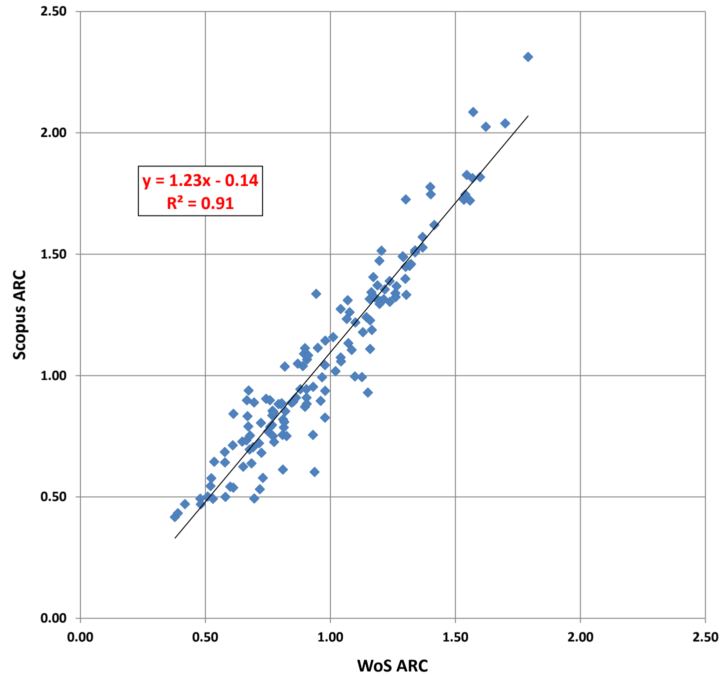

Figure 21 Correlation in average of relative citations (ARC) scores of countries in 2011, in the WoS versus Scopus

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

While the eight-rank increase observed for Israel is difficult to explain, one hypothesis for the nine-rank decrease for China is related to an issue of non-indexed Chinese references in Scopus. Database tables indexing listed references in scientific articles are filled with complex data as they contain full-text versions of these references. In order to match these references back to articles, complex algorithms are used. These rely on combinations of titles, journal names, years, volumes and any other information than can help match publications. The Scopus database is peculiar as it mentions references for which the bibliographic content could not be indexed because they come from materials indexed in characters other than those from the traditional Western Latin character set. In these cases, instead of the more frequent bibliographic information available for most references, the reference is instead only a declaration of the geographical location of the source material (e.g., Chinese source, Russian source, Thai source). This results in these references not being counted as citations to other documents as they could not be matched back onto articles in the Scopus database.

Overall, the reference table for Scopus includes just over 560,000 of these unmatched citations. China is by far the most affected country, with 300,000 citations referring to a Chinese source. Of course, not all these citations might be associated with a Chinese paper as researchers from other countries could provide Chinese characters to cite scientific articles, but it appears safe to presume that the vast majority of these should indeed be related to Chinese papers. Japan (about 124,000 citations) and Russia (about 107,000 citations) are the other two countries most affected by this situation, with a few other countries also being cited as geographical sources, but at much lower levels (e.g., Ukrainian sources [11,000 citations], Korean sources [6,000 citations], Arabic sources [4,000 citations], Greek sources [3,000 citations], Bulgarian sources [1,500 citations], Thai sources [1,000 citations], and Persian sources [600 citations]).

Table VII Average of relative citations of countries in Scopus and the WoS, 2000–2011

Country	Scopus																	WoS		ΔRk
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011		2011		
																Score	Rk	Score	Rk	
Switzerland	1.40	1.50	1.46	1.43	1.45	1.44	1.47	1.49	1.56	1.65	1.64	1.66	1.72	1.74	1.75	1.81	1	1.57	1	-
Singapore	0.97	1.02	1.08	0.92	1.06	1.13	1.16	1.29	1.24	1.48	1.38	1.43	1.57	1.60	1.76	1.75	2	1.40	5	3
Denmark	1.37	1.36	1.39	1.37	1.44	1.46	1.35	1.47	1.53	1.54	1.59	1.58	1.67	1.68	1.72	1.75	3	1.54	2	-1
Netherlands	1.32	1.38	1.36	1.40	1.39	1.38	1.33	1.45	1.42	1.48	1.52	1.59	1.59	1.63	1.71	1.72	4	1.53	3	-1
Belgium	1.17	1.17	1.14	1.16	1.17	1.19	1.21	1.26	1.31	1.41	1.40	1.46	1.51	1.52	1.56	1.62	5	1.42	4	-1
Sweden	1.34	1.31	1.33	1.29	1.28	1.36	1.33	1.35	1.35	1.38	1.41	1.43	1.48	1.52	1.53	1.57	6	1.37	6	-
United Kingdom	1.23	1.26	1.25	1.27	1.30	1.28	1.30	1.32	1.34	1.39	1.40	1.43	1.46	1.47	1.49	1.53	7	1.37	7	-
Australia	1.20	1.19	1.24	1.18	1.18	1.20	1.21	1.21	1.28	1.28	1.32	1.33	1.43	1.41	1.45	1.52	8	1.34	9	1
Israel	1.24	1.22	1.25	1.20	1.18	1.27	1.20	1.21	1.25	1.33	1.34	1.30	1.39	1.36	1.45	1.51	9	1.20	17	8
Norway	1.10	1.10	1.18	1.17	1.17	1.24	1.26	1.31	1.31	1.34	1.33	1.34	1.44	1.37	1.38	1.51	10	1.34	8	-2
Canada	1.30	1.30	1.33	1.29	1.29	1.29	1.23	1.28	1.34	1.35	1.38	1.37	1.43	1.43	1.43	1.49	11	1.29	14	3
Austria	1.26	1.08	1.07	1.14	1.14	1.09	1.07	1.18	1.22	1.24	1.25	1.28	1.28	1.31	1.46	1.46	12	1.32	10	-2
United States	1.41	1.42	1.43	1.42	1.41	1.44	1.41	1.40	1.40	1.42	1.43	1.44	1.44	1.44	1.44	1.46	13	1.32	11	-2
Finland	1.21	1.27	1.25	1.22	1.23	1.25	1.29	1.26	1.18	1.23	1.31	1.34	1.37	1.33	1.45	1.45	14	1.32	12	-2
Ireland	1.05	1.18	1.15	1.18	1.10	1.21	1.12	1.18	1.16	1.25	1.34	1.40	1.40	1.42	1.44	1.45	15	1.30	13	-2
New Zealand	1.28	1.18	1.21	1.07	1.20	1.09	1.07	1.17	1.12	1.20	1.26	1.26	1.30	1.33	1.38	1.39	16	1.24	16	-
Italy	0.96	0.96	1.02	1.01	1.01	1.01	1.05	1.02	1.09	1.17	1.16	1.20	1.25	1.27	1.33	1.37	17	1.19	19	2
Germany	1.07	1.05	1.06	1.07	1.08	1.09	1.09	1.12	1.13	1.18	1.20	1.26	1.25	1.28	1.32	1.37	18	1.26	15	-3
Spain	0.92	0.92	0.96	0.97	0.97	0.97	0.99	1.00	1.05	1.10	1.12	1.16	1.19	1.23	1.26	1.32	19	1.16	21	2
France	1.05	1.04	1.06	1.04	1.04	1.04	1.05	1.07	1.08	1.17	1.17	1.19	1.21	1.22	1.26	1.31	20	1.20	18	-2
Greece	0.85	0.89	0.94	0.88	0.92	0.93	0.95	0.99	1.00	1.06	1.10	1.13	1.15	1.18	1.21	1.23	21	1.07	22	1
Portugal	0.86	0.89	0.88	1.08	0.94	0.99	1.02	0.97	1.07	1.11	1.17	1.17	1.18	1.24	1.15	1.19	22	1.17	20	-2
Saudi Arabia	0.56	0.57	0.62	0.57	0.58	0.61	0.67	0.60	0.62	0.62	0.64	0.75	0.72	0.85	1.00	1.16	23	1.01	23	-
South Africa	0.76	0.77	0.86	0.89	0.83	0.84	0.79	0.84	0.93	1.05	1.06	1.01	1.11	1.06	1.16	1.14	24	0.98	24	-
Rep. of Korea	0.79	0.79	0.82	0.84	0.91	0.89	0.92	0.90	0.90	0.94	0.89	0.92	0.96	1.03	1.07	1.09	25	0.89	29	4
Argentina	0.69	0.76	0.80	0.72	0.73	0.73	0.75	0.82	0.88	0.89	0.97	0.92	1.00	0.95	0.98	1.04	26	0.89	30	4
Thailand	0.85	0.93	0.79	0.80	0.79	0.71	0.76	0.90	0.95	1.03	0.89	0.96	0.88	1.03	0.89	0.94	27	0.90	28	1
Czech Republic	0.61	0.65	0.64	0.63	0.68	0.64	0.63	0.68	0.72	0.74	0.81	0.79	0.84	0.85	0.93	0.94	28	0.98	25	-3
Japan	0.83	0.82	0.84	0.82	0.81	0.84	0.82	0.83	0.85	0.84	0.85	0.88	0.88	0.86	0.87	0.91	29	0.90	27	-2
Mexico	0.75	0.75	0.75	0.76	0.78	0.76	0.73	0.72	0.78	0.78	0.76	0.79	0.82	0.81	0.81	0.89	30	0.81	31	1
Egypt	0.62	0.65	0.67	0.62	0.65	0.61	0.68	0.67	0.69	0.80	0.79	0.81	0.82	0.77	0.83	0.88	31	0.79	32	1
Poland	0.58	0.59	0.62	0.60	0.61	0.61	0.57	0.61	0.64	0.65	0.66	0.73	0.68	0.73	0.76	0.85	32	0.77	33	1
Iran	0.73	0.70	0.92	0.81	0.74	0.74	0.73	0.79	0.73	0.74	0.79	0.83	0.81	0.87	0.90	0.84	33	0.77	35	2
Turkey	0.63	0.64	0.65	0.62	0.71	0.68	0.69	0.69	0.76	0.80	0.79	0.88	0.88	0.87	0.80	0.83	34	0.67	39	5
China	0.58	0.61	0.58	0.63	0.65	0.62	0.70	0.75	0.73	0.70	0.72	0.76	0.78	0.78	0.81	0.83	35	0.98	26	-9
Brazil	0.79	0.73	0.77	0.75	0.79	0.74	0.78	0.77	0.79	0.86	0.77	0.79	0.81	0.79	0.77	0.80	36	0.72	37	1
Malaysia	0.71	0.64	0.73	0.70	0.65	0.73	0.70	0.74	0.77	0.68	0.67	0.74	0.65	0.72	0.75	0.80	37	0.77	36	-1
India	0.60	0.60	0.61	0.60	0.67	0.65	0.67	0.71	0.75	0.79	0.82	0.83	0.82	0.82	0.78	0.75	38	0.77	34	-4
Romania	0.51	0.54	0.57	0.52	0.58	0.55	0.60	0.61	0.67	0.67	0.67	0.65	0.64	0.54	0.59	0.72	39	0.72	38	-1
Russia	0.41	0.46	0.43	0.44	0.44	0.44	0.42	0.48	0.48	0.50	0.54	0.55	0.54	0.51	0.54	0.58	40	0.52	40	-

Note: Fractional counting of citations.
Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

This situation highlights one potential issue with the computation of impact scores: for countries listed above—but especially China, Russia and Japan—impact scores might be slightly underestimated as the current version of the Scopus reference table potentially leaves out citations made to publications from these countries. As noted above, China's relative citation scores are heavily impacted by the change from the WoS to Scopus. However, initial simulations tend to demonstrate that although the inclusion of these citations would increase the ARC of China, the impact would be limited to at most 1%–2%; this finding is based on the assumption that all these unmatched citations would be made to Chinese papers, which is certainly not the case. As such, it appears safe to conclude that the effect of these unmatched citations is limited. Further investigation would be needed to explain the difference in impact scores observed for China. However, the important aspect here is do these differences affect the main conclusions for the performance of China, especially in comparison to the U.S.?

Providing more detail on the performance of both countries in terms of world citation shares, Figure 22 presents trends over time for both countries. As is presented, both databases result in mirror images for these countries, as the decreasing trend for the U.S. and the increasing trend for China are identically replicated in both databases, with curves superimposed on one another. This represents an important finding as it clearly confirms that the findings presented in SEI 2016 regarding the impact of the U.S. and China are most probably real and not artifacts related to the selection of the database.

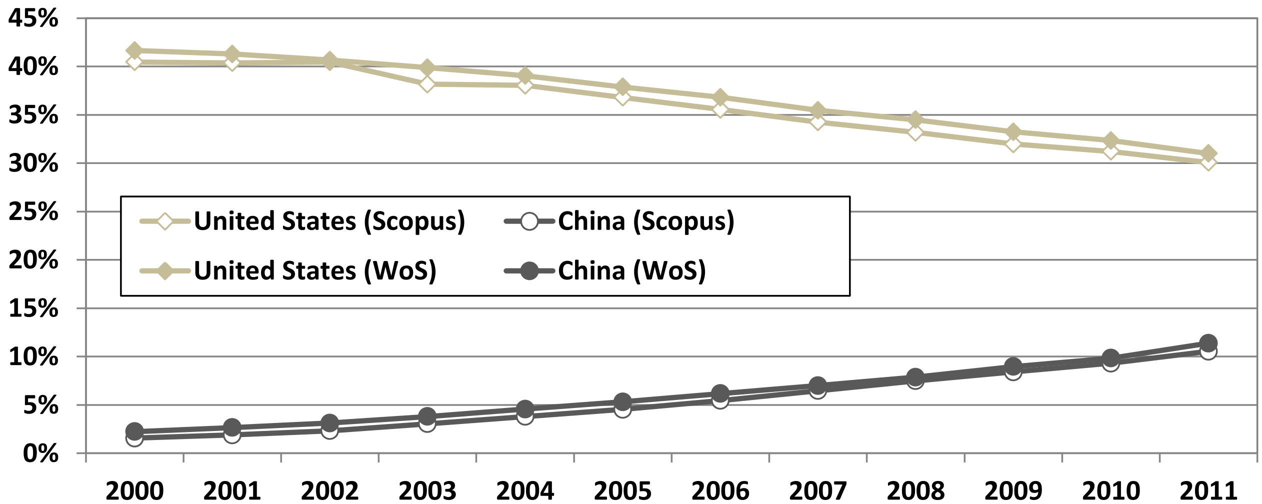

Figure 22 World citation shares of China and the U.S. in the WoS and Scopus, 2000–2011

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

Unsurprisingly, the similar patterns in terms of shares of citations (Figure 22) result in similar patterns for the ARC of both countries (Figure 23): the U.S. ARC score remains mostly stable over time in Scopus and the WoS, while that of China is on the rise in both cases, even almost reaching parity with the world level in 2011 in the WoS. Note that similar results (data not presented) are observed when relative citation scores are normalized by document types, which indicates that neither China nor the U.S. is penalized in terms of impact scores because of the distribution of its output across categories of documents (e.g., journal articles, conference proceedings).

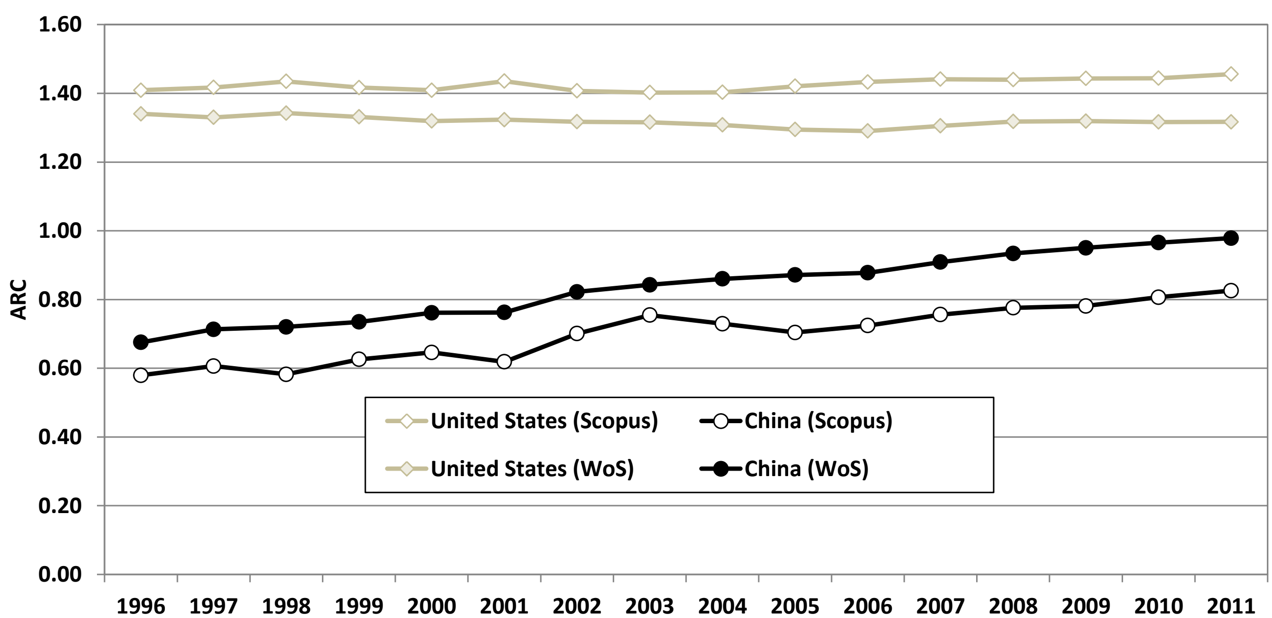

Figure 23 Average of relative citations of China and the U.S., 1996–2011

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the Web of Science (Thomson Reuters)

3 Conclusion

For the 2016 iteration, for the first time in more than 40 years, an entirely different team was selected to compute the bibliometric statistics for the SEI. The 2016 statistics were also computed using the Scopus database for the first time, having historically been computed using data from the scientific publications indexed in the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI), two of the databases contained within the Web of Science (WoS).

This report provides a review of the differences between the tables prepared for SEI 2014 and the new set of tables prepared for SEI 2016. Various avenues were explored to explain the observed differences, including assessing potential variations in the methods of data standardization and in the way the indicators were computed. The report also provides a thorough review of the difference between Scopus and the WoS; however, as the purpose of this report was to compare SEI 2014 and SEI 2016, the results presented here should be used solely in the context of understanding how the change of dataset affects the SEI statistics, not as a comparison of the two commercial offerings.

The version of Scopus used to prepare bibliometric data for SEI 2016 indexes more than 17,000 journals, 3.4 times more than the 5,000 or so journals indexed in the version of the SCI and SSCI previously used. As a result, 22.6 million papers were used for the compilation of bibliometric statistics in SEI 2016, compared to 10.5 million in the previous edition.

A second important difference between Scopus and the SCI and SSCI in terms of impact upon the statistics is that Scopus includes peer-reviewed conference papers. The inclusion of conference papers in SEI 2016 (which were not included in SEI 2014) also has a clear impact on the total coverage, as it results in a notable increase in the share of papers in computer sciences and in engineering. Thus, because of the inclusion of conference papers in Scopus, countries (or other entities) that are specialized in computer sciences and/or engineering will perform relatively better for indicators based on publication counts in SEI 2016. This is the case for China, for example.

These databases, although offering a large representation of the overall pool of papers published, do not provide an entirely neutral representation of the entire population of peer-reviewed journals. Both Scopus and the WoS have a bias toward highly cited journals, toward larger journals, and larger publishers, and a bias for articles written in English and published in English-language journals. Journals also still demonstrate country biases. Researchers tend to publish more frequently in local journals, although this trend is decreasing over time, in part due to the globalization of science (i.e., increased mobility, electronic media and the Internet) and in part due to the concentration of ownership of journals by large editors following takeovers. More than 50% of the journals in Scopus have a strong country bias. Therefore, the selection of journals that are indexed in a database can directly influence the regional bias of bibliometric indicators. For example, compared to the WoS, there are more than 700,000 additional papers in Chinese journals in Scopus, and this yields a substantial increase in the coverage of publications from Chinese authors.

The quality of data recorded in the SCI and SSCI is generally higher than that in Scopus, whereas Scopus indexes substantially more journals than the SCI and SSCI, including more journals from small editors and journals published in languages other than English, as well as book series and the aforementioned conference proceedings. This comes somewhat at the expense of data quality.

Notable variations in data quality between the two data sources include the following:

- While the country is provided for all addresses in the WoS, the country is missing for about 10% of addresses in Scopus.
- The state is provided for all U.S. addresses in the WoS, but is not readily available in Scopus.
- The city and postal code have been parsed in the WoS but not in Scopus.

- Journal names and ISSN have been thoroughly standardized in the WoS, but only partially in Scopus.
- Volume, issue and pages are more standardized in the WoS than in Scopus.
- Scopus contains several documents for which the document type is incorrect and that are erroneously counted as peer-reviewed papers.

In many cases, even given the differences observed between the two data sources, comparisons show that fairly similar and consistent results are obtained between the two editions, at least when comparing statistics computed on large numbers, a finding that coheres with what had been previously demonstrated in the literature. In particular, countries' scientific production is similar when measured in both databases (R^2 = 0.97). Nevertheless, divergences can be observed for some countries. As an example, for the two most active countries—namely, the U.S. and China—the U.S. has more extensive coverage in the SCI and SSCI, while China is more extensively covered in Scopus.

Although important differences are observed in the number of papers by country, what matters more is the relative ranking of these countries. The comparison of countries' output shows that the ranking of the top five countries (the U.S., China, Japan, Germany and the United Kingdom) does not change in SEI 2016 compared to SEI 2014. While the relative rankings do change for most other countries, in general these changes are small and are not relevant in a policy context.

Based on statistics on the number of publications from SEI 2014, which are compiled from the SCI and SSCI, and applying a forecast using exponential trend lines, China's scientific output is foreseen to overtake the U.S. output sometime around 2016 or 2017, and Europe's output in 2018 or 2019. Based on the same statistics from SEI 2016, which are compiled from Scopus data, China reached effective parity with the U.S. in terms of scientific output in 2013. Furthermore, these latest statistics suggest that Europe's output is growing much faster than that of the U.S., and that it may publish twice as many papers as the U.S. as of 2020. China's annual growth is roughly the same in the two databases, at 15%. It is on track to overtake the U.S. and Europe sooner according to SEI 2016 than according to SEI 2014 because China has always had better coverage in Scopus.

When comparing the statistics compiled on U.S. institutional sectors, the results are fairly similar between SEI 2014 and SEI 2016. The academic sector grew slightly faster in Scopus in the most recent years and the share of papers by industry is larger in Scopus, probably because of the importance of conferences and the preponderance of engineering and computer sciences in industrial research, both being more extensively covered in Scopus.

The most important divergence for U.S. institutional sectors between the two editions is observed for the contribution of private nonprofit institutions (NFP). In SEI 2014, this sector's share was not only greater than in SEI 2016, but it slowly increased over the 1999–2011 period. In contrast, in SEI 2016, the NFP's share is lower and has markedly decreased over the same period. The concentration of NFP papers in medical sciences and biological sciences, two fields with a smaller share in Scopus, has been suggested as an explanation for these differences.

Adopting the Scopus database for SEI 2016 has not impacted the statistics at the state level as much as it influenced the comparison of countries. As a group, U.S. states are more homogeneous than the world's countries; in particular, differences in coverage by languages and publishers' location affect the states more or less equally. The patterns of specialization in individual fields of research are also more similar across states than across countries, which means that the enhanced coverage of applied sciences in SEI 2016 is less discriminating at the state level.

Indicators based on co-authorship are highly influenced by the quality of the information about the addresses of authors on each paper. This report shows that the quality of data in Scopus is lower, including the important fact that some author addresses appear to be missing in Scopus. Thus, the annual average of number of addresses per paper recorded in Scopus follows a complex pattern during the 1999–2011 period, which appears to be linked to an inconsistent indexation policy. A highly similar pattern is observed for the evolution of inter-institutional and international collaboration in Scopus. These fluctuating patterns are not observed when using the WoS data. Thus, it is probably safe to assume that the variation observed in co-authorship in Scopus is highly affected by database coverage and data quality, and therefore it is more difficult to analyze the "real" trends in scientific collaboration using data from SEI 2016.

Overall, the numbers of citations received by papers in both databases are similar, following a strong linear relationship, where papers in Scopus tend to be cited slightly more than in the WoS. This could be expected given that Scopus indexes more documents than the WoS, resulting in a higher probability that documents referencing a specific article are indexed in the former database, thus leading to a higher citation score. However, when aggregated, shares of citations received by an entity are nearly the same in the WoS and Scopus. For example, the top 10 countries in terms of citations remain unchanged in both databases, and only a handful of countries move in the ranking, and most by only one rank.

Similar results are observed when using a more robust indicator of scientific impact, the average of relative citations (ARC). In this case, the regression indicates an even higher correlation between ARC scores computed from the two databases. As such, both databases will normally yield similar conclusions when looking at the scientific impact of countries.

A more detailed analysis of the performance of China and the U.S. in terms of world citation shares shows that two databases provide results that are close to mirror images for these countries: the decreasing trend for the U.S. and the increasing trend for China are identically replicated in both databases. This clearly confirms that the findings presented in SEI 2016 regarding the impact of the U.S. and China are most probably real and not artifacts related to the journal selection policies of each database.

In conclusion, this report presents information that should be used to make a more meaningful interpretation of the bibliometric indicators compiled for SEI 2016. The report also confirms that both databases can be used for the production of robust bibliometric data, if the statistics are compiled and analyzed with a careful knowledge of the database coverage. The Scopus dataset used for SEI 2016 is certainly more comprehensive than the SCI and SSCI dataset used in SEI 2014 and the previous editions. However, the use of the SCI and SSCI up to the SEI 2014 edition instead of the full dataset of publications that Thomson Reuters can offer is linked to an historical decision to use the SCI and SSCI, in the interest of obtaining a stable set of statistics. A more complete version of the WoS, including the dataset on conference papers as well as the additional journals in the Science Citation Index Expanded, would be more comparable to Scopus in terms of comprehensiveness. Thomson Reuters plans to increase their offering substantially in the coming years, and if this materializes one could expect to obtain only slight differences between the two databases.

Therefore, in the future, two elements are important to consider for the production of robust bibliometrics for the SEI: the level of representativeness of the sample, and the quality of the metadata. Both databases offer a very large, but non-random sample of the population of peer-reviewed papers published every year. As this report demonstrates, these samples are not unbiased. English-language journals are favored over journals in other languages, some research fields are more extensively covered than others, and so on. The entire population of peer-reviewed papers is not known, and it is thus very difficult to characterize what is present in Scopus and in the WoS. Hence, gaining more knowledge about this population is advisable in order to obtain more representative samples.

Data quality also has a direct impact on the quality of the bibliometric indicators that can be compiled. In this respect, Scopus performs poorly compared to the WoS, but this may improve over time. It is also possible that the quality of Thomson Reuters' product will drop as a result of ongoing content expansion. Therefore, it is recommended that representativeness and data quality be assessed before selecting a database for the compilation of bibliometric indicators, and because the coverage policies of data providers evolve rapidly, this assessment could be performed for each SEI edition.

Annex I – Additional tables

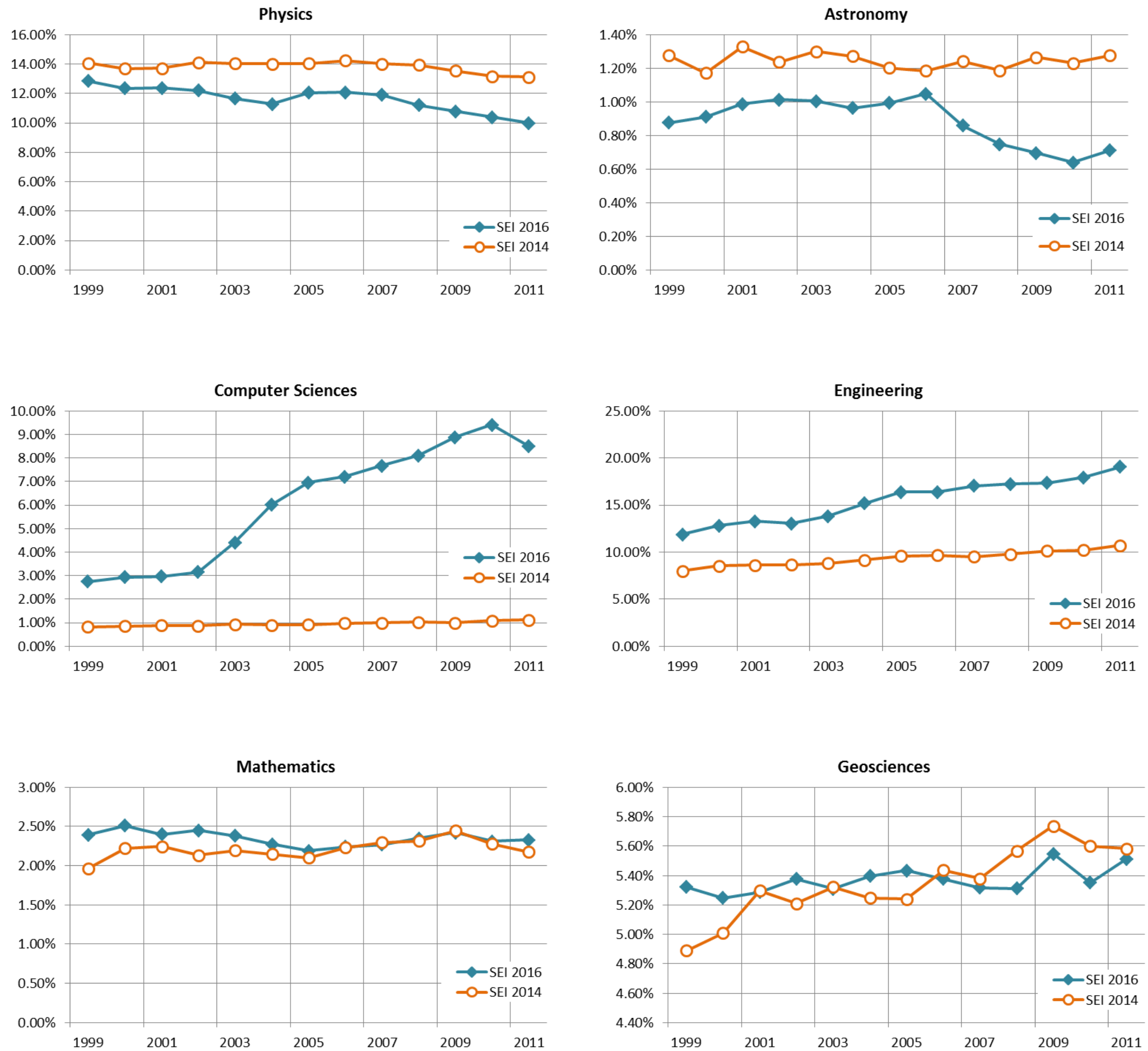

Figure 24 Trends in the share of papers by Caspar field in SEI 2014 and SEI 2016 (1 of 2)

Note: Fractional counting of papers.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

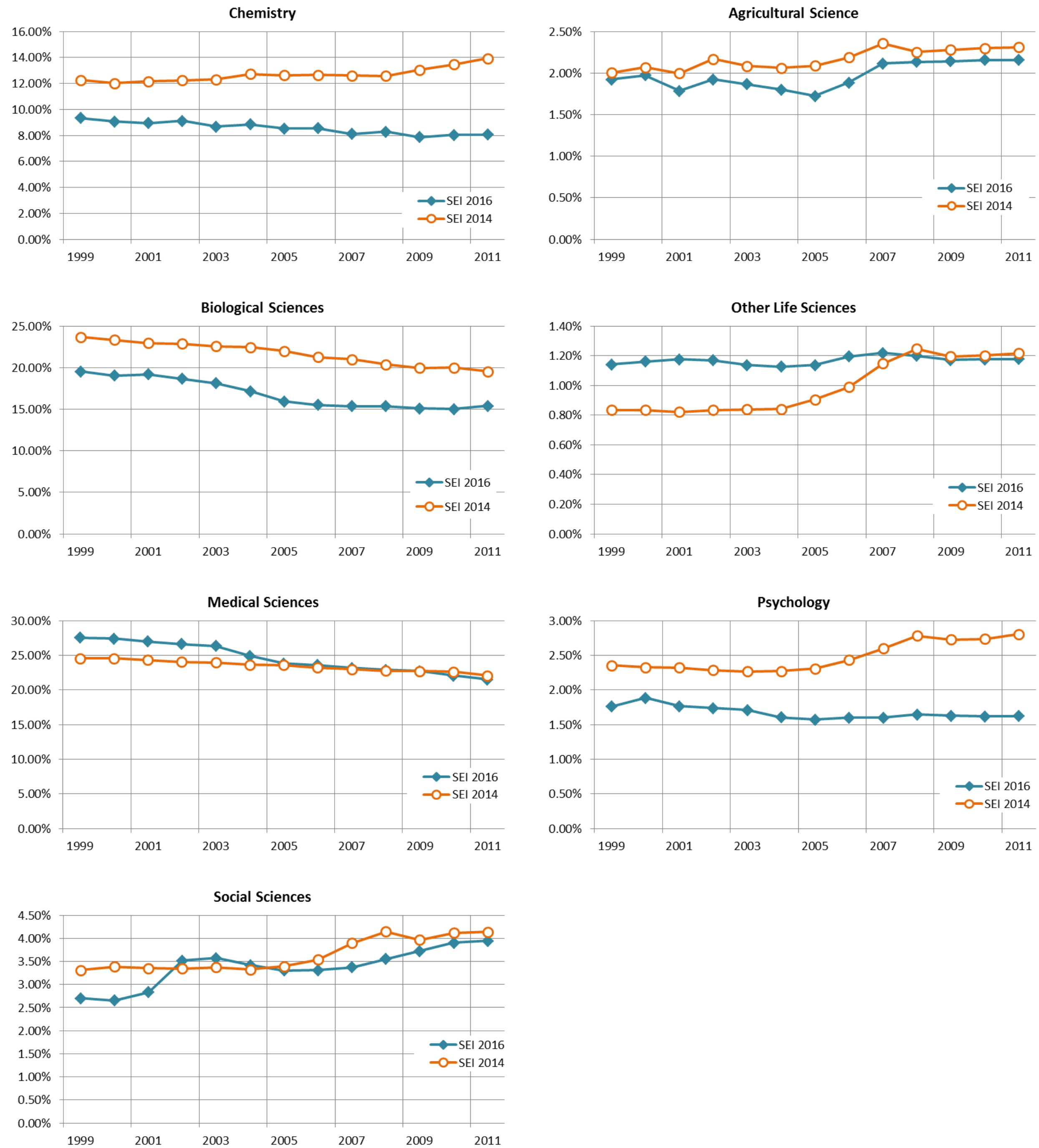

Figure 25 Trends in the share of papers by Caspar field in SEI 2014 and SEI 2016 (2 of 2)

Note: Fractional counting of papers.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

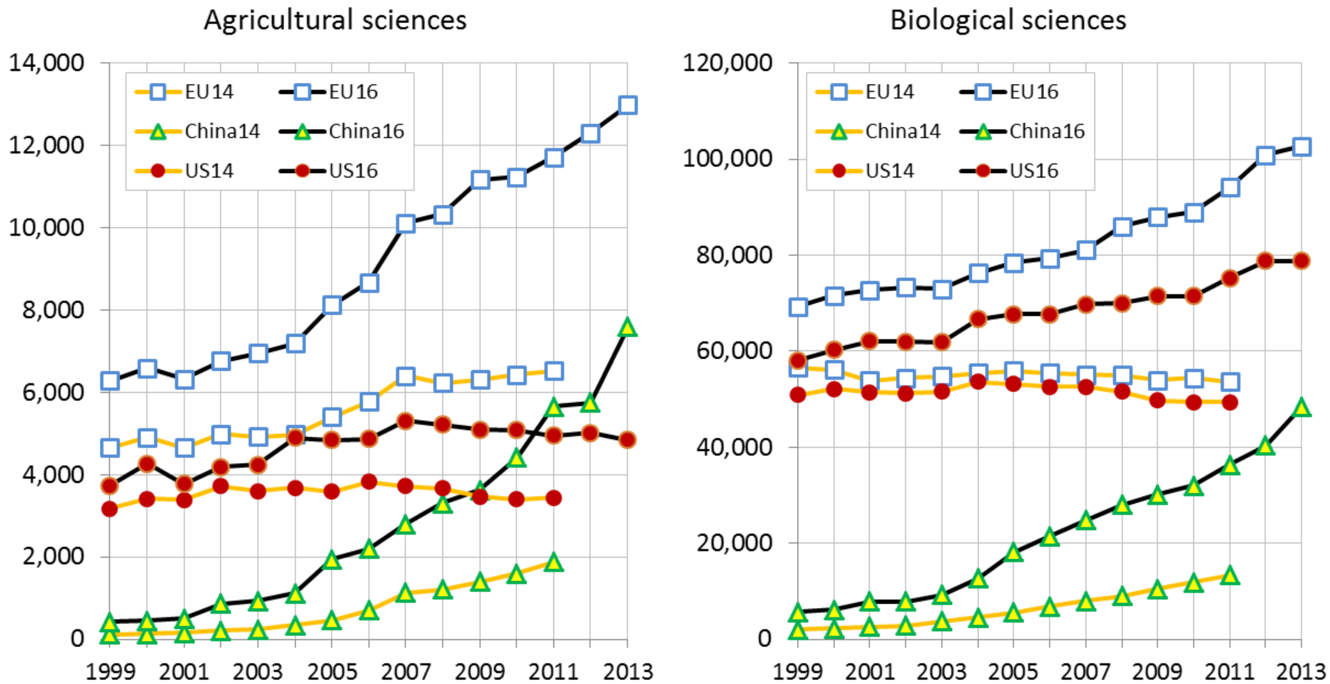

Figure 26 Trends in the number of papers from Europe, the U.S. and China in SEI 2014 and SEI 2016: Agriculture and Biological sciences

Note: Fractional counting of papers. Curves in black = SEI 2016; curves in yellow = SEI 2014.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

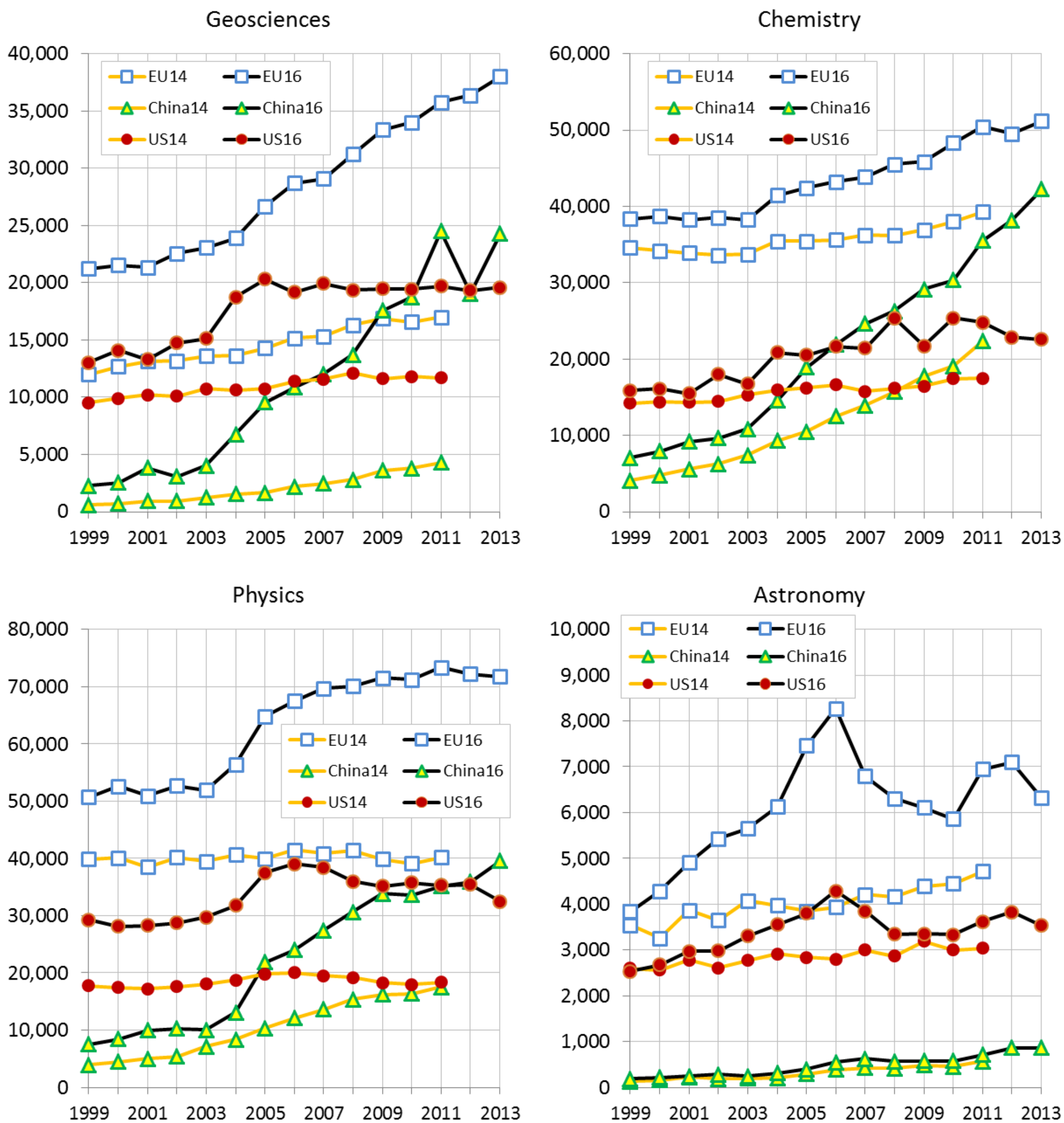

Figure 27 Trends in the number of papers from Europe, the U.S. and China in SEI 2014 and SEI 2016: Geosciences, Chemistry, Physics and Astronomy

Note: Fractional counting of papers. Curves in black = SEI 2016; curves in yellow = SEI 2014.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

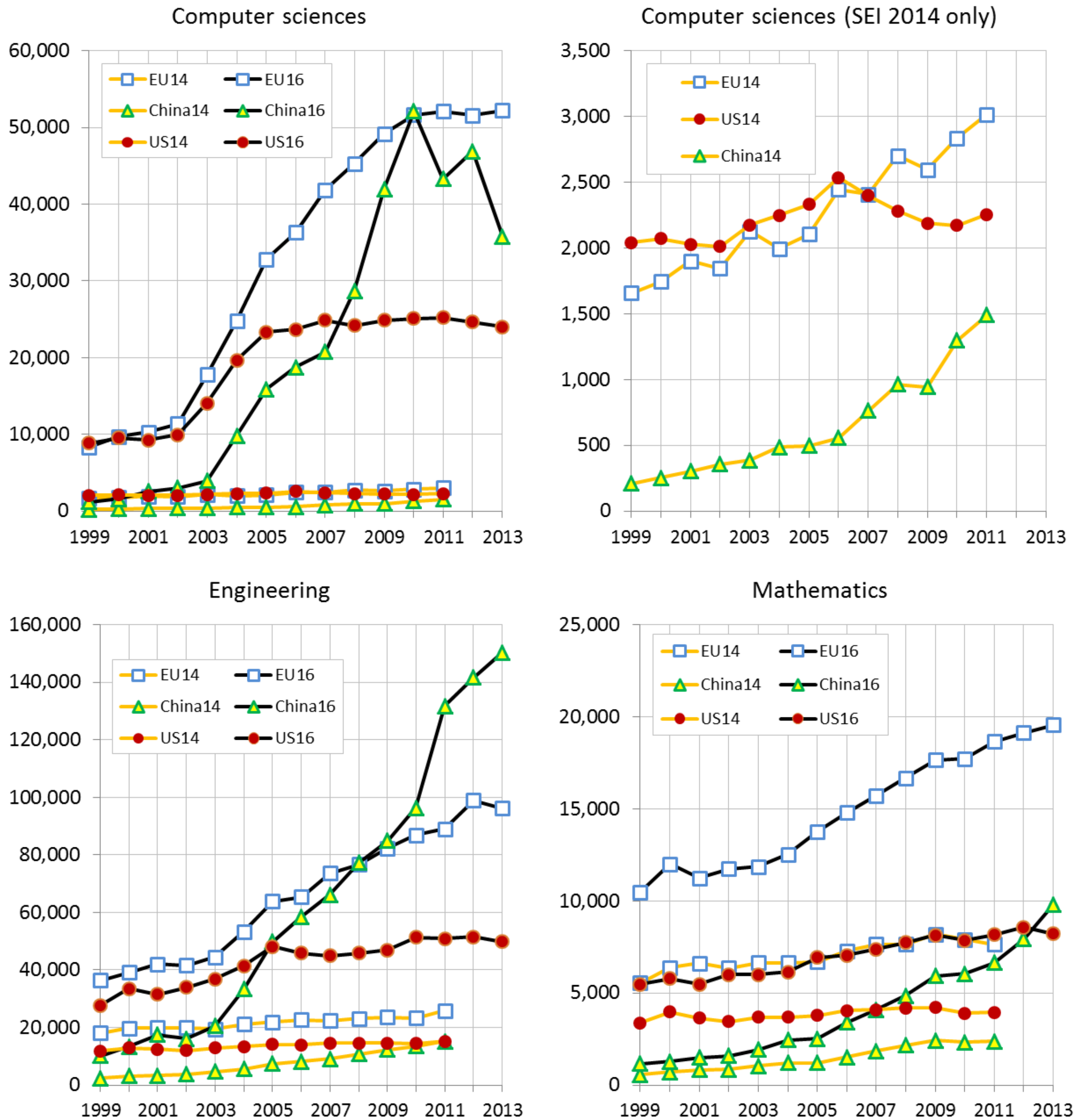

Figure 28 Trends in the number of papers from Europe, the U.S. and China in SEI 2014 and SEI 2016: Computer sciences, Engineering and Mathematics

Note: Fractional counting of papers. Curves in black = SEI 2016; curves in yellow = SEI 2014.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

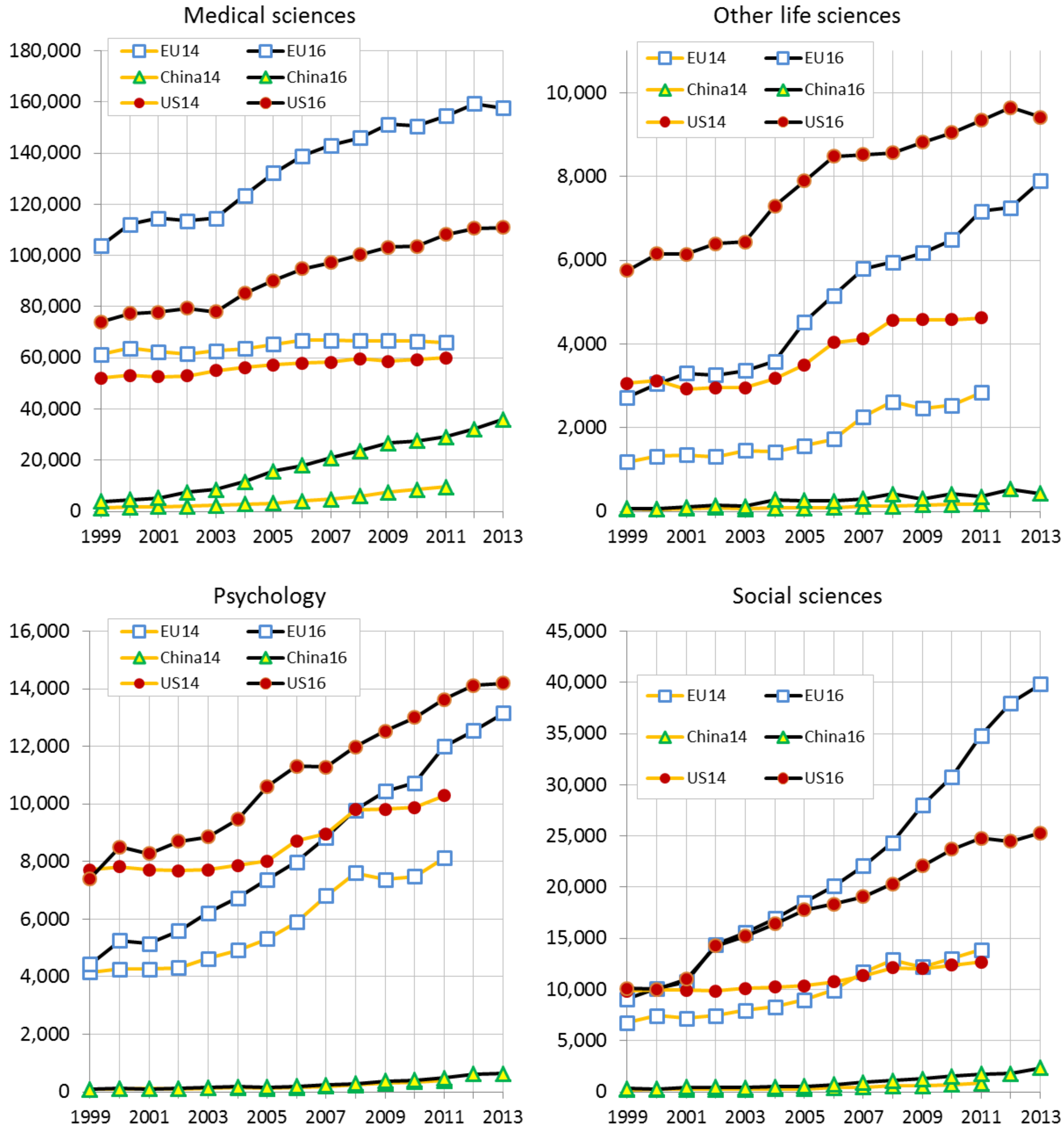

Figure 29 Trends in the number of papers from Europe, the U.S. and China in SEI 2014 and SEI 2016: Medical sciences, Other life sciences, Psychology and Social sciences

Note: Fractional counting of papers. Curves in black = SEI 2016; curves in yellow = SEI 2014.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

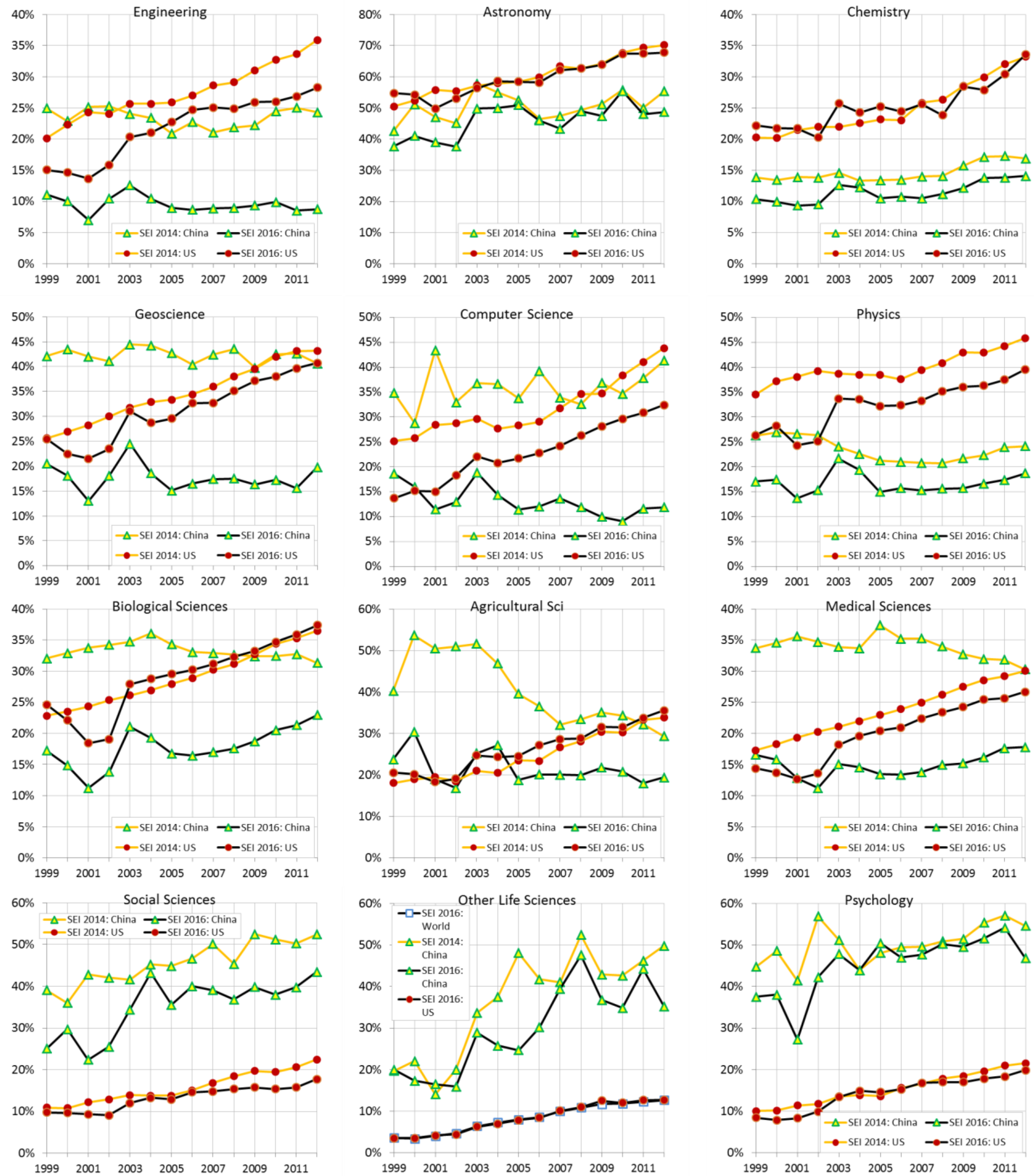

Figure 30 Trends in international collaboration rates in SEI 2014 and SEI 2016, by Caspar field, for China and the United States, 1999–2011

Note: Inter-institutional collaboration rate = proportion of papers from the given country with authors from at least two addresses; international collaboration rate = proportion of papers from the given country with authors from at least one foreign country.

Source: SEI 2016 calculated by Science-Metrix from Scopus; SEI 2014 calculated by the Patent Board from SCI & SSCI

Annex II – Bibliography

Bibliometrics

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342. Retrieved from http://www.akademiai.com/index/34W733263N36H661.pdf

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2008). WoS vs. Scopus: On the reliability of scientometrics. In *Book of Abstracts of the 10th International Conference on Science and Technology Indicators* (pp. 94–97). Retrieved from http://lariviere.ebsi.umontreal.ca/Publications/wos_scopus_vienna.pdf

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, *60*(7), 1320–1326. http://doi.wiley.com/10.1002/asi.21062

Ball, R., & Tunger, D. (2006). Science indicators revisited — Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services and Use*, *26*(4), 293–301. Retrieved from http://epub.uni-regensburg.de/5244/1/ubr11317_ocr.pdf

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271. http://doi.org/10.1007/s11192-008-0216-y

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H.-D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, *3*(1), 27–35. Retrieved from http://econpapers.repec.org/article/eeeinfome/v_3a3_3ay_3a2009_3ai_3a1_3ap_3a27-35.htm

De Sutter, B., & Van Den Oord, A. (2012). To be or not to be cited in computer science. *Communications of the ACM*, *55*(8), 69–75. http://doi.org/10.1145/2240236.2240256

Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, *83*(1), 243–258. http://doi.org/10.1007/s11192-009-0021-2

Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science*, *34*(5), 715–725. http://doi.org/10.1177/0165551507086991

Jacsó, P. (2011). The h-index, h-core citation rate and the bibliometric profile of the Scopus database. *Online Information Review*, *35*(3), 492–501. http://doi.org/10.1108/14684521111151487

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164. http://doi.org/10.1002/asi.21608

Leydesdorff, L. (2012). World shares of publications of the USA, EU-27, and China compared and predicted using the new Web of Science interface versus Scopus. *El Profesional de La Informacion, 21*(1), 43–49. http://doi.org/10.3145/epi.2012.ene.06

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726. http://doi.org/10.1002/asi.v59:11

Meho, L. I., & Sugimoto, C. R. (2009). Assessing the scholarly impact of information studies: A tale of two citation databases—Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *60*(12), 2499–2508. http://doi.org/10.1002/asi.21165

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125.

Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the *h* index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ*, *1*, e183. http://doi.org/10.7717/peerj.183

Moed, H. F., Luwel, M., & Nederhof, A. J. (2002). Towards research performance in the humanities. *Library Trends, 50*(3), 498–520. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/8415/librarytrendsv50i3m_opt.pdf?sequence=1

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. Retrieved from http://econpapers.repec.org/article/eeeinfome/v_3a1_3ay_3a2007_3ai_3a2_3ap_3a161-169.htm

Peknikova, M. (n.d.). Hirsch Index: A comparison of Web of Science, Scopus and Google Scholar. Retrieved from http://www.academia.edu/7741987/Hirsch_Index_a_comparison_of_Web_of_Science_Scopus_and_Google_Scholar

Pudovkin, A. I., & Garfield, E. (2012). Rank normalization of impact factors will resolve Vanclay's dilemma with TRIF: Comments on the paper by Jerome Vanclay. *Scientometrics*, *92*(2), 409–412. http://doi.org/10.1007/s11192-012-0634-8

Vílchez-Román, C. (2014). Bibliometric factors associated with h-index of Peruvian researchers with publications indexed on Web of Science and Scopus databases. *Transinformação*, *26*(2), 143–154. http://doi.org/10.1590/0103-37862014000200004

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. Retrieved from http://eprints.rclis.org/8605/1/Yang_citation.pdf

Zhang, L. (2014). The impact of data source on the ranking of computer scientists based on citation indicators: A comparison of Web of Science and Scopus. *Issues in Science and Technology Librarianship*, (Winter). http://doi.org/10.5062/F4D798CW

General comparison

Abrizah, A., Zainab, A. N., Kiran, K., & Raj, R. G. (2013). LIS journals scientific impact and subject categorization: A comparison between Web of Science and Scopus. *Scientometrics*, *94*(2), 721–740. http://doi.org/10.1007/s11192-012-0813-7

Adriaanse, L. S., & Rensleigh, C. (2011). Comparing Web of Science, Scopus and Google Scholar from an Environmental Sciences perspective. *South African Journal of Libraries and Information Science*, *77*(2), 169–178. http://doi.org/10.7553/77-2-58

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(1), 7. http://doi.org/10.1186/1742-5581-3-7

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506. http://doi.org/10.1007/s11192-010-0185-9

Bosman, J., van Mourik, I., Rasch, M., Sieverts, E., & Verhoeff, H. (2006). Scopus reviewed and compared: The coverage and functionality of the citation database Scopus, including comparisons with Web of Science and Google Scholar. Universiteitsbibliotheek Utrecht/Utrecht University Library. Retrieved from http://dspace.library.uu.nl/handle/1874/18247

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., & Ebrahim, N. A. (2013). A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Social Science*, *9*(5). http://doi.org/10.5539/ass.v9n5p18

De Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing. *Nursing Outlook*, *60*(6), 391–400. http://doi.org/10.1016/j.outlook.2012.04.007

De Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: A longitudinal study. *Scientometrics*, *98*(2), 1547–1565. http://doi.org/10.1007/s11192-013-1089-2

Etxebarria, G., & Gomez-Uranga, M. (2010). Use of Scopus and Google Scholar to measure social sciences production in four major Spanish universities. *Scientometrics*, *82*(2), 333–349. http://doi.org/10.1007/s11192-009-0043-9

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. http://doi.org/10.1096/fj.07-9492LSF

Fingerman, S. (2006). Web of Science and Scopus: Current features and capabilities. *Issues in Science and Technology Librarianship*, (Fall). http://doi.org/10.5062/F4G44N7B

Gavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. *Online Information Review, 32*(1), 8–21. http://doi.org/10.1108/14684520810865958

Ghazimirsaied, S. J., Hemmat, M., Dargahi, H., & Khansari, J. (2011). Self-citation status of Iranian prolific authors in the field of medicine and its impact on the Hirsch Index in Web of Science database. *Payavard Salamat*, *4*(3), 67–76. Retrieved from http://payavard.tums.ac.ir/browse.php?a_id=96&sid=1&slc_lang=en

Goodwin, C. (2014). Web of Science. *The Charleston Advisor*, *16*(2), 55–61. http://doi.org/10.5260/chara.16.2.55

Jacsó, P. (2005). As we may search — Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547.

Jacsó, P. (2012). Grim tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: Reflections on Vanclay's criticism. *Scientometrics*, *92*(2), 325–354. http://doi.org/10.1007/s11192-012-0769-7

Kloda, L. A. (2007). Use Google Scholar, Scopus and Web of Science for comprehensive citation tracking. *Evidence Based Library and Information Practice*, *2*(3), 87–90. Retrieved from https://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/432

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *The Journal of the American Medical Association (JAMA)*, *302*(10), 1092–1096.

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2*(4), 304–316. http://doi.org/10.1016/j.joi.2008.08.001

Miguel, S., Chinchilla-Rodriguez, Z., & de Moya-Anegón, F. (2011). Open access and Scopus: A new approach to scientific visibility from the standpoint of access. *Journal of the American Society for Information Science and Technology*, *62*(6), 1130–1145. http://doi.org/10.1002/asi.21532

Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*(3), 929–942. http://doi.org/10.1007/s11192-012-0783-9

Moed, H. F., Linmans, J., Nederhof, A., Zuccala, A., López Illescas, C., & de Moya Anegón, F. (2009). *Options for a comprehensive database of research outputs in social sciences and humanities* (Research report to the Project Board of the scoping study "Towards a Bibliometric Database for the Social Sciences and the Humanities" set up by the Standing Committees for the Social Sciences and the Humanities of the European Science Foundation (ESF) No. Version 6). Leiden, The Netherlands: Leiden University. Retrieved from http://www.dfg.eu/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/annex_2_en.pdf

Mongeon, P., & Paul-Hus, A. (2014). The journal coverage of bibliometric databases: A comparison of Scopus and Web of Science. École de bibliothéconomie et des sciences de l'information, Université de Montréal. Retrieved from http://www.researchgate.net/profile/Adele_Paul-Hus/publication/268223674_The_journal_coverage_of_bibliometric_databases_A_comparison_of_Scopus_and_Web_of_Science/links/54664afb0cf2f5eb18016c0a.pdf

Pickard, K. T. (2012). The impact of open access and social media on scientific research. *Journal of Participatory Medecine*, *4*, e15. Retrieved from http://www.jopm.org/opinion/commentary/2012/07/18/the-impact-of-open-access-and-social-media-on-scientific-research/

Salisbury, L. (2009). Web of Science and Scopus: A comparative review of content and searching capabilities. *Charleston Advisor*, *11*(1), 5–18. Retrieved from http://www.citeulike.org/user/SRCMethodsLibrary/article/13307571

Sampson, M., McGowan, J., Cogo, E., & Horsley, T. (2006). Managing database overlap in systematic reviews using Batch Citation Matcher: Case studies using Scopus. *Journal of the Medical Library Association*, *94*(4), 461–e219. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1629443/

Schroeder, R. (2007). Pointing users toward citation searching: Using Google Scholar and Web of Science. *Portal: Libraries & The Academy*, *7*(2), 243–248. Retrieved from http://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1007&context=ulib_fac

The Royal Society. (2011). *Knowledge, networks and nations: Global scientific collaboration in the 21st century* (No. 03/11 DES2096). London, UK: The Royal Society. Retrieved from http://www.interacademies.net/File.aspx?id=25069

Vílchez-Román, C. (2014). Bibliometric factors associated with h-index of Peruvian researchers with publications indexed on Web of Science and Scopus databases. *Transinformação*, *26*(2), 143–154. http://doi.org/10.1590/0103-37862014000200004

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–15. Retrieved from http://eprints.rclis.org/8605/1/Yang_citation.pdf

Zhang, L. (2014). The impact of data source on the ranking of computer scientists based on citation indicators: A comparison of Web of Science and Scopus. *Issues in Science and Technology Librarianship*, (Winter). http://doi.org/10.5062/F4D798CW