ARTICLE

Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies

Jeroen Baas[1,3], Michiel Schotten[1], Andrew Plume[1,3], Grégoire Côté[2,3], and Reza Karimi[1]

[1]Elsevier B.V., Radarweg 29, Amsterdam, The Netherlands
[2]Science-Metrix Inc., Elsevier, 1335 Mont-Royal Ave E, Montreal, QC, Canada
[3]International Center for the Study of Research, Elsevier, Radarweg 29, Amsterdam, The Netherlands

Keywords: abstract and citation database, author profile, bibliographic database, bibliometrics, citation linking, Content Selection and Advisory Board, CSAB, data cleaning, data clustering, data curation, data linking, ICSR, institution profile, International Center for the Study of Research, network visualization, ORCiD, quality assurance, research assessment, researcher mobility, science policy evaluation, scientometrics, university ranking

ABSTRACT

Scopus is among the largest curated abstract and citation databases, with a wide global and regional coverage of scientific journals, conference proceedings, and books, while ensuring only the highest quality data are indexed through rigorous content selection and re-evaluation by an independent Content Selection and Advisory Board. Additionally, extensive quality assurance processes continuously monitor and improve all data elements in Scopus. Besides enriched metadata records of scientific articles, Scopus offers comprehensive author and institution profiles, obtained from advanced profiling algorithms and manual curation, ensuring high precision and recall. The trustworthiness of Scopus has led to its use as bibliometric data source for large-scale analyses in research assessments, research landscape studies, science policy evaluations, and university rankings. Scopus data have been offered for free for selected studies by the academic research community, such as through application programming interfaces, which have led to many publications employing Scopus data to investigate topics such as researcher mobility, network visualizations, and spatial bibliometrics. In June 2019, the International Center for the Study of Research was launched, with an advisory board consisting of bibliometricians, aiming to work with the scientometric research community and offering a virtual laboratory where researchers will be able to utilize Scopus data.

1. SCOPUS AS A BIBLIOMETRIC DATA SOURCE

In 2004, Elsevier launched Scopus as a new search and discovery tool (Schotten, el Aisati, Meester, Steiginga, & Ross, 2017). Scopus is an abstract and citation database consisting of peer-reviewed scientific content. At its launch, it contained about 27 million publication records (1966–2004). Since then, the content of the database has grown to over 76 million records at the time of writing, covering publications from 1788–2019, making it among the largest curated bibliographic abstract and citation databases today. Approximately 3 million new items are being added every year. The content in Scopus is sourced from over 39,100 serial titles (with the most recently published content indexed from over 24,500 titles),

an open access journal

Citation: Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

DOI: https://doi.org/10.1162/qss_a_00019

Received: 04 July 2019
Accepted: 26 October 2019

Corresponding Author: Jeroen Baas
j.baas@elsevier.com

Handling Editors: Ludo Waltman and Vincent Larivière

Copyright: © 2020 Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

120,000 conferences, and 206,000 books from over 5,000 different publishers worldwide. Scopus is a curated database, which means that content is selected for inclusion in the database through a rigorous process: Serial content (i.e., journals, conference proceedings, and book series) submitted for possible inclusion in Scopus by editors and publishers is reviewed and selected, based on criteria of scientific quality and rigor. This selection process is carried out by an external Content Selection and Advisory Board (CSAB) of editorially independent scientists, each of which are subject matter experts in their respective fields. This ensures that only high-quality curated content is indexed in the database and affirms the trustworthiness of Scopus. In addition, Scopus being a curated database also refers to the rigorous capturing procedures, publisher agreements, and technical infrastructure in place to source the publications directly from the publishers themselves, ensuring comprehensive, complete, and accurate coverage of the serial content once it has been selected by the CSAB.

Scopus indexes many different elements of scientific publications, obtained from external publishers, such as publication title, abstract, keywords, author names and linked affiliations, references, and drug terms (Berkvens, 2012). Content in Scopus contains publications from scientific publishers from all over the world. Elsevier, the owner of Scopus, is also a scientific publisher. About 9.9% of the serial *titles* (i.e., journals and book series) in Scopus are published by Elsevier (this amounts to an *article* share in Scopus of 17.4% between 2012 and 2018); the other 90.1% of serial titles (and 82.6% of articles, respectively) are produced by an extensive list of global publishers (Figure 1a). In addition, subject coverage of the serial titles in Scopus is quite balanced among the four main subject categories (Figure 1b). Scopus also includes non-English content, as long as an English title and abstract, as well as references in Roman script, are available.

In addition to the fields that are provided in the source data for Scopus, Elsevier further enriches the content using a variety of enhancements; citations provided in the full text are structured and clustered together and, where referencing content that is already indexed in Scopus, linked to the cited Scopus records. This allows users to view the citation count (how many times an article was cited). At the time of writing, the precision for citation linking in Scopus is measured at 99.9% and the recall is 98.3%. This means that in general, references that should be linked to Scopus records are linked in 98.3% of cases; and among all reference links, 99.9% are linked to the correct record.

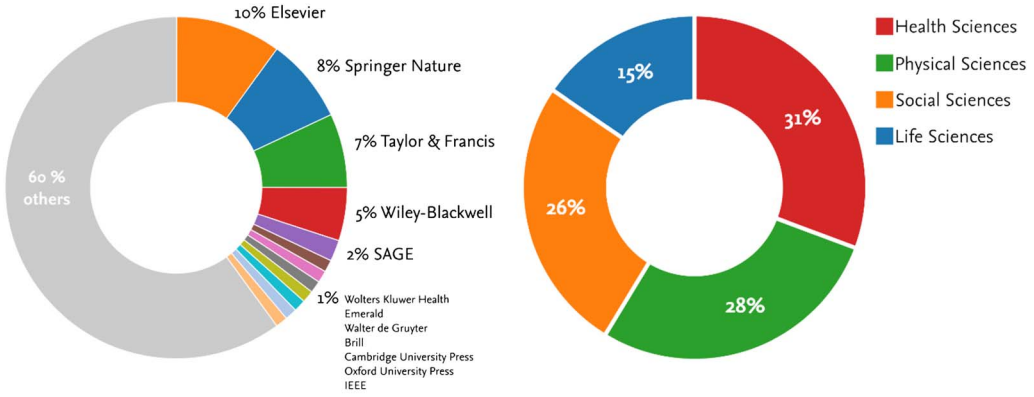


Figure 1. Distribution of publishers (a) and of the four main subject categories (b) of the serial titles indexed in Scopus, rounded to whole percentage points.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

Authorships in the Scopus databases are clustered into publication histories called Scopus (author) profiles. Author profiles are generated using a combination of algorithms and manual curation. Elsevier uses a "gold set" of roughly 12,000 randomly selected authors for quality assessment. This set is updated and expanded annually and is independent of sets used for tuning or training of algorithms. The end-to-end accuracy is measured continuously by several metrics. Moreover, regular spot checks are run on aspects of author profiles, such as canonical names or affiliations. Typically, accuracy metrics are averaged over authors to better represent the typical experience of users. Publications in author profiles have an average precision of 98.1% and an average recall of 94.4%. Both precision and recall are measured based on the best matching Scopus profile for a given "gold set" author. The best match is determined based on the Scopus profile containing the largest *number* of publications for that author (Figure 2). This overlap number divided by the total number of publications by the "gold set" author defines recall (i.e., ratio of publications captured). The same overlap number divided by the total number of publications in the Scopus profile defines precision (i.e., ratio of publications correctly assigned to the author).

A substantial number of author profiles in Scopus have been curated. Curation can be initiated through a variety of sources. A well-established process is through Open Researcher and Contributor ID (ORCID), an open, non-proprietary registry of unique, persistent author identification codes (What is ORCID?, 2018). ORCID is managed by a non-profit organization with the same name established in 2012. Researchers can export, import, or link their publications and curated metadata between Scopus and ORCID. Similarly, researchers can use a feature on Scopus.com called the "Author Feedback Wizard" (AFW) to improve their author profile. Finally, yet another process that results in improved, curated Scopus author profiles is a commercial service offered by Elsevier to subscribers of its "Pure" university administration product, called *Profile Refinement Service*. Pure customers can opt to have their profiles refined upon request or refined every 4 months. Elsevier uses the same service to proactively refine author profiles regardless of Pure subscription whenever needed.

All above efforts combined have led to approximately 1.8 million Scopus author profiles that have been manually enhanced (Scopus index, July 2019). This total has been verified using a "manual curated" flag in the XML data of Scopus author profile records. Nonetheless, we must emphasize that Scopus creates author profiles that are being actively updated by algorithms for all

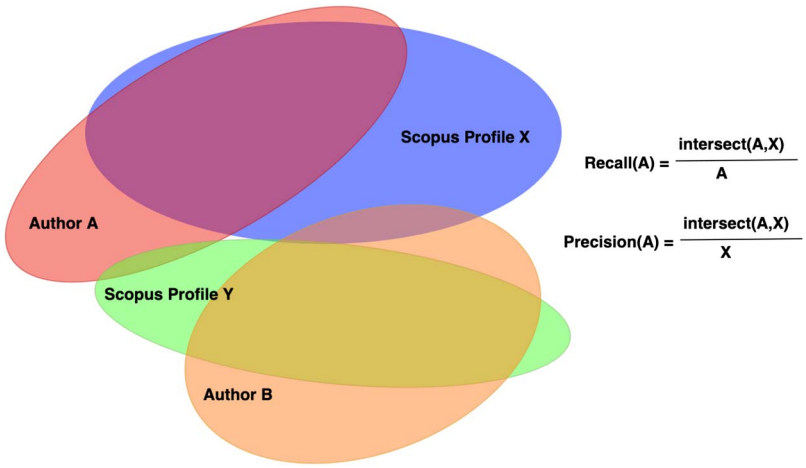


Figure 2. Schematic depicting how precision and recall are calculated for Scopus author profiles. "Author A" and "Author B" represent manually curated "gold set" lists of publications by these authors.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

authorships (of over 76 million publications) covered. Availability of author profiles throughout the corpus enables author level analytics and benchmarking across the database beyond subsample estimations. Moreover, Scopus author profiles are designed to be a complete publication history and so will not remove or hide select publications for personal gain (i.e., negatively impacting publications, such as retraction notices, errata, suspiciously funded). In fact, feedback that Elsevier receives is algorithmically and manually reviewed before changing any existing profile.

An important aspect of the Scopus database is the high coverage and availability of first names, even for relatively old records: from 25% of authorships in 1970–1974 and 52% between 1995–1999 to 82% of authorships in 2015–2019 (Figure 1). This feature strengthens the disambiguation of authors and allows, for instance, gender-based longitudinal analyses that leverage first names (Elsevier, 2017; Lerchenmueller & Sorenson, 2018). Another relevant aspect in the analytical context is the availability of author-affiliation links in publications throughout the database historically. This enables studies dealing with the mobility of researchers, by analyzing the author affiliations and how they change over time.

An important enrichment in the Scopus database is that of institution profiles, allowing different name variants and hierarchies of institutions to be curated in a similar fashion as authors, thereby allowing automated organizing of information where needed (via an advanced, proprietary, and highly accurate institutional profiling algorithm) and manual modification and instruction, where possible. In addition, the full text of articles sourced by Scopus is processed using natural language processing to identify potential references to funding acknowledgements.

To maintain Scopus as a high-quality data source and push the boundaries of quality forward, Scopus introduced internal review processes to constantly monitor preidentified areas of quality focus, such as processing, profile quality, and completeness and accuracy of source data. This allows the content team to identify early trends in the data and to monitor progress on key initiatives to increase quality. For instance, under this program, digital object identifier (DOI) completion rates went up from 87.8% at the start of the program to 99.8% in December 2018. The completeness was measured across a gold record set for which each should have a DOI. Other main focus areas where significant improvements have been made over the past few years, and where continuous investments are being made, are completeness of indexed publication records for the serial titles covered (by weekly comparisons against the CrossRef database), the removal of duplicate MedLine and Article in Press records, the correctness and completeness of citation links (by measuring against a gold set), of the author and institution profiles and publication record metadata (such as document type classifications, publication years, article numbers, country codes, and funding information), as well as improving the timeliness and currency of newly indexed content. These quality review processes employ machine learning approaches, supplemented with manual validation, and concern legacy content (i.e., content already indexed in Scopus) as well as continuously improving and fine-tuning the capturing procedures for newly indexed content.

In addition to expanding and enriching the content, as well as improving the timeliness of the database, a curated database such as Scopus requires re-evaluation of the appropriateness of new and already indexed titles on an ongoing basis. This is needed to exclude poor-quality journals and "predatory" journals and publishers, a relatively recent phenomenon that is a threat to the integrity of science, as well as an increasing challenge to all research publishing stakeholders: authors, editors, researchers, research institutions, funding bodies, research assessment bodies, and governments. To ensure that only the most reliable scientific articles and content are available in Scopus to each of these stakeholders and that the quality of the existing content is maintained, a rigorous process of continuous monitoring and re-evaluation

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

has been installed. This means that titles that have been selected for inclusion in Scopus may be discontinued and no longer indexed going forward. There are three different identification techniques applied, using (a) external feedback (i.e., formally raised concerns about publication standards), (b) heuristics (metrics) to flag underperforming journals, and (c) a machine learning approach to flag outlier behavior, which each lead to titles being tagged for re-evaluation. The ultimate decision to (de)select content lies with the external and independent Scopus CSAB (for full details, please see Elsevier, 2019b; Holland, Brimblecombe, Meester, & Steiginga, 2019). For publications of which the CSAB determines they no longer meet the quality standards for inclusion in Scopus, indexing of new content is discontinued, but content already indexed remains in Scopus, to ensure the integrity of the scientific record as well the stability and consistency of research trend analytics.

2. SUPPORT FOR LARGE-SCALE ANALYSES

Since its official launch in 2004, Scopus has been used globally in many large-scale analyses. There are three types of large studies where Scopus data are used in a central role.

The first group of large-scale analyses deals with national research assessments. The first national assessment supported by Scopus was the Excellence in Research for Australia (ERA) of the Australian Research Council (ARC) in 2010, later repeated in 2012 and 2015. The first edition of the Research Excellence Framework (REF 2014) national assessment in the UK also used Scopus data. The REF 2014 was held by the Higher Education Funding Council for England (HEFCE) and the funding bodies for Scotland, Wales, and Northern Ireland. Up to four research outputs per active researcher were submitted by the UK's Higher Education Institutions (HEIs) and matched to the corresponding records in Scopus for 11 out of 36 "units of assessment" (UoAs; i.e., broad subject areas). For each of these 11 UoAs, Scopus citation counts of the submitted outputs were compared with that UoA's citation benchmarks (also obtained from Scopus as contextual data) and used as an additional assessment criterion by the REF's expert panels, besides peer review of the scientific content of the outputs. Prior to the announcement of the REF 2014 assessment results by HEFCE, the REF Results Analysis Tool provided by Elsevier allowed the UK's HEIs to compare and evaluate their own REF performance across several benchmarks and metrics. More examples of national assessments where Scopus data were used include the 2013–2014 assessment in Portugal held by the FCT ("Fundação para a Ciência e a Technologia"), the ASN ("Abilitazione Scientifica Nazionale") national accreditation rounds (2012–2013, 2016–2018, and 2018–2020), VQR ("Valutazione della Qualità della Ricerca") national assessment in Italy (2012–2013, 2016–2017), and National University Corporation Evaluation (NUCE) national assessment in Japan held in 2016–2017 and to be held in 2020 by the National Institution for Academic Degrees and Quality Enhancement of Higher Education (NIAD-QE).

A second type of analysis supported by Scopus data is government science policy evaluations. The UK department of Business, Innovation and Skills (BIS) commissioned a report that entailed a comparative study of the UK's international research base, in 2011 (Elsevier, 2011) and 2013 (Elsevier, 2013). In 2016, another refresh of this report was issued by the newly renamed Department for Business, Energy, & Industrial Strategy (BEIS) (Elsevier, 2016). More examples of the use of Scopus data for research policy reports include those of the European Research Council (ERC) and other large government bodies, often dealing with program evaluations and research landscape analyses. Many of these evaluations include different data sources of various types (macroeconomic data, for instance), as well as deep qualitative evaluations in which Elsevier works in consortia.

As a further example of this second type of analysis, Scopus data have also been used in the production of bibliometric indicators for the US National Science Foundation's Science and

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

Engineering Indicators (SEI), for the 2016 (National Science Board, 2016) and 2018 (National Science Board, 2018) editions, and will be used in future editions of the SEI reports up to 2022. Scopus data are also the source of bibliometric indicators for the European Research Area in the context of the 2010–2014 study "Analysis and Regular Update of Bibliometric Indicators for the European Commission" (Science-Metrix, 2014) and have recently been selected for the continuation and improvement of this study, now called "Provision and Analysis of Key Indicators in Research and Innovation," for the three coming years.

The third type of analysis is that of university rankings. University rankings are often composed of combinations of evaluations for which only part is a bibliometric resource. Different ranking bodies have variations of subjective and objective data sources to provide ranked lists of universities. These rankings and the media attention they draw provide a platform for academia to engage with the public. Elsevier provides publication output, citation, and international collaboration data from Scopus for each university to organizations in the field of university rankings. These include the World University Rankings and its various derived Regional, Global Subject, Young University, and World Reputation Rankings, as well as the Wall Street Journal/THE College Rankings, which are all issued by Times Higher Education (THE, since 2014); and the World University Rankings and its various derived rankings issued by Quacquarelli Symonds (QS, since 2015); as well as various other regional and subject-specific rankings, such as the Best Chinese Universities Ranking issued by ShanghaiRanking Consultancy (since 2015), Perspektywy in Poland, Maclean's in Canada, National Institutional Ranking Framework (NIRF) in India, the Financial Times Global MBA Ranking, and the Frankfurter Allgemeine Zeitung Economists Ranking in Germany.

The large-scale analyses supported by Scopus are reports in the government and commercial space. In contrast to peer-reviewed scientific studies, where the focus is on access to the data and open reproducibility of results, reports in the evaluation space focus on accountability. Key aspects in those engagements are therefore about how data providers deal with quality concerns, quality assurance, and risk mitigation. For instance, what is the process and timeline in case content is identified as missing?

3. DATA AVAILABILITY FOR RESEARCH

Next to the aim of supporting the academic community with robust results using reliable data in the analyses mentioned, providing access to raw data is an essential component in securing advancement of the bibliometric field. Until 2014, Elsevier supported bibliometricians with data using the Elsevier Bibliometric Research Program (EBRP). The program provided pre-compiled data sets to researchers, after a scientific board reviewed and approved a submitted proposal. Its aim was to enable external research groups or individual researchers in the field of bibliometrics and quantitative research assessment to carry out strategic research using Elsevier data and to present the outcomes in peer-reviewed journal papers and at international conferences.

Since 2014, application programming interfaces (APIs) have taken over the role of providing access to raw data, allowing free use for scientific purposes, such as the text-and-data-mining resources (Elsevier, 2019c) and Scopus APIs for academic research purposes (Elsevier, 2019a). Use of the APIs does not require a Scopus subscription[1]; without a subscription, users will have limited access to basic metadata for most citation records, as well as to basic search functionality. Full access to Scopus APIs is only granted to subscribers of Scopus.

[1] https://dev.elsevier.com/sc_apis.html.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

In addition, Scopus data have been available in bulk for research groups. Research groups working with bulk Scopus data include CWTS (CWTS, 2019), SciMago (Scimago Lab, 2019), DZHW (DZHW, 2019), SciTech Strategies (SciTech Strategies, 2018), and others through tailored agreements that have been established between these groups and Elsevier.

The mission of the International Center for the Study of Research (ICSR) is to advance research evaluation in all fields of knowledge production. To foster this development, the ICSR provides access to a working environment where new ideas and hypotheses can be tested against high-quality, large data sets. This platform, offering a virtual laboratory, will allow researchers to collaboratively develop indicators and methodologies. Elsevier is providing computational access to Scopus data for research purposes on this platform, free of charge. This will also enhance the reproducibility of scientometric studies, by enabling other researchers to verify published research findings using the same data set and methodologies with shared code. Researchers can use the environment for such academic, noncommercial purposes, and access will be organized by the ICSR to review submitted proposals for use of the lab as well as actively reaching out to researchers to collaboratively work on specific research problems. The platform allows researchers to create and extract aggregate derivatives that can be published as part of their work, under the condition that the source of the data is acknowledged. At the moment of writing, this platform is not yet publicly available and will be announced through the ICSR website (International Center for the Study of Research, 2019).

4. EXAMPLES OF STUDIES USING SCOPUS DATA

Scopus data have been used as a source for many types of different bibliometric studies. The different quality properties of Scopus described support different types of analyses.

For instance, there are studies on mobility, using Scopus' unique historic author-affiliation records, such as by Caroline Wagner and Koen Jonkers on international collaboration, mobility, and openness (Wagner & Jonkers, 2017), funding and collaboration (Leydesdorff, Bornmann, & Wagner, 2019), and author (Pina, Barac, Buljan, Grimaldo, & Marušic, 2019) and institutional (Lee, 2012) collaboration networks. Another example of author-mobility analyses can be found in a bibliometric study to measure knowledge transfer (Aman, 2018). The mobility analysis using Scopus author profiles also informs the research policy of governments, such as through the European Commission's Joint Research Center (JRC) report on the rise of China as an industrial and innovation powerhouse (Preziosi et al., 2019).

In addition, Scopus' availability of author first names historically, combined with author profiling, enables studies using author gender assignments: for example, "The gender gap in early-career transitions in the life sciences" (Lerchenmueller & Sorenson, 2018) and "Gender differences in research areas, methods and topics: Can people and thing orientations explain the results?" (Thelwall, Bailey, Tobin, & Bradshaw, 2019). In addition, Scopus author profiles have been used to study the recent phenomenon of hyperprolific authorships (Ioannidis, Klavans, & Boyack, 2018) and for an author database of highly cited researchers (Ioannidis, Baas, Klavans, & Boyack, 2019; Van Noorden & Singh Chawla, 2019).

There are also examples of studies using the full Scopus database to build new algorithms: Richard Klavans and Kevin Boyack developed algorithms on top of the database, resulting in Topics of Prominence (Klavans & Boyack, 2017), which are now prominently displayed in Elsevier's SciVal research performance product (which uses Scopus data as one of its data sources).

In the more traditional sense of bibliometric analysis, there are many studies available around citation analysis and correlations, such as on the influence of highly cited articles

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

on indicators (Thelwall, 2019; Thelwall & Fairclough, 2015), on correlation between citations and Mendeley readership (Maflahi & Thelwall, 2016; Thelwall & Wilson, 2016), on journal usage (Schloegl & Gorraiz, 2010), and studies revisiting bibliometric laws (Thelwall & Wilson, 2014). Scopus data were also used to analyze initiatives in open science, particularly open access (Solomon, Laakso, & Björk, 2013), citizen science (Follett & Strezov, 2015) and new tools in the scientific space, such as ResearchGate (Thelwall & Kousha, 2017). They have been used to evaluate the fate of rejected manuscripts (Bornmann et al., 2009), to investigate potential citation manipulation by reviewers (Baas & Fennell, 2019; Singh Chawla, 2019) and to study the development of multidisciplinarity (Levitt & Thelwall, 2008). At present, Scopus data are used for bibliometric analysis to inform the EU Open Science Monitor (The Lisbon Council, CWTS, & Esade, 2018).

Another form of common analysis performed using Scopus data is around network visualization and spatial bibliometrics (Bornmann & De Moya Anegón, 2019; Bornmann & Waltman, 2011; Leydesdorff & Persson, 2010; Mutz, Bornmann, de Moya Anegón, & Stefaner, 2014) as well as research building new visualization techniques (Leydesdorff, 2010; Mischo & Schlembach, 2018).

5. WHAT CAN WE LEARN FROM SCOPUS DATA, TOGETHER?

In the preceding sections, we have outlined some of the concrete ways in which Scopus data have been used for large-scale evaluative studies (typically at the national or institutional levels) and for exploratory work leading to a plethora of papers on aspects as diverse as topic detection, researcher mobility, and data visualization techniques. But the potential of Scopus to uncover and understand the fundamental forces that drive human knowledge creation through the research endeavor may be limited only by our capability to ask the right questions. How do career paths form and change for individual researchers through space and time? Can we follow people as they develop from "apprentice" to "master" and understand the drift in their topical focus, collaborative patterns, geolocation, and research impact (by citation-based indicators or other means) through careers that may be either very short or very long? Can we identify the conditions that, near the beginning of a research career, predict a long and successful contribution to the knowledge front? And those conditions that foreshadow an early exit from the world of (academic) research? Going beyond Scopus, can we use standardized researcher identifiers, such as ORCID, connected to nonresearch online personas, such as LinkedIn, to pinpoint the exit of trained researchers from publication-centric roles (largely within or adjacent to academia) into careers in organizations in the commercial or charitable sectors? What is the influence of gender, nationality, and early-career mentoring on these outcomes, and how much remains unexplained? Is a career in research more likely the result of persistence or of good fortune—and what does this mean for the development of better and fairer evaluative structures in research? Finally, what are the implications of the answers to these questions for all the actors in research, from educators to public policy experts and from university career advisors to researchers themselves?

This single example shows that those who create and those who use Scopus suffer no lack of imagination to ask challenging questions, and Scopus itself offers a firm base on which to begin seeking answers. The remaining piece of the puzzle is a collective one: How can the bibliometric research community and the creators of Scopus best come together to address these challenges together? In June 2019, the ICSR (International Center for the Study of Research, 2019) was launched, with a wide-ranging brief and the support of an advisory board, including experts in research policy, research evaluation, and bibliometrics, to be a

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

place where a dialogue can happen and research of great interest and importance can be pursued—together. Elsevier, Scopus and the ICSR do not see themselves as *apart from* the world of research but as *part of it*, and this spirit will inform our work for many years to come.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Elsevier colleague Roy Boverhof, who provided the design of the charts in Figure 1.

COMPETING INTERESTS

The authors of this paper are Elsevier employees. Elsevier runs Scopus, which is the database discussed in this article.

REFERENCES

Aman, V. (2018). A new bibliometric approach to measure knowledge transfer of internationally mobile scientists. *Scientometrics*, *117*(1), 227–247. https://doi.org/10.1007/s11192-018-2864-x

Baas, J., & Fennell, C. (2019). When peer reviewers go rogue—estimated prevalence of citation manipulation by reviewers based on the citation patterns of 69,000 reviewers. *ISSI 2019, September 2–5, 2019, Rome, Italy https://www.issi2019.org/*. Retrieved from https://ssrn.com/abstract=3339568

Berkvens, P. (2012). Scopus custom data documentation. Retrieved from https://p.widencdn.net/mrbekb/Scopus_Custom_Data_Documentation_Version9

Bornmann, L., & De Moya Anegón, F. (2019). Hot and cold spots in the US research: A spatial analysis of bibliometric data on the institutional level. *Journal of Information Science*, *45*(1), 84–91. https://doi.org/10.1177/0165551518782829

Bornmann, L., & Waltman, L. (2011). The detection of "hot regions" in the geography of science—A visualization approach by using density maps. *Journal of Informetrics*, *5*(4), 547–553. https://doi.org/10.1016/j.joi.2011.04.006

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. https://doi.org/10.1016/j.joi.2008.11.001

CWTS. (2019). CWTS Journal Indicators. Retrieved from http://www.journalindicators.com/

DZHW. (2019). Competence centre for bibliometrics. Retrieved from https://www.dzhw.eu/en/forschung/projekt?pr_id=484

Elsevier. (2011). International comparative performance of the UK research base—2011. Retrieved from https://www.elsevier.com/research-intelligence/resource-library/international-comparative-performance-of-the-uk-research-base-2011

Elsevier. (2013). International comparative performance of the UK research base—2013. Retrieved from https://www.elsevier.com/research-intelligence/research-initiatives/BIS2013

Elsevier. (2016). International comparative performance of the UK research base—2016. Retrieved from https://www.elsevier.com/research-intelligence/research-initiatives/beis2016

Elsevier. (2017). Gender in the global research landscape. Amsterdam: Elsevier. Retrieved from https://www.elsevier.com/research-intelligence/resource-library/ty/gender-in-the-global-research-landscape

Elsevier. (2019a). Academic research. Retrieved from https://dev.elsevier.com/academic_research_scopus.html

Elsevier. (2019b). Content—how Scopus works. Retrieved from https://www.elsevier.com/solutions/scopus/how-scopus-works/content

Elsevier. (2019c). Text and data mining policy. Retrieved from https://www.elsevier.com/about/policies/text-and-data-mining

Follett, R., & Strezov, V. (2015). An analysis of citizen science based research: Usage and publication patterns. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0143687

Holland, K., Brimblecombe, P., Meester, W. J., & Steiginga, S. (2019). The importance of high-quality content: Curation and re-evaluation in Scopus. Elsevier. Retrieved from https://www.elsevier.com/research-intelligence/resource-library/scopus-high-quality-content

International Center for the Study of Research. (2019). Retrieved from https://www.elsevier.com/icsr

Ioannidis, J. P., Klavans, R., & Boyack, K. B. (2018). Thousands of scientists publish a paper every five days. *Nature*, *561*, 167–169. https://doi.org/10.1038/d41586-018-06185-8

Ioannidis, J., Baas, J., Klavans, R., & Boyack, K. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biology*, *17*(8), e3000384. https://doi.org/10.1371/journal.pbio.3000384

Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2017.10.002

Lee, D. S. (2012). Collaboration network patterns and research performance: The case of Korean public research institutions. *Scientometrics*, *91*(3), 925–942. https://doi.org/10.1007/s11192-011-0602-8

Lerchenmueller, M. J., & Sorenson, O. (2018). The gender gap in early career transitions in the life sciences. *Research Policy*, *47*(6), 1007–1017. https://doi.org/10.1016/j.respol.2018.02.009

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology*, *59*(12), 1973–1984. https://doi.org/10.1002/asi.20914

Leydesdorff, L. (2010). Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for Information Science and Technology*, *61*(2), 352–369. https://doi.org/10.1002/asi.21250

Leydesdorff, L., & Persson, O. (2010). Mapping the geography of science: Distribution patterns and networks of relations among

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024

cities and institutes. *Journal of the American Society for Information Science and Technology*, *61*(8), 1622–1634. https://doi.org/10.1002/asi.21347
Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2019). The relative influences of government funding and international collaboration on citation impact. *Journal of the Association for Information Science and Technology*, *70*(2). https://doi.org/10.1002/asi.24109
Maflahi, N., & Thelwall, M. (2016). When are readership counts as useful as citation counts? Scopus versus Mendeley for LIS journals. *Journal of the Association for Information Science and Technology*, *67*(1), 191–199. https://doi.org/10.1002/asi.23369
Mischo, W. H., & Schlembach, M. C. (2018). A system for generating research impact visualizations over medical research groups. *Journal of Electronic Resources in Medical Libraries*, *15*(2), 96–107. https://doi.org/10.1080/15424065.2018.1507773
Mutz, R., Bornmann, L., de Moya Anegón, F., & Stefaner, M. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level mode. *Online Information Review*, 43–58. https://doi.org/10.1108/OIR-12-2012-0214
National Science Board. (2016). Science and engineering indicators 2016. National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2016/nsb20161/
National Science Board. (2018). Science and engineering indicators 2018. National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2018/nsb20181/
Pina, D. G., Barac, L., Buljan, I., Grimaldo, F., & Marušic, A. (2019). Effects of seniority, gender and geography on the bibliometric output and collaboration networks of European Research Council (ERC) grant recipients. *PLoS ONE*, *14*(2), e0212286. https://doi.org/10.1371/journal.pone.0212286
Preziosi, N., Fako, P., Hristov, H., Jonkers, K., Goenaga, X., Alves Dias, P., … Hristov, H. (2019). China—Challenges and prospects from an industrial and innovation powerhouse. *JRC*. https://doi.org/10.2760/445820
Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics*, *82*(3), 567–580. https://doi.org/10.1007/s11192-010-0172-1
Schotten, M., el Aisati, M., Meester, W. J., Steiginga, S., & Ross, C. A. (2017). A Brief history of Scopus: The world's largest abstract and citation database of scientific literature. In F. J. Cantú-Ortiz (Ed.), *Research Analytics. Boosting University Productivity and Competitiveness through Scientometrics* (pp. 31–58). Boca Raton, FL: Taylor & Francis Group. http://dx.doi.org/10.1201/9781315155890-3
Science-Metrix. (2014). Analysis of bibliometric indicators for European policies. European Commission. Retrieved from https://ec.europa.eu/research/innovation-union/pdf/bibliometric_indicators_for_european_policies.pdf
Scimago Lab. (2019). SJR. Retrieved from https://www.scimagojr.com/
SciTech Strategies. (2018). SciTech Strategies. Retrieved from https://www.scitech-strategies.com/
Singh Chawla, D. (2019). Elsevier investigates hundreds of peer reviewers for manipulating citations. *Nature*, *573*, 174. https://doi.org/10.1038/d41586-019-02639-9
Solomon, D. J., Laakso, M., & Björk, B.-C. (2013). A longitudinal comparison of citation rates and growth among open access journals. *Journal of Informetrics*, *7*(3), 642–650. https://doi.org/10.1016/j.joi.2013.03.008
The Lisbon Council, CWTS, & Esade. (2018). OPEN science monitor—draft methodological note. Retrieved from https://ec.europa.eu/info/sites/info/files/open_science_monitor_methodological_note_v2.pdf
Thelwall, M. (2019). The influence of highly cited papers on field normalised indicators. *Scientometrics*, *118*(2), 519–537. https://doi.org/10.1007/s11192-018-03001-y
Thelwall, M., Bailey, C., Tobin, C., & Bradshaw, N.-A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics*, *13*(1), 149–168. https://doi.org/10.1016/j.joi.2018.12.002
Thelwall, I., & Fairclough, R. (2015). Geometric journal impact factors correcting for individual highly cited articles. *Journal of Informetrics*, *9*(2), 263–272. https://doi.org/10.1016/j.joi.2015.02.004
Thelwall, M., & Kousha, K. (2017). ResearchGate articles: Age, discipline, audience size, and impact. *Journal of the Association for Information Science and Technology*, *68*(2), 468–479. https://doi.org/10.1002/asi.23675
Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839. https://doi.org/10.1016/j.joi.2014.08.001
Thelwall, M., & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972. https://doi.org/10.1002/asi.23501
Van Noorden, R., & Singh Chawla, D. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature*, *572*(7771), 578–579. https://doi.org/10.1038/d41586-019-02479-7
Wagner, C., & Jonkers, K. (2017). Open countries have strong science. *Nature*, *550*, 32–33. https://doi.org/10.1038/550032a
What is ORCID?. (2018). Retrieved from https://support.orcid.org/hc/en-us/articles/360006897674

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00019 by guest on 29 March 2024