TransInformação

ORIGINAL

Editor
Luisa Angélica Paraguai Donati

Conflict of interest
The authors declare that they have no conflicts of interest.

Data Availability
The research data are available on request from the corresponding author.

Received
April 30, 2025

Approved
October 21, 2025

A tool for bibliometric analysis of journals indexed in Google Scholar Metrics and OpenAlex

Ferramenta para análise bibliométrica de periódicos indexados no Google Scholar Metrics e OpenAlex

Edson Mário Gavron[1], Adilson Luiz Pinto[1], Fábio Lorensi do Canto[2], Marcos Talau[3]

[1] Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação. Florianópolis, SC, Brasil. Correspondence to: E. M. GAVRON. E-mail: <edson.gavron@ufsc.br>.

[2] Universidade Federal de Santa Catarina, Reitoria, Biblioteca Universitária. Florianópolis, SC, Brasil.

[3] Universidade Tecnológica do Paraná, *Campus* Curitiba, Departamento Acadêmico de Eletrônica. Curitiba, PR, Brasil.

Article based on the thesis of E. M. GAVRON, entitled "*Google Scholar Metrics e OpenAlex: construção de indicadores para medir impacto e visibilidade de periódicos*". Universidade Federal de Santa Catarina, 2025.

How to cite this article: Gavron, E. M. *et al*. A tool for bibliometric analysis of journals indexed in Google Scholar Metrics and OpenAlex. *Transinformação*, v. 37, e2515501, 2025. https://doi.org/10.1590/2318-0889202537e2515501.

Abstract

Google Scholar Metrics is one of the leading free tools for assessing the impact of academic journals; however, it presents significant limitations, such as the absence of a master list of journals and restrictions on automated data extraction. To overcome these challenges, this study presents the development of the GSM-ALEX_data tool, aimed at the extraction, integration, and bibliometric analysis of journals indexed in both Google Scholar Metrics and the OpenAlex database. The methodology involved the construction of four Python scripts responsible for automating data collection from both platforms, matching records, and structuring an integrated database. The results indicate that GSM-ALEX_data is effective for conducting large-scale bibliometric analyses and providing editorial data, metrics, and publications in an integrated manner. The main originality of the proposal lies in overcoming the limitations of Google Scholar Metrics through its combination with open data. It is concluded that the tool represents a resource for researchers by integrating data from platforms such as Google Scholar Metrics and OpenAlex, thereby expanding the scope available for bibliometric analyses.

Keywords: Bibliometric tool. Google Scholar Metrics. Journal evaluation. OpenAlex.

Resumo

O Google Scholar Metrics é uma das principais ferramentas gratuitas para avaliação do impacto de periódicos acadêmicos, mas apresenta sérias limitações, como ausência de uma lista mestre de periódicos e restrições à extração automatizada de dados. Visando superar esses entraves, este estudo apresenta o desenvolvimento da ferramenta GSM-ALEX_data, voltada à extração, integração e análise bibliométrica de periódicos indexados no Google Scholar Metrics e na base OpenAlex. A metodologia envolveu a construção de quatro scripts em Python, responsáveis por automatizar a coleta de dados nas duas plataformas, realizar

a correspondência entre os registros e estruturar um banco de dados integrado. Os resultados indicam que o GSM-ALEX_data é eficaz para realizar análises bibliométricas em larga escala, fornecendo dados editoriais, métricas e publicações de forma integrada. A principal originalidade da proposta está na superação das barreiras do Google Scholar Metrics por meio da combinação com dados abertos. Conclui-se que a ferramenta representa um recurso valioso para pesquisadores ao integrar dados de plataformas como o Google Scholar Metrics e o OpenAlex, ampliando o universo disponível para análises bibliométricas.

Palavras-chave: *Ferramenta bibliométrica. Google Scholar Metrics. Avaliação de periódicos. OpenAlex.*

Introduction

Google Scholar Metrics (GSM) is a tool used to measure the impact of academic journals. Google Scholar (GS) launched the tool in 2012 and calculates its metrics using citations from all articles indexed in its database, including those not directly covered by GSM (Google Scholar, 2024). Its purpose is to assess the visibility and influence of recent articles by summarizing citations across a wide range of academic publications, thereby guiding authors in selecting where to publish their research. GSM employs the h-index as its core metric, applying a five-year time window and calculating the median h-index (Google Scholar, 2024).

The GS is a search engine that indexes only academic documents. Its crawlers continuously scan university websites, publishers, repositories, databases, and other scholarly sources on the web/ without thematic or linguistic restrictions (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019). This automated indexing approach makes GSM a tool that minimizes selection bias common to commercial databases, as it includes journals from various countries and languages, often covering more regionally focused topics such as the humanities and social sciences (Jacsó, 2012; Leydesdorff; Wouters; Bornmann, 2016; Orduña-Malea; Delgado López-Cózar, 2014; Waltman, 2016).

Another point to highlight is that GS is among the platforms with the broadest coverage of bibliometric data (Martín-Martín *et al.*, 2021). Moreover, due to its free access and user-friendly interface, its use has grown significantly within the academic community (Canto *et al.*, 2022). These characteristics have contributed to GSM's popularity as an alternative for analyzing academic impact through citation metrics.

However, there are several limitations associated with GSM. One of them is the inability to extract data directly from the system, which hinders broader studies on temporal coverage and prevents large-scale analyses (Orduña-Malea; Aytac; Tran, 2019). Another issue is the lack of transparency regarding the number of indexed journals or the absence of a master list (Delgado López-Cózar; Cabezas-Clavijo, 2012). Additionally, there is no verification mechanism for indexed content, which may lead to duplicate records or unjustified exclusion of journals that meet the indexing criteria (Costa; Canto; Pinto, 2020). The lack of *International Standard Serial Number* (ISSN) search functionality further complicates matters, as this identifier could address inconsistencies in record standardization (Costa; Canto; Pinto, 2020).

Conducting large-scale studies using GSM data remains a challenge. Nonetheless, despite its limitations, GSM is still considered a promising tool for bibliometric research (Canto *et al.*, 2022; Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019).

Since GSM does not provide a master list or an API, users must perform journal queries by title. However, this method introduces automation difficulties and requires manual resolution, especially in cases involving homonymous titles or duplicate indexing. These issues stem from lacking additional disambiguating metadata beyond the journal-title.

The literature reports initiatives that use GSM for metric studies involving more than 1,000 journal titles, making manual data collection impractical or overly labor-intensive (Canto *et al.*, 2022; Delgado López-Cózar; Cabezas-Clavijo, 2012, 2013; Orduña-Malea; Delgado López-Cózar, 2014; Pinto *et al.*, 2020).

One solution to automate this process would be utilizing article data available in GSM lists. Until a few years ago, this approach was more complex to implement. However, automation has become feasible with the availability of complete bibliographic data on platforms like OpenAlex.

OpenAlex is a platform that incorporates data from Microsoft Academic. It was created following Microsoft's announcement of discontinuing its service in 2021 and has since established itself as a comprehensive source of bibliographic data (Scheidsteger; Haunschild, 2023; Zhang *et al.*, 2024).

Martín-Martín *et al.* (2021) conducted a study on the coverage of highly cited articles across 252 subject categories. The researchers investigated whether Web of Science, Scopus, Dimensions, OpenCitations, Microsoft Academic, and Google Scholar indexed the articles. Microsoft Academic was identified as the second most comprehensive database, behind only GS.

OpenAlex demonstrates broad coverage across scholarly content, data interoperability through persistent identifiers such as Digital Object Identifier (DOI), open access, and a field structure organized in multiple levels of granularity. These features enable detailed analyses of subfields and support tools such as APIs and snapshots for fast and efficient data access (Hao *et al.*, 2022; Mongeon; Bowman; Costas, 2023; Okamura, 2023). As such, OpenAlex stands out as a robust alternative, particularly for recent publications and interdisciplinary analyses. It offers significant advantages, including open access and integration with multiple sources, demonstrating great potential for research reporting and bibliometric analysis (Rodrigues; Lopes; Batista, 2023; Scheidsteger; Haunschild, 2023; Schnieders *et al.*, 2022).

In this context, a feasible alternative for mitigating GSM's limitations in large-scale studies is to use open bibliographic data, as it enables the integration of GSM information and facilitates enriched analysis. For this reason, OpenAlex emerges as a viable alternative, offering open access to an extensive catalog of scientific articles, authors, and institutions, along with the possibility of downloading the entire database.

Accordingly, this study aimed to develop a tool to extract data from GSM using OpenAlex as a base, allowing information integration and more detailed bibliometric analysis. This is an extension of the work presented in Canto *et al.* (2024), expanding the automation of the GSM data extraction and analysis process and, above all, adding a validation stage for the results based on OpenAlex data.

Methodological Procedures

This study proposes the implementation of automated routines for extracting bibliographic and bibliometric data from the GSM and OpenAlex platforms. To develop these routines, the researchers used the Python 3 programming language, along with libraries such as CSV, JSON, OS, GLOB, GZIP, SYS, Selenium, Datetime, Argparse, Random, Difflib, Natural Language Toolkit (NLTK), Unidecode, DuckDB, and Strftime. Together, these libraries enabled the implementation of the scripts.

To support the development of the routines, the researchers utilized a validated dataset, publicly available in the Zenodo research data repository. This dataset, produced by Canto *et al.*

(2022), comprises information on scientific journals from Latin American and Caribbean countries extracted from GSM. It served as the foundation for executing the automated routines, thereby enabling systematic monitoring and validation of the scripts' accuracy.

The researchers designed the script development to maximize the collection of GSM data, aiming to conduct a comprehensive survey of the content indexed on the platform. The objective was to map, as completely as possible, the journals with bibliometric data available in GSM, enabling large-scale analysis. Some process adjustments were necessary specific approaches, such as working with a predefined list of journals.

This study chose a complete download of the OpenAlex dataset, even though the platform offers tools that simplify data collection and analysis, including a suite of APIs that allow querying its information base as needed. By choosing to use the complete OpenAlex database, data processing became significantly faster than API-based access due to request limitations. The data is compressed and distributed across six directories: Source (Journals, Proceedings, etc.), Work (Articles, Books, etc.), Author, Institution, Concept (Indexing Terms), and Publisher. However, the total size of the uncompressed files easily exceeds 3TB, making it unfeasible to use on computers with limited storage capacity. Therefore, the researchers implemented strategies to process the data in smaller parts.

Another source of data collection was GSM, which does not offer tools for data extraction. Consequently, scraping data from the input list one journal-title at a time was necessary. The alternative adopted for large-scale queries was the use of URLs, allowing for the automation of searches by inserting the journal titles into the search structure. The researchers retrieved the data by accessing the HTML structure and applying field-location techniques to scrape the desired information.

While developing the GSM URL-based data extraction, the researchers observed that Google monitors the volume of requests made to its page. When the system detects excessive access, it activates restriction mechanisms, such as requiring a CAPTCHA to verify user authenticity. If the requests persist, the system may temporarily block the IP address. The study also found that IPs associated with academic institutions, especially when combined with request time randomization, tended to be tolerated and were not blocked by the server. Therefore, to circumvent these limitations, the script was configured to acquire data at randomized intervals between 10 and 30 seconds, using IP addresses from educational institutions.

Tool structure

The data extraction tool developed consists of four main scripts: GSM_search, OpenAlex_Source, OpenAlex_Works, and GSM-ALEX_merge, collectively referred to as GSM-ALEX_data. The GSM_search script was designed to perform searches based on a list of journal titles, while OpenAlex_Source collects information related to the journals. In turn, OpenAlex_Works is responsible for extracting article data, and the GSM-ALEX_merge script performs verification and matching between GSM and OpenAlex records, ensuring integration between the two datasets.

It is important to highlight that the researchers developed the GSM_search script based on the GSM_hdata script by Canto *et al.* (2024), specifically its GSM_hsearch section. Although both scripts perform similar tasks, they were designed with distinct functionalities to meet specific requirements and incorporate several improvements for enhanced automation. The team also developed the remaining three scripts to manage the data acquisition processes between the two sources, GSM and OpenAlex (Figure 1).

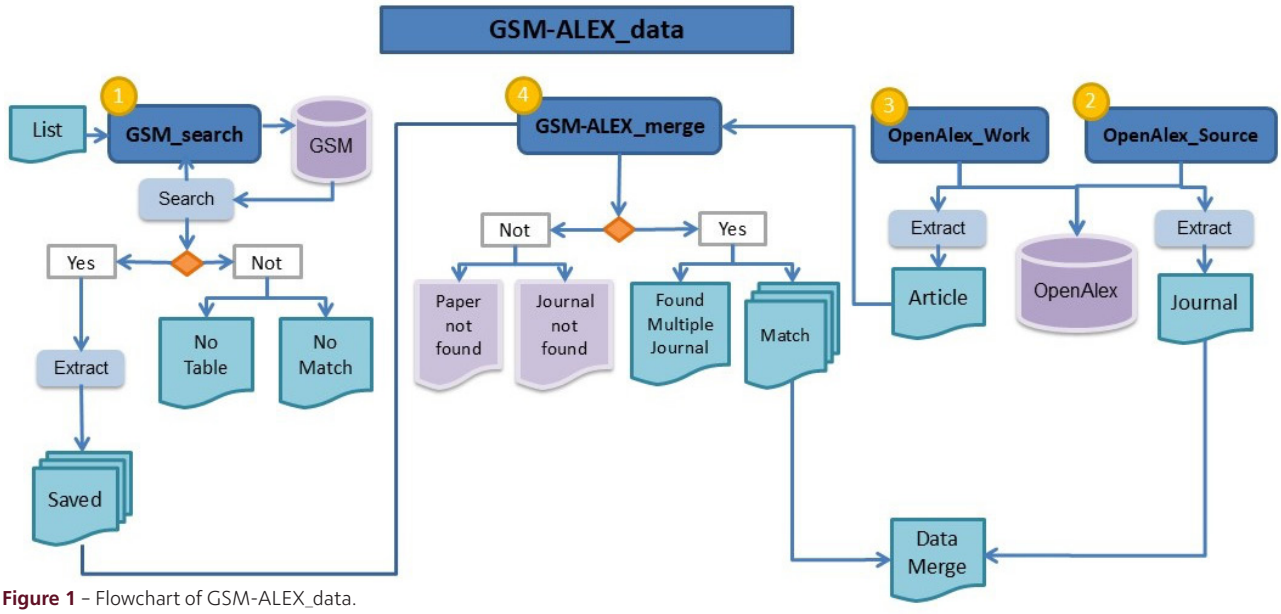

Figure 1 – Flowchart of GSM-ALEX_data.
Source: Prepared by the authors (2025).

The data integration stage (data merge) of the flowchart was structured to ensure consistency between the sets of information. Using a script, the fields illustrated in Figure 3 were collected and saved in a spreadsheet-compatible format. This file was then combined with the output file from OpenAlex_Source, using the journal's unique ID as the connection key. In this way, it was possible to generate a consolidated spreadsheet that brings together the data shown in Figure 3 and the fields detailed in Table 1.

GSM_search

GSM_search is the first component of GSM-ALEX_data and is responsible for collecting essential information from GSM, including the journal title, h5-index, h5-median, and the corresponding URL. It also identifies up to three articles comprising the journal's h5 index. Fewer articles will be returned If the h5 index is lower than three.

The process begins when the script sends a request through the GSM URL, inserting the journal's title into a specific part of the link. For example, if the title is 'The Journal of Finance', the script incorporates this term into the URL to perform the search. The platform may then return a list of data displayed on an HTML table. If it finds no results, it generates an empty table.

URL: scholar.google.com/citations?hl=en&view_op=search_venues&vq=the+journal+of+finance&btnG=

The script received a list of journal titles as input, which could appear in different formats, for instance, with parentheses containing the journal's location or other variations. To address this, the researchers established criteria to adjust the titles in cases where the search returned an empty table.

Initially, the script preserved the title in its original form, except when it consisted of a single word. In such cases, it added double quotation marks to improve search precision. If the search still returned an empty table, the script applied additional strategies. First, it removed numbers enclosed

in parentheses. If that proved ineffective, it excluded the entire content within parentheses. When this exclusion reduced the title to a single word, the script enclosed it in double quotation marks to enhance the search's effectiveness and increase the likelihood of retrieving a valid result.

This title-processing logic was implemented to enhance the accuracy of search results in specific situations while preserving the content within parentheses when necessary. However, this process increases the time required for data collection, as a single title may need to be queried multiple times in GSM.

Once a response with data was obtained, GSM_search performed a comparison to verify whether the input text matched the data in the returned table. This verification was based on four criteria, called Term Match (TM), which evaluate the similarity between the input terms and the results returned by GSM. When a match was confirmed, the system created a file for each title and recorded the TM criterion used for the comparison. This process allowed the identification of which TM produced the highest number of matches between the input title and the GSM results.

The script applied the first criterion based on the percentage of textual similarity. If the similarity reached 90% or higher, it recorded the data under this criterion. It calculated the percentage using the average number of characters in the analyzed titles – approximately 32 characters – with a variation of about three characters. Titles shorter than the average tended to produce higher comparison accuracy, while longer titles could inflate the count and reduce precision.

Nevertheless, this ±3 character variation proved to provide high accuracy in text comparisons and increased match rates. A test requiring 100% equality between texts was too sensitive to slight variations such as punctuation, accents, or formatting, which often resulted in missed matches. Therefore, the script prioritized the 90% threshold, as it offered greater reliability than stricter criteria.

The second criterion incorporated Natural Language Processing (NLP) techniques using the NLTK library to address variations in journal title formatting. The goal was to accommodate common differences such as: (a) Abbreviations: e.g., "Rev." instead of "Revista"; (b) Term substitutions: use of symbols like "&" instead of "AND"; (c) Omission or inclusion of auxiliary words: such as articles and pronouns, which might be inconsistently written or omitted, and (d) Spelling variations: slight differences in spelling or writing style.

This approach eliminated the need to manually define a stopword list, as the NLP model could automatically detect patterns and interpret context without prior manual filtering.

The third criterion was introduced to handle cases where the journal title list included subtitles, but the GSM result table did not. This criterion relied on identifying the shortest part of the text and checking whether it was contained in either the input title or the returned result. The process verified whether text A was included in text B or the other way around, thus addressing challenges related to subtitles.

Finally, the fourth criterion was designed to compare only the letters, disregarding white spaces. This solution addressed situations where the input title contained misspellings or merged two words without the usual spacing.

The script organized the results as follows: when it found a match in GSM, it saved the collected data in individual .txt files, each with a unique name, within a directory named 'saved'

```
tm: 1
Journal txt: the journal of finance
Journal google: The Journal of Finance
h5-index: 104
h5-median: 173
Papers URL: https://scholar.google.com/citations?hl=en&view_op=list_hcore&...
Paper: Do Investors Value Sustainability? A Natural Experiment Examining ...
Paper: Anticompetitive Effects of Common Ownership
Paper: Deviations from Covered Interest Rate Parity
```

Figure 2 – Example of TXT file format from GSM_search.
Source: Prepared by the authors (2025).

In cases where the search returned results that did not meet the established criteria, the script grouped those occurrences into a single file named NOT-match-items-in-table. Conversely, when the search returned no results at all, the script recorded them separately in a specific file named NOT_empty-table.

Applying these criteria allowed maximizing the number of matches identified during the GSM search process. This approach's main objective was to identify as many GSM-listed journal titles as possible that bore some relation to the input titles, thereby broadening the scope of data collection.

OpenAlex_Source

The second procedure developed, OpenAlex_Source, retrieved relevant information for the bibliometric analysis of journals from OpenAlex, enabling the creation of a dataset that complements the information about journals found in GSM. This process resulted in enhanced datasets, allowing for more rigorous bibliometric analysis of the journals.

The information collected included unique identifiers for each journal, such as ISSN and OpenAlex ID, editorial details, title variations, countries, quantitative data such as the number of published articles, bibliometric indicators, and subject classifications of the journals. All the data collected for each journal are listed in Table 1.

Table 1 – Fields extracted from journals in OpenAlex, Brazil, 2025.

Field	Description
ID	Unique identifier link for each journal
issn_l	Journal ISSN
Issn	When there is more than one ISSN, the journal is recorded in this field
display_name	Journal title
Publisher	Publisher
is_oa	Whether the journal is open-access
is_in_doaj	Whether it is indexed in DOAJ
Type	Source type (Journal, Proceedings, etc.)
country_code	Country ISO code
alternate_titles	Title variations
abbreviated_title	Title abbreviations
works_count	Total publications
cited_by_count	Total number of citations for the journal's articles hosted in OpenAlex
h_index	Total h-index
i10_index	i10-index (last 10 years)
x_concepts_display_name	Concepts assigned to the journal based on the concepts of its articles

Source: Prepared by the authors (2025).
Note: DOAJ: Directory of Open Access Journals; ISSN: International Standard Serial Number.

Data collection was carried out using a Python script. The script processes the JSON files found in the *Source* directory, opens them individually, extracts data from the previously identified fields, and then compiles a single CSV file containing the information listed in Table 1 for each journal. It is important to note that more fields are available within the OpenAlex JSON files; however, the selection of fields was based on the specific interests of this research.

OpenAlex_works

The third stage involved the development of a script called OpenAlex_Works. This script extracts information about journal articles. A filter was applied to collect only data from journals published in the last five years (2018–2023), the same time frame used by GSM to compute its h5-index. All journals indexed in OpenAlex published during this period were selected to compose the dataset. This process was developed to cross-reference data extracted from GSM with that from OpenAlex.

There are numerous fields available to describe each article. Therefore, a subset of relevant fields was selected to meet the system's functional requirements. The chosen fields included the article's ID, title, and source type (i.e., the journal in which the article was published). Regarding the article metadata, the fields collected included: title, authorship, document type, and citation metrics. The selected fields are listed in Table 2.

Table 2 – Fields extracted from OpenAlex articles Brazil, 2025.

Field	Description
title	Article title
publication_year	Year of publication
primary_location_id	Journal identifier number
primary_location_type	Source type
primary_location_display_name	Journal name
Type	Document type
authors_names	Author names
cited_by_count	Total number of citations
2yr_cited_by_count	Number of citations in the last 2 years

Source: Prepared by the authors (2025).

The scraping process for OpenAlex_Works differed from that of OpenAlex_Source. Due to insufficient storage capacity, it was impossible to decompress the entire *Works* directory simultaneously. Therefore, the procedure was carried out in stages. After each processing stage, the corresponding file was deleted to free up space and allow the process to continue.

To run the OpenAlex_Works script, a computer equipped with 80 GB of RAM and 20 Intel® Xeon® CPU E5-2420 processors, operating at 1.90 GHz, was used. This configuration enabled the entire procedure to be executed simultaneously without dividing the workload across machines with less memory capacity.

The procedures involving OpenAlex_Source and OpenAlex_Works generated a dataset for in-depth analysis of journals indexed in GSM. The step-by-step, systematized extraction process overcame storage and processing limitations, resulting in CSV files containing the original OpenAlex

fields. These files facilitate data integration and future analyses. This dataset complements GSM data and allows for more detailed bibliometric analyses of journals by research area, country, and metrics from both data sources.

GSM-ALEX_merge

After completing the three previous procedures, combining the datasets and promoting integration between the source data from OpenAlex and GSM becomes possible. The main goal of GSM-ALEX_merge is to compare data extracted from GSM with the information gathered from OpenAlex_Works, verifying their compatibility. The procedure aimed to identify matches between article titles collected from GSM and those indexed in OpenAlex, ensuring the integrity and reliability of the data analyzed.

For this comparison, the DuckDB library was used, enabling an efficient database for cross-referencing information. The matching process was based on article titles, comparing those found in GSM with those from the OpenAlex_Works dataset, ensuring more accurate data integration.

The OpenAlex_Works script was also processed using the same computer with 80 GB of RAM and 20 Intel® Xeon® CPU E5-2420 processors at 1.90 GHz.

The GSM_source script generated GSM data by creating individual files. The script then opened these files and checked for exact title matches in the OpenAlex_Works dataset. When it found a matching title, it created and saved a new file with a unique name composed of the date, time, and part of the journal title. This approach ensures that each generated file has a distinct identifier in .txt format. The output file included information about the articles that presented matches. This data comprises the article title, journal ID, the filename generated by GSM_search, and the associated information in that file.

This structured format facilitates the identification and analysis of compatible data between GSM and OpenAlex_Works, allowing for efficient and organized verification.

The illustration highlights the separation between two generated files: File 1, which contains the data obtained during the GSM_search stage, and File 2, where the articles with matches found in OpenAlex are stored.

In the illustrated example, two matching articles were identified. In such cases, File 2 records the journal ID, the matched article titles, and the related metric data. This organization facilitates the analysis of cross-referenced information.

It is also evident in the illustration that the journal title from GSM (File 1) differs from the title in OpenAlex. This occurred in cases where the journal underwent a title change, demonstrating that the system could handle complex situations such as title discontinuities or replacements.

To improve organization and facilitate analysis, the script created specific files for different scenarios: cases where no article title found a match, cases where at least one article produced a match, and cases where a single article title was linked to more than one journal. In situations with multiple matches, the script also verified whether the journal title from the TXT file matched any of the identified journals. This structure simplified the review process and helped identify potential failures (Figure 3).

```
File 1:
tm: 1
Journal txt: the journal of finance
Journal google: The Journal of Finance
h5-index: 104
h5-median: 173
Papers URL: https://scholar.google.com/citations?hl=en&view...
Paper: Do Investors Value Sustainability? A Natural Experiment ...
Paper: Anticompetitive Effects of Common Ownership
Paper: Deviations from Covered Interest Rate Parity

File 2:
File: /mnt/edson/step1/saved/2024-05-30_06_42_40__thejournal...
Journal txt: the journal of finance
title: do investors value sustainability? a natural experiment ...
primary_location_id: https://openalex.org/s5353659
primary_location_display_name: the journal of finance
cited_by_count: 656
2yr_cited_by_count: 0
title: anticompetitive effects of common ownership
primary_location_id: https://openalex.org/s5353659
primary_location_display_name: the journal of finance
cited_by_count: 469
2yr_cited_by_count: 0
title: deviations from covered interest rate parity
primary_location_id: https://openalex.org/s5353659
primary_location_display_name: the journal of finance
cited_by_count: 394
2yr_cited_by_count: 0
```

Figure 3 – Example of TXT file format generated by GSM-ALEX_merge.
Source: Prepared by the authors (2025).

Tool Validation

The first stage aimed to identify journal titles from a predefined list by checking for matches in GSM using the developed procedure. A list of 688 journal titles was selected, and similarity criteria were established to evaluate textual correspondence. As the script processed the list, titles not found in GSM were manually reviewed, and the script was adjusted to improve identification.

This phase was essential to detect possible flaws and correct the program logic. The list included journals already indexed in GSM and some that were not, providing a test base for verifying the criteria described in GSM_search.

Next, the program was tested with a larger dataset to validate the procedures. This new test involved 3.070 journals hosted in the Zenodo data repository. The list originated from a study on Latin American journals indexed in GSM (Canto *et al.*, 2021). The Zenodo dataset is in spreadsheet format and contains editorial information about these journals, including their titles, as listed in both GSM and Latindex.

Since the validation data refers to 2021, and some journals may have fallen outside GSM's inclusion criteria, updated data had to be collected following the same procedures. Thus, the researchers processed the list of 3.070 journals twice: first using the Gsm_hdata tool by Canto *et al.* (2024) and then using the GSM-ALEX_data tool. This process and its results are illustrated in Figure 4.

Using Gsm_hdata, 2.302 journals (75%) were found among the 3.070 titles. It is important to emphasize that this procedure used journal URLs instead of titles to extract data from GSM, which made the list more accurate and eliminated issues related to homonymous journal titles.

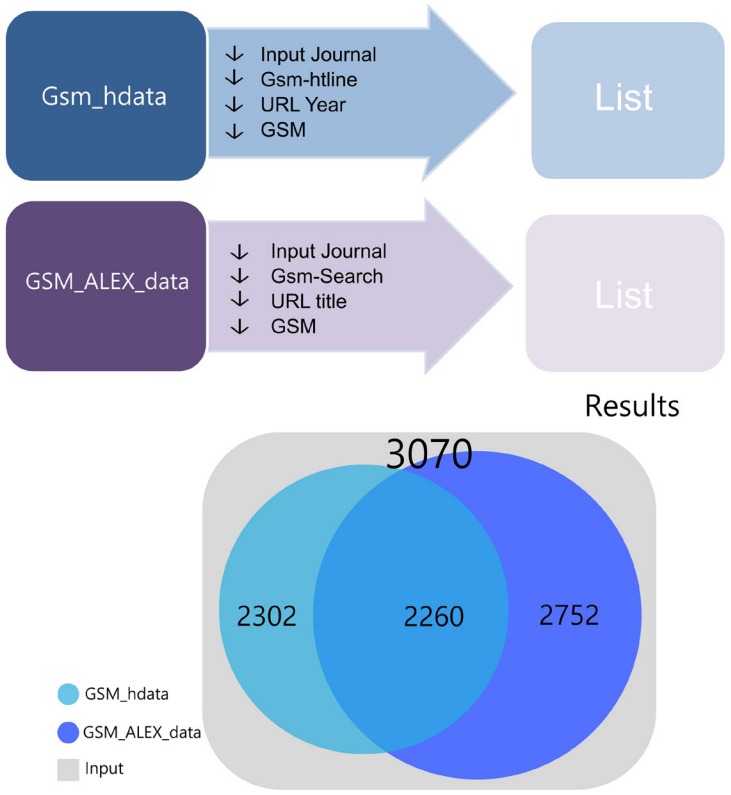

Figure 4 – Workflow of the GSM data collection procedure results.
Source: Prepared by the authors (2025).

The GSM-ALEX_data procedure returned a positive response for 2.752 (90%) of the 3.070 titles searched. Among the 318 unmatched titles, 309 returned no results in the GSM query, possibly because they were no longer indexed. Therefore, GSM-ALEX_data showed an improvement of 450 additional matched titles (15%) compared to Gsm_hdata.

To check whether the titles matched between both procedures, the researchers compared the URLs of each journal retrieved from GSM using Python. Among the 2.752 titles found by GSM_Search, 2.260 had the same URL as those found via Gsm_hdata. Based on this test dataset, the system achieved a 98% accuracy rate. It is also important to note that part of the dataset from Canto *et al.* (2021) was collected manually, whereas GSM_Search is fully automated.

Another procedure in the system involves verifying the output obtained through GSM_Search, serving as a mechanism to ensure that the journal located in GSM is indeed the intended one. This is done by comparing the titles of the articles. Thus, the 2.752 journal titles were submitted to the GSM-ALEX_merge procedure, resulting in 2.421 journals with one or more equivalent article titles between OpenAlex and GSM. This outcome demonstrates that the combination of the GSM_Search and GSM-ALEX_merge procedures yielded a slight performance improvement, surpassing the 2.302 journals identified by Gsm_hdata. Table 3 presents the distribution of the textual comparison criteria (TM) applied during the similarity verification process between the input titles and the records retrieved in GSM.

It can be observed that TM1 accounted for the majority of cases, totaling 2,595 journals, which represents 94.40% of the total. TM2 corresponded to 5.45% (150 journals). As expected, TM3 and TM4 showed a low incidence (0.15% or none), since they are applied hierarchically: records

Table 3 – Distribution of Term Match criteria used for textual comparison, Brazil, 2025.

Criteria	GSM_search		GSM-ALEX_merge	
	n	%	*n*	%
TM1	2595	94,40	2285	94,34
TM2	150	5,45	132	5,45
TM3	4	0,15	4	1,17
Total	2752	100	2421	100

Source: Prepared by the authors (2025).
Note: TM: Term Match.

that do not meet the TM1 and TM2 criteria are reassessed using TM3 and TM4 as a final validation attempt. Consequently, the percentages remain practically mirrored for GSM-ALEX_merge. These results indicate that most matches between input titles and records retrieved in GSM occurred with a high degree of textual similarity, above 90%.

Final Considerations

The tool presented in this study overcomes one of the main limitations of GSM: the absence of a master list containing all indexed journals. This is made possible by automating queries and data collection from GSM, enabling large-scale bibliometric analyses.

Another relevant aspect is that GSM does not provide editorial information about the journals. With GSM-ALEX_data, linking editorial data to journals indexed in GSM via a spreadsheet becomes possible. Furthermore, the tool enables the incorporation of bibliometric indicators from OpenAlex, such as the h-index, total number of citations, and number of published articles, allowing for a more in-depth analysis of the journals.

Among the limitations identified is the time required to extract data from large lists, especially when dealing with more than 100,000 titles. Since the tool must simulate human behavior, data collection can take over a month on a single machine; however, this limitation can be easily overcome by using multiple machines. It is important to note that the research used a dataset of 3,070 titles, covering the entire Latin America and Caribbean region, which was collected in no more than two days.

Since it is a tool that performs scraping, a common limitation in this context is its reliance on the maintenance of consistent patterns in the page structure and security settings. Should any of these patterns be changed, the system will need to be adapted to ensure the continuity of the process.

Another limitation is the volume of data, which demands a computer with compatible RAM and storage capacity, as used in this research. Additionally, when certain information is missing in OpenAlex, such as details about editors or the country of publication, the ability to conduct specific analyses for those journals may be affected.

In light of these limitations, an important gap is identified for future studies, such as adapting the procedures to be carried out via the OpenAlex API, which would allow the process to be applied even on machines with more modest processing capacity, as well as enabling a comparison of these results with those obtained in the present research.

Furthermore, since some data inconsistencies were observed during the procedures, it is essential to conduct a more detailed empirical study to assess the extent to which these inconsistencies may affect the quality of a bibliometric analysis.

The researchers identified improvements to be made to GSM-ALEX_data, such as developing a user-friendly interface for the system, which currently runs without one and therefore limits its use by a wider group of researchers. They also plan to add a criterion that will allow duplicate records to be identified in advance, making it easier to clean the data more efficiently.

Despite these limitations, GSM-ALEX_data proved to be an effective tool for extracting data from GSM and OpenAlex, combining information from both sources and enabling detailed bibliometric analyses of the extracted datasets.

References

Canto, F. L. *et al.* Latin american and caribbean journals indexed in Google Scholar Metrics. *Scientometrics*, v. 127, n. 2, p. 763-783, 2022. Doi: https://doi.org/10.1007/s11192-021-04237-x.

Canto, F. L. *et al.* Gsm_hdata: a bibliometric tool to analyze data from google scholar metrics. *Mobile Networks and Applications*, v. 29, p. 754-761, 2024. Doi: https://doi.org/10.1007/s11036-023-02258-9.

Canto, F. L. *et al.* Latin American and Caribbean journals indexed in Google Scholar Metrics. [Data set]. *Zenodo*, 2021. Versão 2. Doi: https://doi.org/10.5281/zenodo.5704895.

Costa, H.; Canto, F. L.; Pinto, A. L. Google Scholar Metrics e a proposta do novo Qualis: impacto dos periódicos brasileiros de Ciência da Informação. *Informação & Sociedade*, v. 30, n. 1, 2020. Doi: https://doi.org/10.1007/s11036-023-02258-9.

Delgado López-Cózar, E.; Cabezas-Clavijo, A. Google Scholar Metrics: an unreliable tool for assessing scientific journals. *El Profesional de la Información*, v. 21, n. 4, p. 419-427, 2012. Available from: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2012.jul.15. Cited: Dec. 10, 2024.

Delgado López-Cózar, E.; Cabezas-Clavijo, A. Ranking journals: could Google scholar metrics be an alternative to journal citation reports and Scimago journal rank?. *Learned publishing*, v. 26, n. 2, p. 101-114, 2013. Doi: http://doi.wiley.com/10.1087/20130206.

Delgado López-Cózar, E.; Orduña-Malea, E.; Martín-Martín, A. Google Scholar as a Data Source for Research Assessment. *In*: Glänzel, W. *et al.* (org.). *Springer Handbook of Science and Technology Indicators*. Cham: Springer, 2019. p. 95-127. Doi: https://doi.org/10.1007/978-3-030-02511-3_4.

Google Scholar. *Google Scholar Metrics*. 2024. Available from: https://scholar.google.com/intl/en/scholar/metrics.html#overview. Cited: Feb. 21, 2025.

Hao, H. *et al.* Thirty-two years of ieee vis: Authors, fields of study and citations. *IEEE Transactions on Visualization and Computer Graphics*, v. 29, n. 1, p. 1016-1025, 2022. Doi: https://doi.org/10.1109/TVCG.2022.3209422.

Jacsó, P. Google Scholar Metrics for Publications. *Online Information Review*, v. 36, n. 4, p. 604-619, 2012. Available from: https://www.emerald.com/insight/content/doi/10.1108/14684521211254121/full/html. Cited: Feb. 15, 2025.

Leydesdorff, L.; Wouters, P.; Bornmann, L. Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators a state-of-the-art report. *Scientometrics*, v. 109, p. 2129-2150, 2016. Doi: https://doi.org/10.1007/s11192-016-2150-8.

Martín-Martín, A. *et al.* Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science e COCI do OpenCitations: uma comparação multidisciplinar de cobertura por meio de citações. *Scientometrics*, v. 126, p. 871-906, 2021. Doi: https://doi.org/10.1007/s11192-020-03690-4.

Mongeon, P.; Bowman, T. D.; Costas, R. An open data set of scholars on Twitter. *Quantitative Science Studies*, v. 4, n. 2, p. 314-324, 2023. Doi: https://doi.org/10.1162/qss_a_00250.

Okamura, K. A half-century of global collaboration in science and the "Shrinking World". *Quantitative Science Studies*, v. 4, n. 4, p. 938-959, 2023. Doi: https://doi.org/10.1162/qss_a_00268.

Orduna-Malea, E.; Aytac, S.; Tran, C. Y. Universities through the eyes of bibliographic databases: a retroactive growth comparison of Google Scholar, Scopus and Web of Science. *Scientometrics*, v. 121, p. 433-450, 2019. Doi: https://doi.org/10.1007/s11192-019-03208-7.

Orduña-Malea, E.; Delgado López-Cózar, E. Google Scholar Metrics evolution: an analysis according to languages. *Scientometrics*, v. 98, p. 2353-2367, 2014. Doi: https://doi.org/10.1007/s11192-013-1164-8.

Pinto, A. L. *et al.* Periódicos científicos brasileiros indexados no Google Scholar Metrics. *Informação & Sociedade*: *Estudos*, v. 30, n. 4, p. 1-18, 2020. Available from: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/57048. Cited: Mar. 3, 2025.

Rodrigues, D.; Lopes, A. L.; Batista, F. Web of Science citation gaps: an automatic approach to detect indexed but missing citations. Symposium on Languages, Applications and Technologies, 12., 2023, Wadern. *Proceedings* [...]. Wadern: Schloss Dagstuhl, Leibniz Center for Informatics, 2023. Doi: https://doi.org/10.4230/OASIcs.SLATE.2023.5.

Scheidsteger, T.; Haunschild, R. Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *El Profesional de la Información*, v. 32, n. 2, 2023. Doi: https://doi.org/10.3145/epi.2023.mar.09.

Schnieders, K. *et al.* ORCID coverage in research institutions Readiness for partially automated research reporting. *Frontiers in Research Metrics and Analytics*, v. 7, 2022. Doi: https://doi.org/10.3389/frma.2022.1010504.

Waltman, L. A review of the literature on citation impact indicators. *Journal of Informetrics*, v. 10, n. 2, p. 365-391, 2016. Doi: https://doi.org/10.1016/j.joi.2016.02.007.

Zhang, L. *et al.* Missing institutions in OpenAlex: possible reasons, implications, and solutions. *Scientometrics*, v. 129, p. 5869-5891, 2024. Doi: https://doi.org/10.1007/s11192-023-04923-y.

Contributors

Conceptualization: A. L. PINTO. Data Curation: E. M. GAVRON and M. TALAU. Investigation: E. M. GAVRON. Methodology: E. M. GAVRON and A. L. PINTO. Software: M. TALAU. Supervision: A. L. PINTO and F. L. CANTO. Validation F. L. CANTO. Writing – Original Draft: E. M. GAVRON. Writing – Review & Editing: E. M. GAVRON, A. L. PINTO, F. L. CANTO, and M. TALAU.