INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Automatic Detection of Missing Information in the Indexing of Scientific Publications

David Miguel Nunes Rodrigues

Master in Computer Engineering

Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
PhD António Luís Lopes, Associate Researcher,
Iscte - Instituto Universitário de Lisboa

October, 2023

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Automatic Detection of Missing Information in the Indexing of Scientific Publications

David Miguel Nunes Rodrigues

Master in Computer Engineering

Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
PhD António Luís Lopes, Associate Researcher,
Iscte - Instituto Universitário de Lisboa

October, 2023

Acknowledgment

I would like to express my appreciation to everyone who supported me through this phase of my life. The help I received was crucial for me to have a successful academic journey. I want to begin by thanking my family, who always believed in me and made sure I had everything I needed. I also want to acknowledge my friends and the people I met at ISCTE, with whom I shared good times and sleepless nights studying and working on projects.

Lastly, I want to express my gratitude to the teachers at ISCTE, who have always been available and committed to providing us with a strong knowledge base. I would like to give special thanks to my thesis supervisors, Fernando Batista and António Lopes, who worked tirelessly with me throughout the year to produce the best possible research.

Funding: This work is 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.

Resumo

A quantidade de citações que uma publicação científica recebe é uma métrica crucial. Uma publicação pode ser indexada por diferentes bases de indexação de artigos científicos, o que nos permite encontrar citações em falta relativas a essa publicação. Para colmatar esse problema, apresentamos uma solução que deteta automaticamente as citações em falta. Neste projeto, procuramos identificar citações em falta nas bases de indexação Web of Science, Scopus e Google Scholar, além de utilizar o OpenAlex para melhorar a quantidade de citações em falta encontradas.

Durante este projeto, realizámos várias experiências, começando por um protótipo que apenas utilizava 2 bases de indexação (Web of Science e OpenAlex) e depois expandimos a nossa abordagem para incluir o Scopus. Infelizmente, não nos foi possível adicionar o Google Scholar. Ao realizar essas duas experiências, foi possível comparar os dados obtidos no Web of Science antes e depois da inclusão do Scopus, o que nos permitiu avaliar o impacto do acréscimo de uma base de dados na nossa abordagem. Posteriormente, realizamos outra experiência, a fim de avaliar as mudanças que as próprias bases de indexação vão tendo ao longo do tempo.

Depois de analisar mais de 3 000 publicações, detectamos citações em falta em 874 publicações, totalizando 2 212 citações em falta, das quais 1 075 foram detectadas no Web of Science e 1 137 no Scopus. As 1 075 citações detectadas no Web of Science representam um aumento de 54% na quantidade de citações encontradas antes de acrescentar o Scopus à nossa abordagem.

Palavras-Chave: Bases de Indexação, Citações, Web Scraping, Web of Science, Scopus, OpenAlex

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 one or more databases, which are not contributing to a paper's total citation count. To address this issue, we have developed an automated method for identifying missing citations by leveraging multiple indexing databases. In this research, we sought to identify these missing citations in Web of Science, Scopus, and Google Scholar while also utilizing OpenAlex to aid in this process.

Our research journey involved multiple experiments. Initially, we started with a prototype that used only two databases (Web of Science and OpenAlex) and later expanded our approach to include Scopus. Unfortunately, we were unable to incorporate Google Scholar. By conducting these experiments, we were able to compare the data found in Web of Science and gain a deeper understanding of the impact of adding a new database. We also repeated the same experiment one month later to track the changes that occur over time in these databases.

After analyzing more than 3 000 different publications, we successfully identified missing citations in 847 of them, totaling 2 212 missing citations. Out of these, 1 075 were missing from Web of Science, and 1 137 were missing from Scopus. The addition of Scopus to our approach resulted in a 54% increase in the number of missing citations detected in Web of Science, highlighting the significant impact of incorporating this database.

Keywords: Research Databases, Citations, Web Scraping, Web of Science, Scopus, OpenAlex

Contents

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface
CRIS	Current Research Information System
DB	Database
DOI	Digital Object Identifier
GS	Google Scholar
JSON	JavaScript Object Notation
MC	Missing Citations
WoS	Web of Science

CHAPTER 1

Introduction

The number of citations on a research paper is of great value for the researcher and their work, since usually, the more citations a paper has, the higher are the chances of being worth reading and having helpful information. Thus, the authors of the paper can also get prestige from the number of citations of their papers and will be more highly regarded. This is useful not only for the authors to get recognition by their peers, but also for performance evaluation processes in the researchers' institutions that usually include publications' citations as one of the main metrics. This raises the importance of knowing the right amount of citations that any given paper has. Therefore, the problem we aim to solve is to find a way to automatically merge the information from multiple indexing databases so we can get a closer depiction of the real number of citations of a given paper without having to go through the slow and tedious process of manually assessing these differences. We are looking for missing citations in two indexing databases, Web of Science (Wos) and Scopus, while also using OpenAlex[1] to gather extra information we can use to solidify our findings. We also intended to extract information from Google Scholar (GS) but we were unable to do so, which will be explained further within this document.

1.1. Background and Motivation

The scientific community looks at research papers as a backbone of their work. In order to create new knowledge, a thorough research on previous studies needs to be done; finding out what has been tried but failed, or what has been successful is a priority so that time and resources are not spent on what is already available and known.

Although gathering and analyzing previous studies is of the utmost importance, with the huge amount of new research papers that are written every year, it's hard for any researcher to be completely up to date with the new developments in a particular area of expertise.

Given this problem, there is a need to identify which papers should be given the priority. There are different ways of trying to filter and find out the most promising papers of the subject one is studying, like where the paper was published, since the more highly regarded events or journals usually provide the papers of higher quality and substance. In addition, the amount of citations that a paper receives is also a factor that is looked upon, since usually, the more citations a paper has, the higher are the chances of being worth reading and having helpful information. Since one of the metrics in which a research paper is evaluated can be the amount of citations it gets, the authors of said

[1] https://openalex.org/

paper can also get prestige from the number of citations of their paper, and if the authors have a considerable amount of their work with a great deal of citations, then the authors themselves will be more highly regarded.

The job of calculating the number of times a paper has been cited is no easy or simple task. There are multiple citation databases that try to gather this information and present it to the scientific community, however, they don't all classify the same papers as being cited by an article, even if said papers and articles are all indexed in the databases.

Having this discrepancy between databases, makes it hard to have a clear notion of the worth of the work that has been done. Therefore, trying to calculate the right amount of citations a research paper has, is advantageous for everyone and we wished to find and merge the information from all three databases. However, doing this process manually, is tedious and time-consuming, and is not feasible to do so for a great number of research papers every time one is doing a literature review or evaluating a researcher's performance.

1.2. Research Questions

The research questions we wish to address with this work are:

- Is it possible to automatically detect missing citations within different indexing databases?
- What is the impact of adding new indexing databases to the coverage of the process of finding missing citations?

1.3. Goals

The main goal of this project is to build an automated system that consults the citation list of three of the main citations databases, Web Of Science (WoS), Google Scholar (GS) and Scopus, as well as OpenAlex, in order to find different citations between these databases, and if it does, be able to check if these citations are missing or not from the databases. After calculating whether or not there are missing citations, create a report with all the information gathered, where we can see all the citations of an article across all the databases, and the missing citations from each of them, so that afterwards, the researchers can manually verify these missing citations and accordingly warn the specific database of this missing citations. This research and application will have access to Ciência-IUL's API, a system from ISCTE - Instituto Universitário de Lisboa, containing researchers and their respective scientific productions, with the aim of reporting the incorrect citation count of ISCTE's productions. Therefore, all tests and results will use data provided by this system.

1.4. Real and Theoretical Demonstration of the Problem

Now we will present a real instance of this problem, where we manually look and identify missing citations. This information was gathered on the 17 of February of 2023 except for the OpenAlex information which comes from a later date of the 22 of June of 2023.

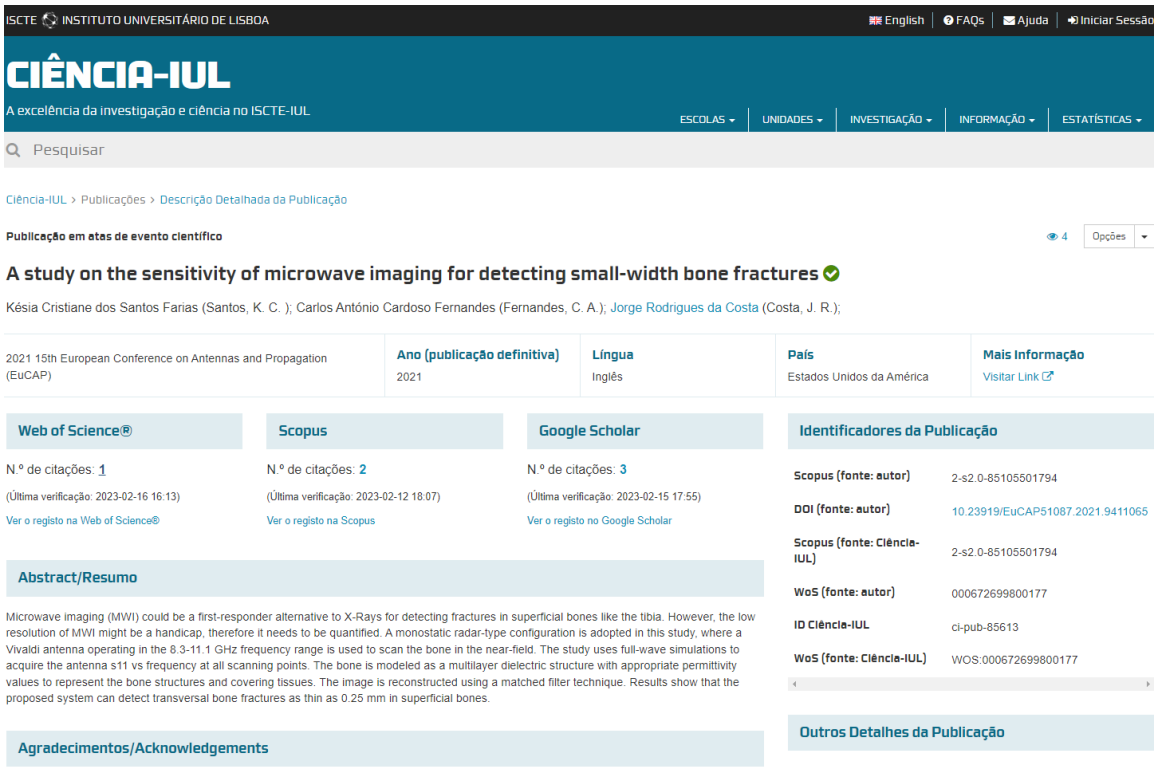


FIGURE 1.1. Ciência-IUL web page screenshot collected at 17/02/2023

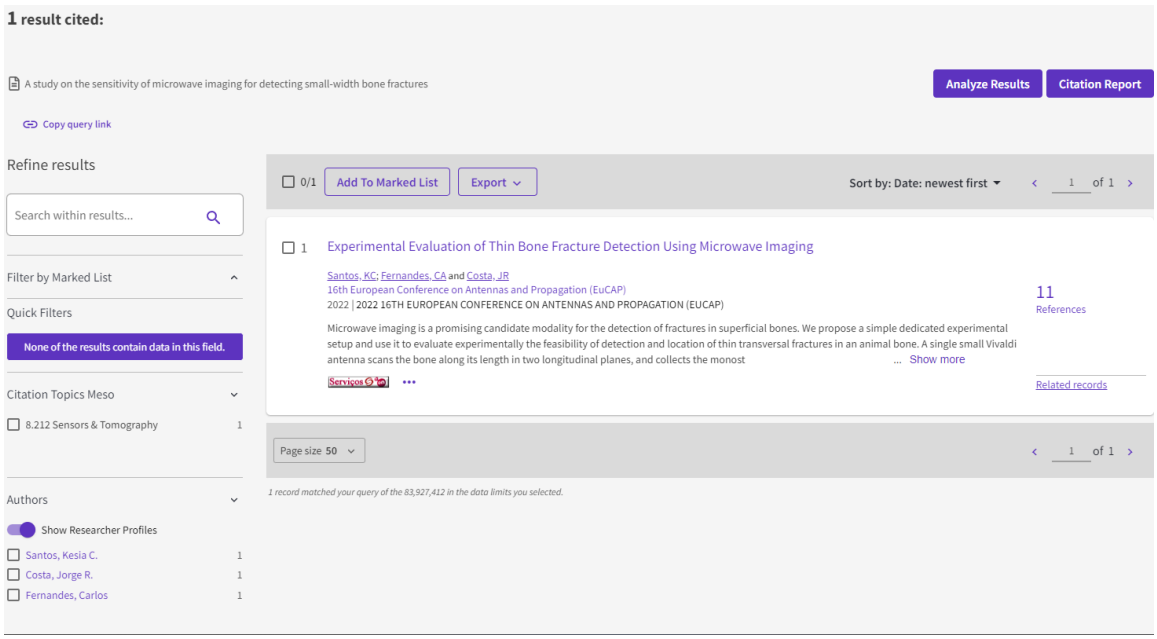


FIGURE 1.2. Web of Science web page screenshot of the citing articles of the paper collected at 17/02/2023

Firstly we go to the Ciência-IUL system as shown in Figure 1.1, where we find the information about the article we are analyzing. From here we can extract a lot of data, but the main information we are looking for is the number of citations from Web of Science, Scopus and Google Scholar. Each of these is a link, for the list of citing publications for each respective database.

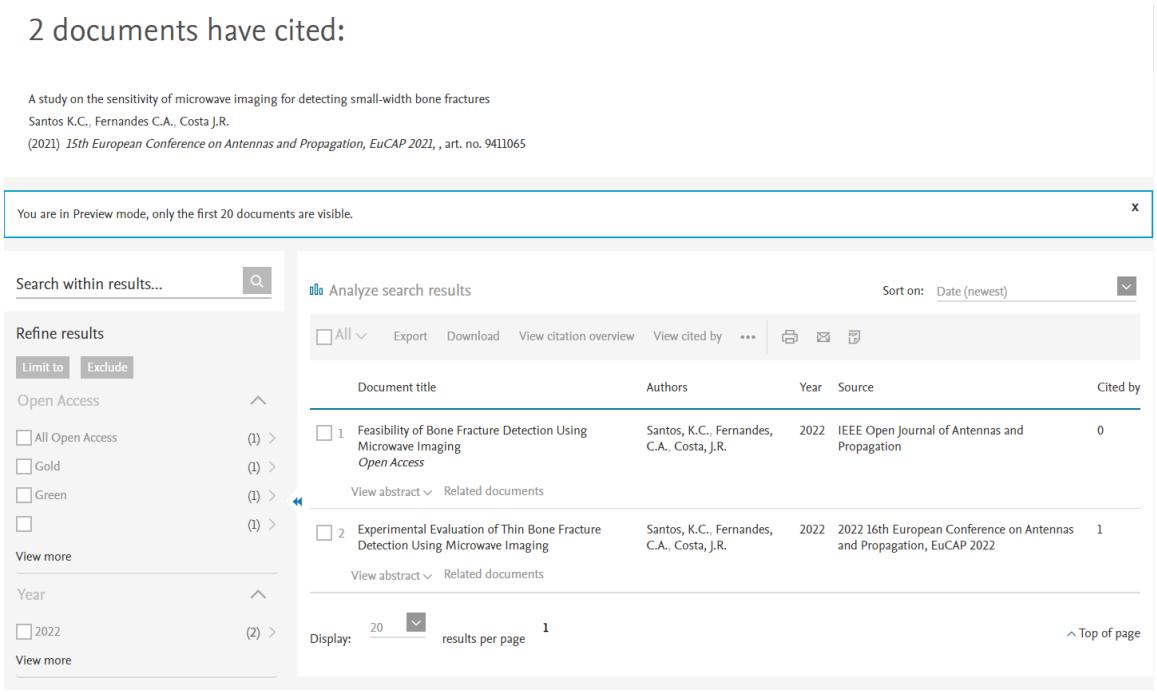


FIGURE 1.3. Scopus web page screenshot of the citing articles of the paper collected at 17/02/2023

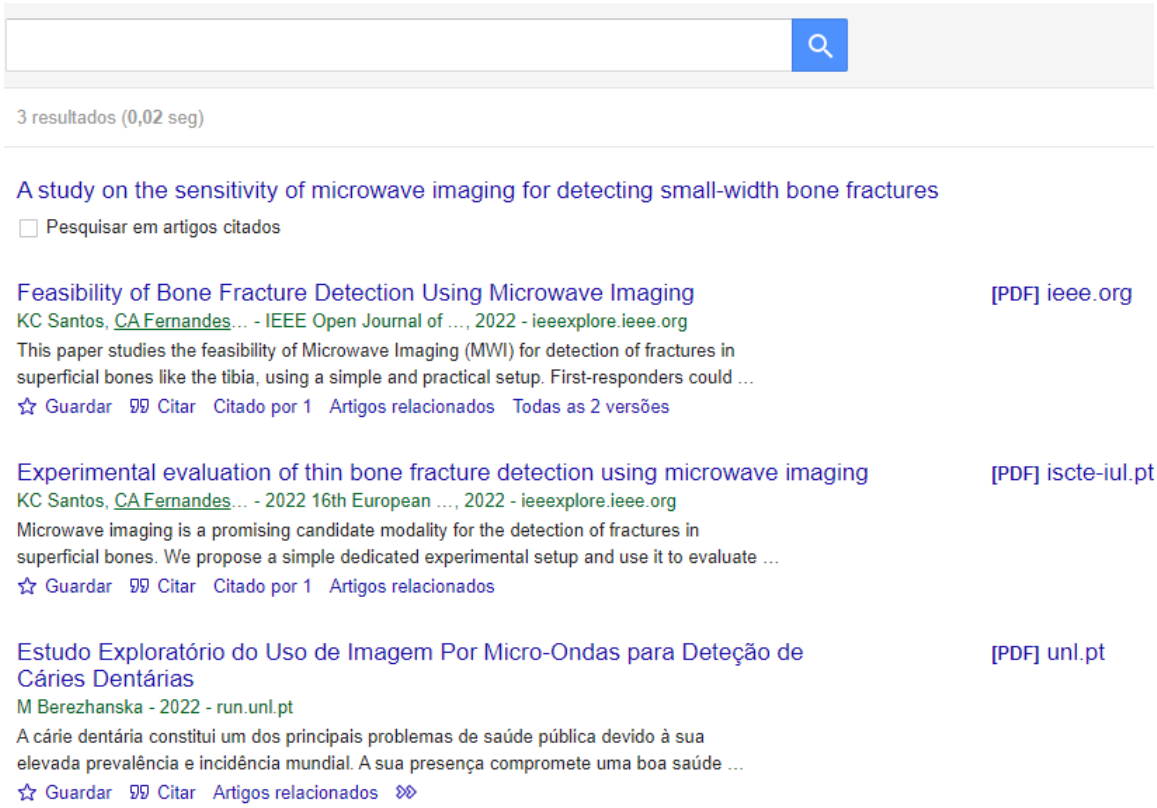


FIGURE 1.4. Google Scholar web page screenshot of the citing articles of the paper collected at 17/02/2023

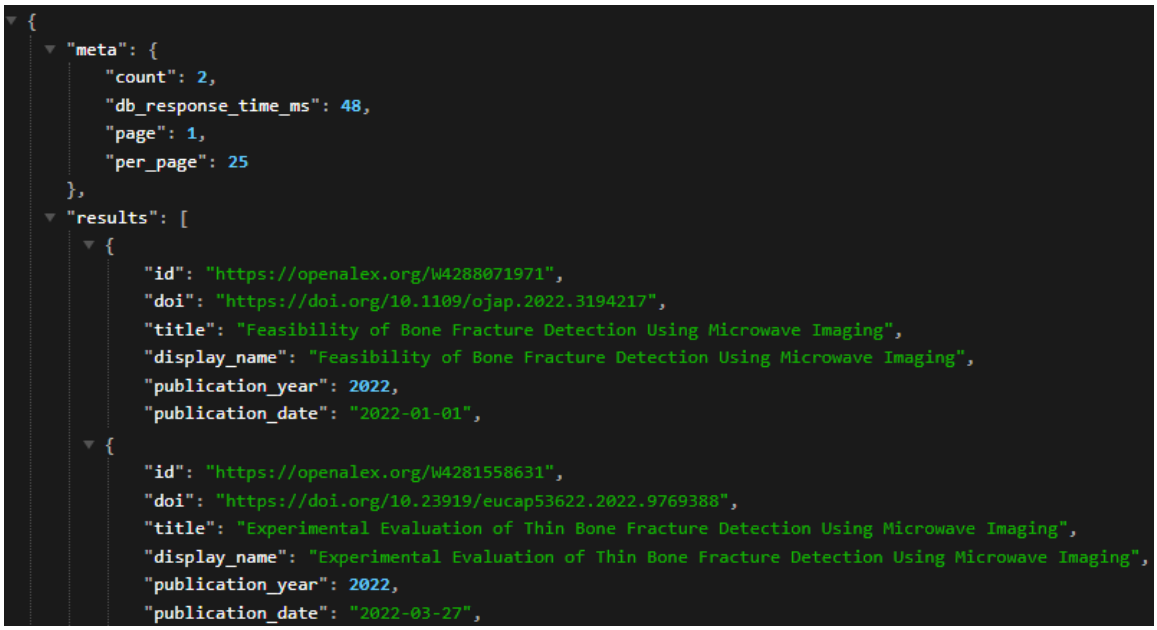


FIGURE 1.5. OpenAlex web page screenshot of the citing articles of the paper collected at 22/06/23

If we click on each of those links, we get sent to the indexing database web site as we can see in the Figures 1.2, 1.3 and 1.4. In this case, every list is different, but there are common papers for each of them. As for OpenAlex, we can access it through the papers DOI, by going to the url `"https://api.openalex.org/works/https://doi.org/<DOI>"` where `<DOI>` is the DOI of the paper we are searching for. A snapshot of a portion of the results from OpenAlex is shown in figure 1.5. In this case, all of OpenAlex's citations were found in the other databases, being the exact same as Scopus citations, so it does not provide any new information.

The only paper WoS found to cite the paper we are analyzing is a paper called "Experimental Evaluation of This Bone Fracture Detection Using Microwave Imaging", and this paper is also present in both Scopus and Google Scholar, therefore it can not be a missing citation in neither of the databases. On the other hand, there is a paper called "Feasibility of Bone Fracture Detection Using Microwave Imaging" which is listed in both GS and Scopus. This means that this article is a potential missing citation for WoS, something we need to confirm.

In order to confirm if it is a missing citation, we must access the WoS web page and query the database looking for this paper, as shown in the Figure 1.6. In this case we queried the database using the paper's title and the query was successful, since the first result on the list of results is the paper we were looking for. Because we found the paper indexed on WoS but the database did not have it in the citations list, we are in the presence of a missing citation.

Lastly there is a citing paper in GS, Figure 1.4, that neither of the other databases found, titled "Estudo Exploratório do Uso de Imagem Por Micro-Ondas para Deteção de Cáries Dentárias". Therefore, once again, we need to check WoS and Scopus for the

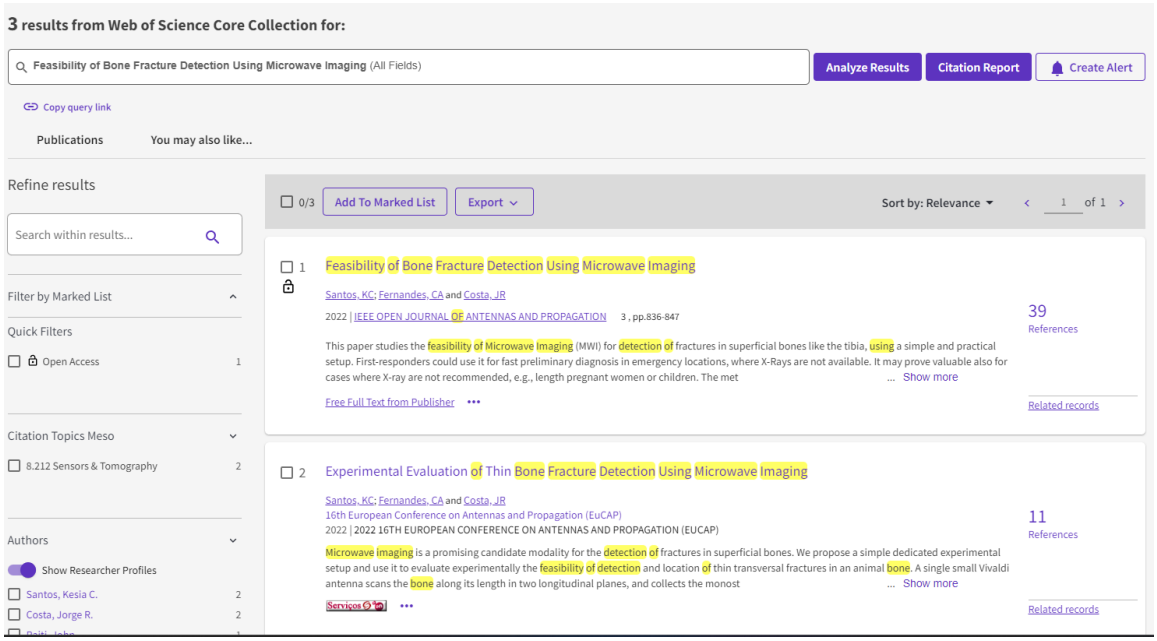


FIGURE 1.6. Web of Science web page of the query results collected at 17/02/2023

presence of this paper in the databases, in order to determine whether or not it is a missing citation. As shown in the Figures 1.7 and 1.8, this citing paper was not found indexed in either of the databases, so there are no missing citations in this case.

For this paper we found a total of 8 citations (3 from GS, 2 from OpenAlex and Scopus, and 1 from WoS) across the databases, where only 3 were different citations. We had 3 potential missing citations (1 in Scopus and 2 in WoS), but only 1 of them ended up being a missing citation.

If now we look at a generic experience and the results of what we wish our program to deliver, we want it to create a report for a given Paper A, like the one shown on Figure 1.9, where we have a column for each of the databases where we are looking for missing citations, and each line corresponds to a different citation for Paper A. Each citing article in this example has been given a generic number to represent it, while in reality we would provide both the DOI and the title of the citing publications. For this paper A we found 6 different citations, from WoS we got the citation list 1, 4 and 6; from Google Scholar we got 1, 2, 3 and 4; and from Scopus we got 1 and 5. In cases like citing article 1, where all databases had the article tagged as citing, we had to do nothing. When we move to article 2, only GS found this citation, which meant that we would have to go to both WoS and Scopus, to confirm if this article was indexed in any of these databases, and because it was indexed in both, it is a missing citation for both. While for article 3, also only GS had it tagged as a citing paper, but after checking WoS and Scopus, and not finding this article indexed in any of them, we can say it is not a missing citation. In this table, every cell with `Missing` or marked with an `x` was a potential missing citation that we had to check for the existence of that article in the respective database, in order to confirm if it was either a `Missing` citation or if it should be marked with an `x`.

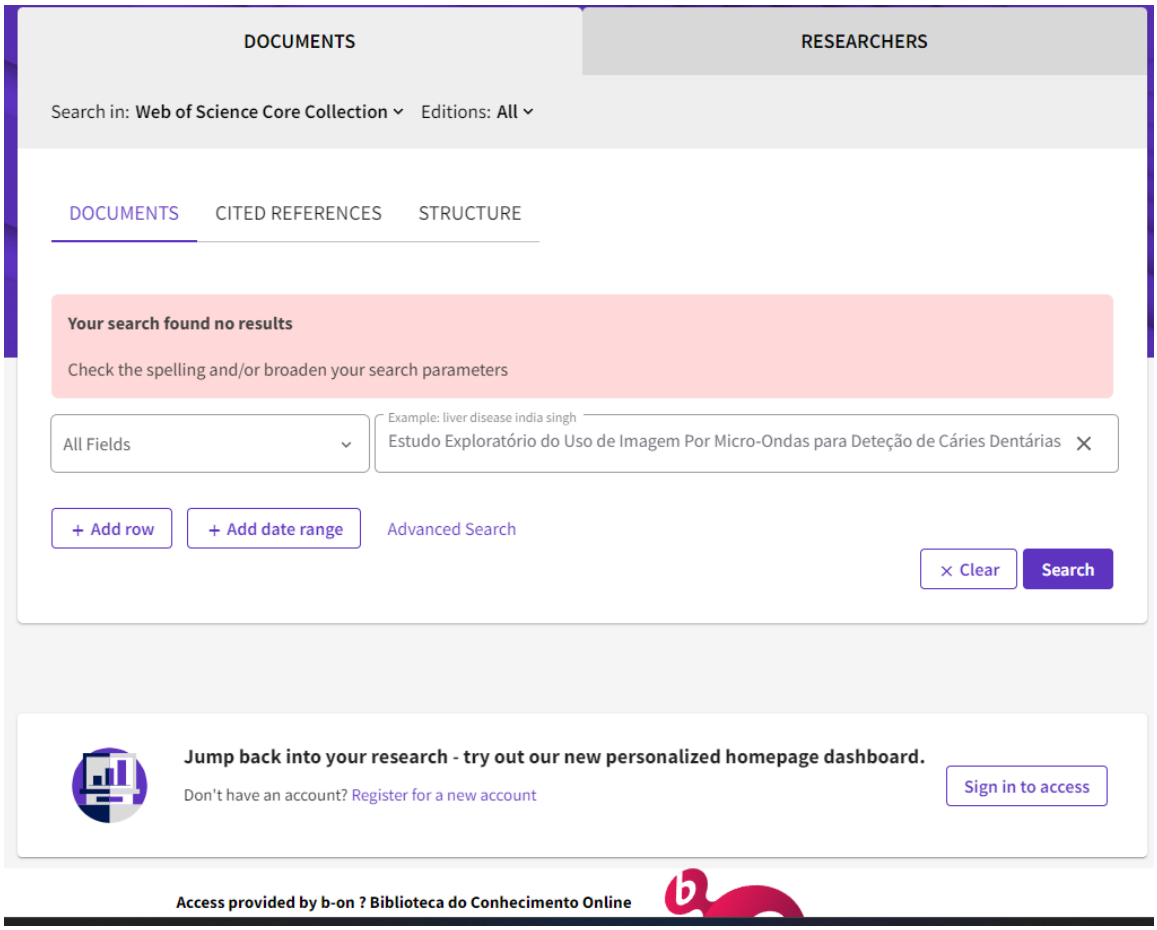


FIGURE 1.7. Web of Science query web page with no results. Collected at 17/02/2023

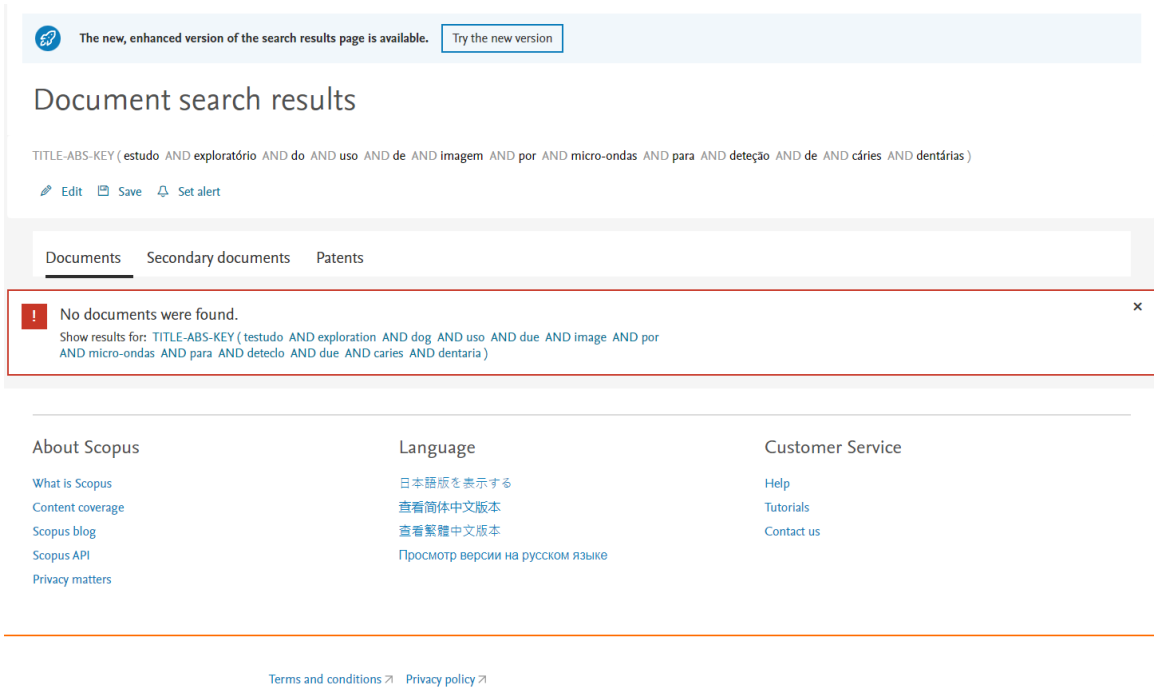


FIGURE 1.8. Scopus web page of the query results collected at 17/02/2023

Paper A Analysis			
Articles	WoS	GS	Scopus
1	✓	✓	✓
2	Missing	✓	Missing
3	×	✓	×
4	✓	✓	×
5	×	×	✓
6	✓	Missing	Missing
Recorded Citations	3	4	2
Expected Citations	4	5	4
Missing Citations	1	1	2
Unique Citations	0	1	1
Missing Percentage	25%	20%	50%

FIGURE 1.9. Example of a report on the missing citations of a paper.

From this report we can gather a lot of information, like the total number of different citations of a paper, which is 6 in this case, one for each entry in the table, we can also get the individual citation count for each database, as well as the expected citation count (citations the database found plus the missing citations). We can find how many citations are only indexed in each database, which we called unique citations. These unique citations could never have been accounted for, if this database was not used in the process. Finally, we can also calculate the percentage of missing citations for each database. For this, we decided to use the amount of citations the database found and the amount of missing citations we found in the database, in order to represent how many citations are missing from the total amount of citations that this database should have.

This process is applicable to any new databases we could add to our approach, to try and search for missing citations in, or just to provide extra new citations that we can search for in the other databases. In the cases of these databases where we do not look for missing citations, like OpenAlex, there is no real need to show them in the report. The only information that adding these databases to the report would give us, is where some missing citations may come from. If the client wishes to see this extra information, it is also possible, although in this example we did not add OpenAlex in order to make it more simple to explain and analyze.

1.5. Methodology

For the elaboration of this project we did not follow any pre-existing methodology structure. Instead we designed a plan we thought that best suited our approach and ideas. Because this approach had already been explored in a manual analysis that was done previously by a professor at ISCTE, we followed the manual approach methodology and transformed it to an automatic approach. As for the dataset, we already had access to the Ciência-IUL database which has information about the papers from the researchers of the ISCTE-IUL institution. Because this process was being built around the idea

of being tested with papers from ISCTE, it was the perfect database for us, since this dataset has the same papers that our application will work it. Our plan started by doing the literature review in order to understand the dimension of the problem at hand, as well as any difficulties we could come across or techniques that could be useful for our approach. After that, we explored the information from the Ciência-IUL database as well as each of the different indexing databases we were going to analyze in order to understand what was the best approach.

Because the web scraping portion of the program seemed to be the most challenging section of the program, we decided to first test our approach on a smaller scale; we would elaborate a program to detect missing citations on the Web of Science database, using only OpenAlex as a baseline. We chose both these databases to start with because OpenAlex is a free tool, accessible by API, providing a lot of information about the articles and citing articles, in a fast and easy way, and we did not want to look for missing citations in it. As for Web of Science, we decided to start with it because we found a way to download the data in a structured file (excel file) which was easily accessible by web scraping, and provided all the data already structured. We also thought that the WoS database would be the one providing the least amount of results, since their source list is smaller than Scopus and GS. This meant that we had to deal with less information, and could test our approach in a smaller and faster scale. With this first prototype of the program, we could run our approach and see the results in order to find what needed to be changed or improved. We also decided that it would be an interesting topic and starting point to write a conference article to report our findings.

After the article was written, we would then expand our approach to encompass the other databases, Scopus and Google Scholar, and once again, we would run the same set of experiments but with the extra data these added databases provide. This way we would also be able to compare the results of the before and after, having a better notion of the impact that adding other databases to our approach would have. These findings would then be presented in a second article, where we could show not only the improvements of our findings in WoS, but also what we manage to gather from the other databases.

CHAPTER 2

State of the Art

As it has been mentioned, doing research on what has already been studied is of the utmost importance, therefore we will also be doing it for this project. Our goal is to get a better understanding of the magnitude of the problem at hand as well as learning what has been done, in order to take approaches or ideas that could help us develop this system.

We started by getting a better understanding of both the importance of citations for the scientific community, and the importance of finding a way to resolve this problem, which is, the incorrect or incomplete citation count in the three main citation databases, Web of Science (WoS), Google Scholar (GS) and Scopus. After getting a grip of the problem and obstacles that face, we look for the best techniques we can use and how to apply them in this project.

2.1. The Importance and Problems of Citations

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 [1].

Therefore, looking only at one citation database can be misleading, since each citation database shows strengths in covering different areas [2], 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 [1]. 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 of the journal to periodically check if the papers are being indexed by the databases [3]. Like Scopus, the other databases also show 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 [2]. Nevertheless, neither database is perfect, and for better results, more than one should be used.

The study conducted by Teplitskiy et al. (2022) [4] shows that the citation count of an article is significantly influenced by a feedback loop: as articles accumulate more citations, they are more likely 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 likely 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 (2006) [5] reports a variety of errors, either by the authors or the databases, in both DOI's and references. Cioffi et al. (2022) [6] tries 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 of 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. Besides wrongfully written DOI's, Franceschini et al. (2015) [7] reports 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, 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% [8]. 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 [9]. 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 [3], since they are able to better find different forms of citations, leaving them less likely to have missing citations, which in turn provides a more complete list of citations. A paper [10] suggests 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 other hand, a great number of articles raise questions about the validity of GS citations. They [10] 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. Finally, a more cautious and intermediate approach is also suggested by another paper [1], 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 [11], tries to find what were the differences of the references 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 other hand, 96.4% of Scopus papers had the same amount of references. They manually analyzed 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.

2.2. Useful Techniques and Tools

This paper [12] 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 potentially missing. 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 [13], analyzed 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.

An advantageous tool we found is OpenAlex, a "fully-open index of scholarly works, authors, venues, institutions, and concepts" [14]. 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 [14]. Using this API it is possible to get easy access to the information from all these other platforms from a single query.

The research conducted by García-Pérez's in [1], through complex queries, tries to get all the articles of a specific author, both in WoS and GS. This work concludes that trying to query GS using author names normally provides very poor results, thus confirming what other authors previously reported. This work also reports the existence of duplicate citations in both databases.

CHAPTER 3

Proposed Approach

In this chapter we describe the dataset we used throughout this project, the proposed architecture of our system, all the external and internal systems we contacted and how they interacted with each other until we achieved the goal of finding missing citations for a given paper. We also have a brief description of our local database.

3.1. Dataset

As it has been mentioned before, the dataset we used was the Ciência-IUL database. This database aggregates data from the researchers of the ISCTE-IUL university, including their research papers. There are approximately 52 000 researcher papers indexed in it, but due to the fact that for our approach we can only analyze papers that have an assigned DOI, we could only work with 12 641 of those papers. Furthermore, some of the information on this database is manually inputted while other is automatic, and both are prone to errors. We have identified cases where for the same publication, there are two different DOIs, and also cases where although only one DOI is provided, it is wrong, since there are DOIs like "-", "00" or "colonia" which are not valid.

As it is shown on Figure 3.1, the vast majority of papers have been written in English, totaling 9505 papers, followed by 1534 papers written in Portuguese. Of the remaining 1602 papers, 1327 did not have any information about their language. As for the country

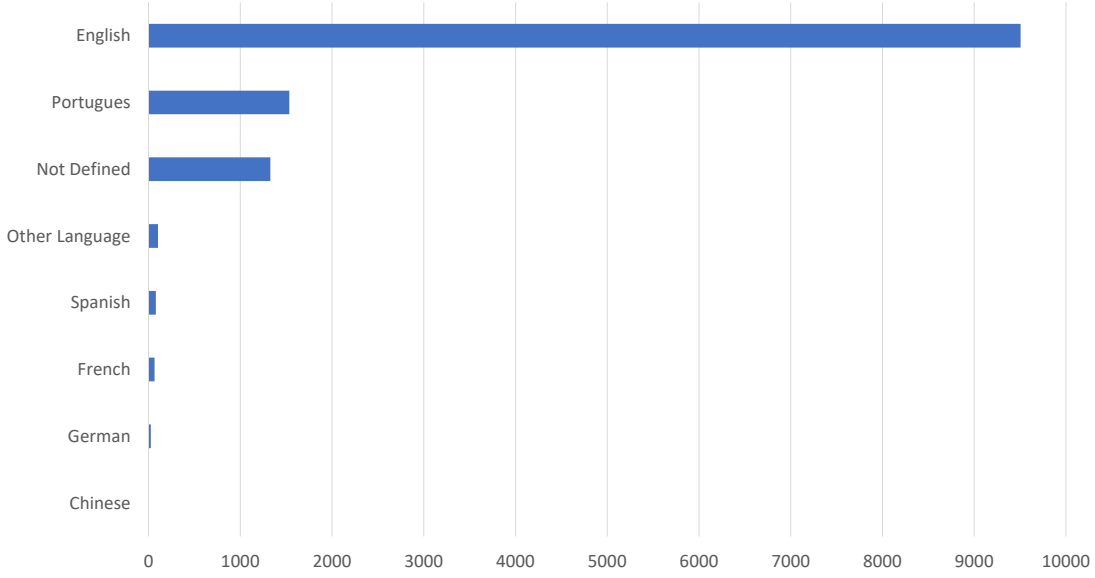


FIGURE 3.1. Number of papers written in each language

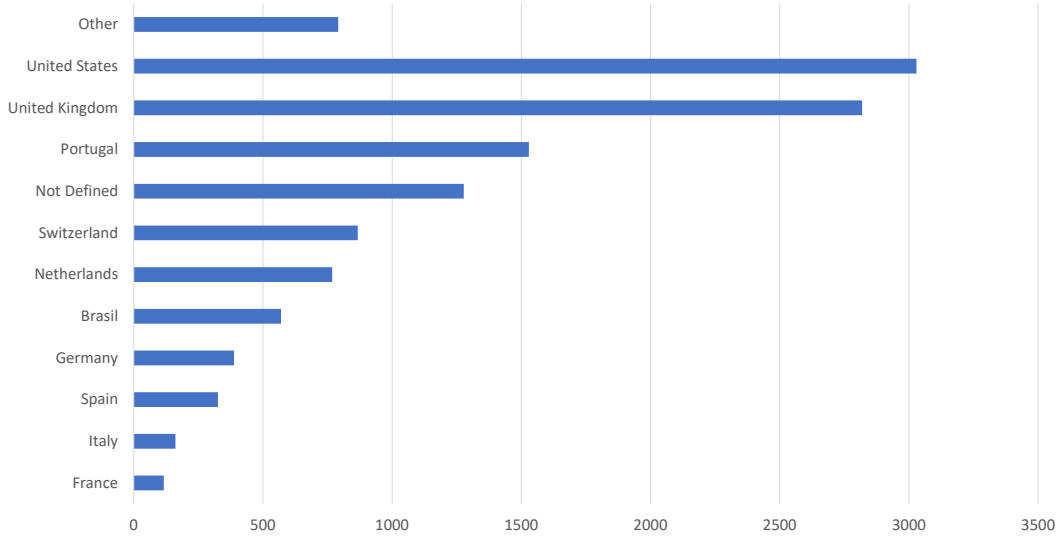


FIGURE 3.2. Number of papers publish for each country

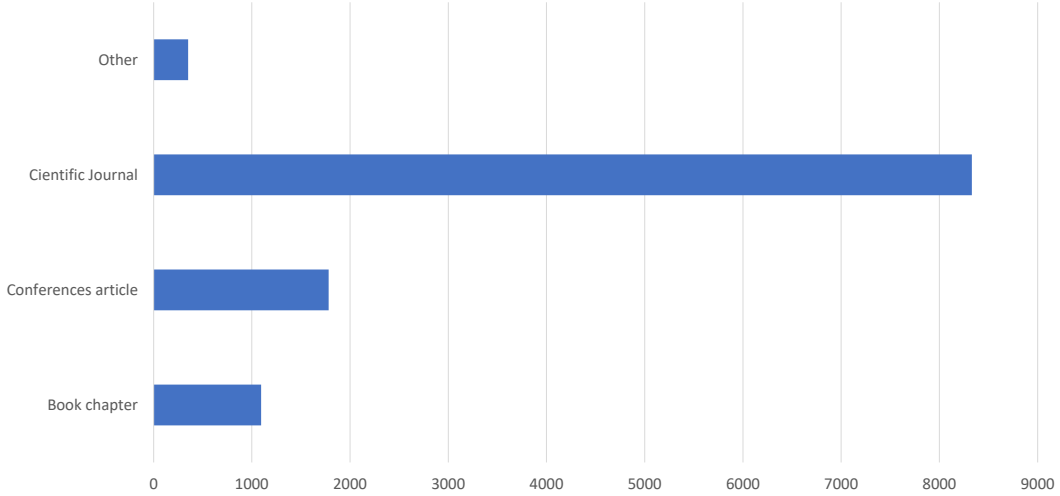


FIGURE 3.3. Number of papers of each type

where they have been published, in Figure 3.2 we can observe that 3029 have been published in the United States of America, closely followed by 2819 in the United Kingdom and 1529 in Portugal. Once again, there are 1277 papers where we do not have the information of where they have been published. Nevertheless, in total, the papers have been published in 69 different countries.

As to where the papers have been published, we can see in the Figure 3.3 most have been published in scientific journals, totalling 8330, while only 1782 were in conferences and 1095 are book chapters. One of the main reasons is that for these three cases,

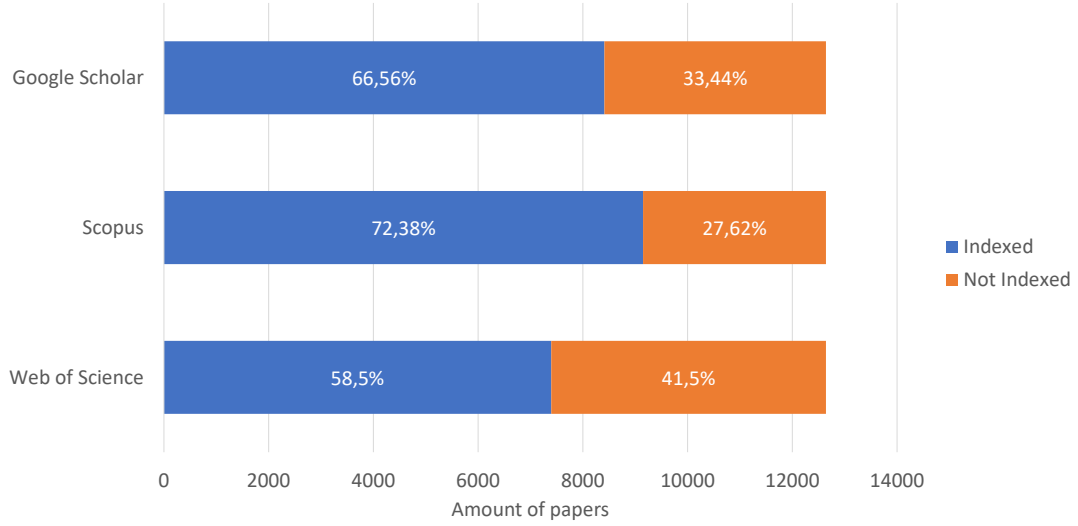


FIGURE 3.4. Number of papers of each Indexed or not in each database

normally papers are assigned DOIs, and since we only work with papers that have DOIs, these stand out.

In Figure 3.4 we can observe how many of the analyzed papers are indexed in each database. Scopus is the database with the most indexed papers (72,38%), while Web of Science only has 58,5% of the papers of our database indexed. This could lead to more missing citations being found in Scopus in total, because there is a higher chance of a paper indexed there, being analyzed, while in Web of Science, there are fewer papers where we can try to find missing citations. In order for us to try to find a missing citation, a paper only has to be indexed in 2 different databases, and OpenAlex can be one of those databases. As for Google Scholar our initial expectations were that they would have the most indexed papers, and also where we would be able to find the most citations for any given paper. But since we were not able to web scrape GS, the amount of indexed papers in this database does not influence the data we get for the other ones.

Finally in Figure 3.5 we can observe that these papers have been published anywhere between 1975 and 2023, but the majority of them have been published in the last decade. Once again, the need for the DOI in the paper can influence these values, since the use of a DOI has been a more common practice during most recent years. We can see that the trend is for more papers to be published each year, except for 2022, where there was a small decrease, most likely because of the effect of the COVID-19 pandemic in the world. Across this project, we will focus on papers from 2015, 2018 and 2021. Because of the time it would take to run the experiments of all 12 641 papers, we decided to look at the papers from only these years. This makes it so we have less data to process, taking less time, but gives us a way to analyze the evolution of the data gathered throughout the time. By choosing these 3 years, we had a sample of 3 074 papers to analyze, which

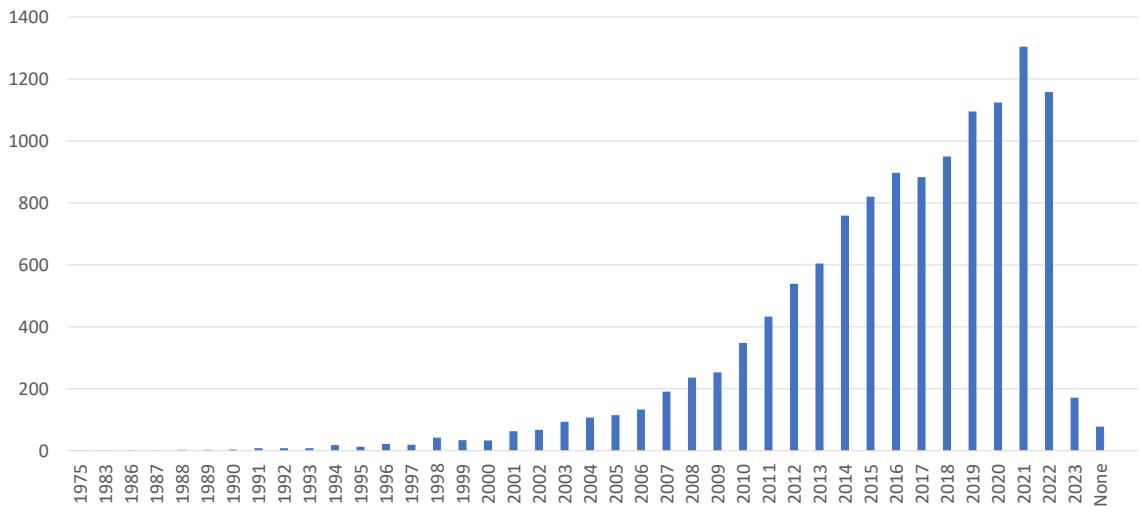


FIGURE 3.5. Number of papers per year

represent around 25% of all the papers we have at our disposal. We decided to focus on these years because we felt that we should focus on the most relevant data we have, and since the university evaluates their researchers' performances from the papers published in the most recent years, it would not be as interesting to see the data from papers that are much older. We also believed that the data from the most recent papers would be more prone to errors, since there was less time for databases to index the papers, and link the citations. Furthermore we had in mind that the COVID-19 pandemic had an impact on the data we were analyzing, therefore, the first year we chose to analyze was 2018, right before the pandemic but still recent. Afterwards, we decided to leave a gap of 3 years before and after, and chose 2015 and 2021. In the year of 2021 there were a couple of years for the scientific community to react and cite the new papers that had been published. As for 2015, it is an older year where the databases have had more time to cement the data, and where we would expect the better results.

3.2. Local Database

Our local database consisted of multiple tables. It was used as a cache, in order to prevent repeating previously done web scraping or processing. It enabled us to get a faster response for a request that has been done previously, as well as having a backup that we could consult in case we wanted to analyze the evolution of any type of data. Because all the data we gather from external sources is prone to changes or updates, we need to be careful with how old was the information we stored in our local database. Even if we had already extracted a citation list of a specific paper, it might not have been up to date anymore in a future analysis and we might have had to extract it once again. The lifetime of each time of information may differ and is adjustable in the configuration file. The lifetimes have not been adjusted yet, and it will be an institutional decision when each type of information is too old and needs to be updated. For all the experiments we

conducted in the elaboration of this thesis, some data has never changed, while other has always been treated as outdated.

The first table we have is a simple one called `Analyzed Publications`. Here we simply stored the DOI of all analyzed publications and the date of the request. This table was consulted when a request was sent to our program, in order to check if we had a up to date report on this publication or not. If it was up to date, then we would consult the `Final Reports` table in order to retrieve the report and send it to the client. In this table every entry was a new entry, in order to have a history of the requests.

The `Final Reports` table stored all the information needed to create a report. The primary key is the DOI of the publication, it also has the date of the creation of the report, a Boolean for each database telling us whether this paper is indexed in it or not, and a JSON file containing the required information to build the report. The data in this table was never updated, and every new report is a new entry, in order for us to analyze the evolution of each paper throughout the time.

There is also a table called `Ciencia-IUL Publications`. In this table we had all the information requested to the Ciência-IUL database about each publication. This table has not been updated during the elaboration of this thesis, in order to not add any new publications and maintain the same dataset for all experiments. The primary key was once again the DOI of the publication, and additional information is also stored, such as the Publication ID, which is the Ciência-IUL unique ID for this paper and a JSON file with all the information that was provided. Each different paper in this table only has one entry, even if analyzed multiple times.

Then we have a table for each of the databases used, Web of Science, Scopus and Google Scholar Citing Publication tables. These tables store the citations of each publication. Each entry of this table is a citation in the respective database. Like all previous tables, there is a date entry to know when the data was inserted. There is a composite primary key, which is the DOI of the publication we are analyzing and the the DOI of the citing publication. There is also a JSON file, which is all the information we have regarding the citing publication. Every time a publication is analyzed, older outdated citations are erased and new entries for the new list of citing papers are created.

Finally, we have the OpenAlex tables. Because we extract both the information about the publication on OpenAlex, as well as the list of citing papers, we have two separate tables for each of these types of data. The `OpenAlex Publications` table has the DOI of the publication as a key, the date the information was gathered and the JSON data with information about this specific paper, and the `OpenAlex Citing Publications` table has the DOI of the publication as a primary key, and a JSON File containing all the citing publications and information about each of them. Again, none of these have repeated information about the same paper, and instead the old data is updated.

Lastly, we had other tables created in order to analyze the data we were gathering, like the `OpenAlex Not Found Publications` table, where we have all the DOIs not found in

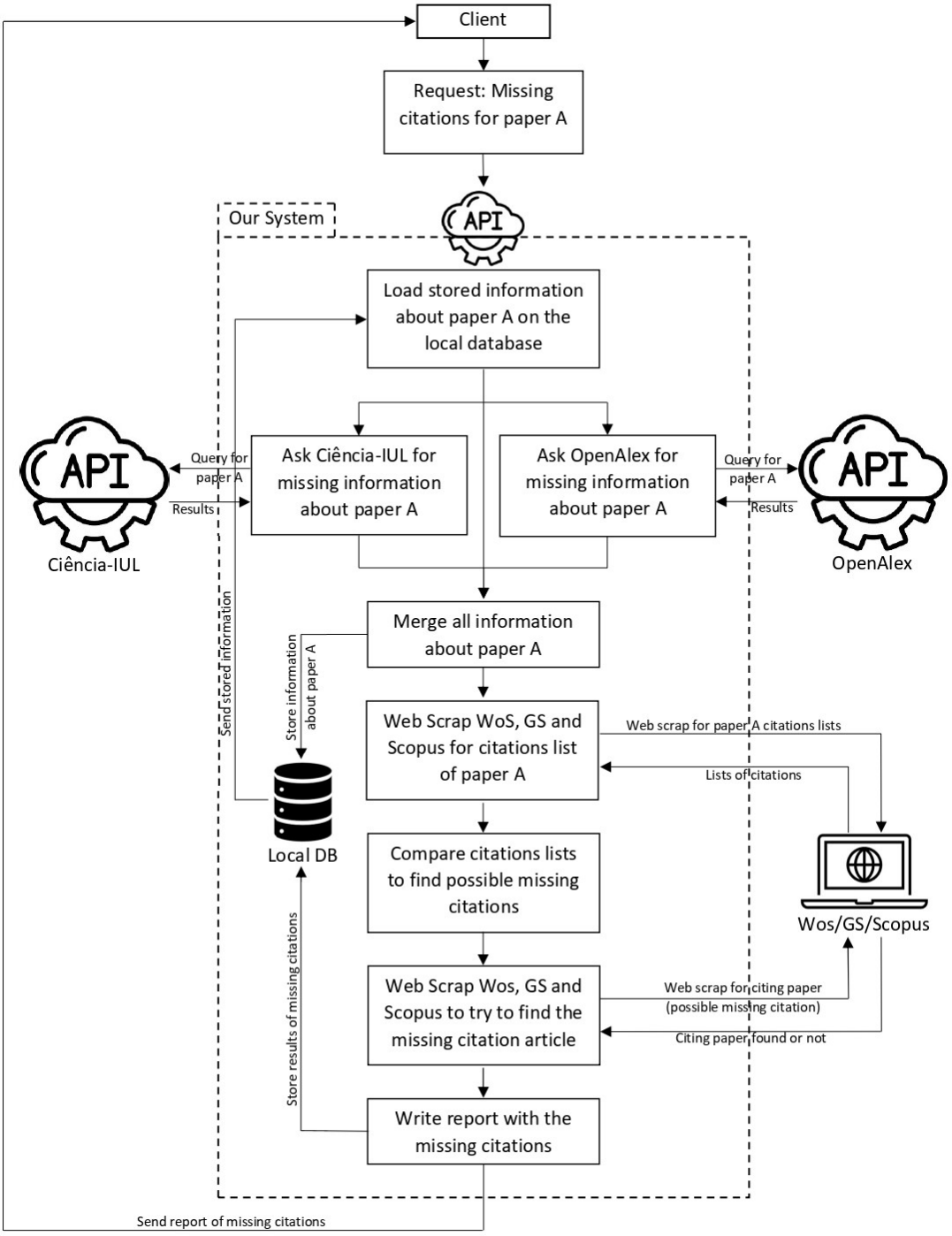


FIGURE 3.6. System Architecture.

OpenAlex, which gives us a way to more easily find invalid DOIs that were given as an input to our program. Not all of the DOIs are going to be invalid, because OpenAlex does not have all the publications indexed in it, but it gave us a smaller sample to search through in order to find errors in the input DOIs, some that could be fixed, others that were out of our scope.

3.3. Proposed Architecture

In order to look for missing citations in an article, we need to go though various steps, described in Figure 3.6, and in order to start this process, all we need is the article DOI. The first step in the verification process is to clean the received data and check if a report about this DOI has been created previously. If it has and we consider that it is up to date, then the report is sent to the client. If it is not up to date, then a new report will be generated.

In order to generate a report, we first want to know the information about this paper on our Current Research Information System (CRIS) API[1]. If a report has been created previously on this publication, then we might already have this information in our local database. Therefore, we search it and check if the information is up to date, if it is not then we will contact the CRIS API for the data we are looking for. After receiving the data, we update the new data in our local database for later use.

The next step is to retrieve the OpenAlex information on the paper. Once again, we check if we already have it in our local db, if not, we query OpenAlex API for their information about this article and store/update the local db.

After retrieving the list of citing publications from OpenAlex, we need the corresponding WoS, Scopus and GS 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. The results gathered from each database are also stored in our local db.

While the information we stored from our CRIS in the local db is more likely to remain the same over an extended period of time, since the publication title, year, or the links to the other databases rarely change, information gathered about the citation lists from OpenAlex, WoS, Scopus and GS receive newer information more often. Therefore, the "lifetime" of the latter information should be reduced when compared to the one from the CRIS, since if a new citation is indexed in one of those databases but we do not extract it and instead use older data on our local db, then it might stop us from finding missing citations in other databases, or we might incorrectly tag this citation as a missing citation.

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

With the list of citing publications from the 3 databases, we can look for possible missing citations in WoS, Scopus and GS. For this we use web scraping to query each database for any possible missing citations. This information is the only data from an external source that we do not store locally, that is because this is the last information we gather before building the report to our client, and if we used this data to build a report, it would result in the same report as the one already stored in our local db.

After detecting all the missing citations in each of the databases, we can create the report we are going to send to the client, and this new report is stored in the local db. We do not update the previous report, in order to be able to see the evolution of the reports along the time. This way we can see for a given publication, if the databases are correcting missing citations, if new citations or missing citations are being added, and the evolution over time of a specific publication.

Through all this process, one of our main goals is not to overload the APIs or repeat processes unnecessarily. Therefore, besides using our local database as a cache, we use delays between every API request or web scraping action to avoid overloading the systems.

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 2 of the databases, because, if we cannot compare the citations list of 2 different databases, no comparison data can be gathered, and no report is created. Also, if the publication has no citations in any database, there is no citations to look for, and no report is created. A diagram with the representation of our system's design can be seen in Figure 3.7.

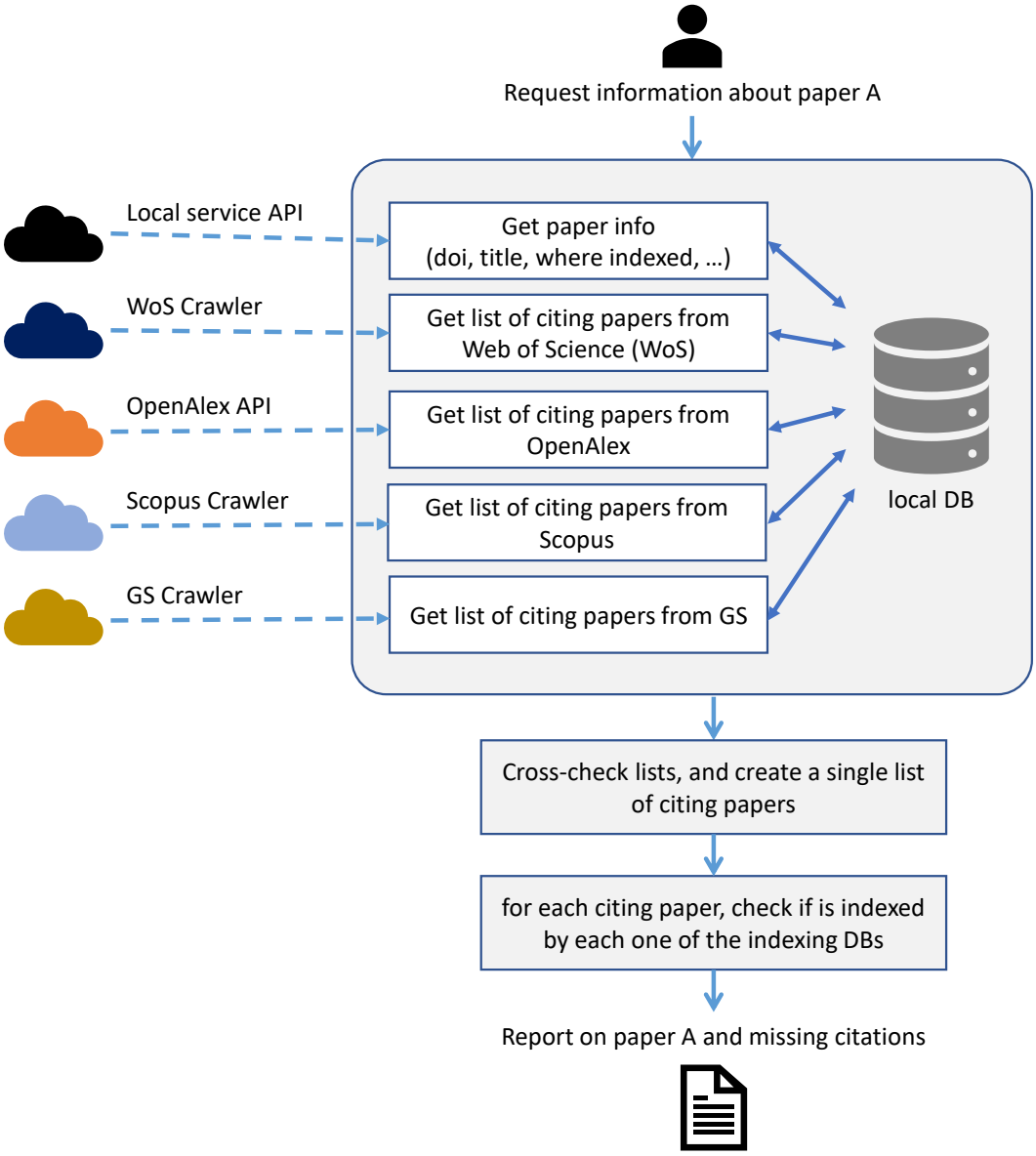


FIGURE 3.7. Overall System Architecture.

CHAPTER 4

First Prototype

In order to first test our approach and see what possible difficulties and problems we would have, we decided to only web scrape Web of Science, therefore we only searched for missing citations in the WoS database by using the OpenAlex database as a baseline.

To first populate our local database, we wanted all the publications we could use to test our approach, mainly their DOIs, so that we could automatically test our approach, instead of manually providing different DOIs. We got this information from the Ciência-IUL API, but there was no direct way to request all of publications they had indexed, therefore we first had to request the API for all the authors, by using the following request: https://ciencia.iscte-iul.pt/api/v2_6/authors/all

After having all the authors in the database, we could then ask for all of their publications by using the author ID in the following request where `author_id` is the id we got from the database: https://ciencia.iscte-iul.pt/api/v2_6/authors/<author_id>/publications/all

After running the previous request for all authors, we had all the publications indexed in the Ciência-IUL database stored in our own database and by looking at the DOIs of the publications we were able to see a lot of different formats. As previously said, some DOIs are wrong, with DOIs like '-', '—', '00', 'colonia', 'desconhecido'. Other DOIs start with tags like 'https://' or 'doi.org/', some start with '/' and others without it, we also have cases of DOIs ending with punctuation or having random blank spaces. In order to have a more even format, the first thing our program does when it receives a DOI is standardize it, leaving only the DOI and not the DOI link or extra punctuation or blank spaces. This way we ensure we are always working with a DOI formatted as '0000/000000000000' and if/when we need it, we can add it to the link 'https://doi.org/'.

4.1. Ciência-IUL and OpenAlex

For both Ciência-IUL and OpenAlex we could get the information via an API request, and for both of them, all we needed was the publication DOI.

In order to request the information on a specific publication in Ciência-IUL we send the following request to the API where `publication_doi` is the doi of the publication we are looking for: https://ciencia.iscte-iul.pt//api/v2_6/publications/by-identifier?id_type=doi&id_value=<publication_doi>

Because Ciência-IUL does not standardize the DOIs stored in it, it could lead us to not finding many of the articles we are searching for, but because the DOI we are providing to our program came directly from Ciência-IUL this is not a problem and in this query we

use the DOI we received as an input without being treated. From this request we receive a lot of information about the paper we are going to analyze. We get some information about authors of the publication, what type of publication it is, the year it was published, the title and other information about the publication itself and the source where it was published. The most important information we get is whether or not the publication is indexed in either Web of Science, Scopus or Google Scholar, and if it is, we get a link for the citation list on the respective database. This allows us to go directly to the web page of this publication, without having to query the web page for it, and without mistakes.

After getting the information from Ciência-IUL, we request the OpenAlex API for information on the publication by using the following query: `https://api.openalex.org/works/https://doi.org/<publication_doi>`

This time around we already use the clean version of the DOI that we received. This way we do not have to verify for each DOI if it has 'https://doi.org/' or just part of it in order to build the request, and we can treat them all the same. From this request we also get a lot of information about the publication and the publishing source, that we can double check with the information from the Ciência-IUL API like the title, the year, what type of publication it is, language, etc. But the most important information we get from it, is the `cited_by_api_url`, which is a request link for all of the citing publications on the OpenAlex database. By requesting this link to the OpenAlex API, we get a list with all the data about each citing paper, their DOIs, titles, publication year, type, etc. The API only sends responses with a maximum of 25 papers, and if the publication has over 25 citations, we need to send a new request, but this time we have to add to the URL `'&page=<number>'`, where `<number>` is the page number we want from the citing list. After extracting the information from all the pages, we have a comprehensive list of citing publications that we can use to search for missing citations in the other databases.

Because during this approach, we will gather a lot of information from different sources, and sometimes from the same publication, that we need to cross reference, we decided to create a standard dictionary to represent a publication, since from some each source the keyword can vary, for example title can be 'title', 'Title' or 'publication_title'. By having these standard representations of the publications, it facilitates the comparison between each publication and allows us to more easily work and expand with new databases. Figure 4.1 shows this standard representation format.

The `db_unique_identifier` enables us to present at the end, exactly what publication we found in each database, whereas if we did not store it, then the client would have to go to the database, and query it via the DOI or title, which might present multiple answers. By providing the unique identifier, which in Ciência-IUL is a number, or in OpenAlex is a 'W' followed by a number, it allows us to unequivocally find the same publication that our system found. The `citing` Boolean allows us to see whether a publication that has been found is already tagged as citing or if it is a missing citation in the different databases.

```
{
    'doi': '<The DOI of the publication>',
    'pub_year': '<The year the publication was published>',
    'pub_title': '<The publication title>',
    'pub_type': '<The publication type (eg. Journal paper,
                 Conference paper, etc)>',
    'db_unique_identifier': '<The unique identifier for
                             this publication in the
                             respective database>',
    'source': '<From which database this information
               was extracted from>',
    'citing': '<Whether this publications is a
               citation or not>',
}
```

FIGURE 4.1. Representation of a publication in a standard format.

4.2. Web of Science

Accessing the WoS database differs from the previous two databases in terms of simplicity. WoS provides only the number of citations and the website link to the citations list, which are the same that we received from Ciência-IUL. To extract more detailed and specific information from WoS, we must rely on web scraping techniques.

Web scraping is a process of extracting data from a web page by accessing the HTML code of the page. In our case, we have specific pages we are looking at, and already got the web link to such pages, making it a web scraping job. On the other hand, web crawling, tries to find or discover links to browse web pages, and then extract the information that they have.

4.2.1. Web Scraping

In this project we used the Selenium framework for python in order to web scrape the databases that we need to retrieve data from. This framework provides a mode that allows the user to see the progress of the web scraping and the changes in the page, by clicking in buttons or navigating to other links, but also has a headless mode, which does the web scraping without having to open the browser, consuming less resources, but still being linked to a browser and user agent.

In order to access the elements of the web pages we were looking at, we used selenium framework function `find_elements(<element_selector>)` where the value of the variable `element_selector` is the HTML selector from the element we want to interact with. We could then use functions like `click()` to click on a button, or `text` in order to get the text from a label, button or text box like it in Figure 4.2. All selectors from the elements we interact with are saved in a configuration file, in order to be easily accessible and changed if any changes in the web pages HTML source code happens.

```
element = find_elements(By.CSS_SELECTOR, <element_selector>)
element.click()
label = element.text
```

FIGURE 4.2. Selenium Framework Element Functions

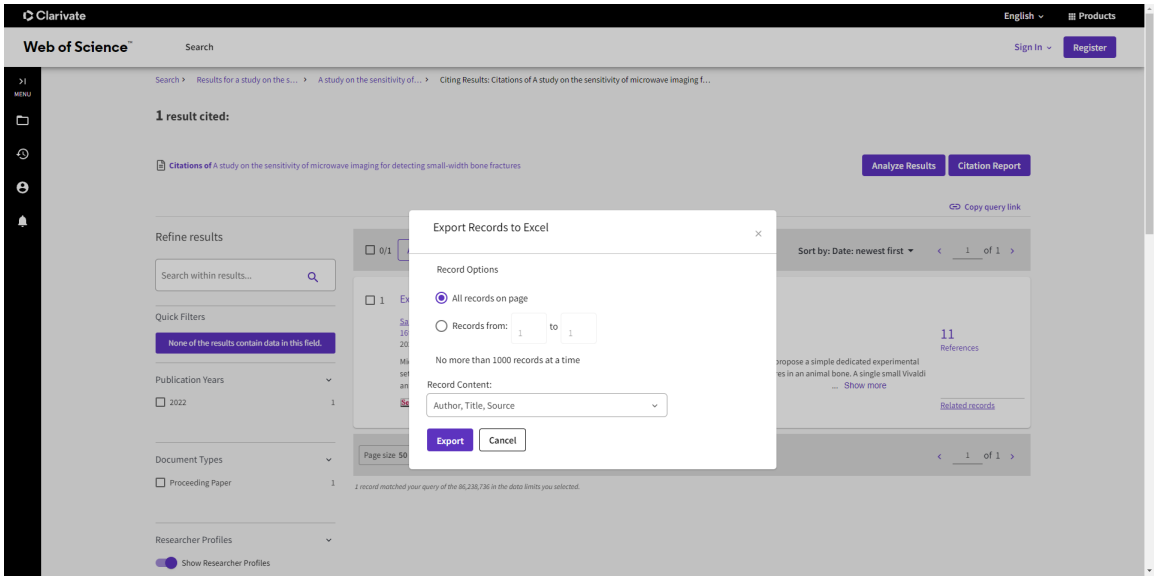


FIGURE 4.3. WoS web page screenshot of the citing articles excel extracting page collected at 10/10/23

During the web scraping portion of the code, we make sure to use delays between every request as to not overload the systems we are communicating with as well as saving all the information gathered in our local database.

In order to web scrape the WoS citation list website, we noticed that adding "(overlay:export/exc)" to the end of the link, meant that instead of being sent to a web page like we have shown before on Figure 1.2, we would be directly sent to a page to download the list of citing papers to an excel file like is shown on 4.3. From here we only have to simulate the click on the export button in order to download the excel file with all the data. This excel has a total of 72 data fields about each of the citing papers, from which the most important for our approach are the 'DOI', 'Publication Type', 'Article Title'. These citing papers are all standardized like shown before, and in the case of the Figure 1.2 paper, the stored data would be a list with just one citing paper which would look like Figure 4.4.

4.2.2. Main Difficulties

Web scraping does not come without its problems, being at the mercy of bugs on the web pages we are searching for data at or changes they can make. During the development of this project the website was down multiple times, either the query web page, the advanced queries, or even the citation lists. Problems like this bring a halt to this process and stop us from continuing web scraping the web page at times.

```
{
    'doi': '10.23919/EuCAP51087.2021.9411065',
    'pub_year': '2021',
    'pub_title' = 'A study on the sensitivity of
                   microwave imaging for detecting
                   small-width bone fractures',
    'pub_type' = 'Conference Article',
    'db_unique_identifier' = 'WOS:000672699800177',
    'source' = 'WoS',
    'citing' = 'True',
}
```

FIGURE 4.4. Selenium Framework Element Functions

One instance of the changes they did during the development of our project was the cookies button, which changed place and name. When this happens, we have to go to the configuration file and update the selector to the new selector of this button.

Another problem that we found in WoS is the lack of consistency in the database relating to the amount of citations of a paper. There were cases where the database said that there were a certain amount of citations to a specific paper but in reality, there were fewer. By accessing the list of citations of a paper, we would come across with less citations than what the database told us. In the most extreme cases they provided us with the number of citations and links to the citation list web page, but in reality, there were no citations. This caused our approach to crash, because the program was expecting it to be a web page to scrap, but it was non-existent, so we had to take measures in order to double check if there were really the citations that WoS told us there were.

In other cases, although the paper should be indexed in WoS, and the API gave us a link to it, we could not find it, which meant that the publication was not indexed in WoS anymore.

This posed a challenge because in either occasion WoS just redirected us to a web page saying "An unknown error has occurred.", which meant that we can not just disregard any message of error, because there should be different approaches to them but we have no way to figure out what situation we are in. If the database does not have the paper indexed in it anymore, we should not look for missing citations in the database. If the paper does not have citations although WoS says it does, then we just tag every citation we find in WoS as a missing citation, finally if the database was down, and did not allow us to consult the data we were looking for, then we should stop the creation of the report, because there will be important data missing.

4.2.3. Detecting Missing Citations

After extracting the list of citing publications from both databases, we need to figure out which citations could potentially be missing from WoS. But because our approach has the objective of working with multiple databases, we needed a scalable solution for

the future. Therefore, we decided to merge all the citing publications from all databases in a list. Like it has been previously mentioned, each publication representation is standardized, and it has a reference to whether it is a citation or not, and the source of this publication.

After having all the publications in a single list, we try to find which ones are actually the same, and group those together. In this case we would expect that each group would at most have 2 publications, one from OpenAlex, and another from WoS, but in reality, the groups can have more citations. As it has been mentioned in Chapter 2, in the research of García-Pérez in [1], they found that some databases have duplicate citations. Our hope is that by grouping the publications this way, duplicate citations will also be caught and grouped together.

In order to figure out whether or not two publications are the same, we have two different approaches. The first one is to check if they both have the same DOI, and if they do, we automatically say they are the same publication. If the DOIs are not the same, we check two more parameters; if they have the same title, and if they were published the same year. The reason we do not rely solely on the DOI, is because sometimes, databases have the wrong DOI for a publication. If this happens, we can still match it by using the title and date of the publication. Furthermore, even if a publication has a DOI, that information is not always in all databases, and if a publication does not have a DOI, we are obliged to use other data in order to find a match.

After matching all the publications and grouping them together, we need to see in what groups there are potential missing citations. For this, we go to each group, and for all citations in a group, we check what databases they are from. If in a group we do not detect a citation from a specific database, then there is a potential missing citation in that specific database. If in a specific group, there are citations from all the databases, then no further work is needed. If the only citation in a group is from WoS, we also do nothing, because we are not looking for missing citations in OpenAlex, but if the only citation in a group is from OpenAlex, then it means that we did not find a matching citation in the WoS citations list. If this citation was not found by WoS, then we have to check WoS, in order to see if this publication is indexed in the database, so we can know if it is a missing citation or not.

As more databases are added to our approach, the process remains the same. If there is a group that does not have a citation from a database where we are looking for missing citation, then there is a potential missing citation that needs to be checked for the missing database of the group.

Now that we know every potential missing citation, we must confirm if this publication is or not indexed in WoS and in order to do this, we must query the database. We start by building the queries, and then we web scrape WoS for each query and analyze the results. At the moment we are building two different queries for each publication; the first query is built by using only the publication DOI: "DO=(<DOI>)", where `<DOI>` is the DOI

of the publication we are searching. The second query is built using both the publication year and title: "TI=(<publication_title>) AND PY=(<publication_year>)", where `<publication_title` and `publication_year` are the title and year of the publication we are searching for, respectively.

Having the queries we want to do to the WoS database, we access the advanced query web page via the link like shown on Figure 1.7. When accessing this page the first thing we must do is to find and click the clear button, because WoS remembers previous queries and if do we not clear the text field, then we will be adding the new query into the previous query. After clearing the text field we can then find the text field element and use the selenium function `send_keys(<Query>)` where `<Query>` is the query we have prepared, to write the query on the web page. Once the query is written we just click on the button `Search` to get the results page.

If no results are found, we will get the message on Figure 1.7, and the web browser will continue to be on the advanced query web page. Therefore we can just check the current web browser URL to know if any results were found, and if not, we can just repeat the steps above to perform a new query. If results were found by WoS, then the results page will behave the same as the list of citing publications, which means that in order to extract the results from the queries, we just need to do the same steps as we have shown in Section 4.2 to extract the citing publications. We add "(overlay:export/exc)" to the end of the URL, and extract the excel file with the results.

After getting the results of the queries from each group of publications, we compare each publication in the results with the respective group, in order to make sure that they are indeed the same publications. We do this by once again matching the DOIs of both publications, or the title and year of the publication. If a match is found, it means that the publication is indexed in WoS, but the citation was not detected, so we add this publication to their respective group. It is added in the standardized form, with the publication information and the `Citing` parameter set to False, so that we know that this publication exists in the database, but it is not citing the paper we are analyzing.

4.3. Building the Report

Now that we have searched for every possible missing citation in WoS, we must build a report with the information we have gathered. The first thing to do is to check in which databases the publication is indexed. In this case, because we are only using 2 databases, this check is redundant, because if the publication was not indexed in both databases, there would be no way to compare 2 different lists of citing publications. Nevertheless, because this approach is built to prepare for the use of multiple databases, this process needs to be done.

In order to know in which databases the publication we are analyzing is indexed, we consult the data gathered from the Ciência-IUL database, and by searching for the publication in OpenAlex as referenced in Section 4.1. After having the list with all the databases where the publication is indexed, we need to access each group of citing

Paper A Analysis		
Citing Articles	WoS	OpenAlex
1	✓	✓
2	Missing	✓
3	Not Indexed	✓
4	✓	✓
5	Not Indexed	✓
6	✓	Not Indexed
Recorded Citations	3	5
Expected Citations	4	5
Missing Citations	1	0
Missing Percentage	25%	0%

FIGURE 4.5. Example of a report created for WoS and OpenAlex for a generic publication

publications we have created in the previous step. Each of these groups is composed of the same citing publications but from different sources which are represented by the parameter `Source`, and also if they are citations that were found by the database or not, represented in the Boolean `Citing`.

For each database where the publication is indexed, we must see if there is, in each group, a citing publication from that database. If we find a citing publication from a database that is indexed, and the Boolean `Citing` is `True`, then it is a citation that the database had already found. If the Boolean is `False`, then it is a missing citation in that database, because we found that this publication was indexed in the database. If there is no publication from this database, then the citing publication is not indexed in the database.

In Figure 4.5 we can see an example of the report that is built, each line represents a different citation, and at the same time, a different group of citing publications that we built in our code. For the first line, the group had both a publication from WoS and OpenAlex with `Citing=True`. While on the second line, there was a group, where once again there were publications from both sources, but for WoS `Citing=False` and for OpenAlex `Citing=True`, which means that WoS did not find this citation, but the publication is indexed in it. The last case is represented by line 3, where only a publication from OpenAlex was present in the group, meaning that this publication was not indexed in WoS.

4.4. Experimental Results

In order to test this first stage of our approach we decided to run an experiment. Like it was mentioned in Chapter 3.1, the process of looking for missing citations in a database

TABLE 4.1. Results per year

	Year		
	2015	2018	2021
Papers analysed	820	950	1304
Papers that met the criteria	446	497	596
Total OpenAlex Citations	10490	10653	4114
Total WoS Citations	8760	9277	3411
Papers with MC	128	140	130
Total MC	233	265	198
Average percentage MC	5,30%	5,10%	9,60%
Average percentage MC in papers with MC	18,50%	18,30%	44,17%
Maximum MC	12	25	25

is time consuming, in particular, the web scraping process. Therefore the experiments were only run on papers from 3 different years, 2015, 2018 and 2021.

As mentioned before, our goal was to find WoS missing citations, using only OpenAlex as a baseline. This means that only papers that are indexed in both these databases are going to be analyzed, and that all missing citations found in WoS come from OpenAlex.

In the Table 4.1 we have a summary of the results for each year of this experiment.

In total, 3 074 papers were analyzed by our system, where 820, 950 and 1 304 are from the year 2015, 2018 and 2021, respectively. For each of the years, only 446, 497 and 596 (respectively) reports were created, giving us a total of 1 539 reports of different papers, which represent half of the papers from these 3 years.

We can see that the values are increasing over time, not only in the number of papers, but also in the number of reports created, especially in 2021, where there was an increase in over 350 papers and almost 100 more reports when compared to 2018, whereas if we compare 2018 with 2015, the increase was only of 120 papers and 51 reports. While the number of reports did not increase as much as the number of papers, that is probably due to the fact that since 2021 papers are much more recent and that being the case, some of them are not indexed in both databases yet or they do not have any citations. We expected that if we did the same analysis one year from now, the data from 2021 would be the one with the most changes.

All the data we are presenting next comes from the papers where a report was created. The total number of citations in WoS for the years 2015, 2018 and 2021 were 8 760, 9 277 and 3 411 respectively. The huge drop in the number of citations in 2021, comes from what has been said before, which is that since the papers have had less time to be cited, the number of citations is going to be lower. The low number of citations, when compared to the other years, also greatly influences the percentages that are presented in Table 4.1.

In 2015, we found missing citations in 128 out of the 446 papers with a total of 233 Missing Citations (MC). This means that 28,70% of the papers had MC, averaging 1,82 per paper. The numbers in 2018 are very similar, with 140 papers out of 497 having MC

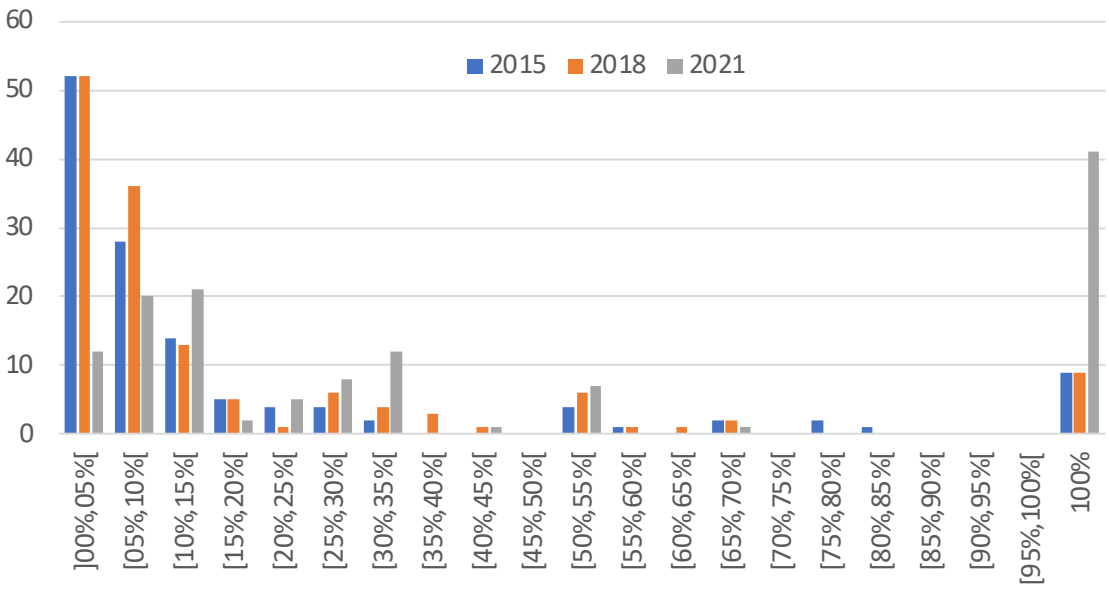


FIGURE 4.6. Distribution of papers according to missing citations percentage per year

(28,17%) and a total of 265 MC, with 1,89 per paper, a slight decrease in the number of papers with MC but a slight increase of MC per paper. Once again the year of 2021 stands out with only 130 papers out of 596 with MC (21,81%) and with 198 MC, with an average of 1,52 per paper. In total, there were 398 papers where we found 696 missing citations.

$$Average\_percentage\_of\_MC = MC/(MC + Citations)$$

In order to calculate the average percentage of MC we chose to use the formula presented above, which will tell us, out of all the citations that a paper should have (the ones the database caught and the ones it missed), how many are missing, giving us results from 0% to 100%. We also counted the average of MC for all the papers, and only for the papers that had MC. Out of all the papers, the percentage of missing citations both in 2015 and 2018 were around the 5%, while in 2021 it was 9,6%. If we look only for the papers with MC in their reports, in 2015 and 2018 we had 18,5% and 18,3% respectively, while in 2021 we got an astounding 44,17%. In 2021 the number shoots up, especially because of the low number of citations in 2021. Once again we argue that these numbers in 2021 are higher because the papers did not have as much time to be cited yet, and the papers that already cited them are still being worked on by WoS, and because the papers are recent, the WoS process is still behind in referencing the citations on the papers.

In Figure 4.6 we present the distribution of papers according to the percentage of missing citations. Although the average number of MC per paper is low (between 1,5 to 1,8), there were some cases where a high number MC were identified as it can be seen in Figure 4.7. In 2015 a paper had 12 MC, and in 2018 and 2021 both had a paper each with 25 MC each. Only 8% of the analyzed papers had over 3 MC, and 7 papers had 6

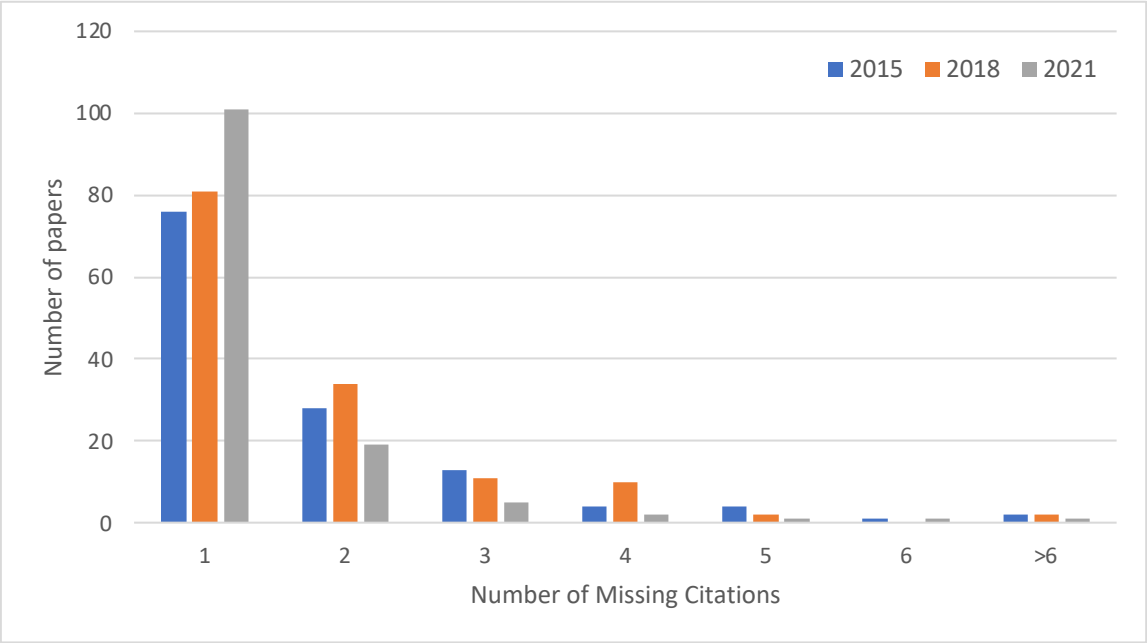


FIGURE 4.7. Distribution of papers according with the amount of missing citations per year

or more MC, showing us that most of them have 3 or less MC. This is also correlated to the fact that there are more papers with less citations and it's more likely to find more missing citations in the midst of a paper with a lot of citations than in a paper without. In most cases, the papers with 100% of missing citations, have either 1 or 2 citations and all of them were missing citations, while papers with more citations can find more missing citations, but they represent a lower percentage in the grand scheme of things.

4.5. Validation of the Reports

In order to have more trustworthy results, we manually checked a sample of the reports that we created. Part of that sample was retrieved at random, but another part was retrieved from reports that stand out because of unusual or unexpected results. We manually checked all reports that had over 3 MC and where those 3 MC represented at least half of the citations of the paper. This is because if a paper had 4 MC out of 100, then it was quite plausible that WoS might be missing a few citations, while if there were 4 MC out of 5 total citations, then something might be wrong, since it means that WoS did not find most of the citations that it should have.

The manual verification followed each of these steps for any given report:

- We had the DOI of the paper that was initially provided to our system, as well as the OpenAlex and WoS web links for this paper. Both were double checked to make sure they related to the same paper.
- For each missing citation identified by the system, we checked the OpenAlex and WoS web links for the citing paper that our system matched, in order to make sure that the OpenAlex publication was the same we found in WoS.

- If the original paper being analyzed was the same in both databases, and all the missing citations were correctly matched with a paper indexed in WoS, then it was a missing citation.

We manually verified 40 random papers, from which 36 had correctly identified their 51 MC, while the other 4 reports had some errors.

In one of these reports with errors, out of 7 MC found, 2 of them were wrongly matched, meaning that the OpenAlex citing publication was not the same as the WoS publication we found through our queries. Upon closer inspection we found that the problem was that these 2 publications had the same DOIs in the different databases, and after a search for the DOI in google, it looked like WoS had the publication with the wrong DOI.

We also found 2 cases, where one of the missing citations was due to OpenAlex saying that the publication was citing itself, and since the publication was indexed in WoS, it was tagged as a missing citation. Therefore we have a total of 2 wrong MC in these 2 reports.

Finally the remaining report was wrong not due to the fact that it incorrectly matched citing papers, but because of an error in the data received by the Ciência-IUL database, where the DOI given to the publication was a DOI of a different publication to the one where the link to the WoS web page of citing papers take us, which led to the comparison of two different articles, in WoS and OpenAlex.

In this manual search, from the 40 analyzed papers, only 10% had some kind of mistake, one of the cases was due to WoS having assigned the wrong DOI to a publication, and the other one was due to the fact that we tagged the publication as citing itself. The latter is an easy fix we can do in our approach, by double checking if a citation does not have the same DOI as the citing paper itself, while the former is a mistake that we cannot work around that easily, since it is a problem in WoS side that we cannot avoid. One way to not have this error, is if we made it so more information besides the DOI had to be right, like the title or the year of the publication. We do not do it like that because we believe that a publication having the wrong DOI is a rare occurrence, and the DOI is the most trustworthy data we have. Finally the mistake from wrong data provided by Ciência-IUL is out of our reach to fix, since the information in this database is provided mostly by the authors of the publications and it is prone to have some errors.

Afterwards, we validated the reports that showed over 3 MC which represent at least 50% of the citations of the report itself. Out of 14 total reports, 13 of them had no errors, correctly identifying 82 MC. The remaining report was wrong due to the fact that the publication we got from OpenAlex, was different from the one in WoS, making our system compare citation lists from two different papers. Therefore this report incorrectly gave us 25 MC. This error came once again from the Ciência-IUL database, which had 2 different DOIs for the same publication, and the DOI being analyzed did not match the one we got from it. In this case, the 2 different DOIs, were provided one by the author and the other one automatically from ORCID (Open Researcher and Contributor ID).

4.6. First Prototype Insights

In general, we are satisfied with the system's performance as it has only encountered a few errors, primarily caused by inaccurate data and not a fault in our approach. Although we can address some of these errors, there are others that are beyond the scope of our approach. As a result, there is a slight chance of miss-classifying a Missing Citation as such. To mitigate the impact of incorrectly classified missing citations, our reports can include links to both databases for any matched papers, allowing for manual verification. This process significantly reduces the effort required compared to starting from scratch, thereby saving researchers a considerable amount of time.

As for wrong data in the Ciência-IUL we mainly came across with invalid DOIs, cases where more than one DOI was provided, or DOIs that did not match the publication, specially in the cases of book chapters, where the DOI was from the book and not the chapter itself, which meant that if we tried to ask OpenAlex, it would provide the list of citing articles of the book and not the chapter.

Because there are mistakes with publications DOIs we also wished to improve the matching of publication, by using extra data like the authors of a paper. But this would have to be implemented carefully, because there are a lot of nuances in these types of data, and we do not want to add False Positives (False missing citations). Using similarity techniques could also be advantageous, both for the Author names, or the title of the publications, in case there were some misspelled words, but once again this would have to be carefully tested.

During the elaboration of this prototype, and running the program for these 3 000 papers, multiple adjustments have been done in order for the program to run more smoothly, specially in the web scraping portion of the job. One of the biggest problems were random errors and exceptions, like the web page crashing, not loading, or being unable to locate an element in the page. These are not common errors, but happen sometimes, and if they happened in the middle of analyzing a report with hundreds of citations, an hour or more time of processing could be lost. Because of this, we also wanted to add a try again option, where if we catch any of these odd exceptions, we would catch the exception, and try again, for a limited number of times so as to not enter in an infinite loop. This way, if while querying a web page a crash happened, the program would once again open the web page, and query it again, and continue from where it left off, instead of crashing down and losing all progress. The delays between each request were also fine tuned in order to not overload the system, but also not expand the time of execution of the program by copious amounts.

Also multiple changes were done to our local database, in order to improve the performance, and to store interesting information. For example, we stopped storing the list of references or related works from OpenAlex, since in some cases there were a lot of them and that took a lot of space. We also had to increase the limits of the sizes of the information that is kept, since if a paper has 2 authors, it is different from one with 50.

The more important information was stored as an individual data entry, while others were JSON (JavaScript Object Notation) files.

Finally, the fact that some of the data collected comes from web scraping web pages, and these are constantly evolving, means that there is also a need to periodically check if there are changes in the web sites, and the impact that it can have on our approach. While in the beginning of this project it was only possible to extract sets of 25 papers at a time from WoS, and navigate throughout each page to extract the whole list, now it is possible to extract all the citations at once. This allows us to get the data faster from a single download, also avoiding multiple requests to the web site.

CHAPTER 5

Inclusion of Scopus and Google Scholar

The next step was to include extra databases, this not only would allow us to find new missing citations in WoS, but also look for missing citations in Scopus and Google Scholar. Because the first prototype was implemented already with the idea of expanding this approach for more databases, the processing of information from both these databases was already done. Therefore we only had to implement the web scraping of both these databases, and fix any bugs that came up after adding them to the program.

5.1. Scopus

When we started working on the web scraping of the Scopus web page, we completed the whole process and after debugging tests in our own browser for one or two web pages, we tried a small set of tests in headless mode for 10 or so different articles to see how it went in a more comprehensive scenario. We quickly started getting unexpected error messages. Because we were in headless mode instead of using our own browser we could not see the web page, so we used a debug tool of Selenium that takes screenshots of the browser so we could try to identify what was going on. What was happening was that at the time, Scopus was in the process of improving their web page, and they had two different web pages, sometimes we would be redirected to the old one, and asked if we wanted to go to the new one, and other times we were directly redirected to the new one. Because we were not dealing with the new web page, when we were sent there, the web scraping failed because all the elements were different and needed new selectors and different interactions. In the Figure 1.8 in Section 1.4 we can see the old page of Scopus and in Figure 5.1 we can see the new one.

We then decided to rewrite the web scraping portion of the Scopus web page, to deal with the new template, and if the link sent us to the old page, since there was an option to be redirected to the new one, we chose that option. We chose to work only with the new one, because we speculated that in the future it would be the only one available. As it is at the time of writing this thesis it seems like the old web page is no longer available, and we are directly sent to the new one.

In order to web scrape Scopus, we need to write two different sets of code, because unlike WoS, the results page of list of citations, and of a query is not the same, therefore it needs slightly different approaches.

To download the citation list which can be seen in the Figure 1.3 we first have to click the `All` button, followed by the `Export` Button. The `Download` button would need an extension to be installed in the browser. After clicking `Export` a pop-up window opens,

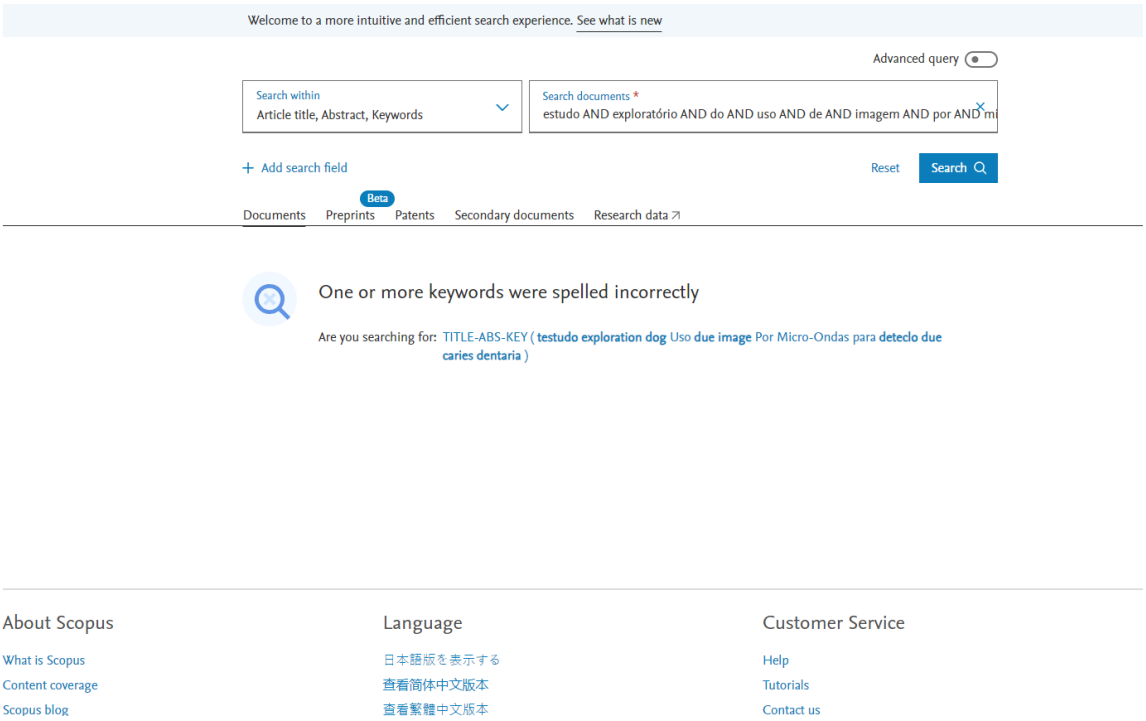


FIGURE 5.1. New web page from Scopus with the results of a query collected at 15/10/2023

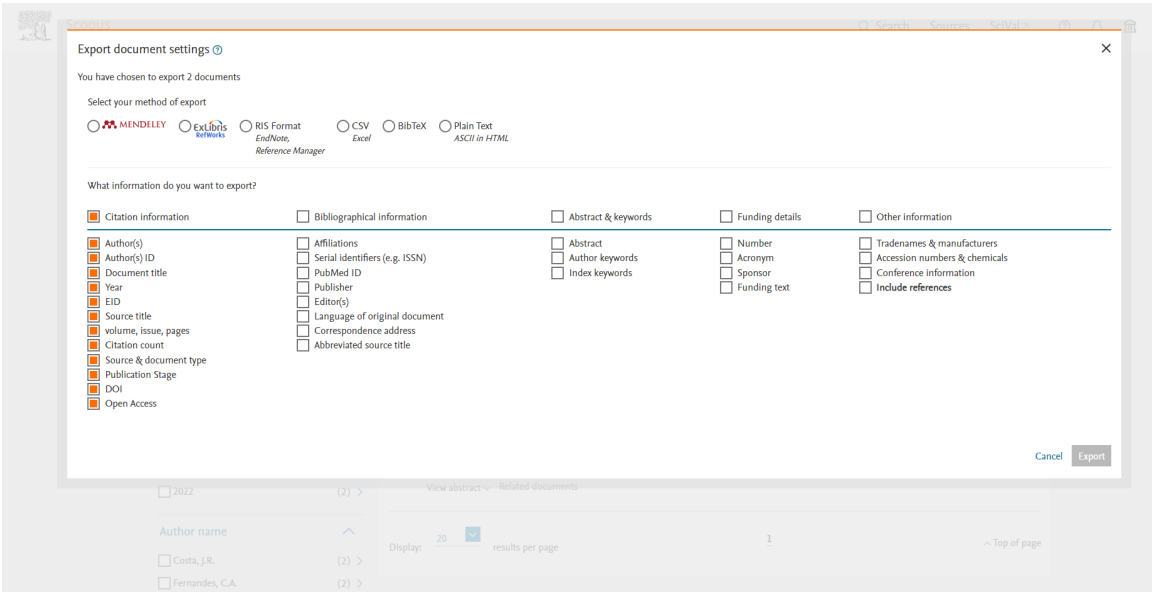


FIGURE 5.2. Pop-up window on Scopus after clicking export collected at 15/10/2023

asking what format and information we wish to download as can be seen in the Figure 5.2. After selecting CSV excel format we click export in order to download the file with the information of all the citing publications. This file is then read by our program, the publications are standardized as mentioned in Section 4.2, in order to be grouped with the citations from other publications and find potential missing citations.

As to querying Scopus for the possible missing citations, we build the same two queries as we did for WoS but with a different syntax; The DOI query: "DOI(<DOI>) where

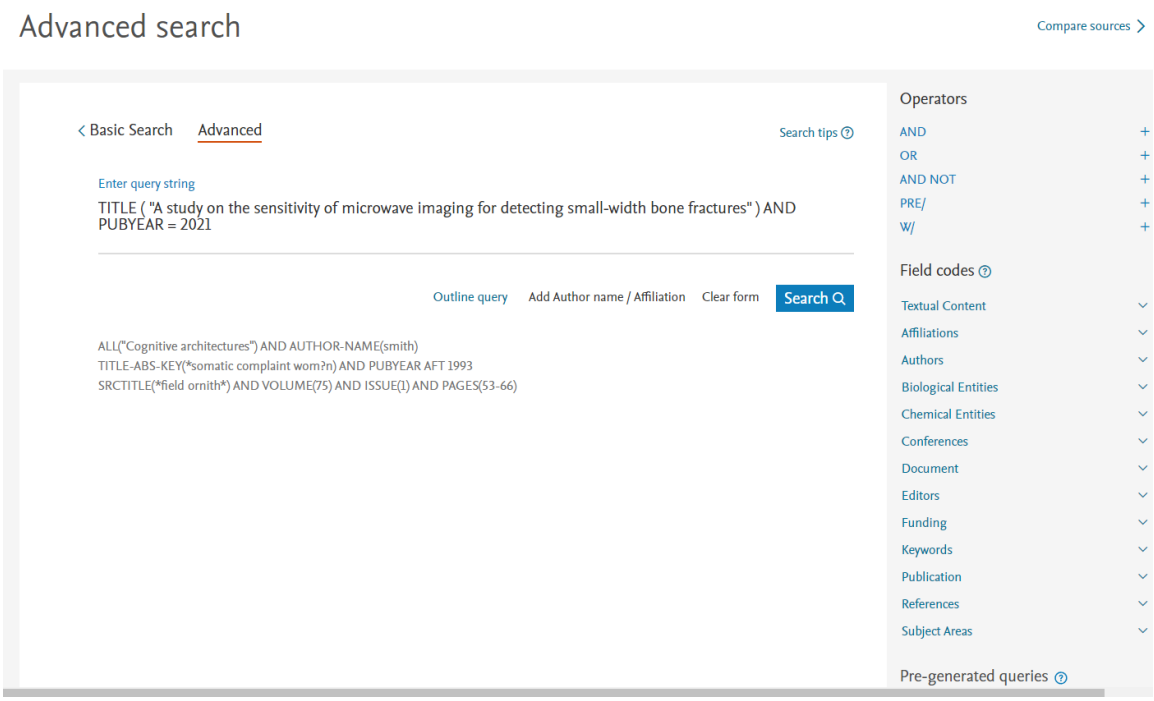


FIGURE 5.3. Scopus Advanced Query web page collected at 15/10/2023

`<DOI>` is the DOI of the publication we are looking for, and the year and title query: "TITLE("<Title>") AND PUBYEAR IS <YEAR>", where `<Title>` and `<YEAR>` are respectively the title and the year of the publication. After we access the Scopus Advanced Query web page shown in Figure 5.3, and click on the `Clear form` button if there is already a query written, next we use the Selenium `Send_keys(<Query>)` function with the query we wish to search for and click the `Search` button.

As it has been mentioned before the web scraping to download the query results is slightly different from the previous one. In this instance, after we click the `All` button, the `Export` button does not open a pop-up window but a dropdown box, where we have to then click the `CSV` option. Only after clicking the `CSV` option, a pop-up window will appear like the one on Figure 5.2, but without the top options of the format of the file, and only with the options for the information we want to download and the `Export` button option on the bottom right. The CSV file that is downloaded is the same as the one from the list of citing publications and once again we standardize the query results, compare them with the publication we were searching for, and if one is a match, then we add it to the group.

5.2. Scopus Difficulties

The biggest setback we faced with Scopus was because of the changes on the web site structure during and after the implementation of this code. First we had the two different sets of web sites, the old web site and the updated version, which made us rewrite the code for the new site. Secondly, there was a change to the HTML code while we were extracting data for our experiment. When running our program for the around 3 000

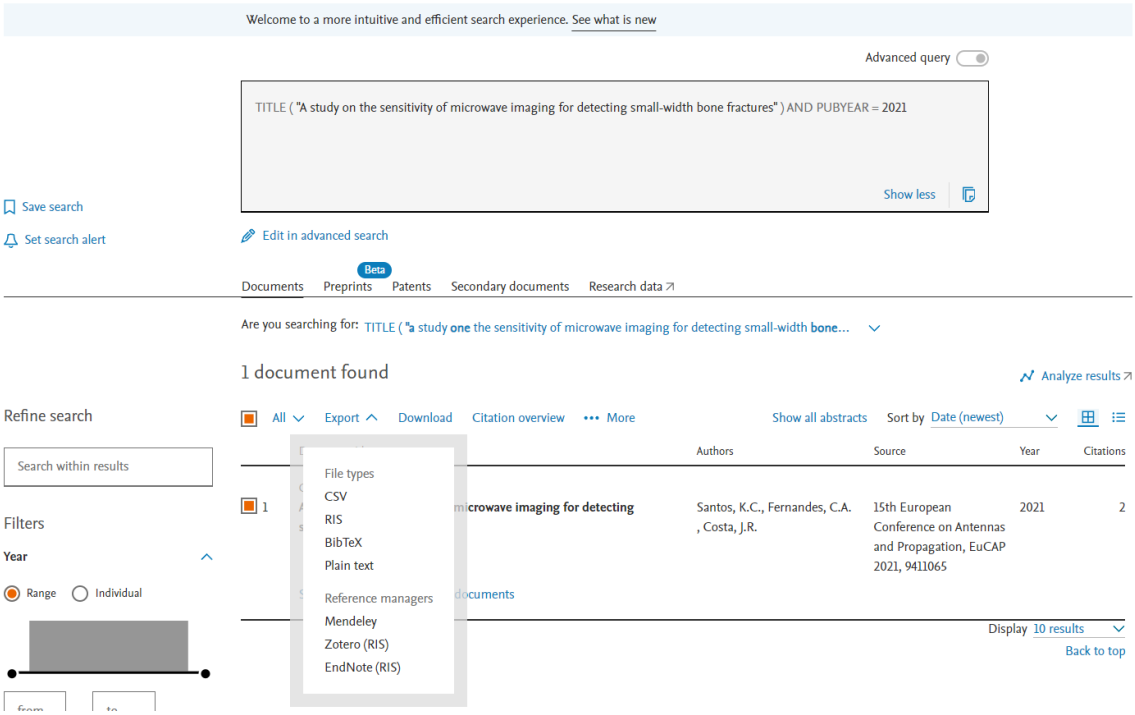


FIGURE 5.4. Scopus web page of the results of a query collected at 15/10/2023

papers to analyze the results, they changed the name of the button `ALL` in the advanced queries results page, which resulted in our program not being able to find any missing citations in Scopus for over two thirds of our dataset.

As a consequence of changing this button name, every time we were searching for a possible missing citation within Scopus, the web scraping did not download any file, and therefore the program concluded that the publication we were searching for was not indexed, when in reality it could be. Because this only happened after we started running our experiment, and we only checked a few of the first reports to make sure the process was working fine (which at the time it was), and did not check periodically during the running of this experiment, we only caught this problem once all the reports were created, and we started analyzing all the results. After fixing this problem, which was done by simply changing the button selector in the configuration file, we had to run this whole experiment again in order to gather the results and be able to write the Journal Paper we intend to submit to the Journal of Information Sciences.

The final problem we faced was how to find if there were no results in the query, because if there are no results for the query, we are sent to the results page nevertheless and are shown an error message. This error message element was always present in the page, but it was invisible if results were found. We ended up checking the sentence "<Number> documents found" as shown on Figure 5.4 where `<Number>` is a number where they tell us how many results were found, and if this number was bigger than 0, then there were results, if we could not find this sentence then no results were found.

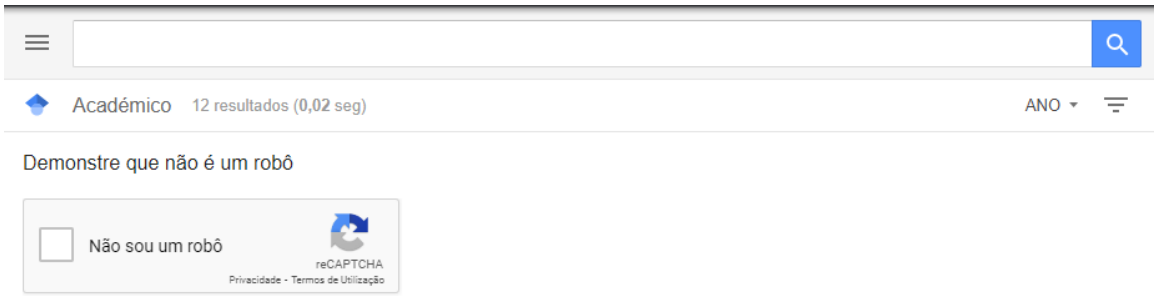


FIGURE 5.5. Google Scholar CAPTCHA collected at 15/06/2023

5.3. Google Scholar

Google Scholar was the database where we were expecting to get the biggest citations lists, and therefore, from where the biggest amount of missing citations in the other databases would come from. This is due to the fact that unlike WoS and Scopus, GS does not have a strict indexing policy, instead it uses web crawlers to index the scientific works they find.

As to web scrape GS, it was the biggest challenge, because unlike the other databases, there is no option to export all results it finds. In order to extract the list of citing publications in Google Scholar, we had to go one by one, to each article in the list, and retrieve their information individually. Another problem is that it has no DOI for any publication, so every citing publication we get, we are only able to work with the title and year, and to search for publications, we also only have that data available. Besides that, the extra information they give about each paper is less than the other databases, and somewhat less reliable according to previous studies. For example, using authors names in queries in GS usually provide bad results, as it has been presented in Chapter 2.

In Figure 1.4, we can see the list of citing publications of a given paper, and in order to extract each citation, we would click in each article in the list `Citar` (Cite) button, which opens a pop-up window, where we click on the `BibTeX` button, which opens another window with just the BibTeX code that we could then extract from the web page. After extracting the data, we go back to the citation list and repeat the process. Each page only shows 10 citations, so if we were extracting a citation list from an article with more than 10 citations, after extracting all the BibTeX from a page, we would have to click on the next page button, and repeat the process.

Because there were so many extra steps to web scrape GS, and it would be very time consuming, we decided to start by running some larger scale tests without searching for missing citations in GS, in order to see how long it would take. This was done because, at the time, we thought that GS would be less useful to try and find missing citations at, and also because they do not have information on articles DOIs, and the queries provide so many results, it would be the most prone to mistakes. While running these tests, we started having unknown errors, and after once again using the screenshot debugging option of Selenium to find them, we saw that our web scraper was being caught by the

TABLE 5.1. Results per year after the inclusion of Scopus

	Year		
	2015	2018	2021
Papers analysed	820	950	1 304
Papers that met the criteria	553	678	758
Total OpenAlex Citations	11 185	12 301	5 308
Total WoS Citations	8 942	9 835	4 118
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

GS anti-bot mechanisms, and they were presenting us with a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) challenge, as shown in Figure 5.5, which our program is unable to solve. Because of this we did not continue web scraping GS, and it was not used for any experiments.

5.4. Experimental Results

Once again we ran an experiment for all papers from 2015, 2018 and 2021 as we have done before. This way, the exact same set of papers were analyzed and we can compare the results with the previous experiment. A summary of the results can be seen in Table 5.1.

When we examined the total number of citations across all three years, Scopus has the highest count, totaling 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 to 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 the more recent work 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

TABLE 5.2. Missing citations averages over the years

	Web of Science			Scopus		
	2015	2018	2021	2015	2018	2021
Total Citations	8 942	9 835	4 118	11 189	12 272	5623
Total Missing Citations	278	390	407	404	349	384
Average percentage MC in all papers	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%

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 1 137, whereas in WoS, we found 1 075 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 to 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 to 2018 but -20 compared to 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:

$$AveragePercentageofMC = MC/(MC + Citations)$$

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 2 citations according to the database, and we find another citation that is missing, it has 33% of missing citations (1/(1+2)).

Analyzing these percentages shown in Table 5.2, WoS experienced a significant increase in missing citations in 2021 compared to 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 to 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.

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

TABLE 5.3. 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 in all papers	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%)		

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 3 analyzed years, we found a total of 2 212 missing citations in both databases combined in the 1 989 reports that were created. This averages 1.11 missing citations for each report.

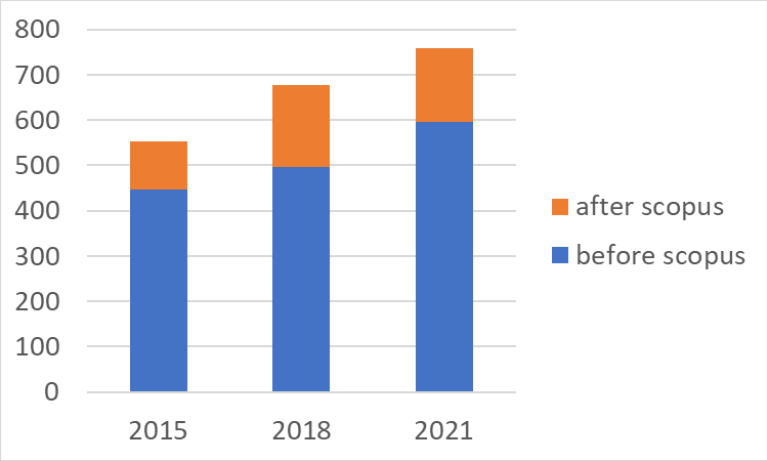
5.5. Comparative results on Web of Science

In section 4.4, we analyzed the results of a similar experiment, focusing on the search for missing citations within WoS while utilizing 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 cataloged in OpenAlex. Consequently, this newly introduced dataset from Scopus had the potential to reveal additional missing citations within WoS.

In this section, our objective is to assess the extent to which this second iteration has uncovered more missing citations within WoS when compared to our prior. The comparison of the previous of results from both studies are presented in Table 5.3

Before we proceed to the comparison of data, 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

FIGURE 5.6. Improvement on WoS on the amount of papers that met the criteria and created reports



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 analyzed it.

Although we analyzed the same number of publications in both studies, this time we generated more reports, increasing from 1 539 to 1 989, a 29% uptick. In Table 5.6 we can see the increase in the number of reports generated by year. 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 analyze it. For instance, if a publication is indexed in WoS but not in OpenAlex, we can still analyze 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. While some of these papers with missing citations may result from new data provided by OpenAlex that was not indexed four months ago, it is more likely that most of them originate from the data provided by Scopus alone. In the Table 5.7 we can see the growth per year of the amount of papers with missing citations.

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,

FIGURE 5.7. Improvement on WoS on the amount of papers with missing citations found

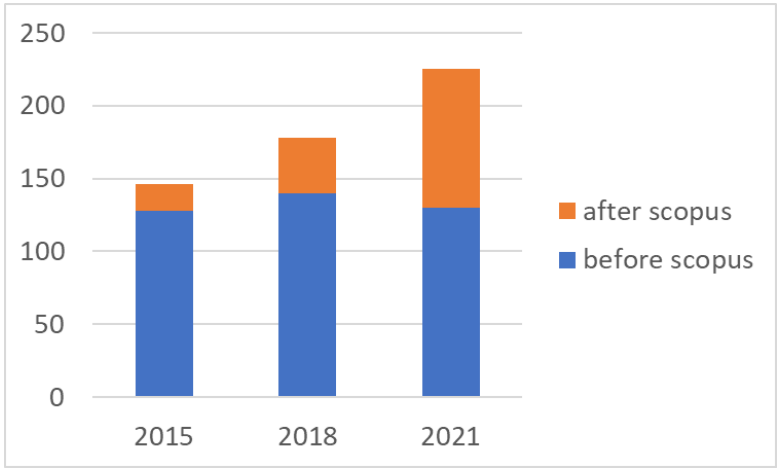
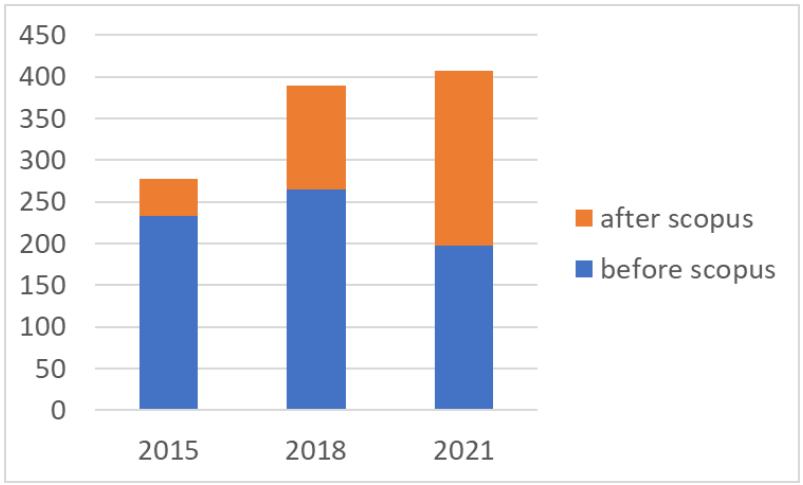


FIGURE 5.8. Improvement on WoS on the amount of missing citations



there were 1 075 missing citations, whereas without Scopus, there were only 696 missing citations, a 54% increase. Taking into account all the papers analyzed, we found 0.53 missing citations per paper that met the criteria with Scopus, compared to 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 71% 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

TABLE 5.4. Difference in the 2021 results on both WoS and Scopus within a 1 month time frame (Before - Experiment started at 21/08/2023, After - Experiment started at 23/09/2023

	Before	After
Papers that met the criteria	758	771 (+13)
Total OpenAlex Citations	5308	5718 (+410)
Total WoS Citations	4 118	4464 (+346)
Total Scopus Citations	5 623	6024 (+401)
WoS Missing Citations	407	315 (-92)
Scopus Missing Citations	384	413 (+29)
Total Missing Citations	791	728 (-63)

in 2021 we were able to more than double the amount of missing citations found as it is show in Table 5.8.

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.

5.6. Temporal Gap Experiment

In the previous section we tried to analyze the improvements that adding a database could bring to our approach, but there is a confounding variable which is the changes that the indexing databases do themselves over time. In this third experiment, we try to explore what changes over time in these indexing databases, without us influencing the results. This time we only analyzed the year of 2021, but it is the one that we expect to change the most, and due to time constraints. The temporal gap between the extraction of the data of the year 2021 from the last section, and the one that we are going to present here is off 1 month.

In Table 5.4 we can see the results of this experiment. Within 1 month time, we were able to generate reports for 13 more papers, meaning that before, those 13 papers were either not indexed in any database or were indexed only in one of them, and now they have been added to a total of at least 2 different databases. All databases had a growth in the amount of citations they found. OpenAlex was the one with the biggest increase, but Scopus is still the one with the most citations. As for WoS, although they have the smaller increase, it represents the bigger portion from all the databases, with an 8.4% increase. The biggest surprise factor was the WoS missing citations, since we found 92 less missing citations, which means that in a space of a month time, WoS managed to

correct almost 25% of their missing citations for 2021. Although we did not expect such a big jump in a month time, it is to be expected that WoS is the one with the most corrections, since as we have seen before, Scopus has the most constant value of missing citations, while WoS had almost double of the average of missing citations in 2021 when compared to previous years, and now slowly that number will decrease to catch up with the values shown in the previous years. On the other hand, in Scopus we managed to find 29 new missing citations that have been supplied by new citations indexed in WoS and OpenAlex. Overall, because of the improvement of WoS, we found fewer 63 missing citations for the papers of the year of 2021.

These results show that the databases are actively indexing more data, which helps explain why over time, it seems to be harder to find missing citations, since the databases have had more time to find what they were previously missing. Nevertheless, they are not able to do a perfect job, and even older papers still have differences in citations that should not exist between databases. Therefore the rate of corrections shown in this experiment in WoS is not expected to keep this rate.

It is also important to note that changes in the source list of indexed journals and conferences in the databases can create abrupt changes within the missing citations found, since some of these publications may no longer be indexed in the database. In march this year, WoS delisted multiple journals from their source list, due to failure to meet WoS' quality criteria. Although there is not an official list or number of removed sources, it is possible for actions like this to create data anomalies.

CHAPTER 6

Conclusions and Future Work

This thesis successfully explores the automation of a process to search for missing citations among multiple databases. We managed to retrieve data from OpenAlex, Web of Science (WoS) and Scopus, and search for missing citations in both WoS and Scopus. Unfortunately due to the fact that Google Scholar (GS) does not allow web scraping in the web site, we were not able to get the data indexed in this database.

Throughout this research work, we did multiple experiments to test our approach and the indexing databases we were analyzing, using as our dataset all publications from the years of 2015, 2018 and 2021 indexed in the Ciência-IUL database which totalled 3 071 different publications. We started with a smaller scale test, using only data from WoS and OpenAlex, where we uncovered a total of 696 Missing citations (MC) in the WoS indexing database. This answers our research question "Is it possible to automatically detect missing citations within different indexing databases?", since we were able to create an application to uncover these missing citations.

Afterwards, in order to answer the question "What is the impact of adding new indexing databases to the coverage of the process of finding missing citations?" we included Scopus in our approach and in order to see the improvements that we would be able to reach in WoS as well as to find MC in Scopus itself. By doing this we were able to uncover 1 075 MC in WoS, an increase of 54% and 1 137 MC in Scopus. It is noteworthy that on average, WoS had a higher average of missing citations in the most recent papers, since in 2021 there are fewer than half the amount of citations than in the other years, but we were able to find more MC than in 2015 and 2018. On the other hand, Scopus managed to have a more even spread of missing citations. It still has on average, more missing citations in 2021, but without it being such a big difference as it is in WoS.

After adding Scopus we also repeated the same experiment, with a 1 month gap for the papers of 2021. This allowed us to ascertain the changes over time in data collected. In this last experiment we managed to find an extra 29 MC for Scopus, but fewer 92 MC for WoS, which is representative of the fact that while Scopus does not have such a big difference within the different analyzed years, WoS has a lot of indexations missing in the year of 2021, and in order to reach the averages that it has presented in 2015 and 2018, more changes have to be done when compared to Scopus. Although this big improvement in one month time was clear for the year of 2021, the analysis done on the years of 2018 and 2015 leads us to believe that this will eventually stagnate, and we will continue to be able to find a good amount of missing citations on the indexing database.

We think that we proved that our approach worked, and the benefits of adding extra databases can have. Although we were not able to gather the data from Google Scholar, we expect that this would have been a huge benefit for our approach. If in the future adding GS can be achieved, or other indexing databases are added to our approach, we believe that the results will get more accurate and substantial. Nevertheless, the databases that are added must be able to provide new data, because if all the citations in the database that are added are already indexed across the databases that we are already working with, then there will not be any improvements. Therefore, before working on adding extra databases, a study on whether or not this new database has meaningful information can be of interest, in order to not waste time on the implementation of new code that will not provide new data, as well as to not slow down the process that is already being done. Although this is an automatic process, it still takes time to analyze each publication. The set of 3 074 papers we were analyzing in each experiment took around 1 month to go through our program, and adding extra databases will slow down this process, especially if it requires web scraping in order to extract the information we are looking for.

The results presented in this thesis (in particular Chapter 4) were used in an article [15] published at the SLATE'23 Symposium on languages, applications and technologies, and a second article, which encompasses the results shown at Chapter 5, has been submitted to the Journal of Information Science and is awaiting approval at the time of writing this thesis.

We believe there are some improvements that can be done to our work, when it comes to the matching algorithms and queries, these could be improved to use other available data like the authors or the source of the publications. Moreover, the inclusion of similarity metrics could help us catch misspelled words in the title and other data. But the use of these techniques also opens up the possibility of False Positive data, where our program could match 2 publications that do not represent the same article. Therefore, these would have to be carefully tested and evaluated in order to help minimize mismatches and ensure the accuracy of our findings.

Bibliography

[1] M. A. García-Pérez, "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," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 10, pp. 2070–2085, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21372

[2] H. F. Moed, J. Bar-Ilan, and G. Halevi, "A new methodology for comparing google scholar and scopus," *Journal of Informetrics*, vol. 10, no. 2, pp. 533–551, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1751157715302285

[3] E. Krauskopf, "Missing documents in scopus: the case of the journal enfermeria nefrologica," *Scientometrics*, vol. 119, pp. 543–547, 2019. [Online]. Available: https://doi.org/10.1007/s11192-019-03040-z

[4] M. Teplitskiy, E. Duede, M. Menietti, and K. R. Lakhani, "How status of research papers affects the way they are read and cited," *Research Policy*, vol. 51, no. 4, p. 104484, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048733322000129

[5] R. A. Buchanan, "Accuracy of cited references: The role of citation databases," *College & Research Libraries*, vol. 67, no. 4, pp. 292–303, 2006. [Online]. Available: https://crl.acrl.org/index.php/crl/article/view/15806

[6] A. Cioffi, S. Coppini, A. Massari, A. Moretti, S. Peroni, C. Santini, and N. Shahidzadeh Asadi, "Identifying and correcting invalid citations due to doi errors in crossref data," *Scientometrics*, vol. 127, pp. 3593–3612, 2022. [Online]. Available: https://doi.org/10.1007/s11192-022-04367-w

[7] F. Franceschini, D. Maisano, and L. Mastrogiacomo, "Errors in doi indexing by bibliometric databases," *Scientometrics*, vol. 102, pp. 2181–2186, 2015. [Online]. Available: https://doi.org/10.1007/s11192-014-1503-4

[8] A. Rivkin, "Manuscript referencing errors and their impact on shaping current evidence," *American Journal of Pharmaceutical Education*, vol. 84, no. 7, 2020. [Online]. Available: https://www.ajpe.org/content/84/7/ajpe7846

[9] I. Tahamtan and L. Bornmann, "What do citation counts measure? an updated review of studies on citations in scientific documents published between 2006 and 2018," *Scientometrics*, vol. 121, pp. 1635–1684, 2019. [Online]. Available: https://doi.org/10.1007/s11192-019-03243-4

[10] A. Martín-Martín, E. Orduna-Malea, and E. Delgado López-Cózar, "Coverage of highly-cited documents in google scholar, web of science, and scopus: a multidisciplinary comparison," *Scientometrics*, vol. 116, pp. 2175–2188, 2018. [Online]. Available: https://doi.org/10.1007/s11192-018-2820-9

[11] N. J. van Eck and L. Waltman, "Accuracy of citation data in web of science and scopus," 2019. [Online]. Available: https://arxiv.org/abs/1906.07011

[12] F. Franceschini, D. Maisano, and L. Mastrogiacomo, "A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 10, pp. 2149–2156, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22898

[13] ——, "Do scopus and wos correct "old" omitted citations?" *Scientometrics*, vol. 107, pp. 321–335, 2016. [Online]. Available: https://doi.org/10.1007/s11192-016-1867-8

[14] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022. [Online]. Available: https://arxiv.org/abs/2205.01833

[15] D. Rodrigues, A. L. Lopes, and F. Batista, "Web of Science Citation Gaps: An Automatic Approach to Detect Indexed but Missing Citations," in *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*, ser. Open Access Series in Informatics (OASIcs), A. Simões, M. M. Berón, and F. Portela, Eds., vol. 113. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, pp. 5:1–5:11. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2023/18519