

Detecting incoherent citation data among three bibliometric platforms: OpenAlex, Scopus and Web of Science

Journal of Information Science

1–12

© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/01655515251330579

journals.sagepub.com/home/jis**David Rodrigues**

Iscte-Instituto Universitário de Lisboa, Portugal

António Lopes

Iscte-Instituto Universitário de Lisboa, Portugal

Instituto de Telecomunicações, Iscte-Instituto Universitário de Lisboa, Portugal

Fernando Batista

Iscte-Instituto Universitário de Lisboa, Portugal

INESC-ID Lisboa, Portugal

Abstract

The number of citations received by a research paper is a vital metric for both researchers and institutions. Various indexing databases share common citations, facilitating cross-database comparison to identify citations missing from multiple databases, which are not contributing to the total citation count of a paper. To tackle this problem, we developed an automated method to detect missing citations by utilising multiple indexing databases. We focus on identifying missing citations in Web of Science and Scopus, using OpenAlex to enhance the process and demonstrate the benefits of cross-referencing databases for more comprehensive citation tracking. We compared the results of a previous experiment where we did not use Scopus. This way, we could measure the impact of adding Scopus to our approach. By using the same data set as before, the papers we were able to analyse increased from 1539 to 1989, and we were able to find a total of 1075 missing citations in Web of Science as opposed to the 696 we found without including Scopus. As for the results in Scopus, we identified 1137 missing citations in the same set of papers, totalling 2212 missing citations found. This outcome proves that these indexing databases cannot accurately detect all citations. Also, for more recent publications, we detected a bigger discrepancy in missing citations found in Web of Science when compared with Scopus. We can also conclude that adding different databases can provide us with better results and a more accurate view of the citation list of a paper.

Keywords

Citation databases; citations; OpenAlex; research databases; Scopus; Web of Science

1. Introduction

The significance of citation counts in research papers cannot be overstated. In academia, the number of citations a paper receives serves as a vital indicator of its value, often closely correlated with its potential to provide valuable insights. Moreover, the authors of highly cited papers garner prestige and recognition within their respective fields. This recognition extends beyond mere peer approval; it plays a pivotal role in the performance evaluation processes employed by research institutions, where citation metrics are a key factor. Consequently, accurately assessing the citation count of any given paper is of paramount importance.

Corresponding author:

Fernando Batista, Iscte-Instituto Universitário de Lisboa, Avenida das Forças Armadas, Lisboa 1649-026, Portugal

Email: fernando.batista@iscte-iul.pt

However, the task of tracking citations is not without challenges, despite the existence of extensive citation databases. These databases, while invaluable, present their own limitations. Each citation database maintains its own unique source list, which naturally leads to differences in the articles and citations they report. This means that some papers indexed in one database may not appear in another, resulting in variations in citation counts across platforms. However, in some cases, discrepancies arise not merely due to differences in indexed sources, but because a citing paper is indexed without properly tagging or recognising the cited paper within the same database. Consequently, relying on a single database may yield incomplete citation data due to both coverage limitations and occasional citation misidentifications within the database. Hence, there is a clear need for a method to automatically identify missing citations across different databases, offering a more accurate representation of a paper's true citation count. Conceptually, 'citation missing' refers to discrepancies where a citing article is present in a citation database but the citation is absent, despite the citing and cited articles being indexed in the same citation database. Operationally, it involves instances where an article appears in multiple databases but fails to register as a citation in at least one of them due to issues like metadata mismatches or incomplete referencing. Our research aims to address this issue by streamlining the identification process, eliminating the time-consuming and tedious manual comparison of discrepancies.

In this paper, we expand on our previous work [1] by adding the Scopus database to our approach, which will enable us to analyse the improvement of adding a new database to our approach, since now we focus on identifying potentially missing citations within both Web of Science (WoS) and Scopus by cross-referencing them with each other and the OpenAlex platform. We selected WoS and Scopus as they are the primary databases employed by our institutions, and improving citation completeness in both is advantageous. OpenAlex's inclusion in our approach offers two key benefits: it is a free tool accessible via an application programming interface (API), simplifying data retrieval and it aggregates data from various sources, potentially yielding a more comprehensive list of citations than other databases.

It is worth noting that we explored the possibility of incorporating Google Scholar into our project; however, it presented challenges due to the complexity of extracting information and the inherent ambiguity in the data it provides, as it lacks DOI information, which is crucial for minimising errors.

The main issue that this article seeks to answer is how the integration of different citation databases can improve the identification of missing citations, thereby providing a more accurate and comprehensive representation of a paper's citation count. Specifically, we aim to address the following research questions:

RQ1. How does the inclusion of multiple citation databases, specifically Scopus, improve the detection of missing citations compared with relying solely on WoS and OpenAlex?

RQ2. What is the impact of missing citations on the overall citation count when considering WoS, Scopus and OpenAlex?

In the subsequent sections, we provide a comprehensive literature review, followed by a detailed step-by-step description of our approach to automatically identify missing citations. Section 4 starts by examining the influence of integrating a new citation database into the process of identifying missing citations. To be specific, we introduced Scopus into a prior study [1], showcasing the enhancements it brought to the outcomes collected from WoS. Subsequently, we delve into the outcomes of this experiment for both databases, conducting a year-by-year analysis in order to produce meaningful comparisons. Finally, section 5 presents the main conclusions and outlines possible future directions.

2. Literature review

Researchers may look for citation counts in articles, either because they are looking for the best articles on specific topics or to evaluate their performance for grants or promotions. One metric that is used to evaluate a researcher's work is the Hirsch's *H*-index, which needs an accurate and error-free list of citations, and the completeness of these records is of the utmost importance for its calculation [2].

The exact algorithms used by Scopus and WoS to identify citations are proprietary and not publicly disclosed. However, several known factors can lead to errors in citation identification. These include incomplete or incorrect metadata (e.g. misspelled author names, incorrect DOIs) and variations in citation formats. In addition, algorithmic challenges, such as disambiguating common author names and handling journal title changes, contribute to mismatches. Time lags in indexing newly published work and backlogs in historical literature can also result in missed citations. While we acknowledge these common limitations of citation databases in this section, our research does not delve into the reasons behind these missed citations. Instead, our primary aim is to detect incoherent citation data among the platforms, rather than to analyse the underlying causes of citation mismatches.

Therefore, looking only at one citation database can be misleading, since each citation database shows strengths in covering different areas [3], which can make a database miss some of the articles that could potentially cite the paper which a researcher is looking at, because there are unique citations for each database that the other ones do not find [2]. For example, in a study from 2006 to 2017 in the journal of *enfermeria nefrologica*, only 50.2% of the papers were indexed by Scopus. It should be the responsibility of both Scopus – to improve their quality control system – as well as the journal to periodically check if the papers are being indexed by the databases [4]. Like Scopus, the other databases also show a lack of coverage in some areas, but all of them are working on enhancing their coverage, and getting as many articles as possible in their databases, showing some improvement over the years [3]. Nevertheless, neither database is perfect, and for better results, more than one should be used.

The study conducted by Teplitzkiy et al. [5] shows that the citation count of an article is significantly influenced by a feedback loop: as articles accumulate more citations, they are more probably to be read, which, in turn, increases the article's impact and fosters additional citations. Also, the readers will be more open to extract meaningful information from them. In addition, it is shown that the sooner a researcher finds an article, the more probably it is to pay closer attention to it, instead of looking for new ones. We can then presume that it is very important that the citation count is not omitted, so that the paper is taken seriously and not pushed aside.

The main reason why citation indexing databases might not correctly find a citation is due to errors. Buchanan [6] reports a variety of errors, either by the authors or the databases, in both DOI's and references. Cioffi et al. [7] try to find and automatically correct wrong DOI's being given to the databases by the authors or being provided by the databases themselves. Although no concrete numbers of the total errors were provided, we can see there are a lot of errors found by the study. Ovid Technologies publisher alone had over 370,000 outgoing citation errors in 2 years. Zhu et al. [8] investigated errors in DOI extraction within WoS and discovered instances where WoS mistakenly replaced the number '0' with the letter 'O', as well as other similar characters such as 'b' and '6' or 'O' and 'Q'. Besides wrongfully written DOI's, Franceschini et al. [9] report that databases sometimes, mistakenly, give the same DOI to different articles, and since the DOI should be unique for each article, these errors can make a difference in bibliometric analysis.

Besides DOI errors, there are also reference errors like authors' names, the conference/journal where it was published or the article title, which are also a problem that can lead to databases missing citations. Through several scientific studies, reports of errors in references can range from 25% to 54% [10]. Some instances of these errors are, for example, several occasions where the date, title and author's name of both Karl Weick's and Walter Benjamin's books were not properly cited [11]. Because the databases react differently to these kinds of mistakes, they also provide different results, and although they find different results, Google Scholar can have an advantage [4], since they are able to better find different forms of citations, leaving them less probably to have missing citations, which in turn provides a more complete list of citations. A paper [12] suggests Google Scholar (GS) should be the main focus while doing bibliometric analysis, since it does not discriminate where the paper comes from and gives an equal chance to every source. On the contrary, a great number of articles raise questions about the validity of GS citations. They [12] defend that GS should not be trusted completely, since they count citations that have not been published by reputable journals and conferences with a publishing format of peer reviews, which makes it easier to inflate the number of citations a paper has. In his 2010 research, Labbé [13] demonstrated an exploit of GS. He showed a method for creating fake scientific documents and having them indexed in GS to artificially inflate the citation counts of other papers. Finally, a more cautious and intermediate approach is also suggested by another paper [2], saying that if GS results are authenticated, they can be a useful tool since a great deal of their citations still come from 'an unquestionably valid but unreachable scholarly sources'.

A follow-up study [14] tries to find what were the differences of the reference lists in Scopus and WoS. While comparing the reference list of around 100,000 papers in both databases, using as a baseline the Elsevier ScienceDirect Article Retrieval API to get the references, they found that WoS had 77.2% of the papers with the same number of references, while 19.3% had fewer references. On the contrary, 96.4% of Scopus papers had the same amount of references. They manually analysed random papers with different results and saw a variety of different reasons why this happened. Since the different databases themselves can extract different references list of a paper, this can also explain why there can be missing citations in some databases, because if they do not extract a reference, or do it wrong, then the paper that it is being cited, won't be found in the database as a cited paper and the citation is lost.

This paper [15] identifies the missing citations of several articles in an automated way. Our work follows a similar methodology to theirs, the main difference being the way we identify if a possible missing citation paper is present in the database where it is missing from. While they check if the paper's publisher is present in the list of indexed papers of the database in question, we aim to look for the paper itself in the database, making sure it really is indexed by it. Because, as stated before, even if the publisher is on the source list, it does not mean that the paper will be indexed in the database. The same researcher team, on a follow-up study [16], analysed the changes over time of the amount of missing citations

for a given set of papers, in order to check if the databases in question corrected the citations, they calculated that in a period of around 14 months, around 13% of missing citations were corrected in both WoS and Scopus.

Visser et al. [17] conducted a study analysing the coverage and citation links among Scopus and four other bibliographic sources: WoS, Dimensions, Crossref and Microsoft Academic. Their research revealed issues of incompleteness and inaccuracy in missing citations across all these sources. Specifically, for WoS, they identified three factors contributing to these shortcomings: missing references from document reference lists, errors in reference extraction and incorrect reference-to-paper matches. Regarding Scopus, Visser et al. found that while the platform contains accurate reference information, it fails to appropriately match these references with the cited papers.

An advantageous tool we found is OpenAlex, a ‘fully-open index of scholarly works, authors, venues, institutions and concepts’ [18]. Using the DOI of a paper, it is possible to get all the information that they gathered about that paper. OpenAlex gets all their data from multiple services, with MAG and Crossref being the most complete ones, but also from ORCID, ROR, DOAJ, Unpaywall, Pubmed, Pubmed Central, The ISSN International Centre and Subject-area and institutional repositories from a plethora of platforms [18]. Using this API, it is possible to get easy access to the information from all these other platforms from a single query.

Due to OpenAlex’s reliance on multiple sources for information aggregation, it is susceptible to receiving inaccurate data. As highlighted by Besançon et al. [19], they uncovered a method to introduce citations that should not exist through metadata manipulation in Crossref, which could later be incorporated into OpenAlex. However, we believe such exploits should be promptly addressed upon discovery and they are probably to affect only a limited number of papers indexed in this bibliometric platform. Therefore, we should not dismiss OpenAlex as a reliable source of information but rather remain vigilant of potential errors it may contain.

In their study, Culbert et al. [20] examined the reliability of OpenAlex by comparing its reference data and certain metadata against those of WoS and Scopus. While analysing a data set comprising over 16 million papers shared across the three databases they came to the conclusion that regarding the number of references, there is comparability among the databases. However, upon scrutinising metadata, they observed that OpenAlex exhibited a higher capture rate of ORCID identifiers, similar levels of Open Access information, but fewer abstracts compared with WoS and Scopus. It is important to highlight that these observations are based on a data set of papers common to all three databases. When considering the entirety of OpenAlex’s database, which includes a larger number of papers, the ratio of references per paper decreases.

3. Proposed method

Our main goal is to automatically find missing citations between different citation databases. In this work, we search for missing citations in both WoS and Scopus, while also using OpenAlex as an extra source of information. We search for citations that are not common to all databases, and with these potentially missing citations, we have to search the database that did not find the specific citation, in order to check if the citing paper is indeed indexed in it. If it is, then it is considered to be a missing citation.

An example of the information we want to find out is shown in Table 1. If paper A is analysed and WoS says that paper A is being cited by the articles 1, 4 and 6, Scopus says that paper A is being cited by the articles 1, 2, 3 and 4, and OpenAlex says paper A is being cited by the articles 1 and 5, there are three citing articles that Scopus and OpenAlex

Table 1. Example of a report for paper A.

Paper A Analysis			
Citing Articles	WoS	Scopus	OpenAlex
1	✓	✓	✓
2	Missing	✓	
3	×	✓	
4	✓	✓	
5	×	×	✓
6	✓	Missing	
Recorded Citations	3	4	2
Expected Citations	4	5	
Missing Citations	1	1	
Missing Percentage	25%	20%	

found and WoS did not, which are 2, 3 and 5, and there are two citing articles that Scopus did not find, which are 5 and 6. The next step is to look for the respective articles in WoS and Scopus, in order to confirm if they are indexed in the corresponding database. And, for example, if article 2 was found to be indexed by WoS but articles 3 and 5 are not, we can assert that article 2 is a missing citation of paper A on WoS. While if in Scopus we discover that the article 6 is indexed, but article 5 is not, then the article 6 is a missing citation in Scopus. Please note that the empty cells in the OpenAlex column represent the fact that we did not search these articles in OpenAlex (since they were already being indexed by WoS and Scopus), hence we cannot confirm or deny that these articles are indexed in the OpenAlex database.

In order to look for missing citations in an article, all we need is its DOI. The first step in the verification process is to clean the received data. Sometimes, the DOI can have extra information, such as the *https* link, the *doi.org* domain, or other invalid characters or blank spaces, so we standardise the DOI, so it is easier to work with it through the remaining steps.

Afterwards, we ask the institution's Current Research Information System (CRIS) API¹ for more information about this paper. From there, we can get the article's title, authors, year, the type of the article (journal, book, conference, etc.), as well as some information about the WoS and Scopus representation of this paper, such as if the article is indexed in any of them, the link for the list of citing publications of our article and the unique identifier (accession number) for this article (WoS ID = WOS: xxxxxxxxxx; Scopus ID = 2-s2.0-xxxxxxx). The CRIS employs periodic and automatic verification processes to maintain the accuracy of the data it contains. Specifically, each paper within the CRIS undergoes regular cheques at least once a week to ensure that it remains up-to-date.

The next step is to query OpenAlex API for their information about this article. Once again, the query is done via the DOI, and all the information gathered about the article from the CRIS is compared with the one gathered from OpenAlex. With OpenAlex we can also get their list of citing publications, which comes with all the information they have about each of these citing publications, namely the DOI, title, publication year and the type.

After retrieving the list of citing publications from OpenAlex, we need the corresponding WoS and Scopus list of citing publications to compare them. Since these databases only provide the link for the page where the information is, we had to extract the information of each citing paper ourselves. We use the link provided by the CRIS in order to retrieve this information which gives us a direct route to the information we are looking for. Nevertheless, it would also be possible to get to the web page with the list of citing publications in other ways. For the WoS list, through the WoS unique identifier, we could build an URL that takes us directly to the WoS page about the paper, and from there we are only one click away from the list of citing publications. For Scopus, the URL is not as simple to build, and the best way to get to the publication would be to try to find the publication via the unique identifier or the DOI and other information we got, but that method is more prone to errors; therefore, if possible, using the link to the page is the better option.

With the list of citing publications from the three databases, we can look for possible missing citations in WoS and Scopus. For this effect, we first merge the three citations lists, in order to get exactly what papers are citing with no repetitions, and in what database they are tagged as citing. In order to create this single list of citing publications we start by standardising the data we have from each database, because the keywords of all these databases are different, since, for example, a publication title can be described as 'Title', 'Article Title' and 'title'. For this reason, we create a dictionary that represents an article citation, which has the following data: 'DOI', 'Publication Year', 'Publication Title', 'Publication Type', 'Database Source', 'Database Unique ID' and a Boolean 'Citing'. Once we have all the data in this standard format, we proceed to compare each article to see if different databases have the same article and create the single list of papers and whether or not they are tagged as citing in each of the databases. This matching of articles is done by first verifying if the DOI is the same, and if it is not, then we compare the article title, year and type. If a paper from this list is tagged as citing in both WoS and Scopus, then no further work is needed, on the contrary, if one of these databases does not have this citing paper, then this publication is flagged as a possible missing citation in the database where it is missing.

Having now all of these publications that are flagged as potentially missing citing publications in their respective databases, we need to check if they are in fact indexed or not in the database. And only if they are, then they are considered missing citations. So once again, we must consult the database in question to extract this information. This step is harder than the previous one, since before, we already had a link to the page of paper we were looking for, while now, we don't have any link of this paper in the database we are looking for. In order to find the paper, we execute two different queries. First, we query using only the DOI of the publication '(WoS query: DO=(doi), Scopus query: DOI(doi))'. If this query does not provide results, we also try to find the publication via the title and publication year '(WoS query: TI=(title) AND PY=(year), Scopus query: TITLE("title") AND PUBYEAR IS year)'. If any of the queries provides results, we extract the results from the database in the same way we extract the list of citing publications, and finally, we can double-check either if the DOI is the same, or if the title, year and publication type are the same. The reason we don't use only the DOI to perform this check is because some publications do not have DOIs (or the database might be

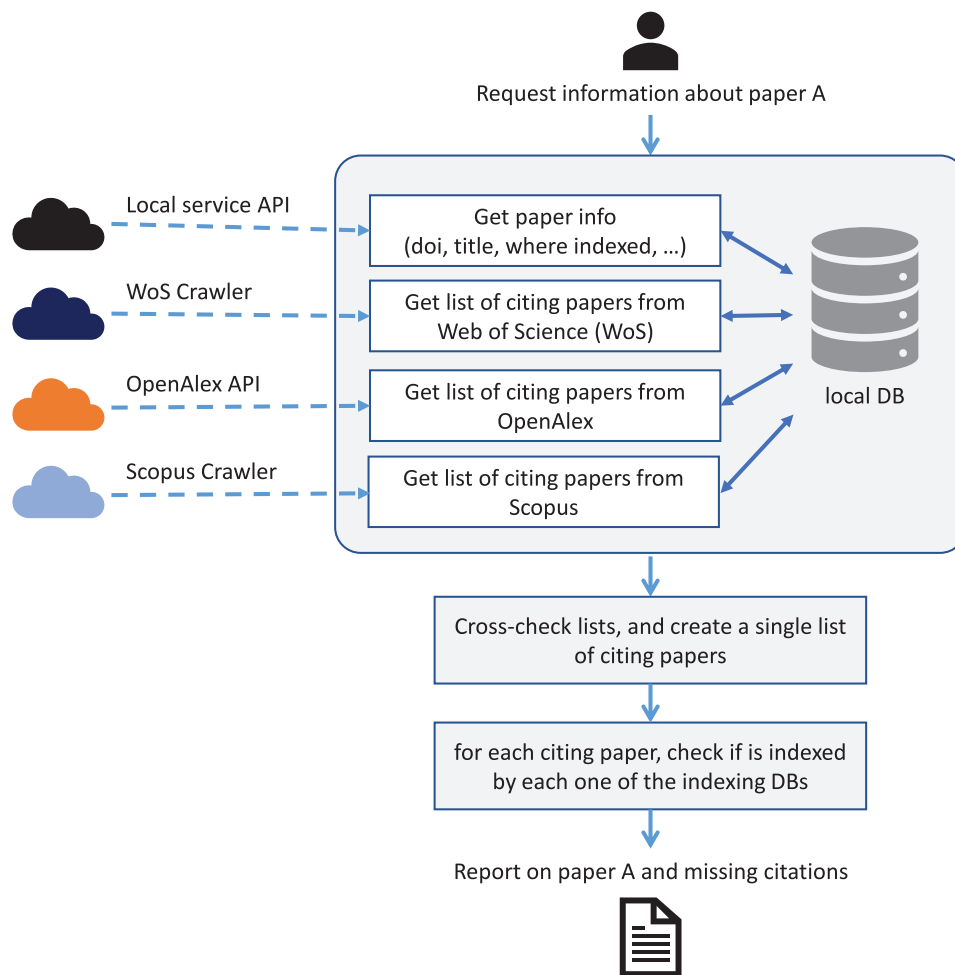


Figure 1. System's design.

missing that information) but also because sometimes there are errors in the assignment of DOIs in databases, as seen earlier in the literature review. Hence, we might not find the correct publication via DOI and do so through the title.

If the queries' results are a match, then we are in the presence of a missing citation, and we tag it as such in the report that it is produced. With every missing citation, we also provide the link of the publication we found that is supposed to be citing our article, so that later we can manually double check any missing citations we might want to validate. These queries are very strict, and specially in the case of the title, if there is an error on a word, or any difference, they will not match, in the future we wish to improve these queries, using similarity techniques that will make it possible to find a match even in the presence of small errors. We would also like to try to expand these queries and use other information we might have on the paper to try to find it in the new database, although this must be carefully tested in order to limit the number of false positives that our application finds.

Through all this process, one of our main goals is not to overload the APIs or repeat processes unnecessarily. With that in mind, we use delays between every request to avoid overloading the systems. Moreover, we have a local database which is used as a cache, so if we are asked to analyse the same publication in a short period of time (the time period of when the data becomes outdated varies from each type of data, and ranges from 1 day to 1 month), we don't have to gather all the information from external sources to provide an answer. Therefore, for the entire process, every time we need any data, we first query our local database, and only if we don't have the data stored locally we will do the process described above in order to get the necessary data.

Finally, it is important to state that in order to create a report for any publication, the publication has to be indexed in at least two of the databases, because, if we cannot compare the citations list of two different databases, no comparison data can be gathered, and no report is created. Also, if the publication has no citations in any database, there isn't anything to look for, and no report is created. A diagram with the representation of our system's design can be seen in Figure 1.

4. Experiments and evaluation

One thing we must always keep in mind is that each database has a distinct source list, resulting in variations in the publications they index. Therefore, each database can also have a different citation list for the same research paper. Nevertheless, there are citations in common that sometimes a database does not find. It is also known that Scopus normally has a higher citation than WoS for most papers, mainly because it has a more extensive source list which leads to more papers indexed in it. The same happens for the papers analysed in this study. As for OpenAlex, the citation counts are very similar to the ones gathered from Scopus as in sheer numbers, but some of the citing publications differ. Nevertheless, just because a database has more citations in it, does not mean that a database with fewer citations cannot provide a missing citation for the one with more citations, justifying why we also use WoS citations to possibly find missing citations in Scopus.

To assess the effectiveness of our proposed approach, we conducted an experiment utilising the papers available within our institution's CRIS as our data set. Out of the 52,000 papers encompassed by the data set, we were only able to incorporate 12,500 of them due to the absence of designated DOI identifiers in the remaining papers. Out of these 12,500 papers, the majority (8330) have been published in scientific journals, while only 1782 were in conferences and 1095 are book chapters.

For our experiment, we adopted a year-based approach, wherein we conducted an analysis of papers published within specific years. We specifically selected the years 2015, 2018, and 2021 for several reasons. First, we sought years that aligned with our research motivation, considering that our university assesses researchers based on publications within the last decade, making it inappropriate to examine papers older than 10 years. In addition, we aimed to examine multiple years not only to gather more data but also to facilitate a comparative analysis to identify any trends or improvements over time. We began with 2018 as our starting point, as it predates the atypical year marked by the COVID-19 pandemic in 2019, making it a more reliable baseline. We chose 2021 as a recent year, allowing sufficient time for publications to influence the scientific community, considering the time it takes for others to read, respond and publish their own work. Finally, we included 2015 to maintain a 3-year interval between 2018 and 2021. A summary of the findings from these 3 years is presented in Table 2.

Table 2. Results per year.

	Year		
	2015	2018	2021
Papers analysed	820	950	1304
Papers that met the criteria	553	678	758
Total OpenAlex Citations	11,185	12,301	5308
Total WoS Citations	8942	9835	4118
Total Scopus Citations	11,189	12,272	5623
Papers with missing citations	231	284	332
WoS Missing Citations	278	390	407
Scopus Missing Citations	404	349	384
Total missing citations	682	739	791
Average citations in OpenAlex	20.23	18.14	7.00
Average citations in WoS	16.17	14.51	5.43
Average citations in Scopus	20.23	18.10	7.42

4.1. Results analysis

When we examine the total number of citations across all 3 years, Scopus has the highest count, totalling 29,084 citations, followed by OpenAlex with 28,794 and WoS with 22,895. While Scopus and OpenAlex have a similar total number of citations, it is noteworthy that when we break down the data by individual years, the number of citations in 2015 and 2018 is quite comparable, but in 2021, there's a more significant disparity. This suggests that Scopus might be better at indexing the most recent citations more rapidly when compared with OpenAlex.

Regarding the evolution of citations over the years, it is natural to observe a decrease in the average number of citations per paper in more recent publications. This decline occurs because there has been less time for these more recent works to be read and cited in subsequent papers.

Regarding missing citations, the number of papers with missing citations has increased over the years. However, when we consider the percentage of papers from each year with missing citations, the increase is relatively modest. In 2015, 41.77% of papers had missing citations, in 2018, it was 41.89%, and in 2021, the year with the highest proportion of papers with missing citations, it reached 43.80%.

When we break down the results for each individual database, Scopus had the highest count of missing citations, totaling 1137, whereas in WoS, we found 1075 missing citations. Notably, the trends in these two databases over the years differ. In WoS, the more recent years had higher numbers of missing citations, increasing from 278 in 2015 to 390 in 2018 (+112) and finally reaching 407 missing citations in 2021 (+17 compared with 2018). Conversely, Scopus showed a decrease from 404 missing citations in 2015 to 349 in 2018 (−55), followed by an increase to 384 in 2021 (+35 compared with 2018 but −20 compared with 2015).

To gain a deeper perspective, we calculated the average percentage of missing citations, first across all papers, and then specifically in papers where missing citations were identified. This calculation used the formula

$$\text{Average Percentage of Missing Citations} = \text{Missing Citations} / (\text{Missing Citations} + \text{Citations})$$

This formula will tell us out of all the citations we think a paper should have in that database (citations the database found plus the citations we found that are missing from the database citation list), what percentage is missing. For example, if a paper has two citations according to the database, and we find another citation that is missing, it has 33% of missing citations ($1/(1+2)$).

Analysing these percentages shown in Table 3, WoS experienced a significant increase in missing citations in 2021 compared with previous years, with 13.18% of citations missing, whereas in 2015 and 2018, the percentages were only 5.29% and 6.26%, respectively. When focusing solely on papers with missing citations, in 2021, an average of 44.4% of their citations were missing, compared with 20.08% in 2015 and 24.23% in 2018. This highlights that, in recent years, finding these missing citations has a more substantial impact on WoS, as they constitute a larger portion of the citations from these papers.

Table 3. Missing citations averages over the years.

	Web of Science			Scopus		
	2015	2018	2021	2015	2018	2021
Total Citations	8942	9835	4118	11,189	12,272	5623
Total Missing Citations	278	390	407	404	349	384
Average percentage MC	5.29%	6.26%	13.18%	5.43%	4.93%	6.52%
Average percentage MC in papers with MC	20.08%	24.23%	44.40%	18.55%	18.28%	24.47%

In contrast, Scopus exhibits a different pattern, with a reduction in average missing citations from 2015 to 2018, followed by an increase in 2021. Scopus does not show a significant spike in 2021. When considering the average of all the reports created, in 2021, Scopus had 6.52% of their citations missing, which is less than half of WoS's rate for the same year. As previously noted, it appears that Scopus has a better ability than OpenAlex to index the more recent publications, and the same trend holds when comparing it to WoS. Since the missing citations in Scopus come from both WoS and OpenAlex, it's expected that they do not provide as many new citations to Scopus, as Scopus provides to WoS.

Over the three analysed years, we found a total of 2212 missing citations in both databases combined in the 1989 reports that were created. This averages 1.11 missing citations for each report.

4.2. WoS improvements with Scopus citation list

In a previous study [1], we conducted a similar experiment, focusing on the search for missing citations within WoS while utilising OpenAlex as our reference point. In this iteration of our research, we expand our scope by not only considering data from OpenAlex but also incorporating the data from Scopus. The inclusion of Scopus has introduced a variety of new citations that were previously undiscovered and not catalogued in OpenAlex. Consequently, this new introduced data set from Scopus has the potential to reveal additional missing citations within WoS. In this section, our objective is to assess the extent to which our current study has uncovered more missing citations within WoS when compared with

Table 4. WoS results before and after the introduction of Scopus.

	WoS results in the experiment without Scopus			WoS results after the integration of Scopus		
	2015	2018	2021	2015	2018	2021
Papers that met the criteria	446	497	596	553 (†24%)	678 (†36%)	758 (†27%)
Papers where Missing Citations were found on WoS	128	140	130	146 (†14%)	178 (†27%)	225 (†73%)
Missing Citations	233	265	198	278 (†19%)	390 (†47%)	407 (†106%)
Average Missing Citation per paper With Missing Citations	1.82	1.89	1.52	1.90 (†5%)	2.19 (†16%)	1.81 (†19%)
Average percentage MC	5.30%	5.10%	9.60%	5.29%	6.26%	13.18%
Average percentage MC in papers with MC	18.50%	18.30%	44.17%	20.08%	24.23%	44.40%
Total Missing Citations	696			1075 (†54%)		

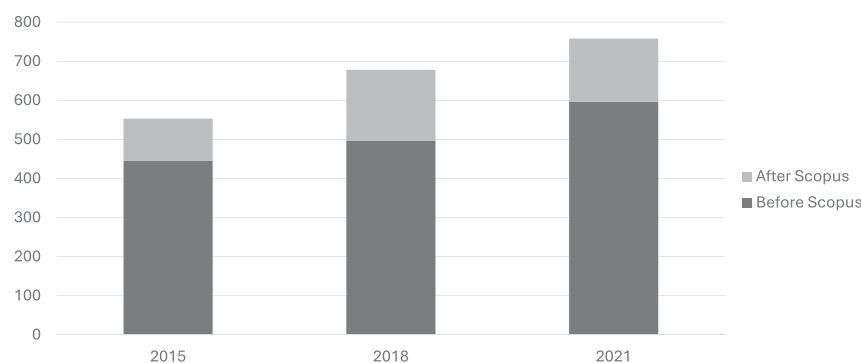
our prior research and the improvement of those findings can be found in Table 4 which data we will now analyse. This analysis will shed light on the impact of incorporating an additional database into our citation search process and the resulting implications for our findings.

Before we proceed to the comparison of data between both studies, it is crucial to acknowledge that there exists a temporal gap of at least 4 months between the data collected from these sources. This difference could mean that the variations in missing citations within WoS cannot be solely attributed to the inclusion of Scopus. Rather, it's plausible that changes have occurred within WoS itself and within OpenAlex over this time frame. These changes might come from the writing of new articles that reference the particular article under analysis, from the introduction of fresh indexations in both databases or from corrections in the old data.

It is also worth pointing out that some of the publications under analysis may not be indexed in either WoS or Scopus. On one hand, if a publication is not indexed in WoS, it becomes impossible to identify missing citations on WoS. On the other hand, if a publication is not indexed in Scopus, we lack the additional data from the Scopus citation list, which means we do not have any extra data on this publication when comparing with the last time we analysed it.

Although we analysed the same number of publications in both studies, this time we generated more reports as it is also shown in Figure 2, increasing from 1539 to 1989, a 29% uptick. This increase in reports is due to the inclusion of Scopus to our approach. Previously, if a publication wasn't indexed in both WoS and OpenAlex, we did not create a report. However, in this new approach, a publication only needs to be indexed in two out of the three databases we consult (WoS, OpenAlex and Scopus) for us to analyse it. For instance, if a publication is indexed in WoS but not in OpenAlex, we can still analyse it as long as it is indexed in Scopus.

Out of these reports, only some of them had missing citations in WoS. While before we had 398 reports with missing citations, we now have 549 reports with missing citations, indicating a 38% increase in the number of papers with missing citations. In Figure 3, we show the number of reports with missing citations for each of the analysed years, before and after including Scopus data, where we can observe a significant increase in reports created for the later year of 2021.

**Figure 2.** Improvement on WoS on the amount of papers that met the criteria.

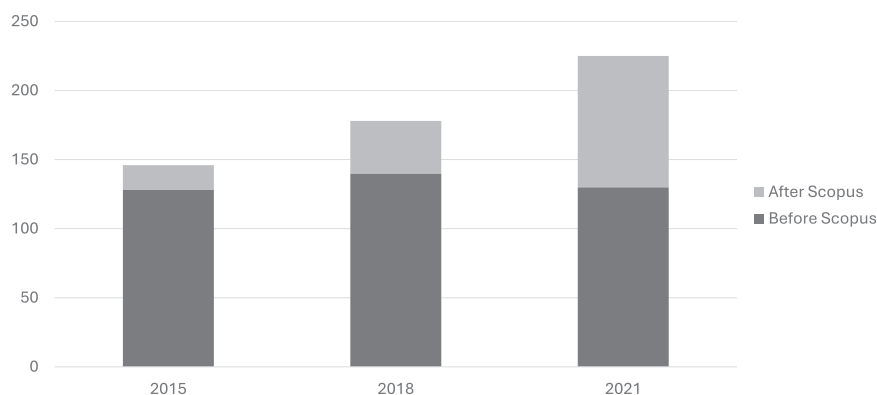


Figure 3. Improvement on WoS on the amount of papers with missing citations.

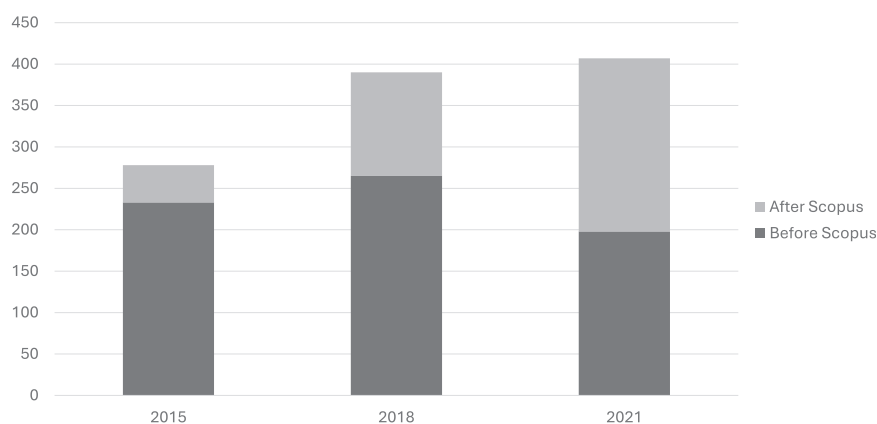


Figure 4. Improvement on WoS on the amount of missing citations.

While some of these papers with missing citations may result from new data provided by OpenAlex that was not indexed 4 months ago, it is more probably that most of them originate from the data provided by Scopus alone.

Comparing the amount of missing citations in both studies, we can see that after integrating Scopus into the process, we were able to find more missing citations. In total, there were 1075 missing citations, whereas without Scopus, there were only 696 missing citations, a 54% increase. In Figure 4, we can also see the breakdown of missing citations for each analysed year, where 2021 again saw the biggest increase – with over double the amount of missing citations when comparing before and after including Scopus. Taking into account all the papers analysed, we found 0.53 missing citations per paper that met the criteria with Scopus, compared with 0.45 missing citations per paper without Scopus, representing an 18% improvement in missing citations per paper.

Looking at the individual years (2015, 2018 and 2021) from these experiments, we can also conclude that Scopus brings far more information, in the most recent years. This suggests that Scopus is better equipped to handle newer data more efficiently than both WoS and OpenAlex, since the amount of papers with missing citations in 2021 increased by 73%, while in 2015 and 2018, only increased 14% and 27%, respectively. Also, the amount of missing citations in 2015 only increased by 19%, while in 2018, it increased by 47% and in 2021, we were able to more than double the amount of missing citations found.

These improvements highlight the value of Scopus and the potential benefits of adding other databases to this process, provided they can uncover citations that others may miss. Whether we use new databases to search for missing citations (like Scopus) or to find missing citations in existing databases (like OpenAlex), the more data we can gather and merge from different sources, the more accurate our results will be. This process not only helps us discover missing citations in each database but also provides a comprehensive view of the total number of citations for a paper, as it aggregates data from multiple providers.

5. Conclusions and future work

This study has provided valuable insights into the dynamics of citation data across different databases, namely WoS, Scopus and OpenAlex. It is essential to recognise that each database maintains its unique source list, leading to variations in the publications they index and the citations they record for the same research papers, which are not considered missing citations, because the citing article must be indexed in the database for the citation to be missing.

The study evaluated the proposed approach by examining a data set of papers from the institution's CRIS, utilising a year-based analysis for the years 2015, 2018 and 2021. This selection of years aligned with the research context, considering the university's evaluation of researchers based on publications within the last decade.

The incorporation of Scopus into our process proved highly beneficial. It unveiled a significant number of previously undiscovered citations, enriching our findings within the WoS database. This inclusion led to an increase in identified missing citations from 696 to 1075 in WoS. In addition, we discovered 1137 missing citations within Scopus itself, resulting in a total of 2212 missing citations across both databases.

A noteworthy observation was the vulnerability of WoS in handling recent years, as evidenced by a substantially higher number of missing citations in 2021 compared with previous years. In contrast, Scopus displayed a more robust ability to index and identify more recent publications.

The incorporation of Scopus into the citation search process has yielded valuable insights into missing citations within WoS. The study's findings highlight the potential benefits of including additional databases to improve the accuracy and comprehensiveness of citation data. By aggregating data from multiple sources, researchers can better understand the true extent of citations for a paper, providing a more comprehensive view of its impact within the academic community.


As for future work, adding other citations databases would be valuable, as it was discussed above, and we believe that adding GS could unearth the most missing citations in both WoS and Scopus. However, it is essential to acknowledge the challenges posed by GS, because this database does not provide the DOI of each publication, and they index many more publications since they do not have a source list, and publications do not have to adhere to the peer review format that WoS and Scopus do. To address these challenges, further enhancements to our matching algorithms, incorporating natural language processing techniques to compare publication titles, authors' names and other relevant parameters, are warranted. This will help minimise mismatches and ensure the accuracy of our findings. In addition, we believe that it would be worth investigating if missing citations are somewhat related with certain elements of the publications, such as specific publishers or journals.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was funded by FCT/MCTES through national funds and when applicable co-funded by FEDER–PT2020 partnership agreement under the scholarship reference Iscte_SIIC/01/2022 and projects UIDB/50021/2020 and UIDB/50008/2020.

ORCID iDs

David Rodrigues  <https://orcid.org/0009-0002-1150-1217>

António Lopes  <https://orcid.org/0000-0003-3045-0304>

Fernando Batista  <https://orcid.org/0000-0002-1075-0177>

Note

1. https://ciencia.iscte-iul.pt/api/v2_6/doc.

References

- [1] Rodrigues D, Lopes AL and Batista F. Web of science citation gaps: an automatic approach to detect indexed but missing citations. In: Simões A, Berón MM and Portela F eds. *12th Symposium on Languages, Applications and Technologies (SLATE 2023), Open Access Series in Informatics (OASIs), volume 113*. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN978-3-95977-291-4, 2023, pp. 5:1–5:11. DOI: 10.4230/OASIs.SLATE.2023.5. <https://drops.dagstuhl.de/opus/volltexte/2023/18519>
- [2] García-Pérez MA. Accuracy and completeness of publication and citation records in the web of science, PsycINFO, and Google Scholar: a case study for the computation of H indices in psychology. *J Am Soc Inf Sci Technol* 2010; 61(10): 2070–2085. DOI: 10.1002/asi.21372. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21372>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21372>

- [3] Moed HF, Bar-Ilan J and Halevi G. A new methodology for comparing Google Scholar and Scopus. *J Informetr* 2016; 10(2): 533–551. DOI: 10.1016/j.joi.2016.04.017. <https://www.sciencedirect.com/science/article/pii/S1751157715302285>
- [4] Krauskopf E. Missing documents in Scopus: the case of the journal *Enfermeria Nefrológica*. *Scientometrics* 2019; 119: 543–547. DOI: 10.1007/s11192-019-03040-z. <https://doi.org/10.1007/s11192-019-03040-z>
- [5] Teplitskiy M, Duede E, Menietti M, et al. How status of research papers affects the way they are read and cited. *Res Policy* 2022; 51(4): 104484. DOI: 10.1016/j.respol.2022.104484. <https://www.sciencedirect.com/science/article/pii/S004873322000129>
- [6] Buchanan RA. Accuracy of cited references: the role of citation databases. *Coll Res Libr* 2006; 67(4): 292–303. DOI: 10.5860/crl.67.4.292. <https://crl.acrl.org/index.php/crl/article/view/15806>
- [7] Cioffi A, Coppini S, Massari A, et al. Identifying and correcting invalid citations due to DOI errors in Crossref data. *Scientometrics* 2022; 127: 3593–3612. DOI: 10.1007/s11192-022-04367-w. <https://doi.org/10.1007/s11192-022-04367-w>
- [8] Zhu J, Hu G and Liu W. DOI errors and possible solutions for Web of Science. *Scientometrics* 2018; 118: 709–718. DOI: 10.1007/s11192-018-2980-7
- [9] Franceschini F, Maisano D and Mastrogiacomio L. Errors in DOI indexing by bibliometric databases. *Scientometrics* 2015; 102: 2181–2186. DOI: 10.1007/s11192-014-1503-4. <https://doi.org/10.1007/s11192-014-1503-4>
- [10] Rivkin A. Manuscript referencing errors and their impact on shaping current evidence. *Am J Pharm Educ* 2020; 84(7): ajpe7846. DOI: 10.5688/ajpe7846. <https://www.ajpe.org/content/84/7/ajpe7846>. <https://www.ajpe.org/content/84/7/ajpe7846.full.pdf>
- [11] Tahamtan I and Bornmann L. What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics* 2019; 121: 1635–1684. DOI: 10.1007/s11192-019-03243-4. <https://doi.org/10.1007/s11192-019-03243-4>
- [12] Martín-Martín A, Orduna-Malea E and Delgado López-Cózar E. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics* 2018; 116: 2175–2188. DOI: 10.1007/s11192-018-2820-9. <https://doi.org/10.1007/s11192-018-2820-9>
- [13] Labbé C. Ike Antkare one of the great stars in the scientific firmament. *ISSI News* 2010; 6: 48–52.
- [14] van Eck NJ and Waltman L. Accuracy of citation data in Web of Science and Scopus. arXiv. Epub ahead of print June 2019. DOI: 10.48550/ARXIV.1906.07011. <https://arxiv.org/abs/1906.07011>
- [15] Franceschini F, Maisano D and Mastrogiacomio L. A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics. *J Am Soc Inf Sci Technol* 2013; 64(10): 2149–2156. DOI: 10.1002/asi.22898. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22898>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22898>
- [16] Franceschini F, Maisano D and Mastrogiacomio L. Do Scopus and WoS correct ‘old’ omitted citations? *Scientometrics* 2016; 107: 321–335. DOI: 10.1007/s11192-016-1867-8. <https://doi.org/10.1007/s11192-016-1867-8>
- [17] Visser M, van Eck NJ and Waltman L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft academic. *Quant Sci Stud* 2021; 2(1): 20–41. DOI: 10.1162/qss_a_00112. https://doi.org/10.1162/qss_a_00112. https://direct.mit.edu/qss/article-pdf/2/1/20/1906541/qss_a_00112.pdf
- [18] Priem J, Piwowar H and Orr R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv. Epub ahead of print May 2022. DOI: 10.48550/ARXIV.2205.01833. <https://arxiv.org/abs/2205.01833>
- [19] Besançon L, Cabanac G, Labbé C, et al. Sneaked references: Fabricated reference metadata distort citation counts. *J Assoc Inf Sci Technol* 2024; 75(12): 1368–1379. DOI: 10.1002/asi.24896. <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24896>
- [20] Culbert JH, Hobert A, Jahn N, et al. Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics* 2025; 1–18. DOI: 10.1007/s11192-025-05293-3. <https://doi.org/10.1007/s11192-025-05293-3>