The Complex Universe of Citation Data for Bibliometric Systems

An Overview Essay

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.

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

15th March 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

Abstract

Citation analysis has been the foundation of bibliometrics and of academic performance measurement for 70 years. Citations are inferred to be a proxy for the significance, importance or respect in which the cited article is held. They are an important proxy measure for academic performance. This creates perverse incentives to game the citation system for personal or institutional gain, and many sophisticated schemes have been devised to create false and dishonestly enhanced citation scores.

Moreover, the universe of citation activity is more complex than is often understood. The major commercial citation systems, SCOPUS and the Web of Science, can create detailed author, article and journal based bibliometric profiles around those journals and other citing publications which are specifically listed and processed in their systems as primary sources. However, there is a large sphere of citation sources which produce citations to primary documents. These are secondary sources. Beyond primary and secondary sources is a global sphere of content whose size is neither known or readily targetable for bibliometric analysis.

In this essay, I explore the complexities of this “bibliometric universe”.

In conclusion, 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.

Introduction

Citation analysis has been the foundation of bibliometrics and of academic performance measurement for 70 years. Eugene Garfield (1925-2017) is widely regarded as the father of the discipline, although his insights were reportedly originally stimulated by the Shephards Citation System. This was a long established methodology for organising US legal case records. A huge and complex literature has grown up around the mathematical and statistical analysis of citations in the academic and research literature. 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.

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 or conference proceedings 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 challenge the mathematics or methodology of bibliometrics. However, it is important that the nuances and significant limitations of the bibliometric system are understood. I seek to consolidate some insights and nuances from general observation of the citation system at work and of the increasing complexities around trust and integrity issues that can be afforded to the citation concept. I also examine the practical efforts to refine the understanding of citation analysis with lessons from the continuing development of the SCOPUS citation system and its quality assurance.

The Core Constraints of the Citation System

There are a number of fundamental issues that limit the utility of the citation methodology. Ultimately, ideas, concepts and the results of academic endeavour must be made available to the general community, from among whom individuals and organisations will ingest the knowledge and intelligence in the original work and turn it into practical outputs of societal impact.

The vast majority of 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. Only a very small proportion of those who are impacted upon 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. Therefore, there may be a substantial disconnect between the content of a paper, its citation receipts, and its practical impact.

Furthermore, citations are context agnostic, so citations which are posted which are critical of the malign content of the cited article will be counted in the same way as are positive and appreciative references. Paradoxically, a “bad actor” article may achieve high citation counts through notoriety. Moreover, in fields where there may be a number of similar papers, subsequent authors may only draw at random on source material for citation purposes to support a particular point. Indeed, the original papers may be misquoted or misrepresented and inappropriately cited.

The Discoverability and Accessibility of Source Material for Citations

It is self evident that articles will only be cited if they are discoverable and accessible to the citing authors. Moreover, even if influential articles are found, read and cited, the citations will only become effective if they are picked up and processed by one or more of the major bibliometric citation systems.

Citations are made in articles within academic journals or other forms of publication, including Book Series, Textbooks, Conference Proceedings or Patent Filings. These are then processed by the bibliometric systems. This creates a problem for the citation system. As part of the quality assurance process, SCOPUS and the Web of Science (WoS), as the major

global citation systems, are necessarily selective of the content which they include in their academic corpus through their expert curation and quality assurance systems.

Primary Source Documents for Citations

The citation-rich documents which are processed by these systems are known as primary documents. They contribute content directly to the bibliometric calculations and to the rewards that flow from them. Therefore, citations which are made in journals or other publications which are not processed by SCOPUS or WoS will not be counted or attributed to the citable article, because they are either inaccessible or simply unknown to the citation systems. This unseen content may hold valuable works and reference material which does not contribute to the bibliometric ecosystem. This challenge is summarised in Figure 1.

The Observable Universe Model of Citation Activity.

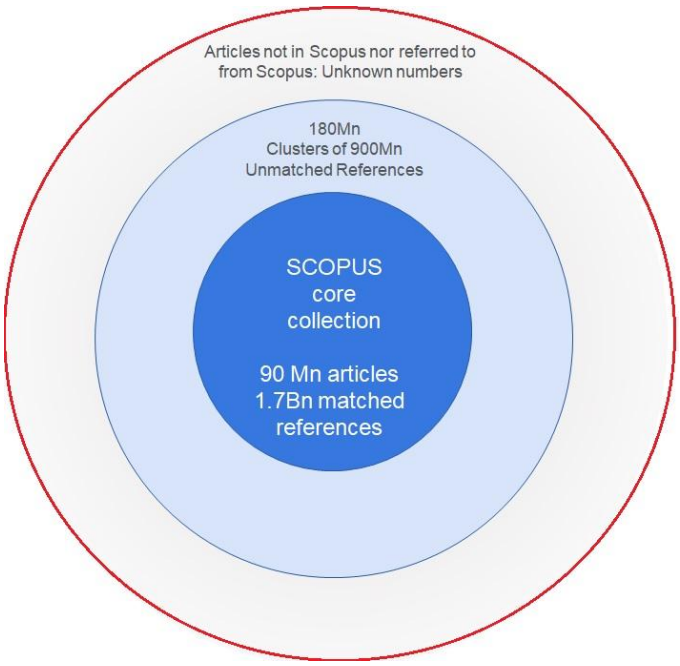

Figure: The universe of citation activity courtesy of Dr R Schrauwen 2023

Dr Rob Schrauwen (Figure 1, personal communication, May 2023) has encapsulated these challenges in an elegantly simple conceptual model which he describes as the Observable Universe of citation data.

This conceptual model is derived from the SCOPUS system but it could be more generally applied to any citation system. The SCOPUS core collection at the centre (the "Corpus") consists of ~100 million quality assured, validated articles from journals, books, conference proceedings and patents, with >1.7 billion validated and matched references.

Outside this core are > ~900 million references which have been organised into 180 million clusters of references which can be matched to sources which are not held within the SCOPUS Core.

Outside these two rings are an "Oort Cloud" of unknown numbers of knowledge sources, producing an unknown number of references which are not captured in any way by the SCOPUS system. However rigorous the search for these sources, they will always be too disparate to permit practical and economic identification, capture and inclusion within the corpus of a major commercial citation system. This 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.

There are many explanations for this exclusion. 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.
- 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.

Further complexities in Citation Data: The significance of Crossref

The Initiative for Open Citations (I4OC) has been 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 I4OC website notes 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, given unrestricted access to bibliographic and citation data in machine-readable form. The present scholarly communication system inadequately exposes the knowledge networks within our literature**.** Citation data are not usually freely available to access, and they are often subject to inconsistent, hard-to-parse licenses, and they are usually not machine-readable"**.**

Crossref was founded in 2000 as a not for profit organisation. It 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.

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 indexed in WoS and Scopus is also available in Crossref. For recent years, 68% of the WoS publications and 77% of the Scopus 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". They commented on the variety of reasons for the incompleteness of the reference data and the DOI allocations. 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.

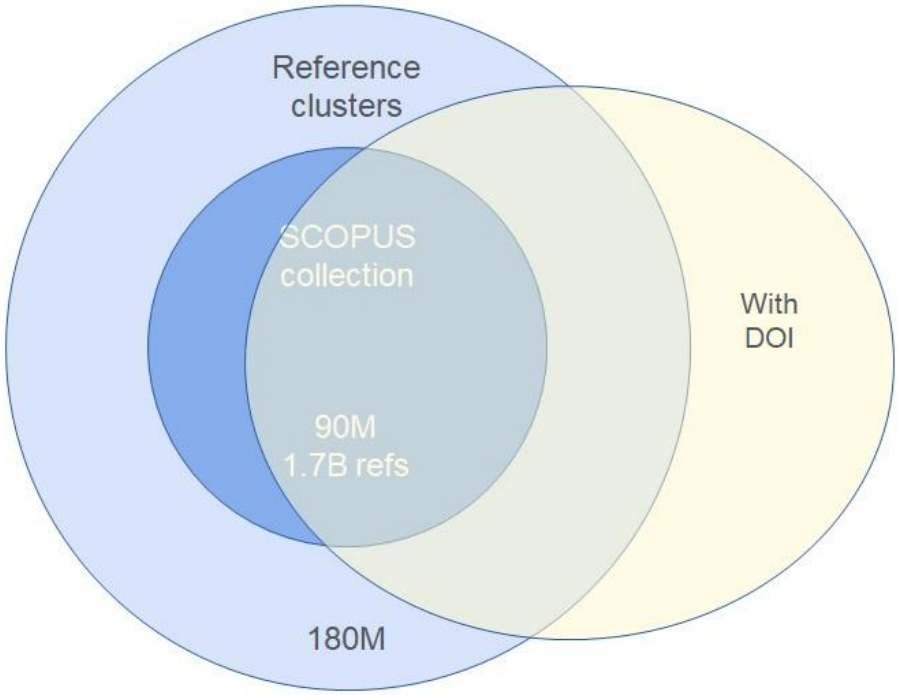

Figure 2. Citation data in the observable universe with the Digital Object Identifier (DOI) overlay of April 2023 (courtesy of Dr Rob Schrauwen)

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. 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.

In the outer reaches of the bibliographic universe lie the well defined collections of the Preprint servers and the global Patent Libraries, whose content has been increasingly incorporated within SCOPUS (Figure 3), along with the Dissertations and Theses collections from Proquest which are now linked to the Web of Science. The Medline collection is independently curated by the US National Institute for Health, but it is fully incorporated within SCOPUS.

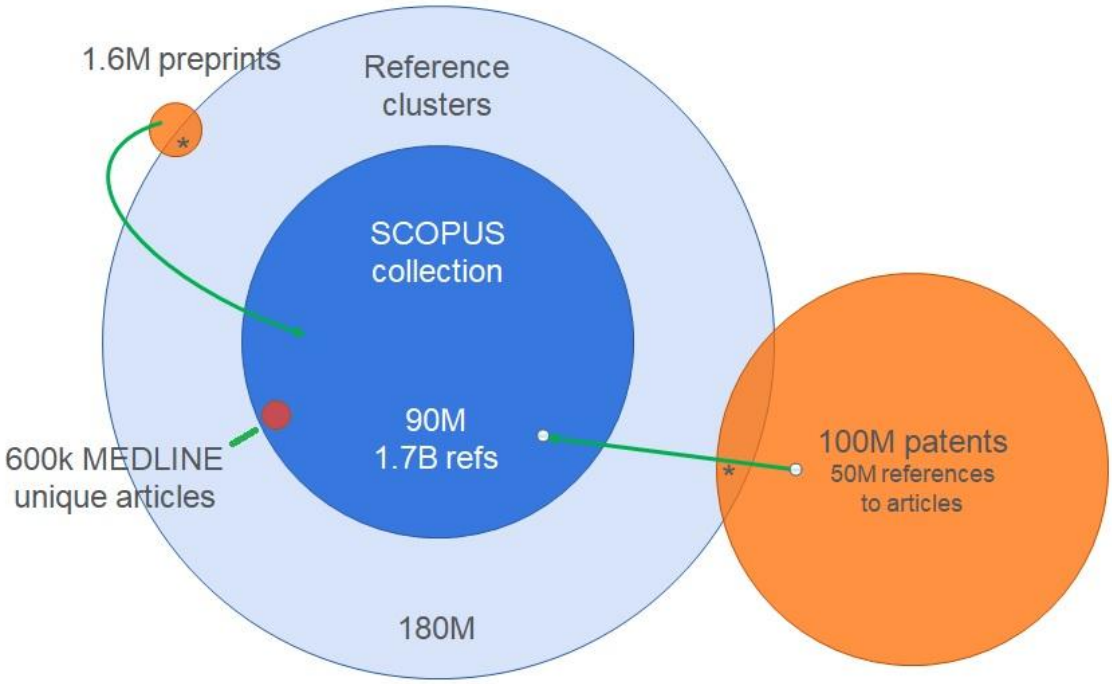

Figure 3. The relationship of Patents, Preprints and Medline-specific articles and references Courtesy of Dr Rob Schrauwen Elsevier

Fraudulent Citation Activity

Citations have acquired critical importance within academic career development, funding decisions and institutional status. It is therefore no surprise that a range of fraudulent models have emerged to take advantage of the complexities of bibliometric science and of computer data systems, so as to boost the citation counts of those who are motivated to do so. Among the many ways of gaming the system to a greater or lesser extent, the following tricks and methods are now well recognised as being particularly malign

Metadata Manipulation

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 article. This criminal 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.

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 cite 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 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)

Citation Planting

This refers to the inappropriate citation of authors on topics which are unrelated to the paper in which they appear.

Irrelevant References are a version of this technique. Fake or hijacked journals are another vehicle for introducing advantageous citations into the bibliometric system.

Editorial and Peer 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.

Citations for Sale

Hazem Ibrahim and colleagues have recently highlighted a further twist on citation malpractice, with the exposure of citation boosting services, which will sell fake citations in bulk (Ibrahim et al 2024). They observe 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 studied 1.6M author profiles from Google Scholar.

Using various measures of citation irregularity, they identified several scientists with suspicious profiles who may have engaged with citation boosting services. They followed the trail to a company which provided citations to one of the suspicious scientists. They generated 20 research articles using a Large Language Model while listing a fictional character as an author, for whom they created a Google Scholar profile and contacted the company to ask them to boost the citations of this profile. They purchased 50 citations, thus proving 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 specifically the safety of Google Scholar in the evaluation of researchers, noting its susceptibility to manipulation.

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).

The Management of Retractions

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.

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

RETRACTED

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], Mayum... Ar...a [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..., Silas F... Eto [l], Dayanne Carla Fernandes [m], Wilson G. Manrique [n], Gabriel Conde [o], ... C. Rosa..., 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... ...d da Universidade de São Paulo, Brazil
[b] Laboratório Integrado de Biociências Translacionais (LIBT), Instituto de Biodiversidade e Sustentab... ...FEM), Universidade Federal do Rio de Janeiro (UFRJ), Macaé, RJ, Brazil
[c] Department of Microbiology, Institute of Biomedical Sciences, University ... Sao Paul...
[d] Department of Cell Biology, Institute of Biomedical Sciences, Universit... ...razil
[e] Reproductive and Molecular Biology Group, Department of Morph... ...Institute ... ...Sciences, S... ...ulo State University, Botucatu, São Paulo, Brazil
[f] Department of Structural and Functional Biology, Institute of Bio... ...of Bot... ...University, SP, Brazil
[g] Department of Preventive Veterinary Medicine, São Paulo S... Uni... ...Brazil
[h] Department of Pharmacology, Center of Research in Infl... ...ry Dis... ...eirão Preto Medical School, University of São Paulo, Brazil
[i] Brasil University, Descalvado, São Paulo, Brazil
[j] Universidade Estadual Paulista Júlio de Mesquita ... ...to de Biociência... ...tamento de Fisiologia, São Paulo, Brazil
[k] Universidade Estadual Paulista Júlio de Mesq... ...ão, ...b, Brazil
[l] Postgraduate Program in Health Sciences, PPGCSA, Federal ... ...ty of Roraima, Brazil
[m] Immunochemistry Laboratory, Butantan ... ...ute, São Paulo, B...
[n] Aquaculture Health Research and ... ...on Group, GRUPESA, Aq... ...ure Health Laboratory, LABSA, Department of Veterinary Medicine, Federal University of Rondônia, Rolim de Moura campus, Rondôn... ...zil
[o] Department of Preventive Vet... ...Medicin... ...Paulo State University, Jaboticabal, Brazil
[p] Department of Cell and Molecul... ...rão Pret... ...ical School, University of São Paulo, Ribeirão Preto, Brazil
[q] Departamento de Bioq... ...Institu... ...ímica, U... ...dade de São Paulo, Brazil
[r] Pontifícia Universid... ...de Mi... ...is, ...
[s] Departamento d... ...cias ... ...leculares, ...de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil
[t] Department ... ...rmacolog... ...d Toxicology, ...ldo Cruz Foundation, FIOCRUZ, Rio de Janeiro, Brazil
[u] Departmen... ...Chemi... ...logy, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Brazil
[v] Laboratory of ... ...Zebrafish Facility, Department of Pharmacology and Toxicology, National Institute for Quality Control in Health, Brasil
[w] Transplantation ... ...ology Lab, Department of Immunology, Institute of Biomedical Sciences, Universidade de Sao Paulo, Brazil
[x] Center for Translati... ...arch in Oncology, Cancer Institute of the State of Sao Paulo, Faculty of Medicine, University of São Paulo, Sao Paulo, Brazil
[y] Laboratory of Peptide B... ...ry, Federal University of São Paulo, Brazil
[z] Laboratory of Human Immu... ...ology, Department Immunology, Institute Biomedical Sciences, University São Paulo, Sao Paulo, Brazil
[aa] Department of Pharmacology, Universidade Federal de Minas Gerais, Brazil
[ab] Graduate Program of Pharmacology, Federal University of Santa Maria, Brazil
[ac] Laboratory of Fish Physiology, Graduate Program of Bioexperimentation and of Environmental Sciences, University of Passo Fundo, Brazil

* Correspondence to: G. Malafaia, Biological Research Laboratory, Goiano Federal Institution – Urutaí Campus, Rodovia Geraldo Silva Nascimento, 2,5 km, Zona Rural, Urutaí, Brazil.
** Corresponding author.
E-mail addresses: guilhermeifgoiano@gmail.com (G. Malafaia), charliesilva4@hotmail.com (I. Charlie-Silva).

http://dx.doi.org/10.1016/j.scitotenv.2021.152345
0048-9697/

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

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 following example of a retracted article 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 (Figure 4).

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, Honorary Research Professor of National Sun Yat Sen University, Taiwan, 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

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

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., 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

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

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

Ibrahim H, Liu, F, Zaki Y, Rahwan Tl, Google Scholar is manipulatable 7th Feb 2024 https://arxiv.org/abs/2402.04607

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

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.

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

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
https://www.sciencedirect.com/science/article/pii/S0048969721074222?casa_token=2e7xLlT0mc0AAAAA:l_s09NJ5nXcJdQoJKGH7YOn0oPdM5PKooNu0sCzH-Bbx14qssrIzyP2dCCdFEyUvZmOfT6k