Essays in the Art and Science of Academic Journal Editing and Publishing

The Complex Universe of Citation Data for Bibliometric Systems

David Anthony Rew, MA MB MChir (Cambridge) FRCS (London)

Honorary Consultant Surgeon to the Faculty of Medicine, University of Southampton, UK and to the Clinical Informatics Research Unit.

Former Editor in Chief, European Journal of Surgical Oncology, 2003 to 2009

Former Member of Council, Committee on Publication Ethics, COPE, 2008-2010

Subject Chair for Medicine to the SCOPUS Content Selection Advisory Board, Elsevier BV, The Netherlands, 2009 to the Present

This is a Working Paper for publication on the ePrint Server of the University of Southampton. It is published with a CC/BY Creative Commons licence for attribution to the Author.

8th June 2025

Correspondence to dr1@soton.ac.uk

Key words

Citation Analysis; Scopus; Web of Science; Citation Fraud; Bibliometric Systems; Citation Source Documents; CrossRef; Metadata Manipulation; Citation Cartels; Coercive Citations; Citation Planting; Ranking manipulation, Article Retractions; Fake Reviews

Contents

Abstract

Citation analysis has been the foundation of bibliometrics and of academic performance measurement for 70 years. Citations are based on the references and information networks which underpin academic writing. They are regarded as a proxy for the significance, importance or respect in which the cited article is held and of academic performance. Citations are an imperfect form of the measurement of the impact of ideas, of individuals and organisations, but they underpin a huge global investment in academic appraisal, performance evaluation and promotion systems.

SCOPUS and the Web of Science (WoS) are commercial citation systems which support this information ecosystem with quality assurance processes. They process selected academic journals, books and other sources into core collections with detailed author, article and journal based bibliometric profiles. These core collections are regarded as **primary sources** for citation analysis.

Beyond the core collections lie a large number of citation sources, which are identified from the primary sources but which are not curated by SCOPUS or WoS. There are known as **secondary sources**.

Outwith the primary and secondary sources is a large volume of uncurated **tertiary content** whose size unknown and which is neither linked nor readily targetable for bibliometric analysis. These spheres of sources can be modelled as a "bibliometric universe of citation activity", which I explore further in this essay.

Citation based career recognition creates perverse incentives to game the citation system for personal or institutional gain. Many sophisticated schemes have been devised to create false and dishonestly enhanced citation scores.

Efforts must therefore be made to educate the global academic community on the benefits and limitations of citation based evaluations; to maximise the trustworthiness of bibliometric data; and to develop methodologies which minimise the opportunities to game the system for fraudulent purposes.

Introduction

References are a long established element of academic writing. They are usually listed in the bibliography at the end of an academic paper or book chapter, and sometimes in page footnotes and margin notes. They establish the history of an idea and they give credit to those who created the knowledge which underpins any new article. References state the authors, the title of the article, the source of the article, and supplementary data which helps the reader to trace the source.

There are well in excess of 1000 million references in the global academic literature, and each reference is de facto a CITATION to another source. Since around 2000, citations have increasingly been enhanced with a unique alphanumeric Digital Object Identifier (DOI) which is often created as a web hyperlink. However, DOI numbers are as yet by no means universally applied.

Bibliometrics is the science of Citation Analysis. It has been at the heart of academic performance measurement since it was first developed by Eugene Garfield (1925-2017) 70 years ago. Garfield is widely regarded as the father of the discipline. He received a PhD in Linguistics from the University of Philadelphia, from where he went on to found the Institute for Scientific Information (ISI). His insights were founded in the Shephards Citation System, which was a long established methodology for organising US legal case records.

Computerisation has substantially advanced the discipline since his early work on citation indices on punch cards and the introduction of the concept of the Impact Factor of a journal. A substantial commercial quality assurance industry has developed around the core concepts. A huge and complex literature exists around the mathematical and statistical analysis of citations in the academic and research literature.

The basic principle of bibliometrics relates to the methodologies and referencing (citing) of an academic article by another article. The reference is inferred to be a proxy for the significance, importance or respect in which the cited article is held. Therefore, more frequently and heavily cited articles will be regarded as being more significant and influential than less cited articles. This in turn will reflect well on the authors of the paper

and their sponsoring institution, and upon the journal, book, conference proceedings or patent in which the cited work was published.

Careers, institutions and businesses have prospered or foundered on the back of bibliometric measures of performance, and the discipline is deeply embedded in academia and the publishing industry. Many new bibliometric measures have been introduced to account for the performance of individual authors, articles and journals and to give greater depth and granularity to Garfield's original concept.

It is not the purpose of this essay to examine the mathematics or the computational methodology of bibliometric measures. However, it is important that the significant nuances and limitations of the bibliometric system are understood by those who use and depend upon it. I am therefore seeking to pass on some of the insights which I have gained from general observation of the citation system at work and from the complexities around trust and integrity which follow from citation dependency. My understanding of citation analysis has been informed by the continuing development of the SCOPUS citation system and by immediate colleagues with a deep understanding of the operational details of citation analysis.

The Limitations of the Citation System

Citations are the metadata of knowledge, but there is a substantial disconnect between the content of an academic source document, its citation receipts, and its practical impact. Most people who learn about original academic work will do so through the filtration of the written or other media, as for example in newspapers, film, and multimedia outputs, which we do not cite on a daily basis. Only a very small proportion of those who are affected by the output of an academic work will revert to the source document, and even then it may not be read in full, let alone cited.

Furthermore, citations are context agnostic. Citations which highlight the malign content of a cited article will be counted in the same way as are positive and appreciative references. Therefore, a "bad actor" article may achieve high citation counts through notoriety. Moreover, reference collections are rarely comprehensive. In fields where there may be a

number of relevant papers, citing authors may only cite the source material selectively to support a particular point. Moreover, the original papers may be misquoted or misrepresented and inappropriately cited.

The Observable Universe Model of Citation Activity.

Source references need to be discoverable and processed by one or more of the major bibliometric calculation systems if they are to have academic impact. This imbues those citation systems which are underwritten by a Quality Assurance (QA) process with considerable influence. SCOPUS and the Web of Science (WoS) select the content which they include in their core collections, and they have invested heavily in the curation of the sources which they include in their data sets to create detailed author, article and journal based bibliometric profiles. Beyond the SCOPUS and the WoS core collections of each system are rings or spheres of citation activity which are less well curated.

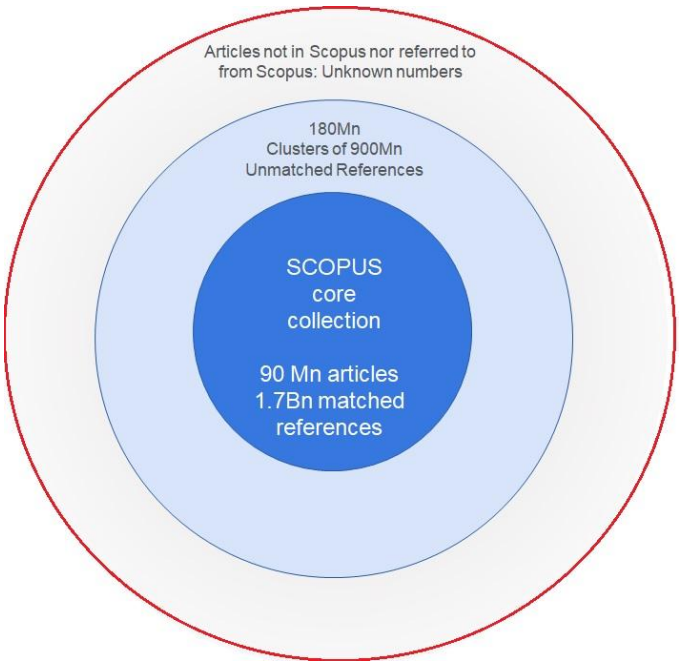

Figure 1: The Observable Universe of Citation Activity, as described by Dr Rob Schrauwen, of the Elsevier Research Data Platform, based upon data from May 2023.

Dr Rob Schrauwen of Elsevier describes a simple conceptual model to illustrate these data relationships, which he describes as the Observable Universe of citation data. (Figure 1,)

Primary Source Collections and Citations

The SCOPUS core collection at the centre (the "Corpus") consisted in 2023 of ~90 million quality assured, validated articles from journals, books, conference proceedings and patents, with 1.7 billion validated and matched references.

The Primary source collections are themselves complex. For example, the SCOPUS Core Collection consists of both active and inactive titles. Inactive titles no longer receive new content because they have ceased publication or because they have been delisted from active processing in consequence of QA concerns. The SCOPUS Core Collection holds around 37000 titles, of which around 23,000 titles are active and around 14,000 titles are inactive.

This is a dynamic data set, as new titles are added or which lapse in print, and as the dynamic SCOPUS re-evaluation system de-selects journals on various grounds. Sources which have ceased publication may nevertheless have a long citation tail, as they may be cited from archival copies long after the publication has gone out of print, and particularly if the source is readily available on the internet.

Such citations are still tracked in SCOPUS and attributed to the source authors, articles and journals. This gives rise to unresolved complexities in the citation systems, as citations from articles in discredited and redacted journals and books may also continue to accumulate within the Core Collections, given the technical challenges of validating individual citations (Cortegiani et al 2020).

Secondary Citation Sources

There is also a large body of sources of identifiable citations, which lie outside the curated content of SCOPUS and WoS. These are known as **secondary sources**. These sources are known about from citations within the primary sources. These citations are catalogued but their source journals, book series or books are not curated in detail by SCOPUS and/or WoS.

Beyond primary and secondary sources is a global "outer sphere" of **tertiary content** whose size is not accurately known or which is not subjected to bibliometric analysis by SCOPUS or WoS. This unseen content may nevertheless hold valuable works and reference material

which does not contribute to the quality assured bibliometric ecosystem. These spheres of primary, secondary and unknown sources can be represented as concentric spheres.

Outside this core were more than 900 million references or Secondary Citations in 2023, which were organised into 180 million clusters of references. Each cluster may be matched to a publication source (book or journal) which is not curated within the SCOPUS Core Collection.

The Tertiary Sphere of Citations

Outside these two spheres is a sphere of unknown numbers of "tertiary" knowledge sources. These create an unknown number of references which are not captured in any way by the SCOPUS system. They will always be too disparate to permit practical and economic identification, capture and systematic quality assurance within the core collection of a major commercial citation system. This lack of total visibility creates a fundamental constraint on the precision of any citation system in terms of measuring the overall societal impact of academic outputs through bibliometric measures.

The spheres of content are continuously evolving. As of June 2025, the SCOPUS core collection contained 101M articles, with more than 1.7Bn matched references and 3.4Bn references overall, at an average of 29.05 citations per article.

Non-Curated Bibliometric Systems

Reference sources are also discoverable by the public search engines, among which are Google and Google Scholar, or on open source data sets such as OpenAlex, which was previously known as Microsoft Academic Graph.

OpenAlex (formerly Microsoft Academic Graph)

OpenAlex is not curated and it draws its data from a range of sources which claims to index *"more than 250M scholarly works from 250k sources, with extra coverage of humanities, non-English languages, and the Global South, and to link these works to 90M disambiguated authors and 100k institutions".*

OpenAlex also enriches the data with topic information, Sustainable Development Goals (SDGs), and citation counts. In terms of our bibliometric universe model, the OpenAlex collection spans the SCOPUS Core Collection, Secondary documents, and the outer sphere of documents and citations which are not processed in SCOPUS or WoS.

Google and Google Scholar

The Google systems also calculate bibliometric data, and Google Scholar creates individual author profiles with citation metrics. The resulting data are widely used in academic evaluations. Dimensions and CrossRef are other systems which contain large numbers (>150M) of source documents However, these systems currently lack the detailed curation, selectivity and QA oversight of SCOPUS and WoS, so they risk a greater proportion of citations from fraudulent sources (Ibrahim et al 2025).

There are many explanations for this large ecosystem of sources which are neither primary nor secondary sources in SCOPUS or WoS. They include:

- journals and sources which have been excluded from the citation systems following failures in the evaluation during the quality assurance and selection processes;
- journals and sources which are eligible for consideration for inclusion in SCOPUS or WoS but which have not been submitted, from where-ever and for whatever reason;
- journals and sources which predate the content coverage of SCOPUS and WoS. For example, rigorous SCOPUS coverage goes back to 1970, but the costs and challenges of sourcing and processing older and historic material have generally precluded their inclusion in the data base;
- the lack of standard structured or indeed any abstracts in older sources;
- sources for which permission has not been obtained for inclusion in the citation systems on commercial or other grounds;
- sources which fall outside the inclusion policies of SCOPUS or WoS;
- sources in languages other than English which have not been tapped.
- private, confidential and non-disclosable sources, for example arising within Government laboratories and research collections, and which relate to sensitive matters such as weapons technology.

The Unique Identification of Individual Citations

In order to process citations efficiently and accurately within citation systems, it essential that each of the >1.7 Bn citations can be reliably and uniquely identified and found by computers and researchers. In the course of curating sources and references, citations have in the past been processed manually by SCOPUS and WoS, and allocated unique proprietary reference numbers. This is essential because an individual reference can be written in many different formats in the presentation of names, in the wording of article titles and in the representation of the source titles. The bespoke proprietary citation metadata in the commercial systems which resulted led to challenges for the wider academic community, and to the creation of a "supplier-agnostic" identification system for citations through CrossRef and the Digital Object Identifier (DOI) system.

The Role of Crossref

Crossref was founded in 2000 as a not for profit organisation for the facilitation of scholarly communication and for the enrichment of publication metadata. Ed Pentz, the first employee of Crossref, recounts the history in a recent website article:

"Crossref was born of the radical changes in the 1990s brought on by the spread of the Internet and development of the World Wide Web and other technologies (HTML, SGML, XML). Everything started moving online, including research and scholarly communications. Our roots go back to 1996 when the Enabling Technologies Committee of the Association of American Publishers put out a call for a persistent identifier system for online content. The Corporation for National Research Initiatives (CNRI) answered the call with the Handle system.

Further work and discussions led to the founding of the International DOI Foundation to develop and govern the Digital Object Identifier (DOI) System which was the application of the Handle System to the digital content space.

Things came together 1999 with the formation of Crossref. ... A prototype project by Academic Press, Wiley, and the DOI-X project, created the technical foundations for reference linking based on centralized metadata and the assignment of DOIs. The prototype was demonstrated at the Frankfurt Book Fair in 1999. Publishers quickly rallied around and

in December 1999 a working group of 12 organizations met and decided to form Crossref as an independent, not-for-profit organization. Crossref was incorporated in January 2000 - as Publishers International Linking Association, Inc. (PILA). The Crossref system went live in June 2000."

Crossref now has more than 20,000 contributing members among publishers, research institutions, Universities, funding bodies, museums, data repositories and other content creators around the world. These include Elsevier (for SCOPUS) and Clarivate Analytics (for Web of Science). Each contributing organisation creates standard Digital Object Identifiers (DOIs) for metadata records that describe and locate their research under Crossref supervision and governance.

The Digital Object Identifier

The Digital Object Identifier (DOI) code is an attempt to create a universal system for the unique identification of any and every citation. The DOI is an alpha-numeric code and hyperlink which is allocated by the publisher at the time that the article is accepted for publication. When applied to academic journal articles, every article which been allocated a DOI can now be readily tracked. The unique DOI is increasingly quoted as part of the reference /citation in various formats.

The DOI system is governed by the DOI Foundation, which was registered as a non-stock membership organisation under the General Corporation Law of the State of Delaware, USA, in October 1997. The Board of DOI represent a range of organisations, including CrossRef and The Publications Office of the European Union. Crossref is now the primary allocation agency for academic journal DOI's.

The DOI Foundation is the registration authority for the ISO standard (ISO 26324) for the DOI system. In turn, it is governed by the Registration Agencies (RAs) which allocate DOI prefixes, and register DOI names. It provides a metadata schema that is associated with each DOI record. Any organization from any commercial, governmental, or not-for-profit sector that is willing to make long-term commitments to the persistence and sustainability of the DOI System can apply to become an RA.

DOI Registration Agencies commit to delivering a reliable and consistent DOI service to their users. They agree to abide with common agreements and policies. They sign an Agreement which grants a set of rights; which stipulates the obligations of the RA and the DOI Foundation; which makes clear the intellectual property rights; and which details change and termination procedures and continuity considerations.

The DOI is a persistent identifier of any physical, digital or abstract object to which it has been allocated. It **is also a dynamic locator which links** to metadata about the object, including the Universal Resource Locator (URL). The persistent feature permits reliable tracking and discovery of the item, through the DOI proxy web servers, even if its web links or database records change over time. There is logical structure to the DOI allocation.

DOI numbers for academic articles

All DOI numbers for academic articles begin with a **10** , followed by a prefix and a suffix which are separated by a forward slash. The prefix is a unique number of four or more digits which identifies the Registration Agency. The suffix is assigned by the publisher to the specific digital object. It can vary in length and structure, and some publishers may use journal abbreviations, ISSN, publication year, or other identifiers.

Importantly, an article is persistently tracked, even if the journal wrapper itself changes. The DOI is therefore not designed to be a permanent journal or source tracker, whose identity and ownership may change over time or even terminate.

There are a number of very significant limitations to the use of the DOI, in that:

- DOi allocations of all types did not take off until around 2015 (see www.doi.org/the-identifier/what-is-a-doi/). Therefore, most citations which predate the creation of the DOI system will not carry a DOI number.

- Many publishers now require the allocation of a DOI to all published manuscripts, but others do not. It is not currently possible to discover the actual numbers, proportions and adherence to the DOI strategy.

In 2023, Turki et al highlighted that "DOI's are (now) important metadata elements in scholarly communications. They observed that while the academic "publishing majors" are heavily invested in the DOI model, many smaller publishers have not yet engaged. One consequence is that DOIs are unevenly used in bibliometric systems. Articles which do not carry DOI numbers will be under-recognised in bibliometric citation systems. (Turki 2023).

Nees Jan van Eck and colleagues explored the relationships between citation data in Crossref, SCOPUS and Web of Science in detail in a blog post in 2018 (van Eck 2018). They noted that at that time, *"a large share of the scholarly literature as indexed in WoS and Scopus is also available in Crossref. For recent years, 68% of the WoS listed publications and 77% of the Scopus listed publications can be matched with Crossref using DOIs."*

They also noted that *"these figures may underestimate the true overlap between the data sources, since matching based on DOIs presents several difficulties, such as missing, incorrect, and duplicate DOIs. To improve matching, publishers and data providers need to work together to offer more comprehensive and more accurate DOI data".* There is of course a substantial cost to such data cleansing, which is borne by the major commercial systems in optimising the integrity of their own data sets.

These observations are summarised in an evolution of Rob Schrauwen's observable universe model, as in Figure 2. In this representation, the overlying DOI oval contained 146 Million articles from CrossRef in May 2023. 17 Million articles also had abstracts, and 37 Million articles had references.

Within Scopus, 70 Million articles had DOIs in 2023. This data highlights the continuing challenges of aligning the SCOPUS (and WoS) content with the wider universe of bibliometric data. However, it also suggests cooperative routes to greater identification and integration with bibliometric sources which presently lie outside the SCOPUS core data set.

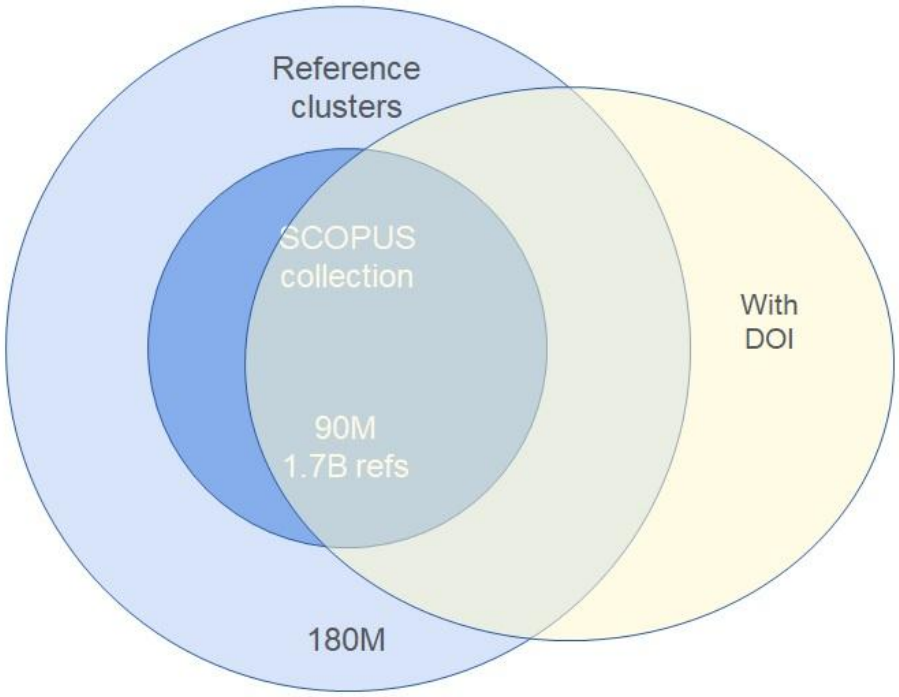

Figure 2. The citation data in the observable universe with the Digital Object Identifier (DOI) overlay on the primary and secondary sources in May 2023 (courtesy of Dr Rob Schrauwen)

The Initiative for Open Citations (I4OC)

Despite the CrossRef and DOI initiatives, it was apparent to many observers that there were still impediments to universal adoption of a unitary citation identification system. The I4OC website noted that *"the present scholarly communication system inadequately exposes the knowledge networks within our literature. Citation data are not usually freely available to access. They are often subject to inconsistent, hard-to-parse licenses, and they are usually not machine-readable".*

The Initiative for Open Citations (I4OC) was launched in 2017. It was established "as a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data"(https://i4oc.org/.). The authors of the I4OC website noted that *"the number of scholarly publications is estimated to double every nine years. Citations and the computational systems that track them would allow researchers to track significant developments in any subject field, if they were given unrestricted access to bibliographic and citation data in machine-readable form".*

I4OC was created following the publication of a leading article in Science by Dalmeet Singh Chawla in April 2017, arising from the eighth Conference on Open Access Scholarly Publishing in September 2016 (Chawla 2017). He noted that:
"...The Initiative for Open Citations (I4OC) aims to make citation data free to all, in partnerships which include the Wikimedia Foundation, the Public Library of Science, and the open-access journal eLife*. So far, the initiative has partnered with 29 journal publishers to enable anyone to access citation data from about 14 million papers indexed by Crossref.*

Initially, only 3% of almost a thousand publishers who were depositing data on Crossref were making citation data openly available. This meant that citation data were available for just 1% of the ~35 million papers on Crossref. That share has risen to more than 40% of Crossref papers as a result of I4OC's efforts... I4OC will allow users to freely access and reuse citation data under CC0, the most liberal copyright license".

As of June 2022, all members of Crossref agreed not limit the distribution of their references, and all deposited references in Crossref are now treated as open metadata. More than 5000 publishers have now submitted to Crossref the references with a Crossref DOI from at least one publication. This list is updated by I4OC every two months.

However, I4OC noted that more than 13,000 other publishers who were depositing publication metadata with Crossref, were failing to submit references with other publication metadata, including publication abstracts, ORCID author identifiers, and funder information. Nevertheless, as of August 2022, the fraction of publications with open references has grown from 1% to 100% out of 61 million articles in Crossref.

The Outer Reaches of the Bibliographic Universe

Efforts to incorporate citations from the lesser navigated outer reaches of the bibliographic universe into the major citation systems continue. These include work on the well defined collections of the major Preprint servers and the global Patent Libraries, whose content has been increasingly incorporated within SCOPUS (Figure 3); and work on the Dissertations and Theses collections from Proquest, which are now linked to the Web of Science.

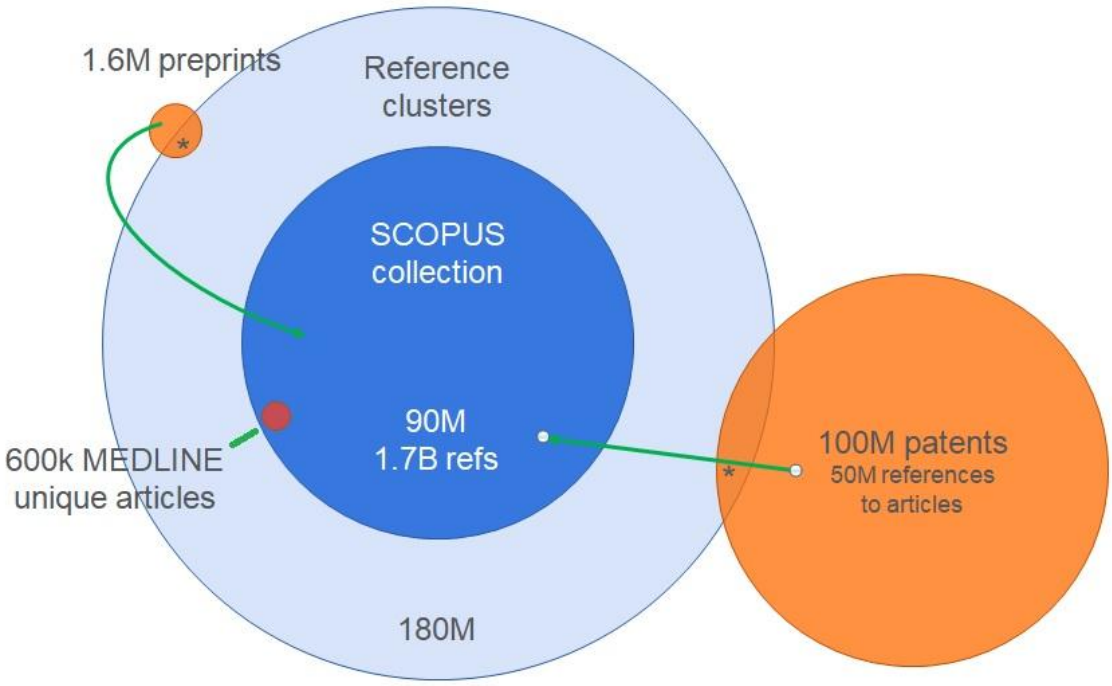

Figure 3. The relationship of Patents, Preprints and Medline-specific articles and references to the SCOPUS Core Collection: (Image and data Courtesy of Dr Rob Schrauwen)

At the centre of the SCOPUS citation universe, the Medline collection of some 600,000 unique articles (in 2023) is independently curated by the US National Institute for Health, but it is fully incorporated within SCOPUS.

Fraudulent Citation Activity and Citation Manipulation

Given that citations have acquired critical importance within academic career development, funding decisions and institutional status, it is no surprise that a range of fraudulent models have emerged to take advantage of the complexities of bibliometric science and of computer data systems to boost the citation counts. Such fraud is generally characterised as Citation Manipulation.

Hazem Ibrahim and colleagues have recently highlighted in detail a range of citation manipulation strategies (Ibrahim et al 2024, 2025). They acknowledge that malpractice can be hard to define, as the line between innocuous and malicious behaviour may be unclear, given the natural tendency of all academics to seek to optimise their personal citations.

They explored the range of data sources which are used to assess citation performance of individuals. Google Scholar was a particularly popular source, along with Web of Science, The US Patent and Trademark Office (USPTO),Microsoft Academic Graph (MAG), and unspecified discipline-specific systems. The curational role of the Content Selection and Advisory Board of SCOPUS was highlighted as contributing to a 96% reduction in anomalously sourced citations when compared to the uncurated Google Scholar system.

They proceeded to study 1.6M author profiles from Google Scholar, noting its “less restrictive indexing protocols”. They noted its indexing of unmoderated personal websites, specific pre-print servers, and systems such as ResearchGate, which seemingly facilitated manipulation practices which were not discernible on other (curated) bibliometric databases.

Anomalous Citation Patterns and Citation Planting

Ibrahim and colleagues identified five anomalous authors whose Google Scholar profiles appeared to be irregular when compared with their peers. They found that anomalous authors were cited from 15 to 45 times in single papers. The anomalies persisted when various explanations were further explored. The inappropriate citations were also unrelated to the paper in which they appeared. This indicated that they had almost certainly been artificially planted in the host article.

Moreover, the planted references were rarely referenced in the main text of the citing paper. They highlighted one citing paper of two pages, which contained one reference in the main text, but which referenced the anomalous author 29 times in its bibliography.

The concepts of the Citation Concentration Index and Citation Concentration Percentage

In consequence of their findings, Ibrahim et al proposed a Citation Concentration (CC) Index, which may also be regarded as a Suspicion Index. This hypothesised index would be based upon the number (n) of papers that cite an author at least (n) times.

They also calculated a Citation Concentration (CC) Percentage. The is the percentage of citations that an author with a CC Index of (n) receives from the papers that cite them at least (n) times. They noted that *"while it is impossible to prove whether a citation has been manipulated, their CC Percentage should capture the proportion of citations to the author that may be manipulated".*

Most authors have a CC index of around one, as they are rarely cited more than once in any one paper. In contrast, for their anomalous authors, CC indices ranged from 23 to 45. Moreover, the highest percentage of anomalous authors were associated with non-peer reviewed papers which were posted on ResearchGate.

The Misuse of PrePrint Servers to Boost Citation Metrics

Ibrahim et al went on to investigate the data on 114 authors in their data set whose Google Scholar profiles appeared to be anomalous. Preprint servers appeared to be associated with high numbers of self-citations. They created a fictional author and 20 papers on "fake news" which were generated by ChatGPT. The articles only included references to each other. They were uploaded to three different pre-print servers.

The fictional author and papers were uncritically indexed in Google Scholar but not in SCOPUS or WoS. Once the fake papers had been removed from the preprint servers, the citations that they had generated persisted in Google Scholar. This work evidenced a route to the fraudulent exaggeration of an author's citation counts.

Citations for Sale and Purchase

Ibrahim and colleagues then exposed citation boosting services, which will sell fake citations in bulk. They observed that *"For such a transaction to take place, it requires three culprits: the researcher who purchases the citations; the researcher who plants them in their own articles in return for a fee; and an individual or company who brokers this transaction".*

They considered that within their cohort of authors with suspicious profiles, one or more may have engaged with citation boosting services for sale and purchase. They followed the trail to a company which provided citations to one of the suspicious scientists. Using their 20 fictional research articles and their fictional author, they created a fictional profile and contacted the company to ask them to boost the citations of this profile through the purchase of 50 citations. They thus established that citations can be bought in bulk for a relatively small fee in a matter of weeks.

They concluded by questioning the reliance on citation metrics when evaluating scientists, and the safety of Google Scholar in the evaluation of researchers. Specifically, they established that high citation counts alone do not identify misconduct, and that unravelling intentional misconduct is extremely challenging.

They note that " *For example, taking the Google Scholar profile of the scientist who is cited 167 times in a single paper p, one cannot discover misconduct without collecting and analyzing all the citations received by that author from the 167 papers which are cited in p. Consequently, those who engage in such practices can easily go unnoticed by the casual observer in browsing Google Scholar profiles."*

Identifying clusters of papers based on bibliographic coupling

Ibrahim and colleagues went on to report that: *After identifying the journal from which our purchased citations originate, we hypothesize that this journal likely contains more citations that were sold in bulk. To understand whether this is the case, we first collected all papers published therein in the first half of 2023. Then, we construct a network where each node in the network represents a paper, and two papers are connected with an edge if they share at least one identical reference...*

As a result, several connected components emerged in this network of papers. In each connected component, they then scanned reference lists for authors who appeared at least 10 times across multiple papers. These authors were identified as potential candidates for having purchased citations. Then, we manually confirm whether the set of referenced papers was consistent among all citing papers. Such consistency is highly improbable by random chance but likely when citations are acquired in bulk. Our investigation unveiled a total of 11 anomalous authors distributed across five distinct connected components of papers. These authors received citations of at least 10 times per paper from multiple sources.

PrePrint Servers and the Scope for Fraudulent Citation Activity

In 2020 and 2021, before the work by Ibrahim et al came to light, the SCOPUS Content Selection Advisory Board debated at length the pros and cons of including the content of PrePrint servers in SCOPUS. We settled on a plan whereby Preprints from arXiv, bioRxiv, ChemRxiv, medRxiv and the Social Science Research Network (SSRN) servers would be accepted and subject to their respective curation policies with coverage from 2017 onwards. As non-peer-reviewed publications, Preprints would not affect existing publication and citation metrics in Scopus.

As stated by Rachel in her corporate blog post (McCullough 2021), the reasons for incorporating PrePrints into SCOPUS were that:

- A preprint is a version of a scholarly paper that precedes publication in a peer-reviewed journal and it acts as an early indication of research.
- Preprints reside on preprint servers, which cover a set of domains and allow for dissemination, laying claim to an idea, and help collect feedback prior to submission.
- In some fields, preprints are the main communication vehicle.
- Preprints differ from Articles-in-Press in that preprints are not peer-reviewed and not accepted for publication in a journal.
- Preprints would enhance Scopus data as an evaluation tool for scholarly output and as decision support tool in terms of researcher performance.
- Preprints provide a more detailed view of researchers and what scholarly work they create over the course of their careers.

Preprints also support a number of use cases, including the identification of collaboration partners; the assessment of a researcher's most recent work and to obtain a more comprehensive portfolio overview; and to allow funding agencies to assess funding applications, monitor project progress, and demonstrate impact in early forms of scholarly output.

The announcement stressed that:
- Preprints are only available for authors who already have a peer-reviewed publication history in Scopus and they are clearly separated from the curated published content.
- Citations to-and-from the preprints, and links with the final version of the article are not captured.
- Therefore, metrics on Scopus, such as publication and citation counts and the h-index would exclude preprint content.
- The version-of-record (published, peer-reviewed articles) would remain the official representation of research in Scopus.

By September 2021, more than 900k preprints had been added to author profiles in SCOPUS. Preprints were deemed as a more valuable signal of research focus than meeting abstracts, which are not a complete and unique record of research and which can drive ambiguity or duplication. Clarivate subsequently announced the inclusion of preprints and a "Preprint Citation Index" to their WoS product portfolio on February 2023.
Ibrahim et al nevertheless noted that *"these measures do not rule out the possibility that citation manipulation affects other bibliometric databases, as authors are free to determine which references to add to their paper, making it impossible to eliminate disingenuously implanted citations".*

They further noted that these issues therefore concern academia as a whole, and that there were two factors that may further exacerbate the issue of citation manipulation, which were:
- The over-publication of special issues, which bypass peer review, and;
- The emergence of generative AI technologies.

They suggested policy and technical adaptations to mitigate citation manipulation, which in turn are closely related to other forms of academic misconduct and "industrialised cheating", including predatory publishing and paper mills. These adaptations include the development of bibliometric databases to include metrics which are designed specifically to track how citations are accumulated by any given scientist, including self citation metrics. Policy improvements include strategies to raise awareness of the complexity of citation malpractice among academic evaluation committees

Metadata Manipulation, "Sneaked References" and Publisher Level Fraud

Metadata manipulation is the strategy of adding references to the metadata of published articles, even though the references do not appear in the overt published lists of references of those articles. This malign activity takes advantage of the way in which bibliometric systems process the data on accepted journals and articles electronically. Besancon and colleagues reported in 2024 that:

"This manipulation exploits trusted relationships between various actors: publishers, the Crossref metadata registration agency, digital libraries, and bibliometric platforms. Extra references are sneaked into the system at Digital Object Identifier (DOI) registration time, resulting in artificially inflated citation counts.

In a case study of three journals from one publisher, they identified at least 9% such references (5978／65,836) which mainly benefitted two authors. These references only exist in metadata registries and they propagate to bibliometric dashboards. They also discovered "lost" references: the studied bibliometric platform failed to index at least 56% (36,939/65,836) of the references which were present in the HTML version of the publications.

Besancon and colleagues observed that:

This manipulation is made possible because Crossref trusts publishers to extract, report, and send them metadata about the publications, including the references. This trust is bound under their membership terms which include keeping metadata accurate and up to-date. ... Effectively, because Crossref is not checking the accuracy of the metadata provided by publishers, this creates a "breach" within the information flow."

They concluded that *"the extent of the resulting distortion in the global literature remains unknown. It requires further investigations and bibliometric platforms which produce citation counts should identify, quantify, and correct these flaws to provide accurate data and to prevent further citation gaming".* (Besancon 2024a)

This sophisticated technical fraud appears to arise at the publisher level. These investigators cited the Indian Open Access publisher Technoscience Academy and a Hindawi article which has since been retracted. Although the root cause for the fraud has not been investigated or explained, the inference is that money passed from the beneficiaries whose citation counts had been inflated to the corrupt publisher.

In the matter of the potential impact of this fraud on SCOPUS, of the three example journals published by *Technoscience Academy*, only one was selected for Scopus coverage and that title was discontinued following internal re-evaluation in 2020. Moreover, in contrast to Crossref and Dimensions, where the "sneaked references" were found, Scopus citation matching is not dependent on the DOI and its metadata.

For Scopus, "*the references are captured from the reference list of the original article and through proprietary algorithms to identify possible citation matches. Therefore, "sneaked references" are unlikely to be captured by Scopus as they do not appear in the original article"* (Meester W. Personal communication).

Besancon and colleagues (2024b) point out that citation manipulation has significant consequences for trust in academic outputs, and for the reputations of those who are caught out. They note that where once the documented manipulations involved modifications of the version of record of the published article available in PDF or HTML by adding references to it, citation manipulation by various actors now occurs in many places and at different times during the life cycle of a scientific publication.

The major bibliometric systems, SCOPUS and Web of Science, take these issues very seriously, both with preventive measures to keep unsafe journals out of their systems, and by rigorous internal processes of data validation and cleansing. Nevertheless, such is the

subtlety and creativity of the bad actors, that the best of contemporary defences are breached and dishonest citation activity enters the data system at the author, article, journal and publisher levels.

Citation Cartels (Rings) and Ranking Manipulations

A citation cartel is a group of academic authors who collude to cite one another's publications in order inappropriately to increase their citation counts and/or those of their employing institutions (Kojaku 2021).

For example, Michele Catanzaro reported for Science journal in 2024, how "*cliques of mathematicians at institutions in China, Saudi Arabia, and elsewhere have been artificially boosting their colleagues' citation counts by churning out low-quality papers that repeatedly reference their work. As a result, their universities now produce a greater number of highly cited math papers each year than schools with a strong track record in the field... The stakes are high—movements in the rankings can cost or make universities tens of millions of dollars... citation manipulation is a symptom of a flawed system of evaluation... Citations and similar metrics are not refined enough to monitor individual performance, and people are always going to find ways to game the system."* (Catanzaro 2024a, b)

In another version of this malpractice, a University that is seeking to manipulate rankings may pay highly cited researchers to claim the University as their affiliation (Ansede 2024).

Irrelevant References are a form of citation which includes "citation stacking". In this malpractice, reviewers or editors request citations without proper justification. It also includes "excessive self-citation," and "Trojan citations" where irrelevant works are cited to boost citation counts.

Editorial and Peer Review Pressure and Coercive Citations

In this version of citation planting, editors and peer reviewers may oblige authors to add references which are favourable to them in exchange for assured publication, or artificially to inflate citation rates and journal impact factors as a condition for publication.

Editor Coercion: Eric Fong and colleagues (Fong 2023) reported that "Some editors seek to inflate their journals' citation count by coercing authors to add citations which reference their journal unnecessarily". They noted that "for coercion to be effective, authors must comply with the editor's demands and add those superfluous citations". They hypothesised that editors might use their publication authority to drive compliance under the threat of rejection of manuscripts of those who do not comply.

Data was collected from a survey of academics with responses from more than 1000 scholars who had been coerced. They found that acquiescence is positively associated with the publication decision, and authors who added the coerced citations reported significantly greater publication success than those who resisted. They also found that authors who acquiesced to coercion reported being more likely to submit to coercive journals in the future and to add superfluous, journal-specific citations before submitting manuscripts.

Reviewer Coercion: Jonathan Wren and colleagues (Wren 2019) recounted the tale in the journal Bioinformatics of a reviewer *"who had requested a large number of citations to their own papers as part of their review. After investigation of their most recent reviews, we found that in every review this reviewer requested an average of 35 citations be added, ~90% of which were to their own papers and the remainder to papers that both cited them extensively and mentioned them by name in the title. The reviewer's phrasing strongly suggested that inclusion of these citations would influence their recommendation to the editor to accept or reject the paper.*

The reviewer was unable to provide a satisfactory justification for these requests and Bioinformatics *has therefore banned them as a reviewer. Our investigation also suggests that the reviewer has behaved similarly in reviewing for other journals. This case has alerted us to how the peer-review system is vulnerable to unethical behaviour, and prompted us to clarify the journal's policy on when it is appropriate for reviewers to request citations to their own work, and to suggest how some of the current weak points in the peer-review system can be mitigated, so that this behaviour can be detected more quickly and efficiently".*

Inaccurate Citations misrepresent the original meaning or intentions of the source paper. As such, they may not be exemplars of deliberate or intentional publication fraud, but they nevertheless undermine the quality of the citation system. Hosseini et al (Hosseini 2020) proposed the development of an online annotation tool which they called "MyCites" as means with which to mark and map inaccurate citations. This would allow ORCID users to annotate citations and alert the authors of the cited and citing articles and the editors of there journals where inaccurate citations were published. Each marked citation would travel with the digital version of the document as persistent identifiers, and would be visible on websites that host peer-reviewed articles.

Fake or hijacked journals and websites are another vehicle for introducing advantageous citations into the bibliometric system. They often involve the creation of fraudulent websites or the cloning (functional duplication) of legitimate journal websites which trick unsuspecting authors into the submission of articles and the payment of publication fees. In the case of hijacked journals, they may duplicate an existing and respected title, or a website, or reactivate a title that has previously been discontinued.

Writing for Retraction Watch in November 2024, Ellie Kincaid reported on "a New hijacking scam that targets Elsevier, Springer Nature, and other major publishers (Kincaid 2024). She wrote that**:**

Until recently, journal hijackers do not appear to have targeted titles from big publishers, in part because their well-known website designs made such clones easy to detect.
Typically, cloned versions of journals' websites are of low quality and don't resemble the recognizable and professional designs of Springer Nature and Elsevier. As described in previous posts, fraudulent publishers would usually copy the ISSN, title and other metadata of niche and university journals in order to avoid identification, and possibly index their unauthorized content in bibliographic databases such as Scopus or Web of Science.

We have catalogued over 300 such cloned journals in the Retraction Watch Hijacked Journal Checker. A small number of these involve major publishers like Springer Nature, Elsevier and Wiley. For example, earlier this year the Journal of Academic Ethics *and* Machine Intelligence Research *which were both published by Springer Nature, were cloned.*

In November 2024, William Black, founder and CEO of PSIref, an online platform aggregating scholarly publication data which offers advertising opportunities for publishers, sent me evidence of a new, more sophisticated scam.

The company "Springer Global Publication" – which is not affiliated with Springer Nature – has published dozens of papers cloning the websites of journals officially published by Elsevier, Springer, the American Medical Association and more. The company had advertised a variety of services on its website, including finding a writer for research papers, editing manuscripts, developing research proposals, analyzing data and managing the peer review process. This collection of services is a classic attribute of a paper mill.

Ellie Kincaid also noted that:
"Although this problematic publisher has registered only 13 journals with Crossref, this type of scam allows the publication of papers in an unlimited number of journals from legitimate publishers. The website of "Springer Global Publication" also lists an additional three fake journals which are not registered with Crossref: Springer Global Journal of Literature & Linguistics, *Springer Global Journal of Humanities and Social Sciences*, and *Springer Global Journal of Economics and Management"*.

The domain of this company was registered on Sept. 18, 2024. Two associated domains, springer.uk.com and sciencedirects.com, were registered on Nov. 11, and Sept. 18, 2024, respectively.Papers published on the cloned websites appear to have content recycled from other sources. The company will most likely offer to publish papers in legitimate and reputable journals, but submitted papers will instead appear on cloned versions of the websites. This new scam represents high-quality fraudulent websites, with a remarkable resemblance to legitimate ones, capable of deceiving even experienced researchers. Be aware!"

Paper Mills are rapidly achieving notoriety as ambassadors for fraudulent publication practice. Simplistically, they are viewed as companies which will generate artificial papers to order, thus contaminating the academic literature with false papers and citations. However, Reese Richardson, Spencer Hong and Luís A Nunes Amaral of Northwestern University(2024), wrote about their investigation for Retraction Watch, noted that: *"the term 'paper mill' may fail to fully capture the diversity and scale of activities overseen by these hydra-like conglomerates. Moreover, the high-level view of these networks that we've unravelled suggests an extraordinary ability to adapt and the capability for aggressive growth – a picture of a resilient enterprise. With so many functionally identical business fronts operating concurrently, those operating the network need not worry if one business is identified publicly, one professional society shuts down or one journal is de-indexed; plenty will remain to fill its place. This is evidenced by the survival of the OMICS Group through a series of loosely connected subsidiaries and spinouts.*

Our quest to map out a full network of associations around one particular business revealed this business was just one head on a hidden hydra, ready to sprout two more if that head was lost. If paper mills and their ilk are as tenacious and robust as this example suggests, we should not settle for half-measures that can be easily evaded. Successfully fighting them will require the kind of large-scale coordination they themselves have displayed"(Richardson 2024).

Misclassification Errors and Skewed Citation Data

Even where there is no deliberate fraud, design challenges in the technology of citation analysis may provide misleading results. Alexey Lyutov and colleagues at Constructor University, Bremen have reported on how imprecise journal and article classification in the scientific disciplines can lead to systematic errors in citation calculations.

They noted that *"misclassified articles have different citation frequencies from correctly classified articles: In the highest 10 percent of journals in each discipline, misclassified articles are on average cited more frequently, while in the rest of the journals they are cited less frequently"* (Lyutov et al 2024).

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

RETRACTED

Toxicity of spike fragments SARS-CoV-2 S protein for zebrafish: A tool to study its hazardous for human health?

Bianca H. Ventura Fernandes [a], Natália Martins Feitosa [b], Ana Paula Barbosa [c], Camila Gas... Bomfim, Anali M.B. Garnique [d], Ivana F. Rosa [e], Maira S. Rodrigues [e], Lucas B. Doretto [e], Dani... Cost..., Bruno Camargo-dos-Santos [f], Gabrielli A. Franco [f], João Favero Neto [f], Juliana Sa... Lun...di [f], Marina Sanson Bellot [f], Nina Pacheco Capelini Alves [f], Camila C. Costa [g], Mayun... Ara...i [g], Letícia F. Rodrigues [g], Camila C. Costa [g], Rafaela Hemily Cirilo [f], Raul Marcelino Colagrande [f], Fran...co ...omes [h], ...fael T. Nakajima [e], Marco A.A. Belo [i], Percília Cardoso Giaquinto [j], Susana Luporini de Oliv... [f], Silas F... ...Eto [l], Dayanne Carla Fernandes [m], Wilson G. Manrique [n], Gabriel Conde [o], ... C. Rosal..., Iris Todeschini [q], Ilo Rivero [r], Edgar Llontop [q], Germán G. Sgro [q,s], Gabriel Umaji Oka [q], Natalia ...anda Bueno [q], Fausto K. Ferraris [t], Mariana T.Q. de Magalhães [u], Renata J. Medeiros [v], Juliana M. ...onça-Gome..., Mara Souza Junqueira [x], Kátia Conceição [y], Leticia Gomes de Pontes [z], Antonio Condin... Neto [z], Andrea C. Perez [ab], Leonardo J.G. Barcellos [ab,ac], José Dias Correa Júnior [ad,ae], Er... Gustavo ...orlass [af], Niels O.S. Camara [w], Edison Luiz Durigon [c], Fernando Q. Cunha [ag], Rafael H. Nóbreg... Glauci... Machado-Santelli [ah], Chuck S. Farah [q], Flavio P. Veras [ai,aj], Jorge Galindo-Villegas [ak], Letícia ... Costa-Lo... ...iago M. Cunha [ai,aj], Roger Chammas [am], Luciani R. Carvalho [an], Cristiane R. Guzzo [c], Guilhe... ...aia [ao,*], Ives Charlie-Silva [al,**]

[a] Laboratório de Controle Genético e Sanitário, Diretoria Técnica de Apoio ao Ensino e Pesqu... da Universidade de São Paulo, Brasil
[b] Laboratório Integrado de Biociências Translacionais (LBT), Instituto de Biodiversidade e Sustentab... (IEM), Universidade Federal do Rio de Janeiro (UFRJ), Macaé, RJ, Brasil
[c] Department of Microbiology, Institute of Biomedical Sciences, University ... São Paulo, Brasil
[d] Department of Cell Biology, Institute of Biomedical Sciences, Universit... São Paulo, Brasil
[e] Reproductive and Molecular Biology Group, Department of Morph... Institute ... Biosciences, São Paulo State University, Botucatu, São Paulo, Brasil
[f] Department of Structural and Functional Biology, Institute of Bio... of Bio... São Paulo State University, SP, Brasil
[g] Department of Preventive Veterinary Medicine, São Paulo State Uni... Brasil
[h] Department of Pharmacology, Center of Research in Infl... Ribeirão Preto Medical School, University of São Paulo, Brasil
[i] Brasil University, Descalvado, São Paulo, Brasil
[j] Universidade Estadual Paulista Júlio de Mesquita ... to de Biociências, Departamento de Fisiologia, São Paulo, Brasil
[k] Universidade Estadual Paulista Júlio de Mesquita Filho, ... Brasil
[l] Postgraduate Program in Health Sciences, FIOCRUZ, Federal ... University of Roraima, Brasil
[m] Immunochemistry Laboratory, Butantan Institute, São Paulo, Brasil
[n] Aquaculture Health Research and ... Group, GRUPESA, Aquaculture Health Laboratory, LABSA, Department of Veterinary Medicine, Federal University of Rondônia, Rolim de Moura campus, Rondônia, Brasil
[o] Department of Preventive Vet... Medicine, São Paulo State University, Jaboticabal, Brasil
[p] Department of Cell and Molecular ... Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brasil
[q] Departamento de Biociências, Instituto de Química, Universidade de São Paulo, Brasil
[r] Pontifícia Universidade ... de Mato Grosso, Brasil
[s] Departamento de Ciências Biomoleculares, ... de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brasil
[t] Department ... Pharmacology and Toxicology, Oswaldo Cruz Foundation, FIOCRUZ, Rio de Janeiro, Brasil
[u] Department ... Chemistry, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Brasil
[v] Laboratory of ... Zebrafish Facility, Department of Pharmacology and Toxicology, National Institute for Quality Control in Health, Brasil
[w] Transplantation ... Biology Lab, Department of Immunology, Institute of Biomedical Sciences, Universidade de Sao Paulo, Brasil
[x] Center for Translati... Research in Oncology, Cancer Institute of the State of Sao Paulo, Faculty of Medicine, University of São Paulo, Sao Paulo, Brasil
[y] Laboratory of Peptide Bi... Chemistry, Federal University of São Paulo, Brasil
[z] Laboratory of Human Immunology, Department Immunology, Institute Biomedical Sciences, University São Paulo, Sao Paulo, Brasil
[aa] Department of Pharmacology, Universidade Federal de Minas Gerais, Brasil
[ab] Graduate Program of Pharmacology, Federal University of Santa Maria, Brasil
[ac] Laboratory of Fish Physiology, Graduate Program of Bioexperimentation and of Environmental Sciences, University of Passo Fundo, Brasil

* Correspondence to: G. Malafaia, Biological Research Laboratory, Goiano Federal Institution – Uruta Campus, Rodovia Geraldo Silva Nascimento, 2,5 km, Zona Rural, Urutaí, Brasil.
** Corresponding author.
E-mail addresses: guilhermeifgoiano@gmail.com (G. Malafaia), charlesilva4@hotmail.com (I. Charlie-Silva).

http://dx.doi.org/10.1016/j.scitotenv.2021.152345
0048-9697/© 2021 Elsevier B.V. All rights reserved.

Figure 4: retraction notice for an article in Science Direct with fake reviews (see text)

Retractions of articles, journals and citations are mandated increasingly frequently when publication malpractice is detected. Sanctions may most easily be applied at the journal level, in which case a journal may be discontinued from a listing in a bibliometric system.

SCOPUS is both responsive to external reporting of fraud and runs regular algorithms ("Scopus Radar") over its entire corpus of journals to detect significant deviations from established publication patterns. Journals whose statistics give rise to concern are then re-evaluated by human subject experts.

When malpractice is detected at the article level, the article may be retracted from the public record and from the bibliometric systems. This should involve removing both the article and the related citations from the system, all be it that this may have a significant, complex and dynamic cascade effect across the wider data set.

Misdemeanours are often discovered after incorporation of a bad item into the data ecosystem by vigilant independent researchers and white knights, whose reports are always followed up and acted upon. However, as evidenced in Besancon's paper, the subtleties of citation crime can be very challenging to detect, requiring the meticulous follow up of suspicions with mathematical, statistical and computing prowess.

The question then arises as to what actually happens once citation fraud is detected. Mrs Tracy Chen of SCOPUS explains that in the matter of discontinued journals:
"SCOPUS relies upon publishers to provide accurate information on journal status and sourcing data. This is not always a timely process. In the matter of article retractions which were published in discontinued journals, retractions with be processed when and where we become aware of them as was the case with a number of Hindawi journals. However, we know that this is not always the case. We aim to continue to record the legitimate content within the scientific record". (Chen T Internal SCOPUS communication)

This raises the question as to whether a journal which has been contaminated with citation fraud should be completely closed, or whether it should be cleansed of the offending content, while protecting innocent content from the axe. Presently, retracted articles are marked as "retracted" but are otherwise not removed from the SCOPUS database. This has the advantage that the articles are appropriately flagged, while still being available for analysis.

Fake reviews and their Bibliometric consequences

The example of a retracted article in Figure 4 illustrates the process when an article is retracted, and the complexities that arise from it. The original paper was published during the covid pandemic with a large number of co-authors.

Following publication in the Elsevier journal Science of the Total Environment, the publisher was alerted to fraudulent reviews, and the following statement was published in the journal:

"This article has been retracted at the request of the Editors-in-Chief. Post-publication, an investigation conducted on behalf of the journal by Elsevier's Research Integrity & Publishing Ethics team determined that two of the reviews for this manuscript were fictitious. Two reviews were submitted under the name of known scientists without their knowledge. The name and fictitious contact details of the reviewers were submitted by the Corresponding Author Guilherme Malafaia during the manuscript submission process.... The Editors-in-Chief have lost confidence in the validity/integrity of the article and its findings and have determined that it should be retracted".

This necessary step nevertheless highlights a number of dilemmas, and it does not necessarily discredit the reported science. It also removes the authorship and contributions of all of the co-authors, whose contributions (unless the article was entirely fake) are now discredited, if not wholly annulled within the bibliometric system.

Caitlin Bakker and colleagues have studied in greater detail the fate of citations from retracted articles. In a paper in the Journal of Clinical Epidemiology in 2024, they noted that: "*Retraction is intended to be a mechanism to correct the published body of knowledge when necessary due to fraudulent, fatally flawed, or ethically unacceptable publications. However, the success of this mechanism requires that retracted publications be consistently identified as such and that retraction notices contain sufficient information to understand what is being retracted and why".*

They investigated how clearly and consistently retracted publications are being presented to researchers, using 441 retracted research publications in the field of public health. Records

were retrieved for each of these publications from 11 resources, while retraction notices were retrieved from publisher websites and full-text aggregators. The identification of the retracted status of the publication was assessed using criteria from the Committee on Publication Ethics (COPE) and the National Library of Medicine. The completeness of the associated retraction notices was assessed using criteria from COPE and Retraction Watch.

2841 article records were retrieved, of which less than half indicated that the article had been retracted. Less than 5% of publications were identified as retracted through all resources through which they were available. Within single resources, if and how retracted publications were identified varied. Retraction notices were frequently incomplete, with no notices meeting all the criteria.

They concluded that the observed inconsistencies and incomplete notices pose a threat to the integrity of scientific publishing and highlight the need to better align with existing best practices to ensure more effective and transparent dissemination of information on retractions (Bakker et al 2024).

In Conclusion

It is clear that there are many complexities in the use of bibliometric systems. These arise both in terms of the practical challenges to the representation of the entire universe of reference generating academic literature within quality assured bibliometric systems, and in citation malpractice for career, financial and/or reputational gain by authors and institutions.

Citations are an imperfect form of the measurement of the impact of ideas, of individuals and organisations, but they represent a huge global investment in professional appraisal systems. These are embedded within the academic evaluation and promotion systems and in the commercial bibliometric information systems which support this ecosystem. Efforts must therefore continue to maximise the trustworthiness of bibliometric data and to develop information exchange systems and sequences which minimise the opportunities to game the system for fraudulent purposes.

Acknowledgements

I am grateful to Professor Julie Cullen of the University of Southampton and to Professor Peter Brimblecombe, recently of National Sun Yat Sen University, Taiwan, and of the University of East Anglia, UK for their reviews of the manuscript.

The observations and resources to which I refer in this essay have been made during my tenure as the Subject Chair for Medicine on the SCOPUS Content Selection Advisory Board. I am grateful to many Board and Elsevier colleagues for contributory observations and discussions on this complex and continually evolving subject over a number of years. The synthesis of this material is entirely of my own volition and should not be construed as representing Elsevier corporate policy in any of the issues which I have discussed.

I am particularly grateful to Dr Rob Schrauwen of Elsevier for his insightful teaching and presentations on various of the topics, and for the use in modified form of a number of his "Robsplainer" diagrams on The Universe of Citation Data.

References

Ansede, M; Dozens of the world's most cited scientists stop falsely claiming to work in Saudi Arabia; El Pais (Spain) Dec 05, 2024 https://english-elpais-com.cdn.ampproject.org/c/s/english.elpais.com/science-tech/2024-12-05/dozens-of-the-worlds-most-cited-scientists-stop-falsely-claiming-to-work-in-saudi-arabia.html?outputType=amp

Bakker CJ, Reardon EE, Brown SJ, Theis-Mahon N, Schroter S, Bouter L, Zeegers MP, The fate of citations from retracted articles. Journal of Clinical Epidemiology,Volume 173, Sept 2024, 111427, ISSN 0895-4356,https://doi.org/10.1016/j.jclinepi.2024.111427.

Besançon L, Cabanac G and Viéville T. Metadata manipulations: Uncovering 'sneaked references' The Conversation July 9th 2024

Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2024). Sneaked references: Fabricated reference metadata distort citation counts. *Journal of the Association for Information Science and Technology*, 75(12), 1368–1379. https://doi.org/10.1002/asi.24896

Catanzaro M. Citation cartels help some mathematicians—and their universities—climb the rankings: Widespread citation manipulation has led entire field of math to be excluded from influential list of top researchers: Science Insider 30 Jan 2024.

Chawla, Dalmeet Singh. "Now free: citation data from 14 million papers, and more might come". Science (6 April 2017): doi:10.1126/science.aal1012

Cortegiani A, Ippolito M, Ingoglia G *et al.* Citations and metrics of journals discontinued from Scopus for publication concerns: the GhoS(t)copus Project *F1000Research* 2020, **9**:415 (https://doi.org/10.12688/f1000research.23847.2)

Eric A. Fong, Ravi Patnayakuni, Allen W. Wilhite, Accommodating coercion: Authors, editors, and citations, Research Policy, Volume 52, Issue 5, 2023, 104754, ISSN 0048-7333, https://doi.org/10.1016/j.respol.2023.104754.

Hosseini, M., Eve, M.P., Gordijn, B. *et al.* MyCites: a proposal to mark and report inaccurate citations in scholarly publications. *Res Integr Peer Rev* **5**, 13 (2020). https://doi.org/10.1186/s41073-020-00099-8Ibrahim H, Liu, F, Zaki Y, Rahwan Tl, Google Scholar is manipulatable 7th Feb 2024 https://arxiv.org/abs/2402.04607

Ibrahim, H., Liu, F., Zaki, Y. *et al.* Citation manipulation through citation mills and pre-print servers. *Nature Sci Rep* **15**, 5480 (2025). https://doi.org/10.1038/s41598-025-88709-7

Kincaid E. Exclusive new hijacking scam targets Elsevier Springer Nature and other major publishers. retractionwatch.com (25th November 2024) https://retractionwatch.com/2024/11/25/exclusive-new-hijacking-scam-targets-elsevier-springer-nature-and-other-major-publishers/

Kojaku, S., Livan, G. & Masuda, N. Detecting anomalous citation groups in journal networks. *Sci Rep* **11**, 14524 (2021). https://doi.org/10.1038/s41598-021-93572-3

Lyutov, A., Uygun, Y. & Hütt, MT. Machine learning misclassification networks reveal a citation advantage of interdisciplinary publications only in high-impact journals. *Scientific Reports* **14**, 21906 (2024). https://doi.org/10.1038/s41598-024-72364-5

McCullough, R. Preprints are now in Scopus! https://blog.scopus.com/posts/preprints-are-now-in-scopus (2021). (Accessed 25th May 2025)

Pentz E. The history of Crossref. https://www.crossref.org/about/history/ (Accessed 8th June 2025)

Richardson R, Hong S and Nunes Amaral LA: hidden hydras uncovering the massive footprint of one papermill's operations: Retraction Watch 1st October 2024 https://retractionwatch.com/2024/10/01/hidden-hydras-uncovering-the-massive-footprint-of-one-paper-mills-operations/

Turki H, Fraumann G, Hadj Taieb MA and BenAouicha M (2023) Global visibility of publications through Digital Object Identifiers. *Front. Res. Metr. Anal.* 8:1207980. doi: 10.3389/frma.2023.1207980

van Eck NJ, Waltman L, Lariviere V, Sugimoto C. Crossref as a new source of citation data: A comparison with Web of Science and Scopus; . Blog Post: Leiden University January 17th, 2018 https://www.cwts.nl/blog?article=n-r2s234

Ventura Fernandes et al. Retraction notice of a paper in Science Direct:
Toxicity of spike fragments SARS-CoV-2 S protein for zebrafish: A tool to study its hazardous for human health? Science of the Total Environment. 813 (2022) 152345

Wren JD, Valencia A, Kelso J. Reviewer-coerced citation: case report, update on journal policy and suggestions for future prevention, *Bioinformatics*, Volume 35, Issue 18, September 2019, Pages 3217–3218, https://doi.org/10.1093/bioinformatics/btz071