ARTICLE

Web of Science as a data source for research on scientific and scholarly activity

Caroline Birkle, David A. Pendlebury, Joshua Schnell, and Jonathan Adams

Institute for Scientific Information, Web of Science Group, 160 Blackfriars Road, London SE1 8EZ, UK

Keywords: authoritative content, bibliometrics, Eugene Garfield, ISI, premier source, selective bibliography, Web of Science

ABSTRACT

Web of Science (WoS) is the world's oldest, most widely used and authoritative database of research publications and citations. Based on the Science Citation Index, founded by Eugene Garfield in 1964, it has expanded its selective, balanced, and complete coverage of the world's leading research to cover around 34,000 journals today. A wide range of use cases are supported by WoS from daily search and discovery by researchers worldwide through to the supply of analytical data sets and the provision of specialized access to raw data for bibliometric partners. A long- and well-established network of such partners enables the Institute for Scientific Information (ISI) to continue to work closely with bibliometric groups around the world to the benefit of both the community and the services that the company provides to researchers and analysts.

1. WEB OF SCIENCE

The Web of Science (WoS) Core Collection database is a selective citation index of scientific and scholarly publishing covering journals, proceedings, books, and data compilations. It is the oldest citation index for the sciences, having been introduced commercially by the ISI in 1964, initially as an information retrieval tool called the *Science Citation Index* (SCI) (Garfield, 1964). The first SCI covered some 700 journals, expanded to 1,573 within two years, and was produced in printed form as a series of volumes presenting bibliographic and citation data in a very small font size. With the rapid growth of the research enterprise in the 1960s, annual volumes of the SCI increased in size and journal coverage. By 1970, around 2,200 journals were indexed, along with four million cited references from these sources.

During these years, a range of innovations and products was introduced by ISI, which citation-indexing pioneer Eugene Garfield (1925–2017) had founded in 1960 (Cawkell & Garfield, 2001; Lawlor, 2014; Lazerow, 1974). The company also produced the Social Sciences Citation Index (SSCI) (1973), the Arts & Humanities Citation Index (A&HCI) (1978), and other indexes covering the chemical (Current Chemical Reactions, Index Chemicus) and proceedings (Conference Proceedings Citation Index) literatures. A citation index for books was launched in 2011 (Adams & Testa, 2012). As technology advanced from the 1960s through the 1990s, other formats and media for distributing and analyzing SCI data—from magnetic tapes, to floppy disks, to CD-ROMs, to standard file formats distributed via the World Wide Web—profoundly changed information access and accelerated bibliometric research based on publication and citation data.

Selectivity in coverage has long characterized the SCI, SSCI, and A&HCI, which were combined and launched on the World Wide Web as WoS in 1997. With the earliest versions of the

an open access journal

Citation: Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. https://doi.org/10.1162/qss_a_00018

DOI: https://doi.org/10.1162/qss_a_00018

Received: 21 June 2019
Accepted: 26 October 2019

Corresponding Author:
Jonathan Adams
Jonathan.Adams@Clarivate.com

Handling Editors:
Ludo Waltman and Vincent Larivière

SCI, journal selection was constrained by cost considerations, including computation and printing. Computing power increased and digital dissemination reduced expenses, but selectivity remained a hallmark of coverage because Garfield had decided early on to focus on internationally influential journals. His decision was informed by Bradford's Law of Scattering (Bradford, 1934) as well as his own research on SCI data that revealed Garfield's Law of Concentration (Garfield, 1971, 1972). Garfield's Law of Concentration generalized Bradford's insights concerning specific fields to all fields of science and demonstrated the existence of a multidisciplinary core set of journals, then as few as 1,000. More recently, the globalization of research has highlighted the relevance of local and regional journals for science that addresses societal needs. The WoS group has deepened its journal coverage, principally through the introduction of the Emerging Sources Citation Index (2015) (Huang et al., 2017; Somoza-Fernandez et al., 2018), to give a more complete coverage of the most influential research while maintaining the balance across subjects and regions that underpins informed search and good analytics.

The coverage of WoS has thus expanded vastly since the inception of the underpinning systems, growing to about 34,000 journals today. This is not directly comparable to the original data set because there have been many mergers, content changes, and deletions as well as extensive additions in most fields. The WoS platform now extends the content of the Core Collection through hosting citation databases of other providers, such as the BIOSIS Citation Index, the Chinese Science Citation Database, the Russian Science Citation Index, and the SciELO Citation Index (for Latin America and Iberia), as well as specialized databases, including Medline, Inspec, KCI—Korean Journal Database—and the Derwent Innovations Index, covering the patent literature. The scope and bibliometric characteristics of the WoS Core Collection and WoS platform are summarized in Table 1.

Visser, van Eck, and Waltman have recently compared different sources of bibliographic and citation data, including WoS (Visser et al., 2019) and—for those who wish to consult it—their analysis provides further information to guide good practice and research use.

2. WOS DATA ENABLED THE DEVELOPMENT OF SCIENTOMETRICS

WoS is not just a catalogue of academic publications. It is a selective, structured, and balanced database with complete citation linkages and enhanced metadata that supports a wide range of information purposes. An early example of the use of SCI data for research is Derek J. de Solla Price's study "Networks of Scientific Papers" (Price, 1965). Price showed how a network of cited references in papers (citations) could be used to describe the structure and dynamics of a research topic, since then often called a *research front*. Sociologists Stephen and Jonathan Cole's work "Scientific Output and Recognition" (Cole & Cole, 1967) is one of the first times that citations were used systematically as a measure of scientific quality or impact (Cole, 2000). This study of physicists sought to show variations in recognition in terms of institutional affiliation, productivity, quality (citations as the indicator), honors, age, and other variables. Citations were found to be highly correlated with peer judgment of "quality." Early research on scientific activity using SCI data typically required many hours of manual labor involving repeated access to different cross-referenced, heavy volumes. Indeed, as one of us can attest (D.A.P.), an early form of research evaluation of individuals made use of a ruler to measure column inches of citations!

WoS was not designed for scientometric analysis. Garfield created the SCI and its sister citation indexes for information retrieval. Use of the data for other purposes, such as research performance evaluation, including rankings, mapping topics and monitoring trends, and investigating aspects of the history and sociology of science and scholarly activity, was of secondary

Table 1. Key Characteristics of data sources built on web of science

	Web of Science *Core Collection*	Web of Science *Platform*
Summary	Citation indexes representing the connections between scholarly research articles found in globally significant journals, books, and proceedings in the sciences, social sciences and art & humanities. The WoS Core Collection is the standard data set underpinning the journal impact metrics found in the *Journal Citation Reports* and the institutional performance metrics found in InCites.	A platform providing access to multidisciplinary and regional citation indexes, specialist subject indexes, a patent family index, and an index to scientific data sets. WoS provides a common search language, navigation environment, and data structure, allowing researchers to search broadly across disparate resources and use citation connections to navigate to relevant research results.
Databases covered	• Science Citation Index • Social Sciences Citation Index • Arts & Humanities Citation Index • Conference Proceedings Citation Index • Book Citation Index • Emerging Sources Citation Index	Citation Indexes include the WoS Core Collection plus the following: • BIOSIS Citation Index • Chinese Science Citation Database • Russian Science Citation Index • SciELO Citation Index • Data Citation Index Subject and regionally specialized indexes: • Biological Abstracts, BIOSIS Previews • CABI: CAB Abstracts and Global Health • FSTA—the food science resource • Inspec • KCI—Korean Journal Database • Medline • Zoological Record Other resources: • Current Contents Connect • Derwent Innovations Index (patents)

Table 1. (*continued*)

	Web of Science *Core Collection*	Web of Science *Platform*
Number of journals	> 20,900 journals plus books and conference proceedings	> 34,200 journals plus books, proceedings, patents, and data sets
Coverage	• Over 75 million records • More than 101,000 books • Over 8 million conference papers	• 155 million records (journals, books, and proceedings) • 39.3 million patent families (> 70 million patents) • 7.3 million data sets
Time period covered	Sciences: 1900–present Social Sciences: 1900–present Arts & Humanities: 1975–present Proceedings: 1990–present Books: 2005–present Emerging Source Citation Index: 2005–present	Journal literature: 1800–present Patents: 1963–present Full cited reference indexing for all WoS Core Collection content Citation indexing for SciELO, Russian Science Citation Index, Chinese Science Citation Index, and BIOSIS Citation Index All content includes times cited for citations from WoS Core Collection and platform Citation Sources
Author indexing	All authors from all publications are indexed. Authors linked to affiliations from 2008–forward.	WoS Core Collection: All authors are indexed for all publications. Other resources: Author indexing varies by resource.

Institution indexing	All author affiliations are indexed. Institution's variants and parent/child relationships are mapped and connected to a preferred institutional name through a manually curated process that is increasingly global in coverage.	Author affiliation indexing varies by collection.
Updating frequency	Daily (Monday through Friday).	Each collection is updated on its own schedule, ranging from daily to monthly.
Citation analysis	Citation counts, and author h-index calculations. "Hot" and "Highly Cited" articles (papers in top percentiles according to year, field and document types) are available from *Essential Science Indicators* integration. Journal Impact Factors and Journal Performance Quartiles are available via *Journal Citation Reports* integration (*JCR* Quartiles available without subscription to *JCR*).	Citation counts, and author h-index calculations. "Hot" and "Highly Cited" articles (papers in top percentiles according to year, field and document types) are available from *Essential Science Indicators* integration.
Controlled vocabulary	No. Keyword fields include Author Keywords, and "Keywords Plus," which are extracted from the titles of Cited Articles. Controlled indexing is provided for institution affiliations (parent/child mapping).	Controlled vocabulary searching is provided for Medline, Inspec, FSTA, BIOSIS, Zoological Record.

interest and importance. Had the database been designed for any of these secondary uses, many data elements would have been collected, indexed, and structured differently. Consequently, research analysis using WoS data necessarily makes use of some features that were designed for information retrieval rather than quantitative analysis.

An example of an "information retrieval" feature is the WoS Subject Categories. These are 254 journal-based categories, each of which represents a specific field or subfield, such as biotechnology & applied microbiology, family studies, medical laboratory technology, or quantum science & technology. This categorical scheme was created and developed to enable information retrieval where a search may be executed or filtered by subject category, and for this reason a journal may be and often is assigned to more than one subject category. For quantitative analysis, however, it would be necessary to adjust counts to avoid duplication of data if collating initially by subject category. Furthermore, contemporary analytic models may focus on topics that draw on parts of multiple categories.

Nonetheless, it may be claimed that without the SCI, the development of scientometrics would certainly have been hampered. For 40 years, almost all advances in our understanding of the global science system and its evaluation and management were based upon these data sources. As Jonathan Cole has noted, "The creation of the *SCI* represents a good case study of how technological innovations very frequently create the necessary conditions for significant advance in scientific fields" (Cole, 2000). Important early applications of our data include the adoption of publication and citation indicators for the first Science Indicators produced by the US National Science Foundation (National Science Board, 1973); their development for this purpose by Francis Narin and his further research on the citation linkage between the patent and scholarly literature (Narin, 1976); the pioneering work of Tibor Braun, András Schubert, and Wolfgang Glänzel of the Information Science & Scientometrics Research Group (ISSRU) of the Hungarian Academy of Science (Budapest), especially on absolute and relative indicators of national research performance (Braun et al., 1985); similar fundamental research on measuring and evaluating the comparative performance of universities and groups of researchers by Anthony van Raan, Henk Moed, and others at Leiden University (Moed et al., 1985); the development of science mapping through cocitation clustering introduced by Henry Small of ISI and Belver Griffith of Drexel University (Griffith et al., 1974; Small, 1973; Small and Griffith, 1974); and investigations of what the ISI data could reveal concerning the sociology of science, pursued by researchers at Columbia University, including Harriet Zuckerman, Stephen Cole, and Jonathan Cole, under the direction of Robert Merton (Cole, 2000; Zuckerman, 2018). With the introduction of the journal *Scientometrics* in 1978, with Braun as founding editor in chief, SCI data and the field itself became inextricably entwined.

Other databases and more recently developed tools that draw on publication and citation data from the WoS Core Collection were explicitly designed for quantitative analysis and research evaluation. These include National Science Indicators (1992), US and UK University Indicators (1995), Essential Science Indicators (2001), and InCites (2010). Typical users of these products have been university research offices, government agencies, and research funding organizations. WoS data continue to be used to search and explore publications on a research topic (such as environmental sex determination: Adams et al., 1987) and compare research activity profiles (for example, that of a country or an institution; see Adams, 2018).

3. ACCESSING WOS DATA TODAY

The WoS group is one of the main business divisions of Clarivate Analytics, a company created from the IP & Science division of Thomson Reuters. The ISI is now a research group

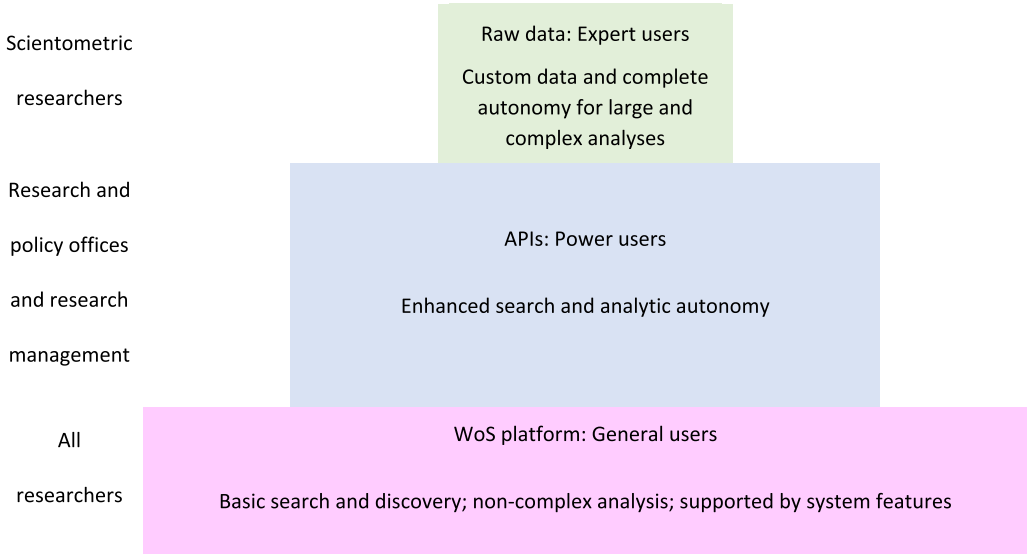

Figure 1. A wide range of use cases apply to WoS data

within the WoS group. It works closely with the company product and data teams as well as carrying out its own research, much of which draws on innovative ideas from bibliometric partners outside the company.

Because we work closely with scientometricians in universities, research funding agencies, and government departments, we know that researchers conducting advanced scientometric studies for new knowledge about science and scholarly communication and dynamics or for policymaking often require a specific set of data in a format that can be analyzed, summarized, and visualized in a different environment. Sometimes the amount of data required can be downloaded from the WoS platform under an appropriate license. In other cases, the amount of data needed may exceed what can reasonably be collected from WoS-related products. In these instances, arrangements for licensed use of custom data may be made.

WoS data are made available to institutions and associated researchers via platforms, APIs, and custom data set delivery (Figure 1). We have long-established relationships with scientometric research groups in universities around the world, enabling us to draw on their knowledge, advice, and innovative capacity and affording us extensive independent testing and feedback on our data systems and quality controls. Since the earliest days of ISI, we have recognized the benefit of these partnerships and have enjoyed the opportunity to collaborate with many partners.

Irrespective of the delivery mechanism, many uses of WoS data are linked to existing institutional subscriptions, in which case data-use entitlement is linked to the product subscriptions. Our institutional and partner pricing reflects volumes of data and frequency of delivery required, alongside a spectrum of use cases from casual to large-scale and commercial data requirements in support of a substantive contract for a third party, such as a national research funding agency.

3.1. Use Cases

The list of possible use cases for WoS data highlighted below is illustrative, not exhaustive. In practice, we invest time to understand each request so that we can come up with a solution

that properly supports our partners and other customers. Each use case may be subject to fees, terms, and conditions that Clarivate deems necessary and appropriate for such use, but whenever possible, the fees associated with academic use are limited to any additional costs incurred.

Basic usage	At the level of individual academic researchers, basic usage is about primary rights to view, use, and copy (download and/or print) information from WoS for individual use and then to distribute and redistribute small ("insubstantial" in the legal parlance) portions of the derived summary information in a nonsystematic (e.g., not a product or service) manner. Creating a reference list for a journal article is an obvious use, but studies may have a more applied purpose as well.
Discovery	Discovery is another fundamental, core function, which entails the ability to query WoS data for research purposes in support of the research process, with appropriate referencing and accreditation. It is what tens of thousands of researchers have done on WoS every day for decades.
Analytics	Basic usage automatically steps up to an applied level for any WoS subscriber by giving the user the ability to download and analyze WoS data for internal use within the subscribing institution and/or internal business operations. That means a user can benchmark a research group against others without needing additional permissions, thereby demonstrating to a university where it should be investing.
Integration (APIs only)	Here additional permissions may be required. This allows WoS data and/or analytics to be integrated into an organization's own application (either internally developed or third party) for its internal use, and it is therefore subject to certain restrictions. If a client does not own or control the application, the third-party provider of the application would require a license or approval from the WoS group.
Public use	Prior agreement is required where the user is planning to make WoS data publicly available, by disseminating data and analysis in reports for an internal website or public-facing site owned, maintained, and controlled by a subscriber (usually the researcher's employer). This would then also allow for the publication of some summary results of analysis of WoS data for noncommercial use.
Commercialization	Commercial licenses are a higher level use case requiring formal agreement, and these usually involve clear licensing of specific data sets, an agreed specification for the way the data will be used, and a time-limited allowance on data use. This provides the ability to produce commercial analysis and reports or incorporate data into a client's application for delivery to a third party. However, although this may seem onerous, it often brings with it additional support for, as an example, the supply of specific data records and metadata in a user-specified format.

The formal terms and conditions of use of WoS data are available on request from the WoS group at the contact email for this article.

3.2. Data Delivery, Volume, and Frequency

Data may be required as "current" at a point in time, or for a historical period, or data may need to be regularly updated to a custom cycle. Because citation counts accumulate over time, and because publication lists grow with new publications, the census date for each data download is significant. Refreshed data sets will have both more records and revised versions of historical records.

Whether delivered in custom data sets or via APIs, WoS offers a suite of updating services, each associated with the delivery tools. Our custom data sets can be updated and re-extracted as frequently as every two weeks, if necessary, or simply annually. Decisions about the preferred cycle will depend on the requirements of the associated research activities and outputs. Our APIs are available in basic, intermediate, and advanced formats, allowing a variety of call speeds and data volumes, again relative to customer requirements and technical competencies.

3.3. Data for Scientometric Research

Basic access for search and discovery is the standard use case for WoS, and this use is permitted for all product formats and delivery mechanisms.

Researchers may face the challenge that standard products and tools do not fulfill the needs of their research project. In such cases we encourage our scientometric partners to contact their designated WoS institutional account manager. Depending on the subscriptions of an institution, we will endeavor to ensure that individual research groups receive adequate and timely access to data that further their academic objectives. We review requests, often with advice and comments from the ISI team, and will respond to the immediate request while engaging with collections managers to ensure ongoing, long-term, and sustainable access to research resources.

WoS is committed to supporting high-quality, innovative research uses of our data. In addition to the search-and-discovery use case, we permit academic analysis and interrogation of the data for research purposes, subject to the implicit agreement that the analyses are for either internal or noncommercial research purposes. We also permit data extracts to create training data sets for use in the development and training of client software applications using mathematical and statistical algorithms that automatically identify patterns in data. All potential uses remain subject to applicable Clarivate terms and conditions, and we expect reasonable and appropriate credit for the output and content in relevant tables and figures in published material.

3.4. Applying for Access to Data

As noted, we encourage scientometric researchers to contact their institutional research data management office in the first instance (usually their institutional library), that typically holds responsibility for WoS subscriptions. They will then be put in contact with their account manager, who has a technical understanding of individual customer holdings.

The key to make this process simple is to be clear about the required data use and research objectives. Usually, we ask scientometric partners to provide a brief (e.g., <500 words) summary outlining the problem they wish to solve. We think that the ability to provide a succinct summary including these elements is a reasonable test of the seriousness of the researcher's

purpose and preparedness. It avoids spending time and effort on trivial, unplanned, and speculative analysis. The summary should specify the data required to address the problem (i.e., why are these the preferred data for that purpose?), an explanation of how progress will be reported, an outline of likely outcomes in terms of academic research and potential policy applications, and a description of the likely deliverables. If a specific data extract or download is required, then charges normally apply to cover company costs and data value. Reductions and waivers can only be considered where the proposed research use is of evident academic significance.

Despite these requirements, in many cases, data use does not actually lead to any major charges. We are asking questions that any proposal to a research funder will already have covered, and it seems reasonable to check that a data request is properly planned, because a pipeline of poorly structured requests obstructs delivery to genuine researchers.

Each request is reviewed by ISI to ensure the continuing appropriate and reasonable use of our data and products. Confirmation of a decision will be made both to the requestor and their institution.

At this stage, exploratory and test projects are often given a go-ahead with no further requirements. If the project is substantive then, subject to agreement, we usually ask our bibliometric partners to do the following:

- Periodically share summaries with ISI, at mutually agreed intervals, of the nature of the work, the data and analyses used in the project, and any outcomes achieved.
- Alert us of any suspected issues, questions, or discrepancies arising from data use.
- At the end of the project, provide ISI with a report on methodology, data use, findings and other observations on data and analytics that arose in the course of the work, including a short publishable section describing the work and key achievements and suitable for a general readership.
- Provide proper attribution to WoS for use of our data, tables and figures, in accordance with any guidelines shared at the outset of the project or as provided by us from time to time.
- Share any publications with ISI prior to submission to a journal (we appreciate the opportunity to correct any misunderstandings about our data before they appear in print).

Researchers may see these requirements as challenging or even bureaucratic. We see the reporting process rather differently. It is the basis for a dialogue between ISI staff and the researchers. We work with the data every day, and we can often help to address unexpected challenges in format or background. We also value alerts about unexpected data problems. We are delighted to learn what the researchers are doing and happy to offer advice and feedback as problems and results emerge.

4. WHO IS USING WOS DATA?

We are very proud that the research activity and publications of the scientometric community have, from the earliest days of our company, informed the content and provision of our data. Although the WoS group and its predecessor organizations, including ISI, have relied on in-house expertise, such as the research of former Chief Scientist Henry Small and our current research staff (e.g., Adams, 2018; Adams et al., 2007; Small, 1973), external researchers have been the source for advice on many improvements in and extensions of our databases and analytic systems, for which we are extremely fortunate and grateful.

As an expression of reciprocal support for the scientometric community, and in appreciation of the benefits derived from this collaborative research, WoS sponsors professional conferences and offers awards and stipends for researchers. For example, we have, for the last three years, sponsored an annual Eugene Garfield Award for Innovation in Citation Analysis, carrying a $25,000 honorarium (http://discover.clarivate.com/garfield-award-2019), and the Clarivate Analytics Doctoral Dissertation Proposal Scholarship (https://www.asist.org/about/awards/doctoral-dissertation-proposal-scholarship/).

4.1. The Garfield Award for Innovation in Citation Analysis

The Garfield Award is intended to attract applications from scientometric researchers around the world. It is particularly focused on proposals for innovative uses of the WoS data, developing novel applications of the citation network and the research indicators derived from it. Since the beginning of the award program in 2017, we have received proposals covering topics including scholarly publishing trends, author contributions and credit, team science, knowledge generation and dissemination, and the science of science.

Jian Wang (Leiden University) was the 2017 inaugural award recipient. Wang's research uses large databases, such as WoS, and advanced statistics to reveal the structures and dynamics underlying science and technology. His recent work explores both the citing behavior of novel research and how this work is cited, contributing to the important discussion of how bibliometric indicators can be enhanced to better identify and assess novel and creative scholarship. Wang is also interested in the translation of science into innovation, and he uses references to scientific literature from within patents to explore this process and examine the conditions needed for successful translation at the individual, team, and network levels (Veugelers & Wang, 2019).

The 2018 award recipient was Orion Penner (École Polytechnique Fédérale de Lausanne) who studies how career success is influenced by the decisions individual researchers make in allocating research effort across multiple topics, or in selecting a new job or mentor (Peterson & Penner, 2014). Penner has been using WoS and patent data for some time, and his Garfield Award submission was an innovative proposal to use natural language processing to measure where and how research papers in WoS might represent a pivot *away* from the topics represented in the paper's cited references, with the aim of identifying individual publications that have played a transformative role in scientific discourse.

In 2019, the Garfield Award recipient was Erjia Yan (Drexel University). Yan's research program is focused on developing and deploying an entity-based knowledge framework to study knowledge production and the diffusion of innovations. To build this framework, he has first been identifying and organizing scientific and technological innovations in large heterogeneous corpora through entity recognition methods. For his Garfield Award submission, Yan proposed to examine proximity factors that contribute to knowledge production and innovation diffusion by systematically integrating WoS and grant data from the U.S. National Institutes of Health to study scientist-level productivity and impact (Li & Yan, 2019).

Through the award program, the recipients may collaborate with ISI researchers on the development of existing and new scientometric approaches, with ISI attentive to the ways in which the research that is being conducted can subsequently be applied by the company to products and services across WoS.

4.2. Other Recent Users

As noted, interaction with academic researchers was part of the origins of ISI and growth of the SCI, and there is continuing use of WoS as a source of data in research publications (Li et al., 2018; Schnell, 2018). Collaboration between ISI staff and leading academic scientometricians has led to recent proposals for new journal indicators (Leydesdorff et al., 2019) and new ways of displaying individual research records (Bornmann et al., 2019). Leading researchers at Ohio State University have long been elaborating the changing global collaborative landscape (Wagner et al., 2019), and work with them has enabled the use of WoS data in a major study on innovation in Chinese research that ISI hopes to support. Although it is beyond the scope of this paper to provide a comprehensive list of all the many academic groups and research projects using WoS, the following examples provide a selection of other interesting use cases.

4.2.1. Bibliometric and scientometric research organizations

Among a wide range of globally leading academic centers and service organizations, if we focus on just one area of northwest Europe, then we see that the Center for Science and Technology Studies (CWTS), Leiden, Netherlands, the Nordic Institute for Studies in Innovation, Research, and Education (NIFU), Oslo, Norway, and the Royal Institute of Technology (KTH), Stockholm, Sweden all use WoS to deliver internal and externally commissioned projects on research and innovation. These organizations have licensed WoS data to populate an in-house database of scholarly literature, enhancing the data through additional normalization and transformation to meet their analytic needs. In addition to project papers and academic articles, they produce annual statistics and R&D indicators at the national and international levels. CWTS has used WoS to develop a publicly available university ranking (Leiden Ranking), and it recently integrated bibliometric data with associated metadata to produce novel analyses on the variation of female authorship across university output. In addition, these groups develop and disseminate new bibliometric indicators and offer advice on good practice for research assessment.

4.2.2. Academic centers conducting cross-disciplinary research

The Northwestern University Institute on Complex Systems (NICO; Evanston, Illinois), the iSchool at Indiana University (Bloomington, Indiana), and the Knowledge Lab at the University of Chicago (Chicago, Illinois) are examples of academic research centers that carry out cross-disciplinary research projects using WoS to study the dynamics of research and innovation and the science of science. In addition to scholarly work, these groups develop and share "big data" analytic tools with the research community. A recent example of work produced at these centers includes a study that analyzed WoS data to study the age of cited references of highly cited papers (Mukherjee et al., 2017).

In a second example, authors affiliated with several of the above institutions and their colleagues used WoS to develop a bibliometric framework for studying mobility of scientists (Robinson-Garcia et al., 2019). Their framework applies a classification of migrant authors and traveling authors to researchers and finds that migrant authors have higher citation impact, despite being less than a third of all mobile authorships. Such large-scale science of science studies require access to the full WoS data set.

4.2.3. Custom extracts for specialized research purposes

As mentioned, for specialized applications, researchers can license access to WoS subsets to meet their research needs. For a study on the relationship between research and development

(R&D) funding and publication productivity for academic chemistry departments in the United States, economists purchased a custom data set of research articles published in a specific subset of chemistry journals and for a specific subset of institutions (Rosenbloom et al., 2015). A Clarivate team representing the WoS worked with the researchers to extract the specific data set for their purposes, thus ensuring they spent their grant funding on the exact data set needed for the study. A custom data set can be cost effective for more specialized research projects, as additional resources are not required to house and maintain the full data set or extract specific data later, and it is easier to apply regular data updates.

As another example of a fit-for-purpose WoS data set, in 2019 several small custom data extracts were provided for a research contest to develop Indicators of Technology Emergence conducted by faculty at the Georgia Institute of Technology (Georgia Tech; Atlanta, Georgia) and researchers at the analytics firm Search Technologies (Porter et al., 2018). Contest participants were challenged to "devise a repeatable procedure to identify emerging R&D topics within a designated S&T domain" and provided with three practice data sets from three difference domains: neurodegenerative and dementia medicine, dye-sensitized solar cells, and smart home. After an initial period to develop their approaches, contest participants were then provided with a fourth data set on an unknown topic and given 10 days to return the results of their method for evaluation by contest organizers. The results of the contest were reported at the 2019 Global Tech Mining conference co-occurring with the 2019 Atlanta Conference on Science and Innovation Policy.

5. CONCLUSIONS

We welcome inquiries from and engagement with researchers with a serious interest in and well-developed proposals for the use of WoS data. The content, structure, and detail of WoS has grown and evolved over more than 50 years, often via beneficial, collaborative enterprise between ISI, its successor companies, and the research community—through search and discovery across many disciplines and through the analytic work of many talented scientometricians. Partnerships between those interested in the content and value of the rich data in WoS continue to be an important part of ISI's work today.

COMPETING INTERESTS

The authors of this paper are Clarivate employees. Clarivate owns the WoS group, which manages the database discussed in this article.

REFERENCES

Adams, J. (2018). Information and misinformation in bibliometric time-trend analysis. *Journal of Informetrics*, *12*, 1063–1071. https://doi.org/10.1016/j.joi.2018.08.009

Adams, J., Greenwood, P. J., & Naylor, C. J. (1987). Evolutionary aspects of environmental sex determination. *International Journal of Invertebrate Reproduction and Development*, *11*, 123–136.

Adams, J., Gurney, K. A., & Marshall, S. (2007). Profiling citation impact: A new methodology. *Scientometrics*, *72*, 325–344.

Adams, J., & Testa, J. (2012). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of ISSI 2011: The 13th Conference of the International Society for Scientometrics and Informetrics*. Leuven: International Society of Scientometrics & Informetrics, pp. 13–18.

Bornmann, L., Haunschild, R., & Adams, J. (2019). R package for producing beamplots as a preferred alternative to the h index when assessing single researchers (based on downloads from Web of Science). *Scientometrics*, in press. https://arxiv.org/abs/1905.09095

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*(3550), 85–86.

Braun, T., Glänzel, W., & Schubert, A. (1985). *Scientometric indicators: A 32-country comparative evaluation of publishing performance and citation impact*. Singapore: World Scientific Publishing Co.

Cawkell, T., & Garfield. E. (2001). Institute for Scientific Information. In E. H. Fredriksson (Ed.), *A century of science publishing* (pp. 151–160). Amsterdam: IOS Press.

Cole, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In B. Cronin and H. Barsky Atkins (Eds.), *The Web of Knowledge* (pp. 281–300). Medford, NJ: Information Today.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition. *American Sociological Review*, *32*, 377–390.
Garfield, E. (1964). Science Citation Index—A new dimension in indexing science. *Science*, *144*(361), 649–654. https://doi.org/10.1126/science.144.3619.649
Garfield, E. (1971). The mystery of transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. *Current Contents*, *No. 17*, 5–6, reprinted in Garfield, E., *Essays of an Information Scientist, 1962–1973*, I (pp. 222–223). Philadelphia, PA: ISI Press.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471
Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science. *Science Studies*, *4*(4), 339–365. https://doi.org/10.1177/030631277400400402
Huang, Y., Zhu, D. H., Ly, Q., Porter, A. L., Robinson, D. K. R., & Wang, X. F. (2017). Early insights on the Emerging Sources Citation Index (ESCI): An overlay map-based bibliometric study. *Scientometrics*, *111*(3), 2041–2057. https://doi.org/10.1007/s11192-017-2349-3
Lawlor, B. (2014). The Institute for Scientific Information: A brief history. In L. R. McEwen and R. E. Buntrock, (Eds.), *The Future of the History of Chemical Information* (pp. 109–126). Washington, DC: American Chemical Society.
Lazerow, S. (1974). Institute for Scientific Information. In A. Kent, H. Lancour, & J. E. Daily (Eds.), *Encyclopedia of Library and Information Science* (vol. 12, pp. 89–97). New York: Marcel Dekker.
Leydesdorff, L., Bornmann, L., & Adams, J. (2019). The integrated impact indicator revisited (I3*): A non-parametric alternative to the Journal Impact Factor. *Scientometrics*, *119*(3), 1669–1694. https://doi.org/10.1007/s11192-019-03099-8
Li, K., & Yan, E. (2019). Are NIH-funded publications fulfilling the proposed research? An examination of concept-matchedness between NIH research grants and their supported publications. *Journal of Informetrics*, *13*(1), 226–237. https://doi.org/10.1016/j.joi.2019.01.001
Li, K., Rollins, J., & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: A selective, dynamic, cross-domain, content-based analysis. *Scientometrics*, *115*(1), 1–2. https://doi.org/10.1007/s11192-017-2622-5
Moed, H., Burger, W. J. M., Frankfurt, J. G., & van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university-research performance. *Research Policy*, *14*(3), 131–149. https://doi.org/10.1016/0048-7333(85)90012-5
Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances*, *3*(4), e1601315. https://doi.org/10.1126/sciadv.1601315
Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons, Inc.
National Science Board. (1973). *Science indicators 1972*. Washington, DC: US Government Printing Office.
Peterson, A. M., & Penner, O. (2014). Inequality and cumulative advantage in science careers: A case study of high-impact journals. *EPJ Data Science*, *3*(1), Article Number 24. https://doi.org/10.1140/epjds/s13688-014-0024-y
Porter, A., Youtie, J., Carley, S., Newman, N., & Murdick, D. (2018). Contest: Measuring tech emergence. *23rd International Conference on Science & Technology Indicators 2018 (STI 2018)*, 1440–1442. https://openaccess.leidenuniv.nl/bitstream/handle/1887/65353/STI2018_paper_232.pdf?sequence=1
Price, D. de Solla. (1965). Networks of scientific papers. *Science*, *149*, 510–515.
Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Lariviere, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, *13*(1), 50–63. https://doi.org/10.1016/j.joi.2018.11.002
Rosenbloom, J. L., Ginther, D. K., Juhl, T., & Heppert, J. A. (2015). The effects of research & development funding on scientific productivity: Academic chemistry, 1990–2009. *PLoS One*, *10*(9), e0138176. https://doi.org/10.1371/journal.pone.0138176
Schnell, J. D. (2018). Web of Science: The first citation index for data analytics and scientometrics. In F. J. Cantú-Ortiz (Ed.), *Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics* (pp. 15–29). Boca Raton, FL: Taylor & Francis.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269. https://doi.org/10.1002/asi.4630240406
Small, H., & Griffith, B. C. (1974). Structure of scientific literatures. 1. Identifying and graphing specialties. *Science Studies*, *4*, 17–40. https://doi.org/10.1177/030631277400400102
Somoza-Fernandez, M., Rodriguez-Gairin, J. M., & Urbano, C. (2018). Journal coverage of the Emerging Sources Citation Index. *Learned Publishing*, *31*(3), 199–204. https://doi.org/10.1002/leap.1160
Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy 48*(6), 1362–1372. https://doi.org/10.1016/j.respol.2019.01.019
Visser, M., van Eck, N. J., & Waltman, L. (2019). Large-scale comparison of bibliographic data sources: Web of Science, Scopus, Dimensions, and Crossref. In G. Catalano, C. Daraio, M. Gregori, H. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th Conference of the International Society for Scientometrics and Informetrics* (vol. 2, pp. 2358–2369). Rome: Edizioni Efesto.
Wagner, C. S., Whetsell, T. A., and Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality & knowledge recombination. *Research Policy*, *48*(5), 1260–1270.
Zuckerman, H. (2018). The sociology of science and the Garfield effect: Happy accidents, unanticipated developments and unexploited potentials. *Frontiers in Research Metrics and Analytics*, *3*, 20 https://doi.org/10.3389/frma.2018.00020