



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS DA EDUCAÇÃO
PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Edson Mário Gavron

**Google Scholar Metrics e OpenAlex: construção de indicadores para medir
impacto e visibilidade de periódicos**

Florianópolis
2025

Edson Mário Gavron

**Google Scholar Metrics e OpenAlex: construção de indicadores para medir
impacto e visibilidade de periódicos**

Tese submetida ao Programa de Pós-Graduação
em Ciência da Informação do Centro de Ciências
da Educação da Universidade Federal de Santa
Catarina, como requisito para obtenção do título
de Doutor em Ciência da Informação.
Orientador: Prof. Dr. Adilson Luiz Pinto

Florianópolis

2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Gavron, Edson Mário

Google Scholar Metrics e OpenAlex : construção de indicadores para medir impacto e visibilidade de periódicos / Edson Mário Gavron ; orientador, Adilson Luiz Pinto, 2025.

154 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação, Florianópolis, 2025.

Inclui referências.

1. Ciência da Informação. 2. Bibliometria. 3. Google Scholar Metrics. 4. OpenAlex. I. Pinto, Adilson Luiz. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

Edson Mário Gavron

Google Scholar Metrics e OpenAlex: construção de indicadores para medir impacto e visibilidade de periódicos

O presente trabalho em nível de Doutorado foi avaliado e aprovado, em 15 de setembro de 2025, pela banca examinadora composta pelos seguintes membros:

Prof. Thiago Magela Rodrigues Dias, Dr.

Instituição: Centro Federal de Educação Tecnológica de Minas Gerais

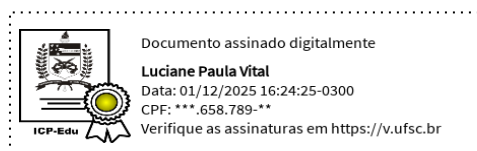
Prof. Alexandre Ribas Semeler, Dr.

Instituição: Universidade Federal do Rio Grande do Sul

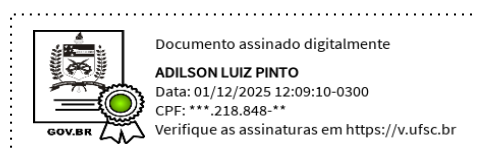
Prof. Washington Luís Ribeiro de Carvalho Segundo, Dr.

Instituição: Instituto Brasileiro de Informação em Ciência e Tecnologia

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutor em Ciência da Informação.



Coordenação do Programa de Pós-Graduação



Prof. Adilson Luiz Pinto, Dr.

Orientador

Florianópolis, 2025.

AGRADECIMENTOS

Gostaria de prestar uma homenagem às pessoas que foram fundamentais para a realização deste sonho.

Agradeço especialmente à minha companheira Shelley, por sua compreensão, pelo ambiente favorável e por toda a dedicação, ajuda, sugestões e contribuições que foram essenciais para o resultado final deste trabalho.

Agradeço também à minha família, que sempre me apoiou e incentivou na construção desta tese, em especial à minha mãe Maria, aos meus irmãos Eva e Evandro, e aos meus sogros Vilmar e Sirlei.

Aos amigos, deixo meu agradecimento pelo apoio e incentivo ao longo da jornada. Ao grupo de estudo, pelas conversas e pelas trocas de conhecimento, com destaque ao Marcos, que dedicou tempo e esforço na criação dos programas utilizados neste estudo.

Sou grato ao meu orientador, Adilson, pela confiança, pela oportunidade de desenvolver este trabalho sob sua orientação e pelo tempo e dedicação investidos ao longo do processo.

Agradeço aos membros da banca pela disposição em avaliar meu trabalho e pelas contribuições valiosas.

À UFSC, pelo suporte acadêmico e pela infraestrutura que possibilitaram a realização da pesquisa.

Por fim, a todos que, de alguma forma, contribuíram para este trabalho, deixo registrado meu sincero agradecimento.

RESUMO

O Google Scholar Metrics apresenta-se como uma alternativa para a avaliação de impacto de periódicos, especialmente por sua ampla cobertura em áreas como Ciências Humanas e Sociais. No entanto, o sistema possui limitações significativas, como a ausência de uma API, a falta de uma lista mestre de periódicos que dificulta sua utilização em análises de larga escala. Diante desse cenário, a presente pesquisa teve como objetivo propor um *framework* (GSM-ALEX_metrics) para análise em grande escala dos periódicos indexados no Google Scholar Metrics, fundamentado na integração com dados bibliográficos abertos do OpenAlex. O método combinou procedimentos de extração automatizada de dados do Google Scholar Metrics com técnicas de integração de metadados provenientes do OpenAlex. Foram desenvolvidos *scripts* para coleta, tratamento e validação das informações. O GSM-ALEX_metrics contemplou, de modo geral, três etapas principais: coleta de dados no Google Scholar Metrics e OpenAlex; cruzamento e enriquecimento com registros do OpenAlex; e análise estatística baseada em múltiplos indicadores. Nos resultados, o GSM-ALEX_metrics se mostrou eficaz para análises bibliométricas em larga escala, integrando dados editoriais, métricas de impacto e informações sobre publicações de forma automatizada. O sistema conseguiu recuperar 25.969 revistas válidas do Google Scholar Metrics. A análise da distribuição do índice h5 mostrou que a maioria dos periódicos se concentra entre 1 e 30, padrão também observado no índice h do OpenAlex. Regionalmente, destacam-se Europa e América, que juntas representam 78% das revistas, sobretudo Estados Unidos e Reino Unido. Essa concentração também se reflete nas métricas de citação, uma vez que cerca de 50% dos artigos não receberam nenhuma citação no período analisado. Em termos de grandes áreas Capes, sobressaem Ciências Humanas, Ciências da Saúde, Ciências Exatas e Biológicas, com maior representatividade em áreas como Medicina, Filosofia, Biologia e Ciência Política. Conclui-se que a integração entre Google Scholar Metrics e OpenAlex, constitui uma abordagem viável para os estudos bibliométricos, por oferecer maior transparência, reprodutibilidade e abrangência analítica. A pesquisa reforça a importância do uso de múltiplos indicadores para captar diferentes dimensões do impacto científico. Dessa forma, o GSM-ALEX_metrics desenvolvido não apenas

supera limitações do Google Scholar Metrics, como também amplia a capacidade de compreender a dinâmica da comunicação científica.

Palavras-chave: Bibliometria. Google Scholar Metrics. OpenAlex. Indicadores de impacto. Ciência Aberta.

ABSTRACT

Google Scholar Metrics presents itself as an alternative for evaluating the impact of scholarly journals, particularly due to its broad coverage in areas such as the Humanities and Social Sciences. However, the system has significant limitations, such as the absence of an API and the lack of a master list of journals, which hinder its use in large-scale analyses. In this context, we aimed to propose a framework (GSM-ALEX_metrics) for large-scale analysis of journals indexed in Google Scholar Metrics, based on the integration of open bibliographic data from OpenAlex. The method combined automated data extraction procedures from Google Scholar Metrics with metadata integration techniques derived from OpenAlex. We developed scripts for data collection, processing, and validation. The GSM-ALEX_metrics framework consisted of three main stages: data collection from Google Scholar Metrics and OpenAlex; cross-referencing and enrichment with OpenAlex records; and statistical analysis based on multiple indicators. In the results, GSM-ALEX_metrics proved effective for large-scale bibliometric analyses, integrating editorial data, impact metrics, and publication information in an automated manner. The system successfully retrieved 25,969 valid journals from Google Scholar Metrics. The analysis of the h5-index distribution showed that most journals are concentrated between 1 and 30, a pattern also observed in the h-index from OpenAlex. Regionally, Europe and the Americas stand out, together representing 78% of the identified journals, especially the United States and the United Kingdom. This concentration is also reflected in the citation metrics, as approximately 50% of the articles received no citations during the analyzed period. Regarding the major CAPES research fields, the most representative areas were the Humanities, Health Sciences, Exact and Earth Sciences, and Biological Sciences, with a higher presence in specific disciplines such as Medicine, Philosophy, Biology, and Political Science. It is concluded that the integration between Google Scholar Metrics and OpenAlex constitutes a viable approach for bibliometric studies, as it offers greater transparency, reproducibility, and analytical scope. The research reinforces the importance of using multiple indicators to capture different dimensions of scientific impact. Thus, the GSM-ALEX_metrics framework not only overcomes the limitations of Google Scholar Metrics but also enhances our ability to understand the dynamics of scientific communication on a global scale.

Keywords: Bibliometrics. Google Scholar Metrics. OpenAlex. Impact indicators.
Open science.

LISTA DE FIGURAS

Figura 1 - Ilustração cálculo do FI	34
Figura 2 - Fórmula do cálculo do CiteScore	36
Figura 3 - Práticas de Ciência Aberta.....	42
Figura 4 - Mapa dos países que constituem o universo da pesquisa	68
Figura 5 - Fluxograma do GSM-ALEX_metrics	75
Figura 6 - Integração das etapas 1,2 e 3 do módulo GSM-ALEX_data.....	79
Figura 7 - Inserção do título na URL	81
Figura 8 - Arquivo TXT Saved GSM_search	82
Figura 9 - Fluxograma do GSM_search	83
Figura 10 - Arquivo TXT salvo.....	87
Figura 11 - Fluxograma do GSM-ALEX-verification	88
Figura 12 - Fluxo do procedimento de coleta dos dados.....	98
Figura 13 - Diagrama de Venn entre Gsm-hdata e GSM-ALEX_verification.....	99
Figura 14 - Síntese do processo de coleta, verificação e integração dos periódicos analisados	102
Figura 15 - Representação dos periódicos por países	105
Figura 16 - Concentração de revistas por grandes áreas da CAPES	106
Figura 17 - Concentração de revistas por área da CAPES	107
Figura 18 – Distribuição da quantidade de títulos de revista por editoras	108
Figura 19 - Mapa dos países com revistas no quartil 1	130

LISTA DE GRÁFICOS

Gráfico 1 – Evolução anual das publicações no período de 2012 a 2025	62
Gráfico 2 – Quantidade de revistas com dados incompletos.....	104
Gráfico 3 - Revistas indexadas por modelo de acesso	109
Gráfico 4 - Diferença textual entre os títulos contidos em ambos os sentidos	113
Gráfico 5 - Diferença textual entre os títulos não contidos em ambos os sentidos .	114
Gráfico 6 - Distribuição do índice h5 das revistas que precisaram de revisão manual	115
Gráfico 7 - Distribuição de frequência do índice h5 do GSM.....	117
Gráfico 8 - Distribuição de frequência do índice h5 da pesquisa.....	118
Gráfico 9 - Distribuição do percentual relativo dos periódicos por faixa do índice h5	120
Gráfico 10 - Distribuição da média de 5 anos.....	122
Gráfico 11 - Distribuição das revistas por percentual de artigos não citados	123
Gráfico 12 - Concentração de citações por artigo com base no índice de Gini	124
Gráfico 13 - Distribuição do índice h5 pelas grandes áreas da CAPES	125
Gráfico 14 - Distribuição do índice h5 por área da CAPES	127
Gráfico 15 – Distribuição do índice h5 por quartil 1	129

LISTA DE QUADROS

Quadro 1 - Princípios FAIR	51
Quadro 2 - Síntese dos procedimentos para alcançar os objetivos do estudo	67
Quadro 3 – Conceitos base do GSM-ALEX_metrics	78
Quadro 4 - Campos coletados pelo etapa OpenAlex_source	84
Quadro 5 - Campos coletados pela etapa OpenAlex_work	85

LISTA DE TABELAS

Tabela 1 - Critérios de comparação entre título de entrada e do GSM	110
Tabela 2 - Concentração de revistas por faixa de similaridade textual	111
Tabela 3 - Distribuição dos periódicos por similaridade abaixo 90%.....	112

SUMÁRIO

1	INTRODUÇÃO	15
1.1	OBJETIVO GERAL.....	17
1.2	OBJETIVOS ESPECÍFICOS	17
1.3	JUSTIFICATIVA	17
2	CONHECENDO AS BASES TEÓRICAS DO ESTUDO.....	19
2.1	DADOS BIBLIOGRÁFICOS.....	24
2.2	BASE DE DADOS BIBLIOGRÁFICAS	26
2.2.1	Fator de impacto.....	33
2.2.2	Eigenfactor Score.....	34
2.2.3	CiteScore.....	35
2.2.4	Scimago Journal Rank.....	37
2.2.5	Índice h.....	38
2.2.6	Não Citados.....	39
2.3	CIÊNCIA ABERTA.....	40
2.3.1	Dados de pesquisa e o Princípios FAIR	47
2.4	GOOGLE SCHOLAR METRICS.....	52
3	PROCEDIMENTOS METODOLÓGICOS.....	63
3.1	CARACTERIZAÇÃO DA PESQUISA	63
3.2	DELINEAMENTO DA PESQUISA.....	65
3.3	UNIVERSO DA PESQUISA.....	67
3.4	COLETA DE DADOS E MATERIAIS UTILIZADOS	69
3.5	GSM-ALEX_metrics (<i>FRAMEWORK</i>)	73
3.5.1	Conceitual	76
3.5.2	GSM-ALEX_data	79
3.5.2.1	GSM_search.....	80
3.5.2.2	OpenAlex_data.....	83
3.5.2.3	GSM-ALEX_verification	85
3.5.5	Tratamento dos dados	88
3.5.6	Analítico	92
4	VALIDAÇÃO DO SISTEMA.....	97
5	RESULTADOS.....	102
5.1	ANÁLISES BIBLIOMÉTRICAS	116

5.2	DISCUSSÃO	131
6	CONSIDERAÇÕES FINAIS	136
6.1	LIMITAÇÕES.....	140
6.2	TRABALHOS FUTUROS.....	142
	REFERÊNCIAS	144

1 INTRODUÇÃO

Os dados de citações permitem a criação de métricas fundamentais para compreender o impacto de um autor, de uma publicação ou de uma fonte de informação em geral. Essas medidas têm implicações diretas em decisões, como a alocação de financiamentos, assinaturas de bases de dados e avaliações institucionais. Entretanto, os principais sistemas utilizados para mensurar o impacto científico global são fechados e controlados por empresas privadas, como, por exemplo, a Web of Science, gerida pela Clarivate Analytics, e a Scopus, administrada pela Elsevier, bases comerciais que cobrem diversas áreas do conhecimento.

Nesse contexto, o Google Scholar Metrics (GSM) surge como alternativa, pois utiliza dados do Google Scholar para gerar suas métricas de impacto. Sua principal vantagem está na abrangência de cobertura, que supera as bases comerciais, incluindo títulos de áreas como Ciências Humanas e Sociais, publicações em diferentes idiomas e revistas provenientes de países emergentes (Martín-Martín *et al.*, 2021; Canto, 2022).

Estudos mostram, por exemplo, que em um conjunto de 2.515 documentos altamente citados em inglês, o Google Scholar recuperou 88% de todas as citações identificadas, muitas das quais não estavam disponíveis em outras bases (Bornmann *et al.*, 2016). Dessa forma, o GSM complementa lacunas deixadas pela Web of Science e pela Scopus, ampliando a representatividade da produção científica em escala global.

Apesar dessas vantagens, o GSM apresenta limitações importantes. Por ser um sistema de indexação automatizado baseado na extração de linguagem natural, ocorrem inconsistências, como a presença de títulos duplicados e a atribuição incorreta de artigos às revistas citadas (Canto, 2022).

Outras restrições incluem a ausência de campo de busca por ISSN, a disponibilização apenas do índice h5 e a inexistência de uma API que permita a extração automatizada de dados.

Além disso, a impossibilidade de realizar análises temporais aprofundadas dificulta o uso do GSM em estudos institucionais ou monitoramentos de longo prazo (Orduna-Malea; Aytac; Tran, 2019). Há ainda incertezas quanto ao número real de

periódicos indexados, já que o Google não divulga publicamente a lista completa de títulos (Gu; Blackmore, 2017).

Nesse cenário, os avanços da Ciência Aberta oferecem caminhos para superar tais desafios. A partir da Initiative for Open Citations (I4OC), lançada em 2016, consolidou-se a obrigatoriedade de que todos os membros do Crossref disponibilizem suas referências de forma aberta e sem restrições de distribuição, transformando esse repositório em um dos maiores conjuntos de metadados bibliográficos do mundo (Initiative for Open Citations, 2024).

Entre as plataformas que se destacam nesse movimento, encontra-se o OpenAlex, além de usar os dados do Crossref, também incorporou os dados do Microsoft Academic e passou a oferecer acesso livre e estruturado a informações sobre artigos, periódicos, autores e instituições (Scheidsteger; Haunschild, 2023).

Além de sua ampla cobertura, o OpenAlex já é utilizado em classificações internacionais, como o CWTS Leiden Ranking, evidenciando sua confiabilidade para análises bibliométricas em larga escala (Zhang *et al.*, 2024). Sua estrutura aberta, padronizada e de fácil utilização também é destacada na literatura como um diferencial, permitindo que pesquisadores iniciem análises complexas em menor tempo e com maior reprodutibilidade (Chawla, 2022).

Assim, observa-se que a união entre o GSM e o OpenAlex pode representar um avanço metodológico significativo para a bibliometria. Enquanto o GSM oferece ampla cobertura de títulos, especialmente em áreas e regiões menos representadas em bases comerciais, o OpenAlex proporciona infraestrutura aberta, dados padronizados e possibilidade de análises avançadas. Integrar essas duas fontes, portanto, permite não apenas superar limitações existentes, mas também ampliar a criação de indicadores e perspectivas analíticas.

Dessa forma, formula-se o problema central desta pesquisa: como viabilizar uma análise em grande escala dos periódicos indexados no Google Scholar Metrics, utilizando dados do OpenAlex, considerando as limitações, a integração e a validação de dados entre essas ferramentas?

1.1 OBJETIVO GERAL

Propor um *framework* de análise em grande escala de periódicos indexados no Google Scholar Metrics com base em dados bibliográficos do OpenAlex que permita a integração e a validação entre essas ferramentas.

1.2 OBJETIVOS ESPECÍFICOS

- a) Desenvolver um processo automatizado para coleta, certificação e integração de dados do Google Scholar Metrics com base nos dados bibliométricos do OpenAlex;
- b) Validar o processo de coleta de dados;
- c) Verificar as características editoriais dos periódicos recuperados do Google Scholar Metrics com base nos dados do OpenAlex;
- d) Analisar os dados bibliométricos dos periódicos do Google Scholar Metrics e do OpenAlex;
- e) Desenvolver indicadores bibliométricos complementares para o *framework*.

1.3 JUSTIFICATIVA

A comunidade científica tem necessidade de múltiplos indicadores para o entendimento do impacto dos periódicos. A utilização de apenas uma métrica pode limitar a compreensão sobre a relevância da revista, uma vez que diferentes indicadores capturam dimensões distintas da produção científica. Como, por exemplo, o Fator de Impacto, embora seja uma medida amplamente empregada e útil, mostra-se insuficiente por si só para representar a complexidade do impacto de um periódico, exigindo o apoio de outras métricas que permitam análises mais abrangentes (Glänzel; Moed, 2002).

Nesse sentido, surgiu uma multiplicidade de indicadores, evidenciando a necessidade de considerar diferentes perspectivas de avaliação, visto que cada métrica fornece uma leitura específica sobre a visibilidade e influência científica (Moed, 2005; Glänzel, 2009).

Esses indicadores têm sido efetivamente utilizados por diferentes países na avaliação de impacto científico, o que reforça sua relevância prática. O índice h5, em especial, é adotado em processos de análise e monitoramento da produção científica em nações como Espanha, Brasil, Colômbia e México (De Filippo; Aleixandre-Benavent; Sanz-Casado, 2020; Alperin; Rozemblum, 2017; Costa; Canto; Pinto, 2020). Esse cenário demonstra que, além do valor acadêmico, os indicadores métricos possuem implicações concretas para políticas científicas e estratégias de avaliação em âmbito internacional, consolidando a importância de estudos que investigam sua aplicação e limitações.

Existe uma série de periódicos que foram identificados exclusivamente no OpenAlex e GSM, sem registro na Web of Science e na Scopus, o que revela uma lacuna significativa na representação da produção científica quando se considera apenas essas duas bases tradicionais. Esse cenário evidencia a importância de integrar diferentes fontes de dados, sendo que somente a combinação de múltiplas plataformas pode oferecer uma visão mais abrangente e representativa do universo dos periódicos científicos.

Até o momento, não se identificam trabalhos que realizem a combinação direta entre os indicadores do GSM e os dados do OpenAlex, de forma automatizada e em larga escala. Essa integração entre GSM e OpenAlex pode adicionar novas possibilidades para análise bibliométrica, permitindo não apenas complementar e aperfeiçoar a métrica do índice h5 do GSM, mas também realizar análises temáticas, regionais e editoriais com maior profundidade. Tal abordagem viabiliza a criação de novos indicadores e ampliaria a capacidade de compreensão sobre o impacto científico, configurando-se como uma contribuição inédita e necessária para a área de estudos bibliométricos.

2 CONHECENDO AS BASES TEÓRICAS DO ESTUDO

Este tópico abordará de maneira concisa a fundamentação teórica sobre o fenômeno citação. O objetivo é discutir as questões que tornam a citação uma prática acadêmica essencial, demonstrando a relação entre a produção acadêmica e o uso da citação.

A produção acadêmica é entendida como uma atividade fim do fazer científico. Esse relato pode ser expresso por alguns meios que servem para comunicação científica, como livros, artigos, teses, anais de eventos, entre outras maneiras de registrar a pesquisa. Apesar de haver meios não formais de relatar a pesquisa, como palestras ou conversas, o enfoque do presente estudo é nos meios formais, em especial as revistas científicas (Targino, 2000).

Sobre o aspecto da área de conhecimento, o estudo está inserido no campo da Ciência da Informação, disciplina que se dedica a examinar o fluxo da informação e investigar os fenômenos relacionados a seu uso e circulação (Borko, 1968). Nesse contexto, a análise de citação torna-se uma ferramenta relevante, pois permite compreender esse fluxo no âmbito da comunicação científica, sendo, portanto, de grande interesse para a Ciência da Informação.

A citação, nesse processo, assume um papel fundamental para compreender as relações que se estabelecem no progresso do conhecimento científico. Isso ocorre porque se parte do princípio de que o conhecimento científico é cumulativo, ou seja, ele avança continuamente, construindo-se sobre descobertas e contribuições anteriores. Dessa forma, a citação permite rastrear o desenvolvimento das ideias e contribui para a legitimidade e credibilidade da produção científica, evidenciando o diálogo entre pesquisadores ao longo do tempo.

Um dos pensadores que expressa em sua teoria esta característica cumulativa do conhecimento científico é Lakatos. O autor propõe uma epistemologia constituída de programas de pesquisa, os quais têm núcleos teóricos bem definidos e aceitos por uma comunidade científica. As teorias, leis e conceitos presentes nos programas de pesquisa funcionam como alicerce para o progresso do conhecimento científico. Com isso, o autor afirma que dentro de um programa de pesquisa, o conhecimento se acumula de maneira produtiva. Significa que novas descobertas se constroem sobre as contribuições anteriores, dentro de uma lógica de validação e

refutação do conhecimento, adicionando novos insights sem descartar completamente as ideias anteriores (Lakatos, 1978).

Para que o progresso científico aconteça, é essencial que ele se apoie no conhecimento acumulado ao longo do tempo, e a citação é a manifestação explícita dessa conexão.

Isso porque, pressupõe que a citação, uma prática essencial em todas as áreas do conhecimento científico, posiciona-se na interface entre o conhecimento passado e o presente, preservando a ética científica na produção de conhecimento (Silveira; Caregnato, 2017).

Sobre o ponto de vista da sociologia da ciência, a citação é um ato de reconhecimento. Pois, para Merton (2013) o conhecimento científico não deve ser restrito, por meio de patentes e embargos autorais como as tecnologias. Dessa forma, o conhecimento deve ser público e a propriedade intelectual do cientista será o reconhecimento e estima pela comunidade.

Merton (2013) tem uma importante contribuição para o desenvolvimento de estudos de citação. Seu pensamento em caracterizar a ciência como uma instituição, que tem normas e valores que devem ser seguidos e transferidos pelos cientistas permitiu estabelecer uma relação entre citante e citado de reconhecimento.

As condutas da prática científica descritas por Merton (2013) são conhecidas como Ethos da ciência. Ethos tem quatro pilares que são:

- a) Universalismo – Ciência é impessoal, independe da identidade do cientista (gênero, etnia, nacionalidade etc.). O conhecimento científico deve ser julgado pelo seu mérito e não por quem o produziu.
- b) Comunismo – O conhecimento científico é considerado um bem público. Aos cientistas cabe o reconhecimento e estima por sua colaboração.
- c) Desinteresse – Os cientistas devem ser imparciais, evitando interesses pessoais que possam comprometer a validade das descobertas.
- d) Ceticismo Organizado – A ciência avança por meio de uma avaliação crítica e minuciosa das ideias e descobertas, o que garante que as teorias sejam testadas antes de serem aceitas.

O Ethos científico descreve princípios éticos que o cientista deve seguir. Esse entendimento é suportado pela ideia mertoniana de que a ciência funciona como uma instituição com normas e valores que os praticantes tendem a seguir. Com isso a citação tem a capacidade de representar a ciência (Romancini, 2010).

Nesse sentido, estudos de citação tornam-se valiosos, pois permitem identificar a relação de reconhecimento entre citante e citado no processo de produção do conhecimento científico.

Esse tipo de estudo permite desvendar:

a relação concebida entre os atos de citação e referenciação, expressam o vigor e a complexidade[...], compreendidos como efeitos resultantes das variadas práticas dos cientistas no domínio da Comunicação Científica. Tais atos percorrem o circuito de apropriação e uso do conhecimento acumulado e disponível, envoltos por elementos objetivos e subjetivos que interligam comunidade e literatura científicas. Como recursos sociais e cognitivos da ciência, as práticas de citações e referências se tornam fenômenos que possibilitam expressar a lógica de produção, organização, disseminação, preservação e utilização de informações que determinam as ações legitimadas e legitimadoras no/do fazer científico (Silveira, 2016).

Estudos de citação pairam sobre dois pontos de vista principais: cognitiva/normativa e social/construtivista. Ou seja, uma citação pode refletir tanto um informe teórico e metodológico de uma contribuição referenciada, quanto um aspecto mais social, com diversas motivações, por vezes bastante idiossincráticas, dos autores que citam o trabalho (Romancini, 2010).

As bases para análise quantitativa de citações são estabelecidas através do reconhecimento do autor, das normas e condutas das práticas científicas sob uma perspectiva normativa, compreendida no uso do conhecimento acumulado. O foco na mensuração possibilita entender a citação como um indicador de impacto, influência ou desempenho de um pesquisador, instituição ou país. Essa abordagem desconsidera a parte semântica da citação, representando apenas a parte sintática dessa prática (Romancini, 2010). Isso demonstra um hiato na forma de analisar, porque não consegue mensurar a qualidade pontual de um trabalho publicado.

Diante das críticas à abordagem normativa, desenvolveu-se uma nova forma de analisar a citação: a abordagem construtivista. Fundamentada em estudos microsociológicos sobre a prática científica, essa perspectiva considera as evidências do caráter plural e persuasivo das citações. Assim, as citações são vistas menos como um componente de um sistema de recompensas e mais com algo complexo e multidimensional (Romancini, 2010).

O construtivismo social entende que a posição atual da ciência resulta de um processo de negociação, no qual diferentes grupos buscam convencer uns aos outros por meio de diversas estratégias de persuasão ao comunicar suas pesquisas. Segundo essa perspectiva, o ato de citar é motivado predominantemente por

objetivos persuasivos, e não necessariamente pelo reconhecimento de mérito aos autores citados (Rousseau; Egghe; Guns, 2018).

A abordagem construtivista sugere que o contexto social em que o autor está inserido pode influenciar suas escolhas de citação, sendo o peso social e psicológico fatores que atuam sobre esse contexto. Além disso, outros elementos, incluindo fatores não-científicos, desempenham um papel na prática da citação. Assim, os motivos pelos quais os autores citam podem variar significativamente entre cientistas (Romancini, 2010).

Bornmann e Daniel (2008) associaram os diferentes níveis de agregação de dados a uma síntese entre as abordagens construtivista e normativa. Eles distinguem entre o “conhecimento local” de nível micro, caracterizado por um contexto particular e potencialmente influenciado por processos sociais (construtivismo), e o “conhecimento comum” de nível macro, que representa resultados amplamente aceitos e que formam uma zona de consenso. No nível macro, especialmente durante as fases de “ciência normal,” o conhecimento, e as citações, tendem a se alinhar mais à teoria normativa, pois os cientistas geralmente aceitam esse conhecimento como base para suas investigações.

Nesse sentido, há uma convergência com a visão de que os estudos de citação devem integrar diferentes orientações da sociologia da ciência, explorando diversas abordagens metodológicas; ainda assim, reconhece-se a dificuldade, senão a impossibilidade, de formular uma “teoria da citação” abrangente (Romancini, 2010).

Assim, mesmo reconhecendo a multiplicidade de fatores envolvidos nas citações, incluindo aspectos sociais, psicológicos, normativos e cognitivos, não há razão para desconsiderar a perspectiva normativa, pois, na elaboração de um artigo, as ações do cientista podem ser consistentes com aspectos de ambas as abordagens.

Há evidências de que as citações se mantêm consistentes e confiáveis como uma medida, principalmente quando são analisadas em aspectos mais amplos, como para instituições, nações ou grupo de revista por área (Aksnes; Langfeldt, 2025). Nesse sentido, a análise das citações de um conjunto amplo de obras, como ocorre na presente pesquisa, configura-se como um indicador de desempenho científico, minimizado pelos efeitos de fatores sociais que podem influenciar o ato de citar.

Tal abordagem justifica-se à luz da perspectiva normativa proposta por Merton, segundo a qual as citações refletem o reconhecimento legítimo de contribuições científicas. Não por acaso, os indicadores bibliométricos baseados em citações têm sido amplamente utilizados na formulação de políticas científicas e na definição de critérios de avaliação de instituições, países e pesquisadores.

Nos anos 1960-70, com a crescente competição entre cientistas e com a redução de fundos públicos para pesquisa, a avaliação científica tornou-se uma prioridade para os formuladores de políticas, que buscaram métodos para embasar decisões e políticas públicas. A análise de citações ganhou força com o surgimento dos índices de citações científicas como o Institute for Scientific Information (ISI), fundado por Eugene Garfield em 1958 (Romancini, 2010). Posteriormente, outros índices surgiram no decorrer dos anos, como Scopus e Google Acadêmico (Mugnaini; Sales, 2011).

Com a criação de índices de citação, permitiu-se criar outros indicadores, que são valores numéricos normalmente utilizados para mensurar uma unidade de análise específica em um determinado contexto. Eles normalmente são compostos por variáveis importantes relacionadas à produção científica, como produtividade (número de trabalhos publicados) e impacto (número de citações recebidas durante um período). Em alguns casos, os indicadores também podem levar em consideração a influência do periódico que faz a citação e a relevância temática entre o periódico citante e o citado (James *et al.*, 2019; Waltman, 2016).

Os primeiros indicadores foram criados com o propósito de medir a influência de periódicos científicos, posteriormente foram utilizados para analisar outros elementos da produção científica, como autores, grupos de pesquisa, laboratórios, departamentos, programas de pós-graduação, regiões e países. O aumento no número de indicadores e de suas aplicações foi impulsionado pela facilidade de cálculo automatizado dos índices por meio de bases de dados científicas (Waltman, 2016).

Os principais indicadores para periódicos incluem o Fator de Impacto e o Eigenfactor, calculados pelo Journal Citation Reports, o CiteScore, calculado pela Scopus/Elsevier, o SCImago Journal Rank, calculado pela SJ&CR com base em dados de citação da Scopus, e o índice h5, calculado pelo GSM a partir de dados de citação coletados pelo Google Scholar (Martín-Martín *et al.*, 2018; Waltman, 2016).

As aplicações desses indicadores consistem em avaliar o impacto da produção científica; ser critério de seleção para financiamento, promoção e prêmios; comparar a produção de diferentes pesquisadores, instituições e países (Waltman, 2016).

A criação de indicadores bibliométricos depende da existência de dados bibliográficos estruturados e organizados de modo a permitir a identificação das relações entre documentos citantes e citados. A seguir, será abordada a origem desses dados e os avanços ocorridos em sua organização, que culminaram na consolidação das bases de dados bibliográficas.

2.1 DADOS BIBLIOGRÁFICOS

Os conjuntos de elementos que descrevem a singularidade de uma obra são conhecidos como dados bibliográficos. Ao elaborar um registro de uma obra, como artigo, livro, tese, entre outras, é essencial incluir informações sobre o autor, título, ano de publicação, editora e outros elementos relevantes. Esses dados são fundamentais para a organização e recuperação de informações em bibliotecas e sistemas de informações. A necessidade de elaborar esses registros surgiu inicialmente nas bibliotecas.

Quando as coleções de informações se tornaram muito extensas para serem organizadas e acessadas diretamente, os bibliotecários desenvolveram registros bibliográficos como substitutos para representar itens específicos do acervo de uma biblioteca. O objetivo desses registros é descrever um documento com detalhes suficientes para identificá-lo exclusivamente entre outros documentos e especificar onde o registro pode ser encontrado dentro de um arquivo de registros bibliográficos (Clarke, 2014).

A evolução dos dados bibliográficos pode ser dividida em vários períodos históricos, cada um marcando importantes avanços na sistematização e organização da informação. Jesus (2021) em seu trabalho apresenta a cronologia desses avanços:

- a) Períodos Remotos: Abrange desde o início da necessidade de sistematização da informação até o desenvolvimento da imprensa. Foi marcado pela criação de registros rudimentares para organizar o conhecimento disponível.

- b) Período Tipográfico: Com o surgimento e estabelecimento da imprensa, este período viu a consolidação de técnicas de impressão e distribuição de informação, terminando no início da formalização dos instrumentos de padronização.
- c) Período Tradicional: Caracterizado pelo início da formalização dos instrumentos de padronização bibliográfica. Este período culminou com várias avaliações críticas desses instrumentos, buscando melhorar a precisão e consistência dos registros.
- d) Período Pré-mecanizado: Começando com a realização da Conferência de Paris e terminando com a Reunião Internacional de Especialistas em Catalogação (RIEC), este período introduziu técnicas mais sistematizadas e organizadas de catalogação.
- e) Períodos Mecanizados e de Metadados: Com a adoção de computadores e da web, estes períodos começaram após a RIEC, implementando práticas que ainda são utilizadas no universo bibliográfico atual. A digitalização e o uso de metadados transformaram a forma como os dados bibliográficos são geridos e acessados (Jesus, 2021).

Com a adoção de computadores e o avanço da web, os dados bibliográficos passaram a ser produzidos e gerenciados de forma mais granular e estruturada, possibilitando maior precisão na identificação de recursos, recuperação de informações e descoberta de registros bibliográficos (Patra, 2017). Essa transição representou uma mudança significativa em relação ao modelo anterior, no qual os registros eram mantidos em papel, restringindo seu uso e interoperabilidade.

O processo de migração incluiu o surgimento de bases em CD-ROM e, posteriormente, com o advento da internet, a transformação dos catálogos em plataformas digitais acessíveis via web, ampliando consideravelmente o alcance e a funcionalidade dos sistemas de informação (Patra, 2017).

A digitalização e interconectividade dos dados permitiram que documentos fossem vinculados por meio de links e citações, viabilizando o desenvolvimento de indicadores bibliométricos. Esse avanço tornou possível mapear a influência e o impacto de trabalhos acadêmicos em áreas específicas do conhecimento, favorecendo a análise da produção científica e das redes de disseminação do saber.

Inicialmente, as bases de dados bibliográficas acessíveis via web eram disponibilizadas mediante assinatura, sendo o Institute for Scientific Information (ISI)

o pioneiro nesse modelo. Posteriormente, surgiram outras plataformas comerciais, como a Scopus, que também oferece métricas e indicadores bibliográficos.

Assim como as publicações científicas, os dados bibliográficos eram, em sua origem, de caráter proprietário. No entanto, com o avanço do movimento de acesso aberto, emergiram iniciativas voltadas à disponibilização pública desses dados, permitindo o surgimento de novas ferramentas construídas com utilização de dados abertos.

Entre essas alternativas destacam-se o Google Scholar, que utiliza informações disponíveis livremente na web, e, mais recentemente, o OpenAlex, que se baseia em dados da Crossref e da extinta Microsoft Academic, oferecendo uma estrutura aberta e acessível para análise e integração de informações bibliográficas.

A seguir, serão apresentadas algumas características das bases de dados citadas que possibilitam a geração de indicadores bibliométricos, com ênfase especial no Google Acadêmico e no OpenAlex, que constituem o foco central deste estudo.

2.2 BASE DE DADOS BIBLIOGRÁFICAS

Para construir esses indicadores, é necessário ter acesso a um conjunto robusto de dados bibliográficos armazenados em sistemas de informação. Esses sistemas armazenam e organizam informações detalhadas sobre publicações acadêmicas, como autor, título, data de publicação, editora e citações, permitindo que os indicadores sejam calculados de maneira precisa e eficiente. A qualidade e a abrangência dos dados bibliográficos são fundamentais para garantir que os indicadores reflitam de forma acurada a influência e o impacto dos trabalhos científicos analisados.

O conceito de indexação de citações foi introduzido em 1955 por Eugene Garfield, que fundou o Institute for Scientific Information (ISI) na Filadélfia, EUA. Em 1961, Garfield lançou o "Science Citation Index" (SCI), esta iniciativa foi a precursora da Web of Science. (Mugnaini, Strehl, 2008).

O Web of Science se tornou uma ferramenta essencial para pesquisadores, bibliotecas e instituições acadêmicas, permitindo a avaliação do impacto da pesquisa, a identificação de tendências científicas e a facilitação da colaboração global. E por muitos anos não teve concorrentes sendo dominante na área.

Atualmente, a Scopus é uma das maiores bases de dados de resumos e citações de literatura acadêmica revisada por pares. Foi lançada em 2004 pela Elsevier. A plataforma foi desenvolvida para oferecer uma ferramenta abrangente e acessível para a descoberta de pesquisa científica, ajudando pesquisadores, bibliotecas e instituições a encontrar informações relevantes e medir o impacto de suas pesquisas.

Essa ferramenta se apresenta como concorrente da Web of Science e desperta interesse tanto de pesquisadores quanto de bibliotecários. Embora a Scopus seja de acesso restrito a assinantes, as métricas produzidas pela plataforma Scimago Journal & Country Rank podem ser consultadas gratuitamente. É importante destacar que os indicadores disponibilizados pelo Scimago são derivados dos dados da Scopus, porém calculados a partir de metodologias e aplicações métricas próprias.

Além disso, em 2004, Google marcou presença no mercado científico com o lançamento do Google Scholar, uma ferramenta que chamou a atenção tanto da comunidade acadêmica quanto do público em geral. O principal objetivo desse serviço era indexar o conteúdo acadêmico e científico disponível na internet, a fim de facilitar a pesquisa e recuperação de dados referentes a título, autor, resumo e palavras-chave, além de permitir o acesso a textos completos de artigos, teses, dissertações, relatórios, anais e outros documentos com origem e estrutura acadêmica (Torres-Salinas; Ruiz-Pérez; Delgado López-Cózar, 2009; Martín-Martín *et al.*, 2018).

A criação do Google Scholar está relacionada com a missão da empresa Google de organizar e tornar a informação acessível a todos. De acordo com o artigo de Gusenbauer (2019), a ideia por trás do Google Scholar foi inspirada na missão geral da empresa de "organizar a informação do mundo e torná-la universalmente acessível e útil". O mesmo autor destaca que o Google Scholar surgiu como uma resposta à necessidade de uma ferramenta de busca acadêmica abrangente, que poderia ajudar pesquisadores e acadêmicos a encontrar informações relevantes e atualizadas em suas áreas de estudo. O Google Scholar foi criado para atender a essa necessidade, oferecendo uma ampla variedade de conteúdo acadêmico em um único local e fornecendo recursos para avaliação de citações e métricas de impacto.

O Google Scholar é um mecanismo de busca acadêmico que utiliza rastreamento automatizado para indexar documentos científicos disponíveis na web.

Seus robôs percorrem continuamente sites de universidades, editoras, repositórios temáticos e institucionais, catálogos de bibliotecas e demais fontes consideradas acadêmicas, independentemente do idioma ou área do conhecimento. Ao contrário de bases como a Web of Science e a Scopus, que se limitam a listas editoriais pré-selecionadas, o Google Scholar adota uma abordagem abrangente e descentralizada, baseada na coleta de dados de fontes abertas e distribuídas pela internet (Jacsó, 2011; Gusenbauer, 2019; Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019).

O Google Scholar indexa documentos de toda a gama de tipos de documentos acadêmicos (livros, capítulos de livros, artigos de periódicos e conferências, materiais didáticos, teses, pôsteres, apresentações, relatórios, patentes, etc.). Ao contrário da natureza seletiva de Web of Science e Scopus, o Google Scholar é dinâmico; por isso reflete o melhor estado da web. (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019). É amplamente utilizado e acessível em todo o mundo. Ele permite buscar e recuperar artigos científicos em várias disciplinas, idiomas e formatos. É possível personalizar as configurações de pesquisa do Google Scholar para obter resultados mais precisos (Torres-Salinas; Ruiz-Pérez; Delgado López-Cózar, 2009)

O processo de indexação do Google Scholar é composto por três etapas básicas: a) compilação de fontes; b) localização de documentos acadêmicos, e; c) indexação. Para que um domínio acadêmico seja indexado com sucesso pelo robô indexador do sistema (Googlebot), é necessário que alguns requisitos técnicos sejam cumpridos, como a autorização de acesso do robô ao site sem bloqueios de segurança que impeçam a coleta de conteúdo, além de uma localização de metadados e arquivos de documentos facilmente identificáveis. Quando o acesso é permitido, a indexação do Google Scholar também alcança a lista de referências dos documentos, o que aumenta a funcionalidade do serviço ao permitir a análise de citações, atividade antes realizada em poucas bases de dados (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019).

Segundo Orduña-Malea, Aytac e Tran (2019) uma das principais vantagens do Google Scholar é sua ampla cobertura de fontes, que incluem não apenas publicações científicas revisadas por pares, mas também outras fontes de informação. Isso significa que o Google Scholar pode ser especialmente útil em

disciplinas onde há uma grande quantidade de informação cinzenta, ou seja, publicações que não são amplamente difundidas ou facilmente acessíveis.

O Google Scholar pode ser uma ferramenta valiosa para complementar outras bases de dados bibliográficas, especialmente em áreas de pesquisa onde a cobertura pode ser limitada (Orduna-Malea; Aytac; Tran, 2019).

No entanto, Torres-Salinas, Ruiz-Pérez e Delgado López-Cózar (2009) comentam algumas limitações do GS como a inclusão de artigos não revisados por pares e outros tipos de publicações não científicas. Também discutem a questão da precisão e qualidade dos resultados de busca do Google Scholar, apontando que é importante ter em mente as limitações da ferramenta ao usá-la para avaliação científica.

Os autores Orduna-Malea, Aytac e Tran (2019) também destacam algumas limitações do Google Scholar em comparação com outras bases bibliográficas, especialmente no que diz respeito à falta de transparência nos critérios de indexação. Essa ausência de clareza torna o sistema mais vulnerável a manipulações, dificultando que especialistas proponham melhorias para aperfeiçoar seus mecanismos de inclusão. Um exemplo desse problema é apresentado por Ibrahim *et al.* (2025), que demonstraram ser possível inflar artificialmente o número de citações para autores no Google Scholar por meio da compra de referências.

Os diversos equívocos de indexação encontrados no Google Scholar (Jacsó, 2011) têm impacto direto na precisão das análises de citação realizadas pelo sistema. Esses erros, aliados à necessidade de uma limpeza e filtragem dos dados, restringem a capacidade do Google Scholar como ferramenta autônoma para avaliação acadêmica sem a necessidade de uma prévia filtragem e processamento dos dados. Além disso, a inexistência de uma opção para baixar os dados de citação torna ainda mais difícil o uso desse sistema para fins de análise de citação (Martín-Martín *et al.*, 2018).

Por essa razão, ao utilizar o Google Scholar em pesquisas, é fundamental considerar que seu processo automatizado de indexação, ao incluir uma grande variedade de documentos, pode gerar inconsistências. Essas características aumentam o risco de recuperação de materiais sem rigor acadêmico, o que pode distorcer a compreensão do tema pesquisado e comprometer a qualidade da análise. (Halevi; Moed; Bar-Ilan, 2017).

As bases de dados bibliográficas discutidas até o momento, Web of Science, Scopus e Google Scholar, não disponibilizam seus dados bibliográficos de forma aberta ou com facilidade de extração, sendo o Google Scholar um exemplo emblemático dessa limitação. Como resultado, a gestão e o controle sobre os indicadores métricos permanecem sob responsabilidade exclusiva de seus mantenedores. Contudo, um dos avanços, impulsionados pelo movimento da Ciência Aberta, tem sido justamente divulgar dados bibliográficos abertos, ampliando as possibilidades de acesso, análise e reutilização dessas informações.

Essa iniciativa começou na 8ª Conferência Open Access Scholarly Publishing realizada em 2016, uma ação para que os dados bibliográficos sejam abertos. Seis organizações (OpenCitations, Wikimedia Foundation, PLOS, eLIFE, DataCite e Centre for Culture and Technology) fundaram a Initiative for Open Citations I4OC.

O objetivo da iniciativa é promover a disponibilidade de dados sobre citações que sejam estruturadas, separáveis e abertas. A partir de 3 de junho de 2022, a iniciativa ganhou força com a obrigatoriedade de todos os membros do Crossref disponibilizar suas referências de forma aberta, sem restrições de distribuição. Isso inclui membros que anteriormente tinham referências limitadas ou fechadas, agora configuradas para serem abertas, tornando todas as referências no Crossref metadados abertos. (Initiative for Open Citations, 2024).

Crossref é uma agência de registro de DOIs (Digital Object Identifier). Quando um editor científico trabalha com o Crossref para registrar um DOI para um documento, o editor fornece metadados desse documento ao Crossref. Esses metadados são então disponibilizados abertamente pelo Crossref. A integridade e a qualidade dos dados disponíveis no Crossref dependem do que os editores oferecem ao Crossref. O próprio Crossref não coleta e enriquece ativamente os dados (Visser; Van Eck; Waltman, 2021).

Por essa característica, o Crossref se tornou uma fonte de dados importante, utilizada por base de dados bibliográficos de dados abertos, como OpenAlex.

OpenAlex ao utilizar o Google Scholar pesquisasx é uma plataforma de dados bibliográficos que proporciona acesso livre a um vasto catálogo de periódicos, artigos científicos, autores e instituições. Esta base de dados é essencial para pesquisadores que buscam transparência e facilidade de acesso a informações acadêmicas.

Com iniciativas como a do OpenAlex, é possível explorar a produção científica de forma abrangente, facilitando a colaboração e a análise de impacto. Além disso, promove a democratização do conhecimento, tornando informações valiosas acessíveis a todos. Esta iniciativa apoia a Ciência Aberta e incentiva a inovação global.

O OpenAlex oferece maior acessibilidade em comparação com bancos de dados que requerem assinatura, como o Web of Science (WoS) e o Scopus. O surgimento dessa plataforma é essencial para promover a comunicação acadêmica e superar as lacunas digitais, tecnológicas e intelectuais que existem entre países e regiões (Zhang *et al.*, 2024).

OpenAlex é uma plataforma que incorporou os dados do Microsoft Academic, oferecendo acesso aberto a informações bibliométricas abrangentes. Isso ocorreu após o anúncio da Microsoft em descontinuar o serviço Microsoft Academic.

Em seu blog, a Microsoft anunciou que o Microsoft Academic, as APIs e os snapshots seriam desativados em 31 de dezembro de 2021. Logo depois do anúncio, a organização sem fins lucrativos OurResearch, com a missão de fornecer um catálogo completamente aberto do sistema de pesquisa global, declarou que preservaria e expandiria o último conjunto de dados completo do Microsoft Academic que inclui periódicos, livros, conjuntos de dados, teses e eventos, excluindo apenas os dados de patentes (Scheidsteger; Haunschild, 2023; Zhang *et al.*, 2024).

O OpenAlex inclui outras fontes de dados como Crossref, Wikidata, ORCID e ROR. Estas fontes são essenciais para garantir a precisão e a interconectividade das informações bibliométricas. Crossref proporciona uma infraestrutura robusta para a gestão de metadados, facilitando a rastreabilidade de citações e a identificação de pesquisadores e instituições. A inclusão de identificadores como ORCID e ROR melhora a visibilidade e a organização dos dados (Chawla, 2022).

O nome OpenAlex, em homenagem à antiga Biblioteca de Alexandria, no Egito, também visa traçar conexões entre esses pontos de dados para criar um banco de dados abrangente e interligado do sistema de pesquisa global, dizem seus fundadores. O banco de dados, lançado em 2022, é uma alternativa gratuita a plataformas baseadas em assinatura como Scopus, Dimensions e Web of Science (Chawla, 2022).

A OurResearch criou o OpenAlex, usando parte de uma doação de US\$ 4,5 milhões da instituição de caridade Arcadia Fund, com sede em Londres. O índice

está atualmente acessível através de uma interface de programação de aplicativos, ou API, que permite realizar pesquisas complexas de até 100 mil requisições por dia (Chawla, 2022).

O OpenAlex é atualizado quinzenalmente, trazendo mais dados de suas fontes. A ferramenta vai um passo além em direção à abertura do Microsoft Academic, porque os dados do OpenAlex estão disponíveis gratuitamente sob uma licença de direitos autorais CCO (licença de domínio público da Creative Commons) para qualquer pessoa usar. Isso significa que se o OpenAlex for descontinuado, qualquer pesquisador poderá continuar de onde o OurResearch parou, em vez de ter que reconstruir todo o banco de dados do zero (Chawla, 2022). Além disso, os usuários têm a opção de baixar e instalar uma cópia completa do banco de dados OpenAlex em seus próprios servidores, basta baixar o banco de dados fornecido (Zhang *et al.*, 2024).

No contexto atual, o progresso na tecnologia de bancos de dados automatizou diversas tarefas, como coleta e limpeza de dados. Embora esta tendência de automação tenha alargado o âmbito da cobertura de dados, também introduziu certos desafios de qualidade dos dados, tais como a precisão e a padronização dos dados. A falta de instituições é uma manifestação específica de problemas de dados que descreve o fenômeno da falta de informação institucional num conjunto de publicações (Zhang *et al.*, 2024). Eles acreditam que fontes de dados abertas, como Crossref e OpenAlex, oferecem oportunidades interessantes para produzir análises bibliométricas de forma totalmente transparente e reproduzível. De acordo com sua declaração oficial, OpenAlex é adotado como uma importante fonte de dados na nova versão do CWTS Leiden Ranking, que oferece informações importantes sobre o desempenho científico de mais de 1.400 grandes universidades em todo o mundo (Zhang *et al.*, 2024).

Outra vantagem destacada na literatura é a facilidade de uso dos dados disponibilizados pelo OpenAlex. Roar Bakken Stovner, pesquisador que estuda padrões de citação, relata que levou apenas cerca de duas horas para começar a trabalhar com o OpenAlex, enquanto o processo de familiarização com o Microsoft Academic Graph (MAG) exigiu aproximadamente uma semana. Segundo o autor, "para pesquisadores que desejam experimentar pequenos projetos por conta própria, o OpenAlex será muito mais fácil de começar" (Chawla, 2022).

Nesse contexto, Scheidsteger e Haunschild (2023) realizaram uma comparação detalhada entre os metadados fornecidos pelo OpenAlex e pelo MAG (Microsoft Academic Graph), avaliando quais campos são mais adequados para análises bibliométricas. Como conclusão, os autores apontaram o OpenAlex como uma alternativa mais apropriada para esse tipo de estudo.

Além disso, Rodrigues, Lopes e Batista (2023) utilizaram o OpenAlex para detectar automaticamente lacunas de citações na base Web of Science. Os resultados evidenciaram que o OpenAlex apresenta vantagens na recuperação de citações mais recentes, o que o torna uma fonte útil para estudos que envolvam atualizações rápidas da literatura.

As ferramentas mencionadas permitem a análise do impacto das publicações científicas por meio de indicadores baseados em citações. No tópico a seguir, serão detalhados os principais indicadores utilizados na avaliação de periódicos, com foco nas métricas derivadas das bases Web of Science, Scopus, Google Acadêmico e OpenAlex.

2.2.1 Fator de impacto

O Fator de Impacto é um indicador bibliométrico conhecido por ser o primeiro a ser desenvolvido e utilizado para avaliar periódicos científicos com base em análises de citações. Criado por Garfield na década de 1960, no Institute for Scientific Information (ISI), o índice foi pensado com o intuito de representar matematicamente a relação entre o número de artigos publicados e o número de citações recebidas em uma lista de periódicos selecionados na época (Garfield, 1972). É calculado como o número médio de citações recebidas por artigo publicado na revista durante os dois anos anteriores. Embora tenha sido projetado para avaliar revistas de ciências naturais e médicas, agora é usado para avaliar periódicos em uma ampla gama de disciplinas (Nederhof, 2006).

O cálculo do Fator de Impacto é feito conforme ilustrado abaixo:

Figura 1 - Ilustração cálculo do FI

$$\frac{\text{Citations in 2021 to items published in 2019 (23) + 2020 (9)}{\text{Number of citable items in 2019 (27) + 2020 (26)}} = \frac{32}{53} = 0.604$$

Fonte: Clarivate (2023).

Além do Fator de Impacto de dois anos, há também uma variação que analisa o impacto em um período de cinco anos (5 Year Impact Factor), que pode ser uma alternativa para disciplinas com desenvolvimento teórico mais lento, visto que o período de dois anos pode ser considerado curto para coleta de citações em tais casos (Nederhof, 2006).

Embora o Fator de Impacto seja o indicador bibliométrico mais utilizado em sistemas de avaliação da produção científica, ele é frequentemente criticado por sua influência excessiva na ciência mundial. Isso ocorre porque a coleta de dados de citação é limitada às publicações indexadas na Web of Science (WoS), o que gera um viés e reflete no índice. Portanto, ele é considerado elitista e limitado a uma pequena parte da ciência mundial, notadamente periódicos anglo-saxões, de idioma inglês, e de aderência temática às ciências duras (Waltman, 2016).

2.2.2 Eigenfactor Score

Desenvolvido em 2007, o Eigenfactor Score foi inicialmente concebido para avaliar o desempenho de periódicos de economia, mas rapidamente foi adotado em outras áreas do conhecimento. Ele é baseado no algoritmo PageRank, criado pelo Google, e utiliza a teoria de análise de redes sociais para medir as relações entre publicações (elementos) e citações (conexões) (Waltman; Van Eck, 2010).

Diferentemente do Fator de Impacto, o Eigenfactor Score mede o desempenho geral de um periódico e não apenas o desempenho médio por artigo. Periódicos altamente influentes em uma determinada área tendem a ter um alto índice nesse indicador (Waltman; Van Eck, 2010).

Eigenfactor Score tem como base o número de vezes que os artigos do periódico publicados nos últimos cinco anos foram citados no ano na base JCR, mas também consideram quais periódicos contribuíram com essas citações para que os

periódicos altamente citados influenciem a rede de revistas menos citadas. As referências consideradas autocitação do periódico são removidas para que as pontuações do indicador não sejam influenciadas por essa prática (Krampl, 2019).

O nES de um periódico médio é igual a 1, enquanto um nES de 5 indica que o periódico é cinco vezes mais influente do que a média da base de dados. A versão normalizada do indicador torna mais fácil a compreensão do resultado, pois em periódicos de baixo impacto, o ES apresenta valores numéricos muito pequenos, como, por exemplo, 0,00017 (Canto, 2022).

2.2.3 CiteScore

O CiteScore é baseado no banco de dados Scopus, que é uma das maiores e mais abrangentes bases de dados bibliográficas disponíveis. Ele cobre uma ampla gama de disciplinas e inclui revistas, livros, conferências e outros tipos de publicações.

O CiteScore é um indicador de desempenho do periódico que foi desenvolvido pela Elsevier. Ele é calculado usando as citações recebidas pelos artigos em um período de três anos e dividindo esse número pelo número total de artigos publicados no mesmo período, incluindo artigos de pesquisa, revisões e outros tipos de conteúdo. O CiteScore é uma métrica de citação abrangente que fornece uma visão geral do desempenho de um periódico ao longo do tempo e é amplamente utilizado por editores, autores e pesquisadores para avaliar a qualidade e o impacto de um periódico (James *et al.*, 2019).

O lançamento do CiteScore foi visto com ceticismo por alguns pesquisadores, que questionaram a necessidade de outra métrica de citação, apesar da existência de outras com a mesma finalidade (James *et al.*, 2019). No entanto, as mudanças realizadas diferenciam o CiteScore do Fator de Impacto, atendendo a algumas das reivindicações da comunidade científica. O aumento da janela de análise dos indicadores tende a favorecer um conjunto maior de disciplinas científicas, principalmente as que possuem um ritmo mais lento e sem impacto de citação imediato como Ciências Sociais e Humanas (Waltman, 2016).

Antes da nova metodologia, o cálculo do CiteScore considerava citações coletadas de qualquer publicação indexada na base, o que incluía periódicos, anais, séries de livros e publicações comerciais. Além disso, todos os documentos de todas

as seções da publicação eram inicialmente inseridos na lista de documentos citantes e citados, incluindo material editorial, cartas, notas (James *et al.*, 2019).

Em junho de 2020, a Elsevier desenvolveu uma nova metodologia de cálculo do CiteScore, aplicável para o índice de 2019. A primeira modificação significativa foi no período de análise, que passou de três para quatro anos. Além disso, todas as citações recebidas nos quatro anos de análise passaram a ser consideradas, ao contrário da metodologia anterior, que considerava apenas as citações recebidas no ano do indicador. A Figura abaixo mostra a nova metodologia de cálculo do CiteScore de 2019 (Elsevier, 2025).

Figura 2 - Fórmula do cálculo do CiteScore



The image shows a digital interface for calculating CiteScore. At the top, it says 'CiteScore 2021' with a dropdown arrow. Below this, the calculation is displayed as: $716.2 = \frac{76,632 \text{ Citations } 2018 - 2021}{107 \text{ Documents } 2018 - 2021}$. At the bottom, it states 'Calculated on 05 May, 2022'.

Fonte: Elsevier (2025).

Uma segunda modificação na metodologia do CiteScore na qual foram incluídos no cálculo documentos de seções revisadas por pares, tais como artigos científicos, artigos de revisão e trabalhos de eventos. Cartas, material editorial, notícias e notas, que geralmente não passam por processos de avaliação por pares, foram excluídos. Isso torna a análise mais robusta, considerando que esses materiais não se tratam, efetivamente, de resultados de pesquisa e não costumam receber citações. Por fim, uma última modificação foi realizada na formatação do número final do indicador, que passou a ser apresentado com apenas duas casas decimais, ao contrário de três como era anteriormente (Elsevier, 2025).

Uma das vantagens do CiteScore em relação a outras métricas de citação, como o fator de impacto, é que ele é mais inclusivo e abrange uma ampla variedade de tipos de documentos. Além disso, a Elsevier disponibiliza gratuitamente o CiteScore de todos os periódicos indexados no Scopus em seu site, permitindo que pesquisadores e estudantes tenham acesso fácil e transparente a essas informações (Okagbue; Silva, 2020).

2.2.4 Scimago Journal Rank

O SCImago Journal & Country Rank é um portal disponível publicamente que inclui os periódicos e indicadores científicos dos países desenvolvidos a partir das informações contidas no banco de dados Scopus.

Os seus indicadores podem ser usados para avaliar domínios científicos, periódicos e países, bem como comparar esses elementos entre si. Os periódicos são agrupados em áreas temáticas (27 grandes áreas), categorias (309 categorias) ou por país. Os dados de citação são extraídos de mais de 34.100 títulos de mais de 5.000 editoras internacionais (Scimago, 2023).

O SCImago Journal Rank, assim como Eigenfactor Score, é um indicador que considera pesos distintos para as citações recebidas, de acordo com o impacto do periódico citante. Ele usa o algoritmo Google PageRank. O SCImago Journal Rank, contudo, possui uma definição matemática complexa, pois o peso de uma citação depende não apenas do impacto da publicação citante, mas também de uma medida da proximidade temática entre as publicações citantes e citadas. Assim, uma citação de um periódico citante deve ser tematicamente mais próxima ao periódico citado (Gonzalez-Pereira; Guerrero-Bote; Moya-Anegón, 2010).

O indicador em questão visa proporcionar uma medição mais precisa do impacto real das publicações, incorporando duas variáveis relevantes do ponto de vista científico que não são consideradas pelo Fator de Impacto. Uma das vantagens dessa formulação é a capacidade de comparar fontes de diferentes campos de assunto, considerando que o número final é normalizado. Além disso, o acesso a esse indicador, disponível na base de dados SJ&CR, é gratuito, ao contrário dos indicadores oferecidos pelos serviços da Clarivate Analytics (James *et al.*, 2019).

No entanto, o SJR é baseado em dados de citação do Scopus, que não cobre todas as revistas em todas as áreas de pesquisa. Sua métrica pode ser afetada pelo número de revistas em uma área de pesquisa. Áreas com muitas revistas podem ter um SJR mais baixo, mesmo que as revistas individualmente tenham um impacto significativo na comunidade científica (James *et al.*, 2019).

2.2.5 Índice h

Em 2005, Jorge G. Hirsch criou o índice h com a finalidade de avaliar o desempenho científico de pesquisadores por meio da relação entre sua produtividade e o impacto de suas publicações. Esse índice é determinado pelo maior número H de trabalhos publicados que receberam pelo menos o mesmo número H de citações. Por exemplo, se um autor publicou 20 artigos e 10 deles receberam pelo menos 10 citações cada, então o seu índice h será 10. O índice h foi proposto como um novo parâmetro para comparar o desempenho científico de pesquisadores que competem pelo mesmo financiamento ou colocação profissional (Hirsch, 2005).

Uma vantagem do índice h é a sua capacidade de combinar em um único número, tanto a quantidade quanto a qualidade, as duas variáveis fundamentais da produção científica. Dessa forma, o valor do índice reflete não apenas o número de artigos publicados, mas também a qualidade desses artigos, medida pelo número de citações recebidas. O índice h estabelece uma medida de desempenho que muitos consideram mais justo do que outros indicadores (Barnes, 2017).

O índice h tem uma vantagem frente aos indicadores supracitados, porque pode ser aplicado a diferentes contextos, como avaliação de pesquisadores individuais, avaliação de departamentos universitários e avaliação de revistas científicas (Bornmann, 2014).

Nos últimos dez anos, o índice h tem ganhado cada vez mais destaque como uma ferramenta para orientar decisões no ensino superior. A utilização de um número simples e facilmente calculável para demonstrar a contribuição de um pesquisador em sua área de atuação tem sido muito atrativa para muitos gestores universitários. Há relatos de que o índice h se tornou a medida quantitativa mais popular da produtividade e impacto de um pesquisador (Barnes, 2017).

No entanto, o índice h apresenta várias limitações e é objeto de críticas na literatura. Em primeiro lugar, sua fórmula de cálculo leva em consideração apenas a parcela de artigos altamente citados, negligenciando aqueles com poucas ou nenhuma citação. Além disso, não indica o número total de citações recebidas, tornando-se inadequado para uma visão abrangente da produção científica, seja em nível individual ou de outras unidades de análise, como periódicos, instituições ou países (Bornmann, 2014; Silva; Dobránszki, 2018).

Um aspecto que parece ser exclusivo da fórmula de Hirsch é a sua tendência a favorecer pesquisadores com carreiras mais longas, mesmo estes com menor impacto, em detrimento de pesquisadores no início de suas carreiras. Isso ocorre porque o valor do índice h nunca pode ultrapassar o número máximo de trabalhos publicados pelo pesquisador (Bornmann, 2014).

Dessa forma, é possível observar que um pesquisador sênior, com 100 artigos publicados, mas apenas 10 com 10 ou mais citações, terá um índice h igual a 10. Enquanto isso, um pesquisador júnior com apenas 10 artigos publicados, todos eles com mais de 100 citações cada, também terá índice h igual a 10. Portanto, nesse caso, o critério quantidade é privilegiado em relação ao critério qualidade (impacto). Como resultado, uma análise baseada somente no valor do índice não permite detectar as diferenças entre esses dois pesquisadores sem uma avaliação mais detalhada de seus currículos (Canto, 2022).

Algumas alternativas foram propostas para minimizar as limitações do índice h . Uma proposta foi a de atribuir maior peso às citações recebidas pelos trabalhos altamente citados, que compõem o índice h . Esse grupo sugere uma variação na forma de cálculo do índice h , com destaque para o índice g . O índice G é definido como o maior número G de trabalhos que, somados, recebem G^2 ou mais citações. Com esse cálculo, o índice g será sempre igual ou maior que o índice h (Egghe, 2005). Outras alternativas como índice h_5 , que faz um corte temporal de 5 anos nas publicações mensuradas, índice i_{10} , que faz um corte utilizando obras com 10 ou mais citações, dentre outras que irão fazer médias ou medianas dos documentos que compõem o índice h .

2.2.6 Não Citados

Os índices de não citação referem-se ao conjunto de obras publicadas que não receberam nenhuma citação dentro de uma determinada janela de tempo (Kozlowski *et al.*, 2024; Schwartz, 1997). Em algumas análises, também podem ser incluídas obras com baixa contagem de citações, ampliando a compreensão sobre a dispersão do impacto científico (Frandsen; Nicolaisen, 2023). Esses documentos representam o que alguns autores descrevem como a "parte invisível da ciência", refletindo trabalhos, mesmo não sendo consenso na literatura, que não contribuíram significativamente para a geração de novo conhecimento (Kozlowski *et al.*, 2024).

Para mensurar esses casos, é comum o uso do percentual de não citados como indicador principal. Essa métrica, quando alta, expressa a concentração de citações em um número restrito de artigos, evidenciando a desigualdade na distribuição do impacto. Estudos como os de Frandsen e Nicolaisen (2023), Kozłowski *et al.* (2024), Schwartz (1997) e Dorta-González *et al.* (2020) aplicaram o percentual de não citados como medida em suas análises.

Além do percentual de não citados, o coeficiente de Gini tem sido utilizado para mensurar a desigualdade na distribuição de citações entre os artigos de um periódico. A escolha desse indicador se justifica pela sua capacidade de capturar a concentração de citações considerando todos os artigos da amostra, proporcionando uma visão mais abrangente sobre a desigualdade de impacto (Kozłowski *et al.*, 2024). Quanto maior a concentração de citações em poucos artigos, maior será o valor do coeficiente de Gini. Por outro lado, quanto mais homogênea for a distribuição das citações entre os artigos, menor será o coeficiente, indicando uma distribuição mais equilibrada.

Foram descritas, portanto, algumas das principais métricas utilizadas na análise de citações em periódicos científicos, algumas das quais serão aplicadas no presente estudo. No tópico a seguir, será abordado um aspecto fundamental que viabilizou esta pesquisa: a disponibilidade de dados abertos para pesquisa científica. Serão apresentados, de forma breve, conceitos relacionados à Ciência Aberta e aos dados de pesquisa.

2.3 CIÊNCIA ABERTA

A Ciência Aberta pode ser vista como um movimento abrangente que promove discussões sociais, políticas, econômicas e outros aspectos, além de práticas científicas transparentes e colaborativas que resultam na democratização do conhecimento científico (Drumond, 2023). Esta abordagem transformadora busca tornar a ciência mais acessível, inclusiva e colaborativa, trazendo benefícios tanto para a comunidade científica quanto para a sociedade como um todo.

Quanto mais claro for o fluxo de trabalho científico, respeitando esses princípios, mais pesquisas poderão ser reproduzidas por outros pesquisadores, independentemente de onde eles estejam. Isso pode ser alcançado por meio de diretrizes de reprodutibilidade, uso e compartilhamento de cadernos de laboratório

abertos, e disponibilização de códigos abertos online. Para o avanço da ciência, é essencial que os trabalhos estejam acessíveis para escrutínio público, ou seja, para serem testados e verificados. Isso inclui estudos de irreprodutibilidade, cadernos de laboratório abertos, fluxos de trabalho abertos, códigos abertos, diretrizes e testes de reprodutibilidade (Araújo, 2021).

A Ciência Aberta é um movimento a partir de outros movimentos, sobre a qual o conhecimento científico deve tornar-se livre para as pessoas utilizarem, reutilizarem, distribuírem sem restrições legais, tecnológicas ou sociais. Esse movimento é composto por diversas frentes que abordam aspectos diferentes, mas complementares, do fazer científico. Um dos movimentos que integram a Ciência Aberta e visto como precursor dessa discussão é o Movimento de Acesso Aberto, que tem como marco histórico a Declaração de Budapeste em 2002. (Araújo, 2021).

Segundo Drumond (2023), há alguns desafios que se associam às questões mais amplas do movimento pela Ciência Aberta. Para Drumond, esses desafios são caracterizados pelo interesse em propriedade intelectual, pela dependência de editoras privadas para a publicação científica e pelas limitações ao acesso e uso de materiais digitais impostas por diferentes licenças. Devido a esses fatores, as primeiras iniciativas do movimento da Ciência Aberta estavam focadas principalmente no fortalecimento do acesso aberto às publicações científicas.

O movimento Ciência Aberta pode ser visto, em um primeiro momento, como uma oposição à mercantilização e privatização do conhecimento científico. Ela visa atender às demandas por novos modelos de publicação. Posteriormente, além de fornecerem os resultados finais das pesquisas científicas, também disponibilizam os métodos completos, anotações, dados coletados e outros materiais que demonstrem como a pesquisa foi conduzida.

Araújo *et al.* (2021) reuniram as principais perspectivas de especialistas brasileiros sobre Ciência Aberta. Sobre esse ponto de vista, ele identificou que a Ciência Aberta é um movimento colaborativo, com foco no uso de tecnologia para o acesso e compartilhamento da pesquisa. As práticas de pesquisa são destinadas a dados abertos, acesso aberto, cadernos de pesquisa abertos, ciência cidadã, software livre e aberto, E-science, ferramentas abertas, revisão aberta, arquivos abertos, crowdsourcing, cultura hacker, curadoria digital, domínio público, publicações ampliadas, recursos educacionais abertos e redes de cooperação, pesquisa aberta. (Araújo *et al.*, 2021).

Recentemente, em 2022, a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) publicou recomendações acerca da Ciência Aberta, em sua 41ª Conferência Geral, Paris, 2021, que está alinhada com os mesmos pontos de vista de autores brasileiros.

De acordo com esta recomendação, a Ciência Aberta pode ser caracterizada como uma estrutura inclusiva que integra diversos movimentos e práticas com o objetivo de disponibilizar conhecimento científico multilíngue de forma aberta, tornando-o acessível e reutilizável para todos. Ela visa ampliar as colaborações científicas e o compartilhamento de informações, beneficiando tanto a ciência quanto a sociedade. Além disso, busca abrir os processos de criação, avaliação e comunicação do conhecimento científico para atores da sociedade além da comunidade científica tradicional. A Ciência Aberta abrange todas as disciplinas científicas e todos os aspectos das práticas acadêmicas, incluindo ciências básicas e aplicadas, ciências naturais, sociais e humanas. Seus pilares principais incluem: conhecimento científico aberto, infraestrutura científica aberta, comunicação científica, envolvimento aberto dos atores sociais e diálogo aberto com outros sistemas de conhecimento (UNESCO, 2022). A figura 3 apresenta algumas práticas que envolvem esses quatro pilares.

Figura 3 - Práticas de Ciência Aberta

1



Fonte: Adaptado de Unesco (2022).

Araujo *et al.* (2021) complementa que a Ciência Aberta se assemelha a um ecossistema que envolve diferentes perspectivas inter-relacionadas, cada uma com suas particularidades. São elas: a) filosóficas: ética, integridade, transparência; b) científicas: inovação, uso, reuso, reprodutibilidade, replicabilidade; c) sociais: rede de colaboração, ciência cidadã, compartilhamento e democratização da informação; d) tecnológicas: padronização, rastreabilidade, interoperabilidade; e) políticas: desenvolvimento de legislações e políticas públicas para promover a Ciência Aberta; f) econômicas: investimento econômico, infraestruturas de comunicação científica e negociações estratégicas de acesso à informação entre países.

É importante destacar que as práticas da Ciência Aberta devem ser acompanhadas de bons princípios. A Unesco (2022) descreve alguns desses valores:

- a) Deve respeitar a liberdade acadêmica e os direitos humanos, promovendo pesquisas de alta qualidade;
- b) Ela deve reunir múltiplas fontes de conhecimento e garantir que métodos e resultados de pesquisas sejam amplamente acessíveis para revisão e avaliação rigorosa, além de proporcionar processos de avaliação transparentes;
- c) A prática científica deve ser inclusiva, sustentável e equitativa, proporcionando oportunidades de educação científica e desenvolvimento de capacidades para todos;
- d) A Ciência Aberta deve garantir equidade entre pesquisadores de países desenvolvidos e em desenvolvimento, promovendo o compartilhamento justo e recíproco de insumos e produtos científicos;
- e) Deve assegurar acesso igualitário ao conhecimento científico para todos, independentemente de localização, nacionalidade, raça, idade, gênero, renda, circunstâncias socioeconômicas, estágio da carreira, disciplina, língua, religião, deficiência, etnia ou situação migratória;
- f) A Ciência Aberta deve incluir uma ampla gama de conhecimentos, práticas, fluxos de trabalho, línguas, resultados e tópicos de pesquisa que atendam às necessidades e ao pluralismo epistêmico da comunidade científica como um todo;
- g) Deve envolver diversas comunidades de pesquisa e acadêmicos, além do público em geral e dos detentores do conhecimento, incluindo povos

indígenas e comunidades locais, e atores sociais de diferentes países e regiões, conforme apropriado.

Com base nas definições, pilares e valores que norteiam a Ciência Aberta, abordados até o momento, observa-se que é algo fácil de se concretizar, essa nova maneira de fazer ciência. Muitas barreiras precisam ser quebradas e ressignificadas, principalmente dos atores envolvidos, para que seja possível ter como resultado uma ciência genuinamente aberta.

No entanto, algumas ferramentas tecnológicas ajudam a facilitar esse processo, pois simplificam o compartilhamento de dados e pesquisas, colaborando substancialmente para a abertura da ciência. Tecnologias baseadas na internet permitem uma representação do processo científico que vai além dos métodos tradicionais e das descrições de análises.

Segundo Araújo *et al.* (2021) há ferramentas disponíveis que são adotadas e utilizadas por pesquisadores; porém, elas ainda não se encaixam perfeitamente nos fluxos de trabalho existentes. Ferramentas como sites de redes sociais, cadernos de laboratório eletrônicos, entre outras, devem ser ampliadas para auxiliar os cientistas nas atividades cotidianas de suas pesquisas (Araújo *et al.*, 2021).

Algumas políticas já existentes colaboraram para que práticas científicas, com base em uma Ciência Aberta, sejam concretizadas. Essas medidas são fundamentais para que haja uma adesão a processos mais transparentes, colaborativos e democráticos. Drumond (2023) reuniu algumas dessas iniciativas internacionais:

- a) União Europeia: Reforma dos sistemas de avaliação da investigação e multilinguismo para aumentar a visibilidade dos resultados da investigação da UE;
- b) Alemanha: Preservação e acesso aberto dos dados científicos financiadas com recursos públicos; Acordos entre cientistas e órgãos públicos em prol do uso responsável dos fundos públicos para ciência; Hospedagem em rede de publicações científicas digitais; Gestão de dados científicos com foco nos requisitos de metadados e interoperabilidade; Plataforma de pesquisa colaborativa; Propostas de alterações em leis que tratam sobre direitos autorais;
- c) Austrália: Compartilhamento de dados; por meio de bases de dados e repositórios acessíveis; Preocupação em agregar valor aos dados

científicos; Estabelecimento de parcerias; incentivo à elaboração do plano de gestão de dados, armazenamento, disseminação e reutilização dos dados.

- d) Canadá: Interlocuções com o governo, com empresas, com organizações filantrópicas com vistas a influenciar decisões políticas e promover infraestruturas de abertura de dados ética e equitativa; desenvolvimento e acompanhamento de políticas públicas que orientem os canadenses quanto à abertura das pesquisas científicas; fornecer plataforma para os canadenses encontrarem e acessarem publicações de acesso aberto;
- e) EUA: Envolve todas as partes interessadas da sociedade civil, agências governamentais e o público americano no desenvolvimento de práticas científicas com as seguintes premissas: transparência, responsabilidade, participação e colaboração;
- f) Holanda: Gerenciamento dos dados científicos; Publicações resultantes de financiamentos pela Netherlands Dutch Research Council (NWO) devem estar em acesso aberto; Módulo de financiamento de práticas de Ciência Cidadã; Novas formas de avaliar pesquisas e pesquisadores;
- g) Portugal: Transparência em todas as etapas das pesquisas científicas; Disponibilização e reutilização de dados científicos; Utilização de ferramentas de colaboração científica;
- h) Reino Unido: Acesso aberto imediato para artigos de pesquisa, revisados por pares, podendo ser publicado em um periódico, por um editor, ou disponibilizado em um repositório, pelo próprio autor; Atribuição da licença Creative Commons (CC BY) para as publicações científicas; Estabelecimento de normalizações técnicas para periódicos e repositórios para facilitar busca, acesso, interoperabilidade e reutilização de artigos;
- i) França: Tornar obrigatória a publicação em acesso aberto de toda produção científica (artigos e dados) resultante de financiamento público; estimular práticas de Ciência Aberta, sobretudo nos programas de pós-graduação.

No Brasil, várias iniciativas têm sido promovidas em prol da Ciência Aberta, destacando-se esforços em diversas frentes para tornar a pesquisa científica mais acessível e transparente. Abaixo algumas dessas iniciativas:

- a) Scielo: Fornece acesso aberto a publicação científica. Uma iniciativa que teve início em 1998 e vem sendo mantidas no Brasil por Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq);
- b) Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT): Fomento a criação de repositórios institucionais em instituições públicas de ensino e pesquisa, por meio do Edital FINEP/PCAL/XBDB Nº 002/2009. (Drumond, 2023);
- c) Lei LEI Nº 12.527, de 18 de novembro de 2011. acesso aberto a Informação;
- d) Instrução Normativa decretos e outros instrumentos legais para a gestão dos dados abertos na administração pública (Drumond, 2023);
- e) Portaria n.º 3.679, de 07 de agosto de 2019: Recria o grupo de trabalho para elaboração de proposta de política nacional para Ciência Aberta, formado por representantes de diversos órgãos como IBICT, CNPQ e CAPES. (Drumond, 2023).
- f) Iniciativas promovidas por instituições científicas, centradas na criação e manutenção dos repositórios institucionais ou bibliotecas digitais. (Drumond, 2023);
- g) IBICT: BrCris reúne e estrutura informações relacionadas a atividades de pesquisa, projetos, publicações, pesquisadores, instituições, financiamentos e demais dados pertinentes ao cenário científico brasileiro. (Dias *et al.*, 2022), e;
- h) IBICT: Projeto Laguna que é a infraestrutura de um lago de dados científicos em acesso aberto. (Carvalho Segundo *et al.*, 2023).

Essas políticas e iniciativas em prol a Ciência Aberta, em especial os dados abertos governamentais, já revelou ser relevantes para sociedade. Um exemplo foi de a sociedade poder acompanhar, praticamente em tempo real, os casos de Covid 19 nos países, e como o os casos diminuiram conforme os números de vacinados aumentavam. Mesmo havendo desconfiança sobre os dados divulgados por alguns países, foi crucial para conhecer o panorama da pandemia mundialmente.

Independentemente de sua origem, se é de pesquisas científicas, de governos entre outras fontes, dados abertos é uma realidade. Mas, para que seu

potencial seja explorado, é imprescindível haver diretrizes que orientem a gestão destes dados. Frisa-se que Ciência Aberta engloba mais do que dados de pesquisa aberto, no entanto, por interesse desse estudo será abordado no tópico abaixo discorrido sobre esse tema específico.

2.3.1 Dados de pesquisa e o Princípios FAIR

Os dados de pesquisa são evidências e insumos coletados, observados, registrados e/ou criados para fins de análise, com o objetivo de produzir resultados de pesquisa em um estudo científico. Eles servem tanto como produto da pesquisa quanto como insumo para futuras pesquisas (Semeler e Pinto, 2019).

Podem também ser entendidos como registros factuais, como escores numéricos, registros textuais, imagens e sons, utilizados como fontes primárias para a pesquisa científica e são comumente aceitos pela comunidade científica como essenciais para a validação dos resultados da pesquisa. Um conjunto de dados de pesquisa representa de forma sistemática e parcial o tema que está sendo investigado (Monteiro; Lucas, 2019).

Os dados de pesquisa são insumos comuns no fazer científico há muito tempo. Recentemente, facilitado por avanços nas Tecnologias de Informação e Comunicação, observou um crescente interesse em tornar público os dados de pesquisa. O acesso aos dados facilita a avaliação da pesquisa, bem como, surgimento de novos estudos e aprofundamento conhecimento já existentes.

Um exemplo de pesquisa, que pode ser beneficiado com o acesso aos dados brutos, são as que sintetizam conhecimento, como revisões sistemáticas, em especial as que utilizam de metanálise. O acesso a esses dados permitirá um estudo mais detalhado, pois além de analisar estatisticamente os resultados dos artigos, poderá também refazer os estudos aumentado em número e heterogeneidade a amostra.

O movimento de abertura de dados científicos começou a emergir a partir das agências de fomento à pesquisa, tanto públicas quanto privadas. No Brasil a FAPESP é um exemplo, pois passou a exigir, pelo menos, a elaboração de um plano de gestão de dados (Drumond, 2023).

A cultura de compartilhamento de dados científicos ainda está em construção e enfrenta diversos desafios, incluindo custos inerentes, como a geração

e manutenção de identificadores, aquisição e manutenção de hardware e software, atualização e manutenção de infraestrutura física, além de treinamento e capacitação de pessoal (Laender *et al.*, 2020).

Além das barreiras de infraestrutura técnica e tecnológica que as instituições de pesquisa precisam superar para a abertura de dados, também existem dificuldades culturais entre os cientistas, como a resistência de alguns pesquisadores em compartilhar seus dados por medo de perder a prioridade de suas descobertas (Santos; Guanaes, 2018).

O auxílio financeiro às pesquisas, oferecido por agências de fomento, não é o único incentivo para a abertura dos dados de pesquisa. No caso dos periódicos científicos, suas políticas que incentivam ou exigem o compartilhamento de dados têm implicações significativas no número de pesquisadores que realmente disponibilizam seus dados. Drumond (2023) fornece a ideia que é improvável que os pesquisadores compartilhem dados na ausência de requisitos formais ou incentivos específicos.

Para entender a dinâmica da abertura dos dados de pesquisa, é crucial considerar as diferentes escalas de produção de dados científicos. A Big Science envolve grandes projetos colaborativos que naturalmente demandam altos custos, investimentos e infraestruturas bem definidas, tanto tecnológicas quanto de pessoal. Por outro lado, a Small Science refere-se a pequenos laboratórios e pesquisadores individuais ou pequenos grupos em universidades e institutos de pesquisa, que desenvolvem um grande número de projetos científicos. Este conjunto de dados originado na Small Science é metaforicamente chamado por alguns estudiosos de "cauda longa da ciência", pois gera uma quantidade expressiva e heterogênea de dados, que são mais difíceis de compreender, manipular e arquivar, enquanto os dados provenientes da Big Science seriam a cabeça da ciência (Sales; Sayão, 2019).

Na etapa de planejamento da pesquisa, o pesquisador deve considerar as questões relacionadas aos dados que serão coletados ou gerados. É essencial documentar essas informações para manter um registro das ações relacionadas aos dados, que podem ser atualizadas conforme necessário. Para isso, existe o Plano de Gestão de Dados. Embora o Plano de Gestão de Dados e os metadados ainda não sejam exigências generalizadas para o depósito em repositórios de dados, sua

importância é destacada por abordar, desde o início da pesquisa, como os dados serão tratados (Drumond, 2023).

Na perspectiva dos dados abertos, discute-se tanto as formas de armazenar esse material quanto como disponibilizá-lo para uso e reuso, levando em consideração os padrões dos dados abertos. Os dados abertos estão integrados nos processos intermediários da comunicação científica, focando nos dados coletados durante a pesquisa. Esses dados, independentemente do meio em que foram inicialmente empregados, podem fornecer diferentes insights à medida que outros pesquisadores os reutilizam. A ideia central é que os dados de pesquisa não pertencem ao periódico onde a pesquisa foi publicada, mas sim à comunidade científica. Isso envolve, portanto, ações e iniciativas relacionadas a Big Data, revistas de dados abertos, padrões de dados abertos, uso e reuso de dados abertos (Araújo *et al.*, 2021).

Nesse sentido, quando um estudo é iniciado com a premissa da Ciência Aberta, o pesquisador deve estar ciente de que, desde o planejamento da pesquisa, os dados devem ser tratados como um ativo a ser gerado e posteriormente disponibilizado em acesso aberto. A gestão dos dados científicos começa antes da coleta e, quando orientada pelos princípios FAIR, torna-se fundamental para a descoberta, compartilhamento e reutilização desses dados (Drumond, 2023).

FAIR é um acrônimo para Findable, Accessible, Interoperable e Reusable, que surgiu a partir de uma reunião realizada em janeiro de 2014 em Leiden, nos Países Baixos, intitulada "Jointly designing a data FAIRPORT". Esses princípios são considerados fundamentais para a descoberta, acesso, compartilhamento e reutilização de dados de pesquisa, além de promoverem a interoperabilidade global. Eles estabelecem diretrizes para a publicação de recursos digitais, conjuntos de dados, códigos e objetos de pesquisa de acordo com os princípios FAIR (Wilkinson *et al.*, 2018).

Nesse contexto, os princípios FAIR são considerados guias fundamentais para a gestão de dados de pesquisa. Esses princípios têm sido amplamente divulgados por comunidades internacionais nos últimos anos e, recentemente, chegaram ao Brasil, onde estão sendo adotados como mais uma abordagem de estudo e aplicação pelos gestores de dados (Henning, 2018).

No entanto, esses princípios só começaram a ser amplamente reconhecidos como pilares das boas práticas de gestão de dados após sua publicação no

periódico *Nature - Scientific Data* em 2016. Foi então que a Comissão Europeia passou a exigir planos de gestão de dados alinhados ao FAIR para todos os projetos financiados no âmbito do H2020 (Programa Horizon 2020 da Europa). Desde essa exigência, diversas agências de financiamento na Europa e nos Estados Unidos passaram a solicitar a adesão a esses princípios, tornando-se uma prática irreversível (Henning; Moreira, 2020).

Observa-se que cada letra do acrônimo FAIR corresponde a um conjunto específico de requisitos que exigem conhecimentos técnicos e estruturais distintos. Esses princípios devem ser vistos não como um padrão, mas como um guia para a preparação dos dados de pesquisa para reutilização, sob condições claramente descritas para pessoas e máquinas (Henning, 2018).

Wilkinson fez a primeira publicação em um periódico desses princípios em 2016, divulgando de forma mais detalhada o que cada letra do acrônimo expressa. No quadro 1 é apresentado esses princípios.

Quadro 1 - Princípios FAIR

Acrônimo	Detalhes
Findable	E1. metadados recebem um identificador globalmente exclusivo e persistente E2. os dados são descritos com metadados ricos (definidos por R1) E3. os metadados incluem clara e explicitamente o identificador dos dados que descrevem. E4. metadados são registrados ou indexados em um recurso pesquisável
Accessible	A1. metadados são recuperáveis pelo seu identificador usando um protocolo de comunicação padronizado A1.1 o protocolo é aberto, gratuito e universalmente implementável A1.2 o protocolo permite um procedimento de autenticação e autorização, quando necessário A2. Os metadados são acessíveis, mesmo quando os dados não estão mais disponíveis
Interoperable	I1. Os metadados utilizam uma linguagem formal, acessível, partilhada e amplamente aplicável para representação do conhecimento. I2. metadados usam vocabulários que seguem os princípios FAIR. I3. metadados incluem referências qualificadas a outros metadados.
Reusable	R1. metadados são ricamente descritos com uma pluralidade de atributos precisos e relevantes. R1.1. metadados são liberados com uma licença de uso de dados clara e acessível. R1.2. metadados estão associados à proveniência detalhada. R1.3. metadados atendem aos padrões da comunidade relevantes para o domínio.

Fonte: Adaptado de Wilkinson *et al.* (2016).

Os princípios FAIR, que promovem a descoberta, acessibilidade, interoperabilidade e reutilização dos dados, são fundamentais para a prática de dados abertos. Muitos desses critérios estão intrinsecamente ligados à arquitetura das ferramentas utilizadas para hospedar e gerenciar esses dados. Repositórios de dados são ferramentas que podem incorporar esses princípios, pois sua arquitetura é projetada para garantir que os dados sejam facilmente encontrados, acessíveis, interoperáveis com outros sistemas e reutilizáveis. Portanto, a implementação de repositórios de dados alinhados aos princípios FAIR é essencial para alcançar os objetivos de dados abertos e maximizar o valor e a utilidade dos dados de pesquisa.

Com base no exposto, a arquitetura de um repositório que segue os princípios FAIR deve ser projetada para atender aos usuários humanos por meio de uma interface web, garantir a interoperabilidade com outros sistemas e permitir a indexação de seus metadados por motores de busca. O acesso aos metadados e às bases de dados deve ser controlado para possibilitar a disseminação escalonada, considerando os diferentes níveis de sensibilidade dos dados. Os metadados devem seguir padrões que facilitem a integração, descoberta e descrição das bases de dados (Drumond, 2023).

Importante destacar que esses princípios vêm justamente para minimizar o desafio em lidar com dados de pesquisa. Esses dados, por sua natureza variada, podem incluir diferentes formatos, estruturas e tipos de informação, o que complica a padronização e a uniformidade dos metadados. Cada conjunto de dados pode ter requisitos específicos de descrição, dessa forma padrões universais são fundamentais à interoperabilidade entre sistemas. Essas complexidades tornam o processo de criação de metadados uma tarefa desafiadora, que demanda tempo, recursos e um alto nível de expertise para garantir que os dados sejam bem descritos, acessíveis e reutilizáveis.

No entanto, dados bibliográficos, que também são dados de pesquisa, principalmente quando provenientes de listas bibliográficas de publicações científicas, têm uma vantagem. Há vários anos esses dados são objeto de interesse das áreas Biblioteconomia e Ciência da Informação, como resultado, padrões de metadados para dados bibliográficos já estão consolidados. Isso facilita a integração, descoberta e reutilização desses dados.

2.4 GOOGLE SCHOLAR METRICS

O GSM foi aberto ao público em abril de 2012 como um ranking de publicações acadêmicas, calculados a partir dos dados do Google Scholar. Essa ferramenta fornece uma maneira fácil de identificar as publicações mais influentes (periódicos, anais, artigos) e pesquisadores (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019).

O Google Scholar representou uma alternativa interessante às outras bases de dados bibliográficas como Web of Science e Scopus. Isso porque o GSM cria rankings de periódicos com base no índice h, uma possibilidade frente ao Fator de Impacto, o Scimago que apresentam outras métricas (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019).

O índice de impacto utilizado no GSM para as publicações é o h5, que corresponde ao índice h de Hirsch com um período de cinco anos. O período de análise inicial foi de 2007 a 2011, com citações coletadas até abril de 2012. O sistema também calcula a mediana h5, métrica que corresponde à mediana do número de citações dos artigos que integram o índice h5 (Delgado López-Cózar; Cabezas-Clavijo, 2012; Orduna-Malea; Delgado López-Cózar, 2014).

A partir da página inicial do GSM, é possível explorar as publicações indexadas de três maneiras: listas por idioma, listas por categorias/subcategorias de assuntos e pesquisa pelo título da publicação. A lista das publicações por idioma, onze idiomas, destaca-se por apresentar as 100 publicações de maior impacto. Lista por categoria funciona exclusivamente para o idioma inglês. O resultado máximo recuperado para lista por categoria e pesquisa por título é de 20 publicações.

Cada publicação indexada no GSM tem seus trabalhos citantes identificados para compor o índice h5 (h5-core), cujos registros são apresentados no formato tradicional do Google Scholar, com informações de autor, título do trabalho e da publicação e ano. A inclusão de publicações no GSM é feita de forma automatizada, desde que atendidos os critérios mínimos de pelo menos 100 trabalhos publicados no período de avaliação e pelo menos uma citação recebida, ou seja, índice h5 maior que 0 ($h5 > 0$) (Canto, 2022).

Ainda há incerteza em relação ao escopo e à abrangência do GSM, ou seja, o número total de publicações indexadas. Isso ocorre porque a Google não disponibiliza dados quantitativos sobre o sistema nem uma lista completa de publicações (Gu; Blackmore, 2017).

Mesmo não sendo conhecido com exatidão sua totalidade, o GSM se destaca por sua cobertura mais abrangente nas grandes áreas como as Ciências Humanas e Sociais, publicações além do inglês e países emergentes. Isso torna o sistema relevante no processo de avaliação científica, preenchendo uma lacuna deixada pela JCR e Scopus na geração de indicadores de impacto (Bornmann *et al.*, 2016).

Um dos benefícios mais evidentes do GSM é a sua gratuidade. Outra vantagem do uso da ferramenta é a facilidade de uso, pois todas seguem um mesmo padrão de interface e recursos de navegação. Os usuários já estão familiarizados com esses recursos a partir do uso do tradicional motor de busca. Além disso, o reconhecimento da marca contribuiu para um bom posicionamento dos serviços oferecidos pela Google (Jacsó 2012; Ordunã-Malea; Delgado López-Cózar, 2014).

A falta de transparência é apontada como um problema decorrente da não divulgação do número de periódicos indexados ou de uma lista geral (master list), (Delgado López-Cózar; Cabezas-Clavijo, 2012). Ademais, é possível que periódicos que atendam aos requisitos técnicos e de impacto não sejam indexados sem razão

aparente ou à falta de padronização de entradas por título pode levar à duplicação de registros no sistema (Costa; Canto; Pinto, 2020).

A indisponibilidade de um recurso de busca por ISSN, considerando a existência das variações de escrita e abreviações possíveis para o título de um periódico é uma falha da base. Segundo Costa, Canto e Pinto (2020) o uso do ISSN poderia solucionar o problema da falta de padronização dos registros como observado na sua pesquisa.

Além disso, a limitação de exibir apenas 20 periódicos por categoria temática ou por busca de título dificulta a realização de análises métricas comparativas. Também não são divulgados os critérios utilizados para a classificação das categorias de assunto das publicações indexadas, nem se esse processo envolve alguma decisão humana ou se ocorre exclusivamente de forma automatizada. Outro aspecto negativo é a impossibilidade de calcular o índice h5 com a exclusão das autocitações, conforme apontado por Delgado López-Cózar e Cabezas-Clavijo (2013).

Ademais, a impossibilidade de extração de dados diretamente do sistema, o que dificulta a realização de estudos mais amplos sobre a cobertura temporal impede análises mais amplas e dificultam a mensuração da produção científica em nível institucional (Orduna-Malea; Aytac; Tran, 2019).

A seguir, são descritos os trabalhos publicados entre 2012 e 2025 que abordaram o uso do Google Scholar Metrics em estudos relacionados a periódicos acadêmicos. A seleção dos estudos foi realizada por meio das bases Web of Science e Scopus, sendo incluídos apenas os trabalhos que efetivamente utilizaram o GSM para análise de periódicos, com a exclusão daqueles que abordavam o tema sobre outros aspectos.

O trabalho de Jacsó (2012) foi um dos primeiros a realizar uma análise aprofundada das funcionalidades e limitações do GSM. O autor destacou os aspectos técnicos da ferramenta, incluindo o cálculo do índice h5 e a organização dos rankings por área temática e idioma. A pesquisa enfatizou as inconsistências nos dados, a cobertura irregular das revistas e a carência de documentação técnica detalhada, apontando fragilidades que afetam a confiabilidade do GSM como fonte para avaliação de periódicos.

Ainda em 2012, Delgado López-Cózar e Robinson-García questionaram a inclusão de repositórios institucionais nos rankings do GSM, o que contrariava o foco

original da ferramenta, que era avaliar periódicos científicos. O estudo criticou a metodologia do GSM ao apontar a falta de critérios claros na seleção de fontes e a dificuldade de replicabilidade dos resultados. A análise trouxe à tona a preocupação com a mistura de diferentes tipos de fontes no sistema, o que comprometeria a comparabilidade entre os rankings de revistas (Delgado López-Cózar; Robinson-García, 2012).

Em 2013, Jacsó retomou o tema, agora propondo a necessidade de personalização das categorias temáticas dentro do GSM para atender melhor às necessidades dos usuários finais. O estudo envolveu a análise de 134 periódicos, com uma comparação adicional de 50 outros títulos, explorando como a segmentação das áreas poderia ser mais ajustada aos interesses específicos dos pesquisadores. A pesquisa destacou a limitação do GSM em oferecer classificações refinadas e adaptáveis, sugerindo melhorias para aumentar sua aplicabilidade em avaliações bibliométricas personalizadas (Jacsó, 2013).

Túñez López, Mateos Martín e Pablos Coello (2013) abordaram o GSM dentro de uma discussão mais ampla sobre as mudanças no paradigma da avaliação científica. Embora o foco principal tenha sido a evolução das métricas digitais e de acesso aberto, o GSM foi citado como um exemplo emergente de fonte alternativa de mensuração de impacto, especialmente por sua gratuidade e fácil acesso em comparação com bases comerciais.

No mesmo ano, Delgado López-Cózar e Cabezas-Clavijo (2013) realizaram uma análise comparativa entre o GSM, o Journal Citation Reports (JCR) e o SCImago Journal Rank (SJR). O estudo destacou que, apesar das limitações metodológicas do GSM, como a ausência de critérios claros de indexação e a fragilidade na normalização dos títulos de periódicos, a ferramenta oferecia uma alternativa de baixo custo para avaliação de periódicos. A pesquisa também estimou que o GSM cobria cerca de 40.000 periódicos.

Cabezas-Clavijo e Delgado López-Cózar (2013) focaram na subárea biomédica, comparando o índice h de periódicos presentes no GSM com os índices disponíveis na Web of Science e Scopus. O estudo revelou variações significativas entre os diferentes sistemas de indexação e questionou a confiabilidade do GSM para avaliações de impacto, especialmente para revistas de medicina intensiva. A extração dos dados foi realizada manualmente, evidenciando também as

difficultades operacionais do uso do GSM para análises bibliométricas mais complexas.

Em 2014, Orduña-Malea e Delgado López-Cózar realizaram um estudo longitudinal sobre a evolução dos rankings do GSM, analisando dados de aproximadamente 1.000 periódicos em dez idiomas diferentes. O trabalho explorou a estabilidade e o crescimento dos rankings ao longo do tempo, apontando a necessidade de melhorias na transparência e atualização do sistema. O método de extração consistiu em consultas de manuais ao site do GSM, com análise comparativa entre edições anuais da plataforma (Orduña-Malea e Delgado López-Cózar, 2014).

Khurshid (2014) aplicou o GSM para comparar o desempenho de periódicos sauditas de odontologia, medicina e farmácia, considerando 36 periódicos. O autor destacou a utilidade do GSM para pesquisadores de países em desenvolvimento, ressaltando, porém, limitações quanto à cobertura e à categorização temática dos periódicos indexados.

Em 2015, Spiroski analisou o impacto científico de instituições biomédicas macedônias, utilizando o GSM como uma das fontes de dados. O autor selecionou seis periódicos da Macedônia indexados na plataforma e comparou suas métricas com outras fontes bibliográficas. O estudo evidenciou o papel do GSM como ferramenta acessível para pesquisadores de países com menor representação em bases tradicionais (Spiroski, 2015).

Packer (2015) investigou indicadores de foco nacional nas pesquisas publicadas em periódicos de saúde pública da América Latina. O estudo utilizou o GSM como fonte complementar para identificar os 245 periódicos analisados, com destaque para 10 títulos da subárea de saúde pública. A pesquisa reforçou o potencial do GSM para apoiar análises regionais de produção científica, mesmo com suas limitações metodológicas.

Sims *et al.* (2016) exploraram o uso do GSM para identificar 27 periódicos de medicina de emergência com o objetivo de avaliar a promoção de registros de ensaios clínicos nessas publicações. O trabalho reforçou o papel prático do GSM como ferramenta para seleção de periódicos em revisões sistemáticas e estudos de políticas editoriais.

Walters (2016) conduziu uma análise sobre fontes de informação e indicadores bibliométricos para avaliação de periódicos na área de ciências sociais.

O GSM foi incluído como uma das plataformas examinadas, com foco nas vantagens e limitações para a avaliação de revistas dessa área. O estudo destacou a cobertura ampla, mas também apontou as dificuldades na obtenção de dados detalhados para análises aprofundadas.

Repiso Jiménez-Contreras e Aguaded (2017) analisaram revistas ibero-americanas de educação indexadas tanto na base SciELO quanto no GSM. O estudo utilizou o índice h5 do GSM como um dos indicadores para medir o impacto dessas revistas, evidenciando a sobreposição e divergência entre os periódicos listados nas duas fontes. A abordagem combinou dados de ambas as plataformas para avaliar a visibilidade e o impacto das publicações da área educacional.

Walters (2017) aprofundou a discussão sobre rankings de periódicos baseados em citações, incluindo o GSM como um dos objetos de análise. O autor avaliou aspectos metodológicos essenciais, como cobertura de periódicos, critérios de inclusão e impacto das diferenças entre os principais sistemas de ranqueamento, incluindo GSM, JCR e SJR. A abordagem crítica destacou as limitações de cada ferramenta e reforçou a necessidade de cautela ao comparar métricas provenientes de diferentes fontes.

No mesmo ano, Gu e Blackmore (2017) buscou consolidar atributos essenciais para a caracterização de periódicos acadêmicos na era digital, integrando dados de cinco grandes bases bibliográficas: Ulrichs, JCR, SJR, Cabells e GSM. O GSM foi utilizado como uma das fontes para a variável de impacto acadêmico, por meio do índice h5, compondo, junto com JCR e SJR, o conjunto de indicadores empregados na construção do atributo "Scholarly Impact". A inclusão do GSM visou ampliar a cobertura de periódicos analisados, considerando sua disponibilidade aberta e seu alcance em diferentes áreas e subáreas do conhecimento.

Camacho *et al.* (2019) apresentaram uma metodologia para avaliar o impacto de periódicos latino-americanos por meio da integração de dados provenientes do Scopus, GSM e Latindex. A proposta consiste em um indicador que reflete o grau de influência dos artigos publicados em periódicos regionais na comunidade científica. Essa abordagem permite a organização e a classificação dos periódicos indexados no Latindex, oferecendo uma representação mais clara do impacto desses veículos no contexto da produção científica da América Latina.

Lotbiniere-Bassett, Riva-Cambrin e McDonald (2018) utilizaram o GSM como critério para selecionar os 20 principais periódicos de neurocirurgia com base no

índice h5, com o objetivo de analisar políticas de conflitos de interesse. O estudo adotou uma abordagem transversal, destacando a aplicação prática do GSM para a escolha de amostras em pesquisas sobre políticas editoriais e transparência em publicações científicas.

Ahmad *et al.* (2019) investigaram a influência científica de periódicos sauditas, utilizando o GSM para identificar os títulos com maior índice h5 entre 2012 e 2017. A análise destacou o crescimento da produção científica saudita e apontou o GSM como uma ferramenta útil para pesquisadores interessados em avaliar a visibilidade e o impacto regional das publicações.

Niforatos, Narang e Trueger (2020) analisaram conflitos financeiros de interesse em periódicos de medicina de emergência, utilizando o GSM para identificar as principais revistas da área. A seleção foi baseada no ranking de índice h5, evidenciando mais uma vez a utilização prática do GSM para a escolha de periódicos em estudos de políticas editoriais e integridade científica.

Costa, Canto e Pinto (2020) analisou o impacto dos periódicos brasileiros de Ciência da Informação com base na metodologia proposta pelo novo Qualis, utilizando como indicadores o Fator de Impacto (JCR), o CiteScore (SJR) e o índice h5 do GSM. O trabalho destacou o uso do GSM como alternativa viável para avaliação de periódicos sem cobertura em bases tradicionais, embora tenha identificado inconsistências na classificação preliminar do Qualis e apontando limitações, como a necessidade de tratamento cuidadoso dos dados extraídos do Google Scholar por meio do Publish or Perish.

O estudo de Pinto *et al.* (2020) teve como objetivo mapear e caracterizar os periódicos científicos brasileiros indexados no GSM. Utilizando um *script* automatizado complementado por busca manual, os autores identificaram 1.063 periódicos nacionais presentes no GSM, equivalentes a 55% dos títulos consultados na base UlrichsWeb. O GSM foi utilizado como fonte para obtenção do índice h5 e da mediana h5, permitindo descrever o impacto dessas publicações, com destaque para a predominância de periódicos das grandes áreas de Ciências Humanas e Sociais e para a baixa concentração de periódicos com h5 superior a 10. O estudo também apontou limitações e inconsistências do GSM, recomendando cautela em seu uso para fins avaliativos formais.

O estudo de González-Pardo, Repiso e Arroyo-Almaraz (2020) realizou uma análise das revistas ibero-americanas da área de Comunicação, focando em sua

presença e características em diferentes bases de dados, incluindo o GSM. O GSM foi utilizado para identificar a cobertura e o impacto das revistas, sendo a segunda base com maior número de periódicos indexados, atrás apenas do Latindex. A pesquisa comparou a representatividade dos periódicos por país, periodicidade, antiguidade e volume de produção, destacando o perfil editorial majoritariamente vinculado a instituições educacionais e com frequência semestral.

A pesquisa de Diaz *et al.* (2021) comparou diferentes métricas de impacto em periódicos da subárea de psiquiatria, incluindo o GSM como uma das ferramentas analisadas. A pesquisa discutiu a correlação entre o índice h5 do GSM e outras métricas tradicionais, reforçando o papel complementar da plataforma para avaliação de periódicos em contextos multidimensionais de análise de impacto.

O estudo de Walters (2022) apresentou um guia para docentes das subáreas de Administração e Economia sobre as principais fontes e métricas utilizadas na avaliação de periódicos acadêmicos. O GSM foi abordado como uma das bases de dados utilizadas para análise de impacto de periódicos, sendo destacado por sua ampla cobertura e pelo fornecimento de indicadores como o índice h5 e a mediana h5. No entanto, o autor ressaltou limitações importantes do GSM, como a ausência de dados em nível de artigo para a maioria dos periódicos e o fato de suas métricas serem dependentes do tamanho da publicação, recomendando cautela em seu uso para avaliações institucionais ou individuais.

O estudo de Koli, Kulkarni e Shetty (2022) teve como objetivo analisar os fatores que influenciam o tempo entre a conclusão de estudos, a submissão, a aceitação e a publicação de artigos em periódicos de ciências médicas. O GSM foi utilizado como uma das principais fontes para classificar os periódicos por nível de impacto (alto, médio, baixo e predatório), com base no índice h5 e na mediana h5. O GSM foi especialmente importante para avaliar periódicos não indexados em bases tradicionais, sendo aplicado também na diferenciação de periódicos predatórios, aos quais foram atribuídas pontuações zero quando não apresentavam dados no sistema.

O estudo de Canto *et al.* (2022) teve como objetivo analisar a cobertura de periódicos latino-americanos e caribenhos no GSM, identificando a quantidade de títulos indexados por país e comparando-os com outras bases como Scopus, Journal Citation Reports e SciELO. O GSM foi utilizado como principal fonte para mapear o impacto das publicações, por meio do índice h5 e da mediana h5, obtidos via busca

automatizada com *script* Python. O trabalho destacou o GSM como uma alternativa relevante para análise de impacto na região, apesar de suas limitações quanto à recuperação e filtragem de dados e à ausência de controle bibliográfico.

O estudo de Canto *et al.* (2023) teve como objetivo analisar a autocitação de periódicos ibero-americanos a partir do índice h5 do GSM. Utilizando a ferramenta bibliométrica *Gsm_hdata*, os autores investigaram 4.049 periódicos indexados simultaneamente no Latindex e no GSM, recalculando o índice h5 sem autocitações para avaliar seu impacto. O GSM foi a fonte principal para extração do índice h5 e identificação das autocitações, permitindo calcular taxas por país e área temática. Os resultados indicaram que a autocitação média foi baixa e que a exclusão prévia das autocitações no cálculo do índice h5 não se mostrou necessária, embora o monitoramento de casos extremos seja recomendado.

O estudo de Costa *et al.* (2023) teve como objetivo analisar os níveis de correlação entre diferentes indicadores de impacto aplicados a periódicos da América Latina e Caribe. O GSM foi utilizado para extrair o índice h5 por meio de um *script* automatizado em Python, compondo o conjunto de indicadores analisados ao lado de métricas como Impact Factor, CiteScore e SJR. Os autores avaliaram a correlação entre esses indicadores, observando que o índice h5 do GSM apresentou boa relação com outras métricas, apontado como um indicador adequado para avaliação de periódicos na região.

O estudo de D'Almonte *et al.* (2024) teve como objetivo verificar o uso do índice h na classificação de periódicos brasileiros da área de Comunicação e Informação. Para isso, foram extraídos, via GSM, os dados de 20 revistas, incluindo o índice h5, total de citações e metadados como autoria, URL e ano. O GSM foi a fonte utilizada para medir o impacto dos periódicos e calcular as taxas de autocitação. A análise evidenciou problemas como erros de metadados em duas revistas, o que levou os autores a sugerirem a complementação do uso do índice h com avaliações qualitativas para uma análise mais justa da produção acadêmica.

Johnson *et al.* (2024) utilizaram o GSM para selecionar 100 periódicos da subárea de otorrinolaringologia para um estudo comparativo sobre políticas de compartilhamento de dados. O trabalho utilizou Scimago Journal & Country Rank como fonte para seleção de amostras e complementou com o uso dos GSM.

Harlianto e Harlianto (2024) investigaram pagamentos da indústria de saúde aos membros do conselho editorial de periódicos de medicina de emergência,

utilizando o GSM para identificar as revistas com maior índice h5 na área. O estudo destacou possíveis conflitos de interesse relacionados ao financiamento de editores por empresas do setor.

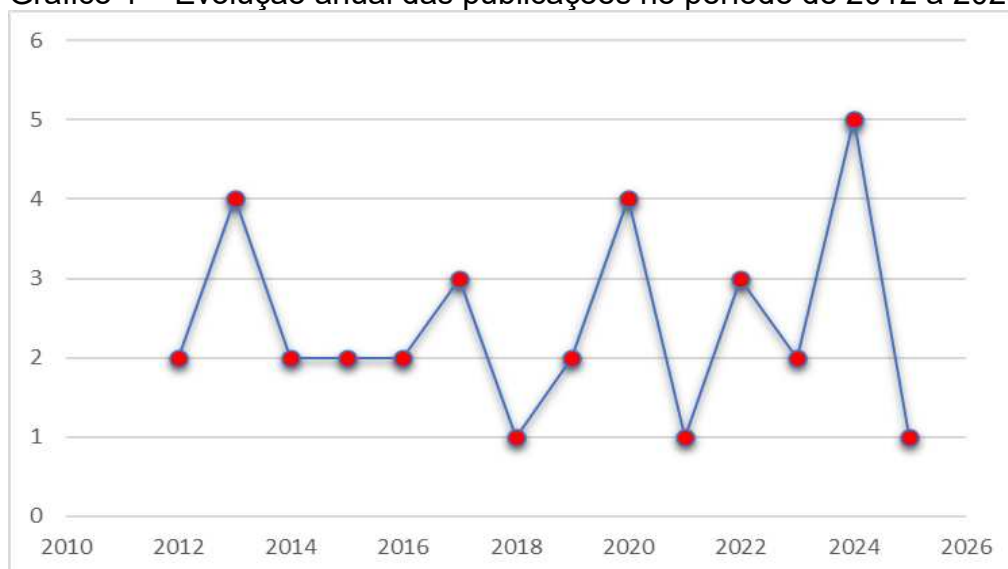
Em 2024, Canto apresentou a ferramenta "Gsm_hdata", desenvolvida para facilitar a análise de dados extraídos do GSM. O estudo descreveu o processo de coleta e organização dos dados do GSM, além de demonstrar a aplicação da ferramenta em análises bibliométricas. A proposta visou superar as limitações de acesso e manipulação dos dados do GSM, facilitando o uso por pesquisadores da área (Canto *et al.*, 2024).

Ainda em 2024, Hagood *et al.* analisaram a adesão de periódicos da medicina de emergência a diretrizes de reporte e registro de ensaios clínicos, novamente utilizando o GSM para selecionar os títulos mais influentes com base no índice h5. A pesquisa reforçou a utilidade prática do GSM como ferramenta de amostragem para estudos de políticas editoriais (Hagood *et al.*, 2024).

Hughes *et al.* (2025) avaliaram lacunas na transparência de políticas editoriais de periódicos biomédicos, utilizando o GSM para identificar os principais periódicos da área com base em seu desempenho no índice h5. A análise abordou tópicos como a adoção de políticas de conflito de interesse, registro de ensaios clínicos e transparência editorial, destacando o papel do GSM como fonte inicial para identificação de revistas de alto impacto.

Complementando esse panorama, foram identificados 34 artigos recuperados nas bases Web of Science e Scopus. A escolha dessas duas fontes para o levantamento bibliográfico justifica-se por seu caráter multidisciplinar, pela ampla cobertura internacional e pela criteriosa curadoria dos periódicos indexados, fatores que contribuem para assegurar a qualidade das publicações e minimizar a inclusão de periódicos predatórios. O Gráfico 1 apresenta a distribuição desses artigos por ano de publicação.

Gráfico 1 – Evolução anual das publicações no período de 2012 a 2025



Fonte: Dados da pesquisa.

Em síntese, os estudos revisados demonstram a utilização do GSM como fonte de dados para análise de periódicos acadêmicos ao longo dos anos. Apesar das limitações metodológicas frequentemente apontadas, como a falta de controle bibliográfico, inconsistências nos metadados e restrições na recuperação de dados, o GSM tem se consolidado como uma alternativa relevante, especialmente para pesquisadores de regiões com menor representação em bases comerciais tradicionais. Sua ampla cobertura e acessibilidade têm permitido desde análises descritivas de cobertura e impacto até estudos mais avançados sobre autocitação, correlação entre métricas e políticas editoriais. Esses trabalhos evidenciam o papel do GSM na ampliação das possibilidades de avaliação de periódicos, contribuindo para o avanço das pesquisas bibliométricas em contextos nacionais e internacionais.

3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo apresenta a caracterização e delineamento adotados para o desenvolvimento da pesquisa. Busca-se explicitar de forma clara e objetiva para garantir a transparência e a reprodutibilidade, atendendo aos critérios científicos de rigor e consistência.

3.1 CARACTERIZAÇÃO DA PESQUISA

A pesquisa em questão adota teorias e leis da Ciência da Informação, com ênfase nos fundamentos dos estudos métricos da informação, visando solucionar problemas concretos. Trata-se, portanto, de uma investigação aplicada, pois busca que os resultados sejam aplicados na prática para resolver situações reais (Marconi, 2018).

Além disso, conforme observado por Santaella (2010), a principal motivação da pesquisa aplicada reside em contribuir para a resolução de problemas específicos, utilizando conhecimentos já disponíveis. Contudo, os efeitos dessa aplicação podem ir além da simples resolução imediata, proporcionando uma ampliação do entendimento sobre o problema e levantando novas questões a serem exploradas em pesquisas futuras.

No que se refere aos métodos de procedimento, destacam-se as diversas opções que o pesquisador pode adotar em função do problema e do objeto de estudo. Entre essas opções estão os métodos histórico, estatístico, comparativo, tipológico, etnográfico e experimental, entre outros, conforme explicitado por Lakatos e Marconi (1992).

O fenômeno a ser estudado é a prática acadêmica de referenciar conhecimentos anteriormente publicados, na qual é possível quantificar os eventos e aplicar o método estatístico para representar o fenômeno. Conforme Lakatos e Marconi (2007), esse método possibilita descrever quantitativamente a sociedade, permitindo comprovar relações entre fenômenos e obter generalizações acerca de sua natureza, ocorrência ou significado.

Nesse contexto, a pesquisa busca detalhar fenômenos, situações, contextos e eventos, especificando as propriedades, características e perfis dos objetos analisados. O objetivo é coletar ou medir dados de modo a identificar variáveis

relevantes, o que, segundo Hernández Sampieri, Fernández Collado e Baptista Lucio (2013), constitui característica essencial da pesquisa descritiva. Volpato (2007) destaca que a novidade deste método reside na forma como o fenômeno é descrito, ou seja, por meio de um prisma ou abordagem até então não realizada.

Como já destacado, o método descritivo visa aprofundar o conhecimento sobre a natureza do fenômeno, sua composição e os processos que o constituem ou nele ocorrem. O pesquisador busca compreender e interpretar a realidade sem exercer influência que a modifique, concentrando-se na descoberta e observação dos fenômenos para descrevê-los, classificá-los e interpretá-los (Rudio, 1986).

Outra característica da pesquisa descritiva é identificar as possíveis variáveis que podem estar relacionadas ao objeto principal, por meio de medições e coleta de dados alinhados ao objetivo do estudo (Michel, 2009; Hernández Sampieri, Fernández Collado e Baptista Lucio, 2013).

As pesquisas descritivas de natureza quantitativa, como é o caso do presente estudo, fundamentam-se na quantificação como meio para garantir a validade das generalizações, adotando um modelo único de investigação inspirado nas ciências naturais (Chizzotti, 2003).

A análise dos dados permitirá formar métricas capazes de identificar padrões de ocorrência e, assim, explicar fenômenos. Pretende-se, com o estudo, “[...] organizar, sumarizar, caracterizar e interpretar os dados numéricos coletados” (Martins e Theóphilo, 2009, p. 107), os quais serão estudados com base em técnicas da estatística descritiva.

Os dados analisados são provenientes de citações encontradas no GSM e na base OpenAlex, ambas reconhecidas por fornecer métricas acadêmicas. Esse tipo de dado, segundo Vanti (2002), constitui objeto de investigação da bibliometria, caracterizando a presente pesquisa como um estudo bibliométrico.

Uma característica comum nas pesquisas quantitativas é a apresentação dos resultados em gráficos, tabelas ou índices, que permitem ao pesquisador analisar e relatar de forma mais objetiva e clara o fenômeno estudado. O presente estudo pretende expor os dados coletados utilizando esses recursos, com o intuito de facilitar a compreensão e potencializar a análise crítica do objeto de estudo.

3.2 DELINEAMENTO DA PESQUISA

No que se refere ao delineamento da pesquisa, este estudo adota como referência o método utilizado na tese defendida por Canto (2022). Na referida tese, foi proposto um *framework* composto por três dimensões: teórica, tecnológica e analítica. Neste tópico, serão discutidos os aspectos em que as pesquisas apresentam convergências, bem como suas principais divergências.

A dimensão teórica é a que apresenta maior alinhamento entre os dois estudos, uma vez que ambos compartilham o mesmo objeto de investigação. Dessa forma, conceitos e definições consolidadas, como os relacionados a índices e métodos estatísticos, são comuns a ambas as pesquisas. As distinções, entretanto, incluem o enfoque na autocitação presente na pesquisa de Canto (2022), enquanto o presente estudo cria outras métricas como artigos não citados, índice de Gini e índice h5.

Quanto à dimensão analítica, a presente pesquisa concentra-se nas métricas de citação, sem considerar os fatores sociais que as influenciam, abordagem adotada por Canto (2022), que desenvolveu um protocolo para análise de autocitação em periódicos indexados no GSM. Na parte analítica foram verificados a relação entre as métricas de citação com as não citadas, média de citação e relação entre os índices h e h5 do OpenAlex e o GSM.

No que se refere à dimensão tecnológica, a tese apresentou uma ferramenta com quatro procedimentos complementares. O primeiro é *Gsm_hsearch*, responsável por coletar os títulos periódicos do GSM. O *Gsm_hscite* é a funcionalidade destinada à identificação e análise de autocitações. Já o *Gsm_hcitation* tem como objetivo a extração de dados de citação dos artigos que compõem o índice h5. Por fim, o *Gsm_hline* é o recurso utilizado para coletar informações referentes às edições anteriores do GSM (Canto, 2022).

Dentre os quatro procedimentos mencionados, a presente pesquisa apresenta resultados semelhantes aos gerados pelo *Gsm_hsearch*. Trata-se de um aprimoramento dessa ferramenta, uma vez que o novo código incorpora comparações e tentativas de busca que não eram realizadas anteriormente. Além disso, a estrutura de arquivamento foi modificada, adotando o salvamento em arquivos individuais.

Adicionalmente, foi criado um *script* específico para validação dos dados extraídos do GSM, o qual verifica se os títulos dos artigos das revistas possuem correspondentes na base OpenAlex por meio do diretório *Works*.

Assim, do ponto de vista técnico, os *scripts* se complementam e permitem uma integração para a coleta automatizada dos dados. A integração das abordagens viabiliza estudos longitudinais, avaliações de autocitação e análises bibliométricas dos periódicos de forma mais completa.

Importante destacar que ambos os estudos propõem um *framework*, no entanto seus objetivos são diferentes. A seguir, no quadro 2, são apresentado de forma sucinta os procedimentos adotados para alcançar cada objetivo específico definido, sendo que uma descrição mais detalhada será fornecida ao longo do texto.

Quadro 2 - Síntese dos procedimentos para alcançar os objetivos do estudo

Objetivos específicos	Procedimentos
Desenvolver um processo automatizado para coleta, certificação e integração de dados do Google Scholar Metrics com base nos dados bibliométricos do OpenAlex	Foi desenvolvido um sistema para extrair dados de revistas e artigos indexados no OpenAlex. O sistema realiza a coleta e o arquivamento estruturado dos dados, permitindo a criação de conjuntos de dados, obtido via OpenAlex e outro via extração do GSM. Esses conjuntos são então comparados automaticamente, possibilitando a certificação dos dados coletados do GSM.
Validar o processo de coleta de dados.	A validação do processo de coleta de dados foi realizada a partir da replicação do procedimento proposto por Canto (2022), com o script <i>Gsm_hdata</i> . Foram aplicadas rotinas de verificação de consistência, identificação de falhas de coleta e comparação de títulos entre extrações, assegurando a confiabilidade dos dados obtidos.
Verificar as características editoriais dos periódicos recuperados do Google Scholar Metrics com base nos dados do OpenAlex.	As informações editoriais das revistas foram extraídas do OpenAlex e integradas aos dados coletados do GSM, possibilitando uma análise das revistas com foco em suas características editoriais.
Analisar os dados bibliométricos dos periódicos do Google Scholar Metrics e do OpenAlex.	Integrar os dados do GSM aos do OpenAlex por meio do identificador, campo ID, da revista, possibilitando a realização de análises bibliométricas combinadas entre as duas bases.
Desenvolver indicadores bibliométricos complementares para o <i>framework</i>.	A partir dos dados do OpenAlex, foram elaboradas métricas complementares considerando o total de artigos recuperados pelo procedimento <i>OpenAlex_data</i> .

Fonte: Elaborado pelo autor.

3.3 UNIVERSO DA PESQUISA

O universo da pesquisa foi definido a partir dos registros de periódicos indexados na base de dados OpenAlex. A obtenção da lista de periódicos utilizados ocorreu por meio do *download* dos dados disponibilizados pela plataforma em fevereiro de 2024. Esses dados estavam compactados e organizados em seis conjuntos distintos: *Sources* (periódicos, anais, séries, etc.), *Works* (artigos, livros, entre outros), *Authors*, *Institutions*, *Concepts* (termos de indexação) e *Publishers*.

Para compor a lista de periódicos, foram extraídos os arquivos pertencentes ao diretório *Sources*. Os registros estavam segmentados e distribuídos em múltiplos arquivos dentro de diferentes pastas, o que exigiu uma etapa de compilação. Para

isso, foi desenvolvido e aplicado um *script*¹ com o objetivo de consolidar os registros em um único arquivo no formato CSV.

O total inicial identificado foi de 250.686 registros, englobando periódicos, anais de eventos, séries de livros, entre outros tipos de publicação. Com o intuito de refinar o conjunto de dados, foram aplicados filtros para selecionar apenas os registros classificados como periódicos (Journal) que contivessem ISSN e informação sobre o país de origem. Registros que apresentavam ausência dessas duas informações foram desconsiderados.

Após a filtragem, obteve-se um conjunto final de 103.212 registros de periódicos, os quais constituem a base utilizada no *framework* proposto por este estudo.

A aplicação de uma seleção por revista com ISSN justifica-se pela relevância dessa informação na identificação das publicações, especialmente quando há a necessidade de cruzamento de dados com outras fontes, como a Web of Science. Já o critério relacionado ao país permite mapear a distribuição geográfica dos periódicos, fornecendo subsídios para análises regionais e mapeamento de possíveis concentrações editoriais.

Figura 4 - Mapa dos países que constituem o universo da pesquisa



¹ Python versão 3

Fonte: Elaborado pelo autor

O mapa ilustra a distribuição geográfica das 103.212 revistas, indicando, por meio da intensidade das cores, a concentração de periódicos indexados por país, quanto mais intensa a cor, maior o número de publicações registradas.

3.4 COLETA DE DADOS E MATERIAIS UTILIZADOS

A coleta de dados desta pesquisa foi realizada a partir de fontes abertas disponibilizadas pelo OpenAlex, por meio de consultas automatizadas ao GSM via URL, além da utilização de dados provenientes do estudo de Canto (2022), que serviram como base para validação do procedimento de extração. Os dois primeiros constituem os conjuntos de dados centrais da análise, enquanto o terceiro foi empregado como parâmetro de conferência e consistência.

No caso do OpenAlex, duas abordagens são possíveis para a obtenção dos dados: o uso da API pública ou o *download* integral de sua base de dados. Neste estudo, optou-se pela segunda alternativa, que, embora exija maior capacidade de processamento, 376 GB de dados compactados, permite uma maior autonomia e velocidade na coleta. Essa limitação pode ser contornada por meio da adaptação do script para utilização da API, caso se deseje uma abordagem mais leve e incremental.

Para a construção do primeiro conjunto de dados da pesquisa, foram utilizados os registros bibliográficos provenientes do diretório *Sources*, que reúne informações de periódicos, anais, séries e outras publicações. Esses registros encontram-se organizados em múltiplos diretórios, contendo arquivos fragmentados.

A fim de viabilizar o tratamento e a análise sistemática dos dados, foi desenvolvido um *script*². Essa rotina de coleta de dados realiza a leitura automatizada dos arquivos JSON, extrai os campos relevantes e consolida os registros em um único arquivo no formato CSV. Esse formato foi escolhido por sua compatibilidade com ferramentas de planilha eletrônica, facilitando a visualização, a manipulação e a análise subsequente dos dados bibliográficos extraídos tanto por humanos como por máquinas.

² Python versão 3 Bibliotecas: CSV, JSON e OS.

Cabe destacar que os registros bibliográficos do OpenAlex são compostos por diversos campos. No entanto, nem todos apresentam aderência aos objetivos desta investigação. Por esse motivo, procedeu-se à seleção criteriosa dos campos, desconsiderando, por exemplo, informações como o link para o Wikidata, que não contribuem diretamente para os fins analíticos do estudo.

Com base nos dados extraídos, foi elaborado um arquivo TXT contendo os títulos dos periódicos identificados. Essa lista constituiu a base de entrada para a etapa de busca no GSM, servindo como referência para verificar a presença dos respectivos títulos na base e permitir o cruzamento entre os dados obtidos no OpenAlex e aqueles recuperados no GSM.

A partir da lista de periódicos elaborada com base nos dados do OpenAlex, foram desenvolvidos dois *scripts* para viabilizar a coleta e certificação dos dados extraídos do GSM. O primeiro *script* foi responsável por realizar a extração automática de informações diretamente do GSM, recuperando, para cada título de periódico consultado, os seguintes elementos: título conforme apresentado no GSM, índice h5, mediana do índice h5, URL da revista, e, quando disponíveis, os três artigos mais citados listados pelo sistema. O segundo *script* teve o objetivo de realizar a certificação dos dados coletados, por meio do cruzamento com os registros previamente extraídos do OpenAlex. Essa etapa teve como finalidade verificar se o periódico identificado no GSM correspondia efetivamente ao título da lista de entrada, validando assim a confiabilidade da correspondência entre as bases.

Ambos os *scripts* utilizados nesta etapa da pesquisa resultam de uma parceria informal entre este pesquisador, vinculado à Universidade Federal de Santa Catarina (UFSC), e o Prof. Dr. Marcos Talau³, da Universidade Tecnológica Federal do Paraná (UTFPR), responsável pela elaboração dos códigos voltados à automação dos procedimentos descritos.

Ainda, *scripts* auxiliares foram desenvolvidos pelo pesquisador com apoio da ferramenta ChatGPT, visando realizar tarefas pontuais, como compilação de dados distribuídos em múltiplos arquivos CSV, contagem de caracteres, entre outras operações de suporte ao processamento dos dados.

³ Lattes: <http://lattes.cnpq.br/4522157219119759>.

O primeiro *script* (GSM_search) foi implementado em Python⁴. A coleta de dados foi realizada em ambiente web, por meio do navegador Mozilla Firefox, com automação do processo de inserção dos títulos das revistas no campo de pesquisa do GSM.

Cabe destacar que o GSM disponibiliza URLs específicas para consulta, organizadas conforme o idioma do país de acesso. Para garantir maior abrangência e padronização linguística, foi adotado o link correspondente à versão em inglês, considerando sua prevalência como idioma de publicação científica em periódicos acadêmicos.

Para evitar bloqueios por parte do servidor do GSM durante a execução das consultas automatizadas, o GSM_search foi configurado com um temporizador aleatório entre as requisições, simulando o comportamento humano na navegação. Essa estratégia foi fundamental para manter a estabilidade do processo de raspagem ao longo do tempo, minimizando o risco de detecção como atividade automatizada.

Considerando a lentidão decorrente dessa medida, a execução do programa foi distribuída em três computadores distintos, operando simultaneamente com IPs diferentes. Esse método permitiu ampliar a capacidade de coleta sem comprometer a integridade do processo ou violar as restrições impostas pelo servidor. Com essa abordagem, foi possível concluir a coleta completa dos dados no GSM em um período aproximado de 20 dias.

Os dados recuperados foram armazenados em arquivos de texto individuais, agrupados em um único diretório, sendo que cada arquivo recebeu uma denominação única, composta por um fragmento do título do periódico, seguido pela data e hora da consulta, garantindo assim o controle e rastreabilidade.

Nessa primeira etapa foi construído um conjunto de dados coletados do GSM. No entanto, esses poderiam conter inconsistências, seja pelo procedimento adotado ou por falhas de indexação do GSM. Como por exemplo os casos de títulos homônimos ou similares, gerando possíveis duplicidades, ou revistas indexadas erroneamente pelo processo de indexação do GSM.

Para minimizar o risco de inconsistências e garantir que os dados extraídos do GSM correspondessem, de fato, a periódicos presentes no OpenAlex, foi

⁴ Python versão 3, Bibliotecas SYS, Selenium, Datetime, OS, Argparse, Random, DiffliB, NLTK e Unidecode.

desenvolvido um segundo *script* (GSM-ALEX_verification) com a finalidade de comparar os títulos dos artigos recuperados no GSM com os registros bibliográficos disponíveis no OpenAlex. Essa etapa de certificação exigiu a estruturação dos dados em um formato mais eficiente, uma vez que a comparação direta utilizando arquivos CSV revelou-se demasiadamente lenta, o que comprometeria a viabilidade do estudo.

Como solução, foi adotado o uso da biblioteca DuckDB, um sistema de gerenciamento de banco de dados relacional otimizado para alto desempenho em consultas analíticas sobre grandes volumes de dados. O DuckDB permite a execução de consultas SQL diretamente em arquivos, utilizando a memória virtual da máquina, sem a necessidade de carga prévia em bancos tradicionais, o que se mostrou particularmente adequado ao contexto da pesquisa.

Com base nessa abordagem, foi desenvolvido o GSM-ALEX_verification, gerado em linguagem Python⁵, cuja principal função consistiu em comparar os dois conjuntos de dados, os artigos extraídos do GSM e os registros de artigos do OpenAlex, gerando como resultado uma lista de revistas certificadas.

O GSM-ALEX_verification realizou também consultas relacionais entre os artigos raspados do GSM e os artigos indexados no OpenAlex, identificando correspondências entre comparação textual dos títulos. Quando uma correspondência era encontrada, a revista associada era considerada validada e os dados eram registrados em um arquivo de saída.

Os dados do OpenAlex utilizados nessa etapa foram provenientes do diretório *Works*, que reúne registros bibliográficos de artigos, livros e outros tipos de produção científica. Para realizar a extração dos dados do *Works*, foi utilizado um *script* (OpenAlex_data) em Python⁶ para coleta de alguns campos dos registros dos artigos publicados entre 2018 e 2023, período correspondente ao intervalo considerado pelo GSM para o cálculo do índice h5. Essa extração resultou em um subconjunto do diretório *Works*, composto exclusivamente pelos dados considerados pertinentes à análise proposta.

O diretório *Works* constitui o maior conjunto de dados do OpenAlex, com aproximadamente 343 GB compactados e cerca de 249 milhões de registros. Devido ao seu volume, a extração completa exigiria mais de 3 TB de espaço em disco. Para

⁵ Python versão 3, Biblioteca DuckDB, as bibliotecas SYS, GLOB, CSV e Strftime.

⁶ Python versão 3, Bibliotecas CSV, GZIP, JSON, GLOB e OS

contornar essa limitação, adotou-se uma estratégia baseada na descompactação sequencial de subdiretórios, seguida do processamento imediato e exclusão dos dados temporários, o que demandou elevada capacidade de memória RAM.

Para viabilizar esse procedimento, foi utilizado um supercomputador com 80 GB de memória RAM e 20 núcleos de processamento, equipados com processadores Intel® Xeon® CPU E5-2420 a 1.90 GHz, o que possibilitou a execução da tarefa de forma contínua e eficiente. Para fins de escalabilidade e menor custo computacional, a API do OpenAlex representa uma alternativa viável ao processamento integral da base local.

Como resultado, obteve-se uma base consolidada de aproximadamente 32 milhões de registros de artigos, armazenada em um arquivo CSV, a qual foi utilizada como base de comparação com os dados coletados no GSM por meio do GSM_search.

3.5 GSM-ALEX_metrics (*FRAMEWORK*)

A presente pesquisa utiliza um *framework* híbrido, denominado GSM-ALEX_metrics, estruturado a partir da combinação de componentes conceituais, técnicos e analíticos, com o objetivo de orientar tanto a automação da coleta de dados no GSM quanto a análise estruturada dessas informações. A escolha por esse formato justifica-se pela natureza aplicada do estudo.

De acordo com a classificação apresentada por Macedo e Souza (2022), o *framework* híbrido integra elementos conceituais, teóricos e práticos, sendo indicado quando a pesquisa busca simultaneamente compreender um fenômeno e propor soluções operacionais. Leshem e Trafford (2007) reforçam que o *framework* conceitual constitui um componente essencial em pesquisas acadêmicas por fornecer coerência teórica e lógica interna ao estudo, funcionando como um "mapa do território" investigado. Essa estrutura permite organizar os conceitos centrais, estabelecer suas inter-relações e orientar a formulação dos objetivos e métodos de análise.

Complementarmente, Tomhave (2005) define *framework* como uma estrutura fundamental que articula pressupostos, valores e práticas, incluindo diretrizes para implementação. Essa definição ressalta sua utilidade como ferramenta aplicada, sobretudo em projetos voltados ao desenvolvimento de

sistemas. O autor ainda diferencia *frameworks* de modelos e metodologias, situando-os como estruturas com flexibilidade operacional, mas com maior aplicabilidade prática do que modelos puramente teóricos.

O GSM-ALEX_metrics desenvolvido nesta pesquisa é composto por quatro módulos interdependentes, que abrangem desde a coleta automatizada até a análise bibliométrica e a visualização dos resultados, permitindo uma abordagem sistematizada e adaptável à investigação de periódicos indexados no GSM e OpenAlex.

O Módulo 1, conceitual, delimita os principais conceitos e variáveis analíticas que sustentam a pesquisa, como índice h, média de citações e percentual de não citados. Este módulo estabelece a base conceitual que orienta as demais etapas do GSM-ALEX_metrics, garantindo consistência na coleta e na interpretação dos dados.

O Módulo 2, GSM-ALEX_data, é responsável pela extração dos dados diretamente da interface do GSM. A coleta é realizada por meio de técnicas de *web scraping*; verificação dos dados por meio do cruzamento entre os títulos dos artigos coletados via GSM e os títulos dos artigos registrados no OpenAlex; e integração dos dados coletados no GSM com os campos extraídos do OpenAlex, como ISSN, área temática, país da revista, editora e status de acesso aberto.

O Módulo 3, Tratamento dos dados, que engloba a consolidação, incluindo deduplicação, normalização e exclusão de registros inconsistentes. Os dados provenientes dos módulos anteriores são integrados em arquivos unificados (CSV), prontos para compartilhamento e visualização. Este módulo também contempla a geração de representações gráficas, como gráficos, tabelas dentre outros modos de representar os dados coletados.

O Módulo 4, Analítico aplica aos dados estruturados para o cálculo de indicadores de impacto e distribuição científica. Entre os indicadores considerados estão: índice h e h5 (do GSM e do OpenAlex), média de citações por artigo e percentual de artigos não citados. A análise também contempla segmentações por país, área do conhecimento e editora, permitindo identificar padrões de concentração e visibilidade dos periódicos indexados.

Abaixo será apresentado o fluxograma do GSM-ALEX_metrics, com propósito de representar visualmente os procedimentos propostos.

Figura 5 - Fluxograma do GSM-ALEX_metrics



Fonte: Elaborado pelo autor

Conforme ilustrado na imagem, o GSM-ALEX_metrics apresenta uma estrutura híbrida, evidenciada pela organização modular de sua arquitetura. O módulo 1 representa o aporte teórico da pesquisa, responsável pela fundamentação conceitual que sustenta todas as etapas subsequentes. Os módulos 2 e 3, concentram-se na dimensão prática e técnica, abrangendo desde a coleta automatizada de dados até o tratamento, validação e estruturação das informações obtidas. Por fim, o módulo 4 contempla a etapa analítica, na qual os dados estruturados são interpretados por meio de indicadores bibliométricos.

Antes de apresentar com mais detalhes dos módulos que compõem o GSM-ALEX_metrics, é necessário detalhar o funcionamento das fontes e justificar a adoção das abordagens propostas. O desenvolvimento dos softwares para o *framework* demandou um entendimento sobre a estrutura e os mecanismos de operação do OpenAlex (que tem os dados estruturados) e em especial o GSM, que não é amigável para coleta de dados em larga escala. O estudo do funcionamento do GSM foi para orientar e definir os recursos computacionais necessários, estimar o tempo envolvido no processo e alinhar as estratégias adotadas aos objetivos da pesquisa.

- a) O GSM não tem API disponível, sendo os dados recuperados por meio da sua estrutura em HTML, com recursos para localizar os campos que se pretende coletar e isso dificulta o procedimento e o torna mais moroso. Não obstante, caso o HTML seja alterado, o *script* precisará ser

atualizado porque o Xpath pode não corresponder aos dados selecionados;

- b) O GSM monitora as requisições de consulta em sua página e, ao identificar um volume excessivo de acessos, ativa mecanismos de restrição, como o uso de CAPTCHA para verificar se o usuário é humano. Caso as requisições persistam, o sistema pode bloquear temporariamente o acesso do IP. Observou-se, contudo, que IPs vinculados a instituições de ensino possuem uma tolerância maior antes de serem bloqueados pelo GSM. No caso da pesquisa, não ocorreram bloqueios no procedimento.
- c) O layout da página é bastante simples, apresentando poucas informações. Basicamente, exibe uma tabela com, no máximo, 20 resultados, contendo o título, o índice h5 e sua mediana, além de links para os artigos que compõem o índice h5. A inclusão de informações adicionais, como o ISSN ou outra característica única que identifique o periódico tornaria o processo de coleta mais eficiente.
- d) Não há recursos disponíveis do próprio GSM que facilitem o processo de coleta ou análise dos dados em larga escala. Existe uma categorização temática que lista algumas revistas, no entanto nem todos os periódicos recebem essa classificação, e a lista fica restrita a 20 resultados por consulta, não permitindo expandir a análise para um grupo maior.
- e) A possibilidade até aqui encontrada para uma busca em larga escala no GSM, que permitiu uma análise maior do que tem indexado na ferramenta, foi por meio da sua URL. Por esse procedimento é possível realizar a busca pelo título da revista de forma automatizada.

Dessa forma, a adoção de um *framework* híbrido em módulos permite integrar aspectos conceituais, técnicos e analíticos, garantindo a consistência dos dados e replicabilidade do processo. O detalhamento que envolve os quatro módulos do GSM-ALEX_metrics é apresentado nos tópicos abaixo.

3.5.1 Conceitual

O Módulo Conceitual constitui o alicerce teórico e interpretativo do *framework* proposto, sendo responsável por delimitar os principais conceitos,

variáveis e fundamentos que orientam tanto a etapa de coleta quanto a análise dos dados. Os conceitos aqui sistematizados foram selecionados com base em sua relevância para a compreensão da dinâmica de visibilidade, circulação e impacto de periódicos científicos indexados no GSM.

A citação é entendida como um ato social que, quando quantificado, pode representar uma forma de reconhecimento acadêmico. Sua análise permite mensurar o alcance e a influência de uma publicação por meio de técnicas de estatística descritiva, as quais servem de base para a construção de diversos indicadores bibliométricos. Entre eles, destaca-se o Fator de Impacto, bem como o índice *h*, que busca mensurar simultaneamente a produtividade e o número de citações, refletindo o equilíbrio entre quantidade e relevância científica.

A seguir, será apresentado um quadro com os principais autores e conceitos de referência que sustentam a fundamentação teórica deste módulo, compondo a base sobre a qual se estrutura o GSM-ALEX_metrics desenvolvido nesta pesquisa.

Quadro 3 – Conceitos base do GSM-ALEX_metrics

Conceito	Definição	Fonte
Abordagem Normativa	Citação representa reconhecimento e mérito.	Merton, 2013
Índice h	Relação entre o número de documentos publicados e o número de citações recebidas, baseada em uma proporcionalidade entre produtividade e impacto.	Hirsch, 2005
Área temática	Os padrões de citação são influenciados pelas características da área. As medidas de citação em geral, não são, sem normalização, apropriadas para comparações entre campos.	Glänzel; Moed, 2002
Não citados	Publicações não citadas são aquelas que, após sua publicação, não receberam nenhuma citação dentro de um determinado período de observação.	Egghe e Rousseau (1990)
Amplitude	Medida de dispersão que indica a diferença entre o maior e o menor valor de um conjunto de dados.	Barbetta, 2002
Média aritmética	Valor médio obtido pela soma de todos os valores dividida pelo número de observações	Rousseau; Egghe; Guns, 2018
Mediana	Valor central de um conjunto de dados ordenado.	Rousseau; Egghe; Guns, 2018
Percentis	Valores que dividem os dados em 100 partes iguais.	Rousseau; Egghe; Guns, 2018
Correlação	Medida entre dois conjuntos de variáveis de acordo com os valores por elas assumidos. Spearman.	Rousseau; Egghe; Guns, 2018
Índice de Gini	Na bibliometria permite identificar assimetrias, como a predominância de poucos periódicos, responsáveis por grande parte da produção ou impacto científico de um campo de pesquisa.	Rousseau; Egghe; Guns, 2018
Distribuição dos dados	Refere-se às características visuais e estatísticas do modo como os dados estão distribuídos em relação ao eixo horizontal.	Barbetta, 2002
Variável aleatória contínua	Aplica-se a um intervalo dos números reais que não conseguimos enumerar seus possíveis resultados, por estes formarem um conjunto infinito.	Barbetta, 2002

Fonte: Elaborado pelo autor.

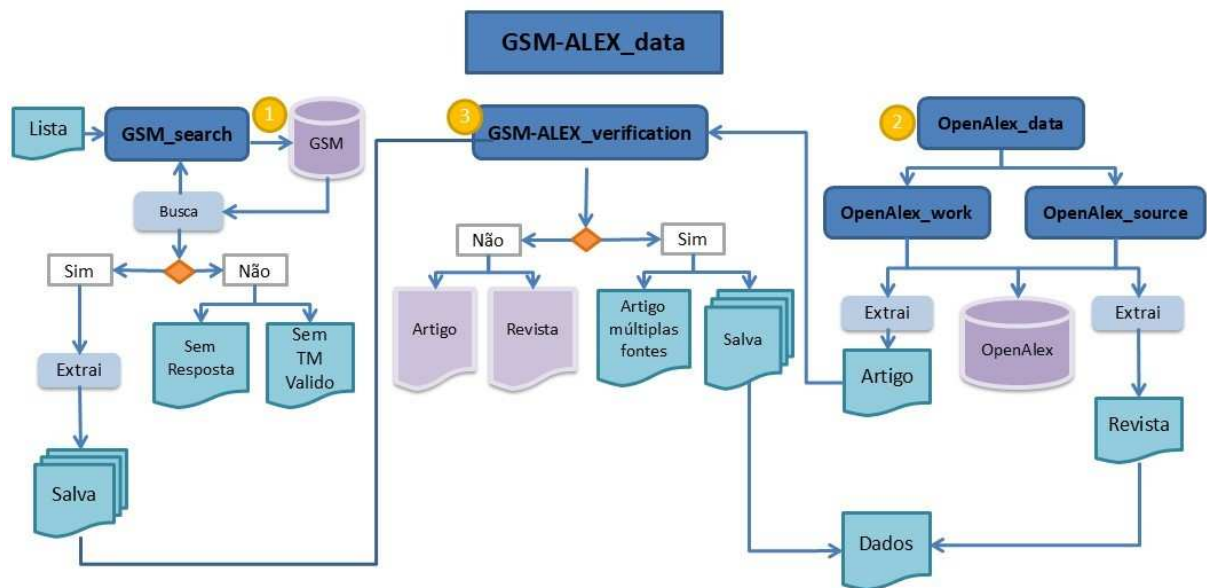
Ao reunir conceitos, este módulo oferece o suporte necessário para interpretar os dados coletados e orientar as análises subsequentes com base em referências consolidadas na literatura.

3.5.2 GSM-ALEX_data

O módulo GSM-ALEX_data constitui a etapa inicial do GSM-ALEX_metrics proposto, sendo responsável pela extração, verificação e integração dos dados utilizados nas análises subsequentes. Trata-se do componente operacional do processo, encarregado pela coleta automatizada das informações que fundamentam a pesquisa.

O procedimento tem início com o GSM_search (1), dedicado à raspagem dos dados diretamente da interface do Google Scholar Metrics (GSM). Em seguida, o OpenAlex_data (2) realiza a extração dos registros provenientes dos diretórios *Sources* e *Works* da base OpenAlex. Por fim, o GSM-ALEX_verification (3) é executado para verificar a correspondência entre os títulos dos artigos coletados no GSM e aqueles indexados no OpenAlex. A seguir, apresenta-se um fluxograma que sintetiza as etapas e os procedimentos empregados neste módulo.

Figura 6 - Integração das etapas 1,2 e 3 do módulo GSM-ALEX_data



Fonte: Dados da pesquisa.

Na imagem, é possível visualizar como os processos estão inter-relacionados dentro do módulo GSM-ALEX_data, evidenciando a articulação entre suas etapas. A fase GSM-ALEX_verification exerce um papel central ao estabelecer a conexão entre os dados extraídos pelo GSM_search e os registros presentes no OpenAlex_data. Além de verificar a existência de correspondências entre os títulos dos artigos coletados no GSM e aqueles indexados no OpenAlex, essa etapa

também adiciona o identificador (ID) da revista proveniente do OpenAlex. Esse procedimento permite integrar as informações coletadas em cada etapa, resultando em um arquivo consolidado no formato CSV.

Nesse arquivo, o campo 'TM' indica o critério utilizado para a inclusão da revista na amostra, enquanto 'Título da revista' corresponde à lista de entrada original. O campo 'GSM título' registra o nome do periódico conforme coletado na interface do GSM, e 'Papers URL' armazena o endereço eletrônico da página da revista nessa plataforma. Os campos 'Paper1', 'Paper2' e 'Paper3' contêm os títulos dos três primeiros artigos extraídos do GSM, e 'File1' identifica o nome do arquivo salvo localmente para cada revista.

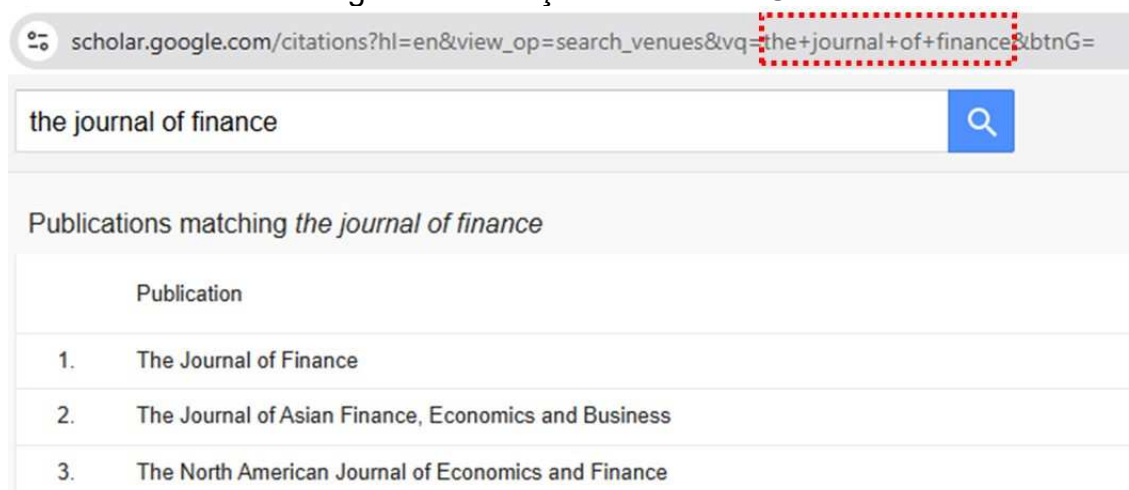
Na sequência, os campos 'Title1', 'Title2' e 'Title3' referem-se aos títulos dos artigos correspondentes localizados na base OpenAlex, enquanto 'Primary_location_id1', 'Primary_location_id2' e 'Primary_location_id3' registram os respectivos identificadores das revistas associadas. Por fim, os campos 'Primary_location_display_name1', 'Primary_location_display_name2' e 'Primary_location_display_name3' apresentam os títulos padronizados das revistas conforme indexados no OpenAlex. Abaixo será descrito o funcionamento de cada etapa envolvida no módulo GSM-ALEX_data.

3.5.2.1 GSM_search

O GSM_search é responsável por realizar a consulta automatizada de títulos de periódicos no GSM. Por meio desse procedimento, viabiliza-se a coleta em larga escala de dados diretamente da interface do GSM, permitindo a extração de informações como o título da revista no GSM, índice h5, mediana do índice h5, URL do periódico, e, quando disponíveis, os três artigos mais citados que compõem o índice h5.

A coleta é automatizada e realiza consultas através da URL do GSM com o título da revista. O título da revista é inserido dinamicamente em um segmento da URL, simulando uma consulta realizada pelo usuário. Por exemplo, para o periódico The Journal of Finance, o GSM_search insere esse título diretamente na URL. A resposta retornada é uma página HTML estruturada em tabela, da qual as informações são extraídas por meio de técnicas de web scraping. A Figura 6 ilustra esse processo de inserção e exibição dos resultados na interface HTML do GSM.

Figura 7 - Inserção do título na URL



Fonte: Adaptado pelo autor do GSM, 2024.

O arquivo de entrada do GSM_search é composto por uma lista de títulos de periódicos, um por linha, totalizando mais de 100 mil registros. Para garantir a continuidade do processo em caso de interrupções, o GSM_search possui um sistema de memória, permitindo a retomada da coleta a partir do último título processado.

Diversas estratégias foram implementadas para minimizar erros de busca relacionados à variação nos títulos. A entrada foi padronizada, priorizando caracteres alfanuméricos, mas mantendo parênteses, hifens e o símbolo “&”. Quando o título era formado por uma única palavra, aplicavam-se aspas duplas, recurso que demonstrou aumentar a precisão da busca sem comprometer o resultado.

Nos casos em que a busca não retornou resultados, o GSM_search realizava ajustes específicos no título, principalmente no tratamento de informações entre parênteses (como datas ou localizações). Primeiramente, removiam-se os números dentro dos parênteses; caso persistisse a ausência de resultados, todo o conteúdo entre parênteses era excluído. Após a modificação, caso o título se reduzisse a uma única palavra, as aspas duplas eram novamente aplicadas.

Embora essas modificações aumentem a chance de correspondência, também ampliaram o tempo necessário para a coleta. Para evitar bloqueios por parte do servidor do GSM, foi implementado um temporizador randomizado entre as requisições, variando entre 10 e 30 segundos. A rotina de coleta incluía ainda a definição de horários aleatórios para início, pausas ao longo do dia e término das atividades, simulando o comportamento humano.

Ao obter um resultado, o GSM_search comparava o título retornado pelo GSM com o título da entrada, com base em quatro critérios denominados TM (Term Match):

- a) TM1: O primeiro critério estabelecido baseou-se no percentual de similaridade entre os textos. A escolha desse percentual foi baseada na média de caracteres dos títulos analisados, que tem em torno de 32 caracteres, com uma variação aproximada de 3 caracteres que correspondem a 90% de similaridade.
- b) TM2: O segundo critério, baseado em Processamento de Linguagem Natural (PLN), utiliza a biblioteca NLTK, e reconhece abreviações (“Rev.”/“Revista”), substituições (&/AND), palavras auxiliares ausentes e pequenas variações de grafia. Essa abordagem eliminou a necessidade de criar uma lista de *stopwords*.
- c) TM3: O terceiro critério é para os casos de títulos com subtítulos. Ele verifica se um dos textos (entrada ou resultado) estava contido no outro, resolvendo casos de títulos com ou sem subtítulo.
- d) TM4: O quarto critério foi desenvolvido para comparar exclusivamente as letras, ignorando os espaços entre palavras. Essa abordagem foi adotada para lidar com situações em que o título foi indexado com palavras juntas, sem o espaço em branco para separá-las.

Se algum dos critérios era atendido, os dados da revista eram salvos em arquivos TXT individuais, nomeados com um identificador exclusivo (parte do título, data e hora) e armazenados no diretório “saved”. A Figura 7 apresenta a estrutura desses arquivos.

Figura 8 - Arquivo TXT Saved GSM_search

```
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&venue=veq_zTAgrk0J.2023
Paper: Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows
Paper: Anticompetitive Effects of Common Ownership
Paper: Deviations from Covered Interest Rate Parity
```

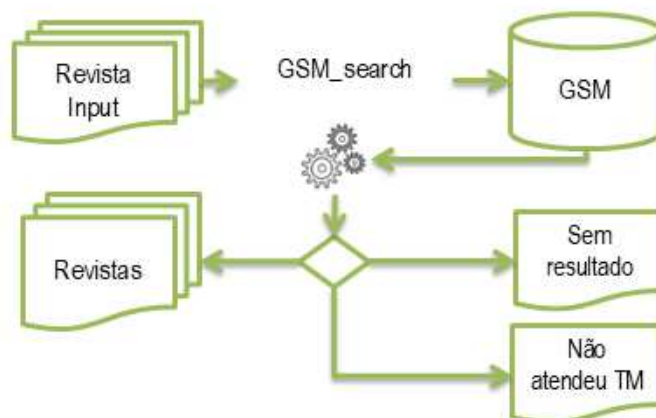
Fonte: Dados da pesquisa

Outras duas situações foram tratadas separadamente: busca sem resultado, quando o GSM não retornou nenhum dado; e para resultado não validado, quando o GSM retornava algum dado, mas nenhum critério de TM foi atendido. Para ambos os

casos, foram criados arquivos específicos para facilitar a conferência desses casos futuramente.

Esses critérios e procedimentos foram aplicados com o propósito de ampliar a identificação de títulos de periódicos no GSM, contornando as limitações impostas pelas variações textuais na comparação entre registros.

Figura 9 - Fluxograma do GSM_search



Fonte: Elaborado pelo autor.

3.5.2.2 OpenAlex_data

O OpenAlex_data tem como finalidade enriquecer os dados coletados no GSM, permitindo uma análise bibliométrica mais aprofundada das revistas e assegurando a identificação precisa de cada periódico, minimizando ambiguidades com títulos semelhantes. Essa etapa é composta por dois subconjuntos de procedimentos: OpenAlex_source e OpenAlex_work, ambos voltados à coleta de informações específicas, extraídas dos diretórios correspondentes da base OpenAlex.

No procedimento OpenAlex_source, foram extraídos dados referentes aos periódicos. O objetivo foi identificar atributos que possibilitem a caracterização editorial das revistas, como título, ID, ISSN, editora, país, status de acesso aberto, índice h, entre outros. O Quadro 4 apresenta os campos extraídos e suas respectivas descrições:

Quadro 4 - Campos coletados pela etapa OpenAlex_source

Campos	Descrição
ID	Link identificador único de cada revista
issn_l	ISSN da revista
Issn	Quando há mais de um ISSN, a revista fica registrada nesse campo
display_name	Título da revista
Publisher	Editora
is_oa	Se a revista é de Acesso Aberto
is_in_doaj	Se a revista está indexado no DOAJ
Type	Tipo da fonte (Revista, Anais...)
country_code	Código ISO do país.
alternate_titles	Variações dos títulos
abbreviated_title	Abreviações dos títulos
works_count	Total de Publicações
cited_by_count	O número total de citações dos artigos das revistas hospedado no OpenAlex
h_index	Índice h total
i10_index	Índice h10 10 anos
x_concepts_display_name	Conceito atribuído para a revista com base em conceitos dos seus artigos

Fonte: Elaborado pelo autor

A coleta foi realizada por meio de um *script* desenvolvido em Python⁷, capaz de percorrer todos os arquivos presentes no diretório Source, estruturados no formato JSON. O OpenAlex_source reconhece os campos de interesse e realiza a extração seletiva das informações, gerando ao final um arquivo único consolidado no formato CSV, que servirá como base para análises posteriores e cruzamento com os dados extraídos do GSM.

O segundo procedimento, denominado OpenAlex_work, teve como objetivo coletar dados sobre os artigos publicados e indexados no OpenAlex, limitando-se ao período de 2018 a 2023, intervalo correspondente à janela temporal usada pelo GSM para cálculo do índice h5. A finalidade foi ampliar as possibilidades de análise por meio do cálculo de métricas como média aritmética, mediana e percentual de artigos não citados, permitindo maior profundidade nas análises sobre os periódicos, bem como cruzar as informações sobre os títulos dos artigos coletados do GSM com OpenAlex, fazendo a verificação da revista.

Assim como na etapa anterior, os arquivos analisados encontram-se no diretório Works, organizados em diversos arquivos JSON. Cada registro de artigo contém metadados detalhados tanto sobre o próprio trabalho quanto sobre a revista

⁷ Python versão 3.

na qual foi publicado, o que permitiu extrair dados relevantes de ambos os níveis. Quadro 5 apresenta os campos coletados:

Quadro 5 - Campos coletados pela etapa OpenAlex_work

Campos	Descrição
Title	Título do artigo
publication_year	Ano publicação do artigo
primary_location_id	Número de identificação da revista
primary_location_type	Tipo da fonte
primary_location_display_name	Nome da revista
Type	Tipo de obra
authors_names	Nome dos autores
cited_by_count	Número de citações total
2yr_cited_by_count	Número de citação dos últimos 2 anos

Fonte: Dados da pesquisa.

A extração foi feita com um *script*⁸ com funcionalidades semelhantes às aplicadas no OpenAlex_source. Entretanto, devido ao volume de dados, o diretório *Works* que ultrapassa 300 GB de dados compactados. Foi necessária a implementação de um procedimento sequencial, no qual cada arquivo era processado, em seguida, excluído da memória para liberar espaço de para continuidade do processamento.

Para viabilizar esse processamento intensivo, utilizou-se um supercomputador com 80 GB de memória RAM e 20 processadores Intel® Xeon® CPU E5-2420 a 1,90 GHz. Isso gerou ao final um arquivo único consolidado no formato CSV.

Os dados extraídos foram organizados em dois arquivos distintos: um contendo a lista de artigos e outro a lista de periódicos identificados na base OpenAlex. Essa separação facilitou o uso dos dados no módulo GSM-ALEX_data, sobretudo na etapa de certificação e análise métrica.

3.5.2.3 GSM-ALEX_verification

O GSM-ALEX_verification corresponde ao procedimento de validação dos dados coletados por meio do GSM_search. Para essa finalidade, foi utilizado um conjunto de dados contendo os títulos dos artigos extraídos do OpenAlex, especificamente por meio do procedimento OpenAlex_work.

⁸ Python versão 3.

Esse procedimento tem por objetivo comparar os artigos coletados no GSM com os artigos indexados no OpenAlex, a fim de verificar a compatibilidade entre os registros. O processo consistiu em identificar correspondências exatas entre os títulos dos artigos, considerando apenas os caracteres alfanuméricos (letras e números), com a exclusão de sinais de pontuação e outros caracteres especiais.

Foi adotado um critério de 100% de similaridade como parâmetro de validação, justificando-se pela existência de títulos com pequenas variações que, apesar de semelhantes, podem representar publicações distintas. Além disso, cada revista no GSM pode apresentar até três artigos mais citados.

Dessa forma, a validação da revista foi considerada positiva se ao menos um desses artigos encontrasse correspondência no OpenAlex. Assim, a presença de pelo menos um título coincidente com o OpenAlex era suficiente para garantir a associação entre os registros das duas bases. Essa estratégia visou equilibrar rigor e viabilidade, reconhecendo as limitações naturais de cobertura e possíveis variações na indexação de artigos entre as fontes.

Para a execução desse procedimento, utilizou-se a biblioteca DuckDB, que possibilita a criação de um banco de dados leve e de alto desempenho, ideal para o cruzamento de grandes volumes de dados diretamente na memória. A correspondência entre os dados foi realizada com base na comparação entre os títulos dos artigos do GSM e os registros do OpenAlex.

O processamento do GSM-ALEX_verification exigiu alto poder computacional e, por essa razão, foi realizado no mesmo supercomputador com 80 GB de memória RAM e 20 processadores Intel® Xeon® CPU E5-2420, com frequência de 1,90 GHz.

O processo consistia na leitura dos arquivos gerados pela etapa GSM_search e na verificação dentro do conjunto de dados do OpenAlex_data da existência de correspondência entre os títulos. Quando identificada uma correspondência, era gerado um arquivo TXT contendo as informações relevantes, nomeado com data, hora e parte do nome da revista, garantindo assim a singularidade e rastreabilidade de cada resultado.

O arquivo gerado em TXT contém informações sobre os artigos que apresentaram correspondências, incluindo seu título, o ID da revista, e o nome do arquivo gerado na primeira etapa GSM_search. Esse formato organizado facilita a

identificação e análise dos dados compatíveis entre o GSM e o OpenAlex_data. Abaixo apresenta-se uma ilustração do formato de arquivo.

Figura 10 - Arquivo TXT salvo

```
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_op=list_hcore&venue=veq_zTAgrk0J.2023
Paper: Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows
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.txt

title: do investors value sustainability? a natural experiment examining ranking and fund flows
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
```

Fonte: Dados da pesquisa.

A ilustração permite identificar a separação entre o File 1, que contém dados originados na etapa de GSM_search, e o File 2, onde são armazenados os artigos que possuem correspondência no OpenAlex. No exemplo apresentado, dois artigos foram encontrados; nesse caso, o File 2 registra o ID da revista e o título correspondente de cada artigo e dados métricos do artigo.

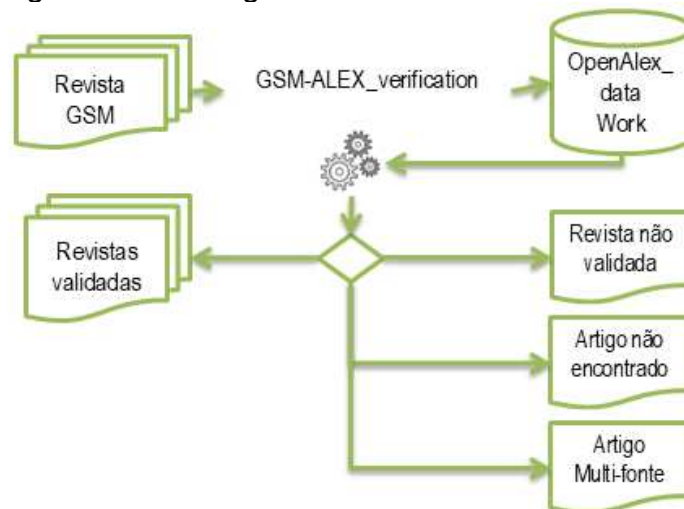
Para melhorar a organização dos procedimentos e facilitar a análise dos casos, foram criados arquivos específicos para diferentes situações: casos em que nenhum título de um artigo encontrou correspondência, situações em que ao menos um artigo apresentou correspondência e casos em que o título de um artigo estava associado a mais de uma revista. Nos casos em que há múltiplos correspondentes, foi verificado também se o título da revista do arquivo TXT era igual a algumas das encontradas. Essa estrutura foi projetada para simplificar a conferência e a identificação de possíveis falhas no processo.

Com o intuito de melhorar a organização dos resultados e facilitar a conferência de inconsistências, foram criados arquivos separados para diferentes cenários:

- a) casos em que nenhum dos três artigos da revista no GSM apresentou correspondência;
- b) situações dos artigos que não encontraram correspondência, mas pertencem a revistas validadas por meio de ao menos um dos três títulos analisados;
- c) e casos em que o título de artigo estava vinculado a múltiplas revistas.

Nessa última situação, foi realizada uma verificação adicional para conferir se o título da revista informado no arquivo GSM_search era compatível com algum dos registros encontrados no OpenAlex. Essa estrutura foi projetada para simplificar a conferência e a identificação de possíveis falhas no processo. Abaixo são apresentados os procedimentos do GSM_search

Figura 11 - Fluxograma do GSM-ALEX-verification



Fonte: Elaborado pelo autor.

A ilustração acima sintetiza graficamente a lógica do procedimento para certificação final das revistas com base na comparação com os registros do OpenAlex e do GSM.

3.5.5 Tratamento dos dados

O Módulo de Tratamento dos Dados tem como objetivo consolidar, padronizar e organizar as informações utilizadas pelo GSM-ALEX_metrics,

assegurando a integridade da base final de análise. Essa etapa é essencial para eliminar duplicidades, corrigir inconsistências, alinhar formatos de entrada e estruturar os dados de forma compatível com os procedimentos estatísticos e analíticos subsequentes. Além disso, contempla a geração de arquivos organizados em formato CSV, adequados tanto para o processamento automatizado por máquinas quanto para análise em planilhas eletrônicas.

Para definir o percentual de similaridade entre os títulos de periódicos do módulo GSM-ALEX_data, foi criado um *script*⁹ que contava o número de caracteres por título, calculando a média dessa quantidade. Com base nisso, estabeleceu-se um limite de 90% de similaridade, o que equivale, em média, a uma diferença de até três caracteres. Essa margem foi adotada para permitir pequenas variações textuais entre os registros, sem comprometer a precisão da correspondência.

Durante essa etapa, identificou-se que os títulos dos periódicos apresentavam variações de escrita, com uso de ideogramas, parênteses e diferentes convenções tipográficas. Para avaliar o impacto dessas variações na recuperação de dados no GSM, foi selecionada uma amostra de 10 casos para testes manuais. Constatou-se que a melhor estratégia era manter o título conforme registrado no OpenAlex, realizando ajustes pontuais apenas quando necessário, especialmente no tratamento de elementos entre parênteses, como país de origem ou datas, que prejudicavam a busca.

Também se observou que, no caso de títulos compostos por apenas uma palavra, o uso de aspas duplas ao redor do termo aumentava a precisão da recuperação.

O tratamento dos dados teve início com um mapeamento dos elementos disponíveis no OpenAlex, a fim de compreender a estrutura e os campos relevantes para os objetivos da pesquisa. A partir dessa análise, foi elaborado um arquivo CSV contendo atributos essenciais dos periódicos, como ID, título, país, ISSN, editora, status de acesso aberto, número de trabalhos publicados, número de citações, índice h, índice i10 e o conceito temático mais representativo. A lista de títulos desse procedimento foi utilizada como entrada para a coleta automatizada no GSM.

Os arquivos gerados por cada *script* foram projetados para facilitar a identificação e a conferência dos dados coletados. Para isso, foram previamente

⁹ Python versão 3.

mapeadas as diferentes situações que poderiam ocorrer durante o processo de coleta, e as saídas dos programas foram desenvolvidas para contemplar essas possibilidades.

Os arquivos foram estruturados no formato TXT, com campos definidos, de modo a permitir a recuperação automatizada das informações de interesse e converte-las em formato CSV.

Os dados gerados pelo módulo GSM-ALEX_verification, armazenados em arquivos TXT, foram integrados ao arquivo CSV resultante da coleta no OpenAlex_source. Essa integração foi realizada por meio do campo `primary_location_id`, que serviu como chave de correspondência. Quando havia correspondência entre os IDs, os dados do GSM, como *h5 index*, *h5 median*, *journal_google* e Paper URL, eram incorporados ao arquivo principal. O resultado foi uma base consolidada contendo apenas revistas validadas, com dados provenientes das duas fontes.

A estruturação para análise dos dados foi inicialmente realizada com o suporte de planilhas eletrônicas, que permitiram centralizar as variáveis, aplicar filtros, fórmulas e procedimentos de organização.

O processo de limpeza de duplicados nos dados bibliométricos foi realizado em duas etapas. A primeira etapa consistiu na identificação e exclusão de registros duplicados com base na URL, que representa o endereço eletrônico único de cada artigo. Para cada conjunto de registros que compartilhavam a mesma URL, foi realizada uma comparação de similaridade textual entre os campos com títulos dos periódicos. A escolha da linha a ser mantida seguiu um critério de prioridade, dando maior peso à correspondência igual de títulos da revista de entrada com proveniente da coleta do GSM, quando não atendido foi utilizado os títulos provenientes da validação por conferência artigos.

Na segunda etapa, a limpeza de duplicados concentrou-se na coluna `primary_location_id`, responsável pela identificação única de cada periódico na base OpenAlex. Foram selecionados todos os registros em que o mesmo ID aparecia em mais de uma linha. Para cada grupo de duplicatas, foi comparado o nome do periódico constante nas colunas com seus respectivos títulos. A abordagem em duas fases permitiu uma filtragem dos registros, combinando métodos automáticos¹⁰ de

¹⁰ Python versão 3.

detecção de duplicatas com decisões baseadas em similaridade textual, sempre buscando preservar a fidelidade aos títulos originais dos periódicos e garantir a integridade dos dados para as análises bibliométricas subsequentes.

A partir do conjunto de dados estruturados, foram aplicadas métricas bibliométricas como média de citações, mediana e percentual de artigos não citados, calculadas com base nos artigos indexados no OpenAlex_work, utilizando como chave o identificador presente no campo `primary_location_id`. Em razão do elevado volume de dados, que inviabilizou o uso de softwares de planilhas eletrônicas tradicionais devido à limitação na carga de arquivos com milhões de registros, foi desenvolvido um *script*¹¹ específico em para realizar o processamento dessas métricas. Esse *script* foi aplicado no módulo arquivo CSV do módulo OpenAlex_work, e permitiu a contagem de artigos não citados, bem como o cálculo da média aritmética e mediana para cada periódico.

A categorização temática das revistas foi baseada nos conceitos atribuídos automaticamente pelo OpenAlex via Wikidata. Cada revista recebe múltiplos conceitos, ranqueados por um score de relevância. Foi adotado o conceito com maior *score* como critério para classificar a revista dentro das nove grandes áreas do conhecimento da CAPES: Ciências Exatas e da Terra (CET), Ciências Biológicas (CB), Engenharias (Eng), Ciências da Saúde (CS), Ciências Agrárias (CA), Ciências Sociais Aplicadas (CSA), Ciências Humanas (CH), Linguística, Letras e Artes (LLA) e Multidisciplinar (Mul). Essa classificação também permitiu realizar análises em nível de área, considerando a aderência temática de cada periódico.

As revistas também foram distribuídas em quartis, considerando as grandes áreas do conhecimento CAPES e suas respectivas áreas. Esse agrupamento foi realizado com base nos valores do índice h5 extraídos do GSM.

Por fim, todos os *scripts* utilizados neste módulo foram construídos com o apoio de inteligência artificial generativa, por meio da ferramenta ChatGPT, que auxiliou na formulação dos códigos e na automatização dos processos.

¹¹ Python versão 3.

3.5.6 Analítico

O Módulo Analítico compreende a etapa de análise dos dados extraídos pelos diferentes módulos que integram a GSM-ALEX_metrics, fundamentando-se nos conceitos apresentados no Módulo 1. A análise envolve a aplicação de indicadores bibliométricos disponibilizados pelo GSM e OpenAlex, bem como a formulação de métricas adaptadas ou desenvolvidas especificamente a partir dos fundamentos teóricos estabelecidos.

No que se refere à interpretação da citação dentro da comunicação científica, adota-se a abordagem normativa, em que a citação é entendida como um indicador de mérito e reconhecimento acadêmico, conforme a concepção de Merton (2013).

Os indicadores bibliométricos, como o índice h e o número de citações, foram tratados como variáveis quantitativas sendo discretas, em sua maioria, por representarem contagens, mas analisadas sob a perspectiva de variáveis contínuas. Essa decisão se justifica pela influência de comportamentos acadêmicos e fatores externos imprevisíveis que afetam sua distribuição, e também pela natureza quase contínua que o índice h assume em grandes conjuntos de dados, nos quais pequenas variações (como entre $h = 10$ e $h = 11$) podem representar mudanças no impacto científico de um periódico. Tal abordagem permite a aplicação de métodos estatísticos fundamentados na variabilidade dos dados, como testes de correlação, análises de distribuição, que pressupõem variáveis em escala intervalar ou contínua. Além disso, contribui para uma análise mais sensível e integrada com outros indicadores bibliométricos contínuos, como médias e percentis de citação, preservando a consistência metodológica do estudo.

A análise estatística foi estruturada em duas abordagens complementares. A primeira contempla a avaliação direta dos dados brutos por revista, incluindo o número total de artigos publicados, total de citações recebidas, quantidade de artigos não citados e a mediana das citações. Esses dados representam a dimensão quantitativa acumulada da produção científica e permitem observar a performance absoluta de cada periódico.

A segunda abordagem consiste em uma análise descritiva baseada em métricas derivadas, como a razão entre o número de citações e o número de artigos publicados (média de citações por artigo), o índice h5 e indicadores extraídos

apenas dos artigos efetivamente citados. Esses indicadores representam uma avaliação relativa da distribuição do impacto, e são utilizados para fins comparativos e de ranqueamento entre os periódicos.

Ao combinar essas duas perspectivas, uma baseada em volume total e outra em indicadores normalizados, a análise permite diferenciar revistas com alta produção, mas baixo impacto médio, daquelas com menor volume de publicações, mas maior eficiência em gerar citações. Essa abordagem integrada oferece uma visão mais equilibrada e precisa do desempenho dos periódicos.

Nesse contexto, o índice h5, extraído do GSM, foi adotado como métrica central de impacto e passou a ser comparado a outros indicadores obtidos do OpenAlex, como forma de aferir a consistência entre diferentes fontes. Para aprofundar a análise, as revistas foram agrupadas em nove grandes áreas do conhecimento da CAPES e 21 áreas, permitindo segmentações temáticas mais refinadas e adequadas à diversidade disciplinar presente na amostra.

Embora o índice h avalie conjuntamente produtividade e impacto, ele não capta a distribuição desigual das citações, como a concentração em poucos artigos altamente citados. Por isso, foram integradas outras métricas descritivas, como:

- a) Média aritmética, que representa o número médio de citações por artigo, considerando também os artigos não citados. Apesar de suscetível a outliers, é útil para mensuração geral de impacto e foi utilizada na construção de indicadores derivados para cada revista em janela de cinco anos;

$$\text{Média}_{5a} = \frac{\text{Total de Citações nos Últimos 5 Anos}}{\text{Total de Artigos Publicados nos Últimos 5 Anos}} \quad (1)$$

- b) Mediana, que oferece uma medida mais robusta contra valores extremos e revela o centro da distribuição de citações;
- c) Amplitude, utilizada para verificar a dispersão entre os artigos mais e menos citados, identificando outliers e variação de impacto;
- d) Percentual de não citados, que fornece uma dimensão crítica da visibilidade acadêmica, complementando métricas de impacto positivo;

Também foi calculado o Índice de Gini, empregado para medir a concentração de citações, destacando revistas altamente dependentes de poucos artigos de alto impacto, a fórmula aplicada foi:

$$G = \frac{2 \sum_{i=1}^n i \cdot x_i}{n \cdot \sum_{i=1}^n x_i} - \frac{n+1}{n} \quad (2)$$

Onde:

n: número total de elementos na lista (artigos);

x_i : valor da i -ésima observação (número de citações do artigo i), ordenada em ordem crescente;

i : é o índice da posição do item na lista ordenada de valores

Σ : somatório.

Além dessas medidas, as revistas foram distribuídas por quartis dentro de suas respectivas grandes áreas e áreas, com base no índice h5 e na mediana das citações. O objetivo dessa categorização foi situar cada periódico em relação à sua temática, por causa da heterogeneidade disciplinar. A visualização por boxplot permitiu compreender melhor as faixas de impacto e padrões de concentração de desempenho entre as categorizações.

Com o intuito de compreender o comportamento das variáveis e sua distribuição, foram elaborados histogramas para observar a simetria dos dados. A partir disso, foram escolhidas técnicas apropriadas de correlação: Pearson para distribuições simétricas e lineares; Spearman para distribuições assimétricas.

As principais análises de correlação incluíram:

- a) Índice h5 do GSM com índice h5 gerado com os dados do OpenAlex para identificar a consistência do impacto das revistas nas duas bases;
- b) Índice h5 do GSM com o índice h disponível no OpenAlex com o intuito de verificar a tendência de manutenção de altos níveis de impacto pelas mesmas revistas ao longo do tempo;
- c) Total de artigos publicados pelas revistas com índice h5 do GSM, examinando a relação entre volume e reconhecimento;

- d) Acesso aberto e impacto, avaliando se as revistas com modelo de publicação aberto têm desempenho diferenciado nas métricas de citação em relação a revistas pagas.

A média de citações, índice h e percentual de não citados, também foram usados como fim de comparação entre revistas de acesso aberto e fechado.

Por fim, os dados e resultados foram representados graficamente por meio de ferramentas como Tableau e Flourish, visando facilitar a interpretação visual e a comunicação científica dos achados.

Para assegurar a qualidade e a consistência dos dados coletados durante a extração de informações, foi realizada uma análise dos casos de baixa similaridade textual entre os títulos das revistas. O objetivo principal dessa etapa foi identificar possíveis inconsistências, erros de associação e variações linguísticas que pudessem comprometer a integridade das análises bibliométricas.

Inicialmente, foi calculado o índice de similaridade textual entre os títulos recuperados do GSM e os nomes das revistas identificadas no OpenAlex. Para isso, utilizou-se o *script*¹², que permite comparar o grau de semelhança entre duas cadeias de texto, gerando um índice de similaridade com valores entre 0 e 1. Os casos com índice inferior a 0,9 foram selecionados para uma análise mais refinada.

Na etapa seguinte, foi desenvolvido um *script*¹³ com função de detectar o idioma dos títulos presentes em ambas as colunas. Esse *script* aplica técnicas para classificar o idioma de pequenos trechos de texto, que teve como finalidade distinguir casos em que a diferença observada entre os títulos fosse decorrente de tradução. Foram criadas duas novas colunas no banco de dados para registrar o idioma detectado em cada fonte: uma correspondente ao título recuperado do GSM e outra ao título do OpenAlex.

Com base nas informações de similaridade textual e de idioma, outra função do *script* foi realizar a categorização dos casos, utilizando as seguintes classes:

- a) *Idioma diferente*;
- b) *Diferença por subtítulo*, quando a discrepância estava relacionada à presença ou ausência de subtítulos nos registros;

¹² Python, versão 3, com a biblioteca difflib.

¹³ Python, versão 3, com a biblioteca Langdetect.

- c) *Possível erro - Outra revista*, em situações em que a similaridade entre o texto de entrada utilizado na busca, e o título do OpenAlex, foi inferior a 0,5 indicando possível associação incorreta, e;
- d) *Revisar manualmente*, para os casos que não se enquadraram nas categorias anteriores, mas cuja similaridade era considerada insuficiente para validação direta.

Com a conclusão dessas etapas de verificação, análise de similaridade, detecção de idioma e aplicação de filtros de elegibilidade, encerra-se o módulo analítico do GSM-ALEX_metrics proposto nesta pesquisa.

4 VALIDAÇÃO DO SISTEMA

Na etapa inicial, buscou-se verificar a correspondência entre os títulos de uma listagem pré-definida com os registros disponíveis no GSM. Para isso, foi selecionado um conjunto com 688 títulos de periódicos e definidos critérios de similaridade textual. Durante o processamento da lista, os títulos foram revisados, e os casos não identificados permitiram ajustes incrementais na lógica do sistema para aumentar sua eficácia na identificação.

Essa fase foi essencial para detectar inconsistências e refinar o funcionamento da rotina de busca. A lista original incluía títulos já presentes no GSM, além de outros não indexados, o que contribuiu para o aprimoramento dos critérios implementados na ferramenta de busca GSM_search.

Em seguida, o sistema foi submetido a um volume mais amplo de dados, com o intuito de validar os procedimentos estabelecidos. Esse segundo teste utilizou uma amostra de 3.070 periódicos disponíveis no repositório de dados Zenodo, oriunda de uma pesquisa sobre revistas latino-americanas indexadas no GSM (Canto *et al.*, 2021). Os dados estavam organizados em uma planilha eletrônica, contendo informações editoriais dos periódicos, além de seus respectivos nomes no GSM e no Latindex.

Considerando que os dados de validação se referem ao ano de 2021, e que o GSM anualmente atualiza sua lista revistas, tornou-se necessário realizar nova coleta com dados novamente, mantendo o mesmo método. Vale destacar que o autor teve acesso ao sistema utilizados por Canto *et al.* (2023).

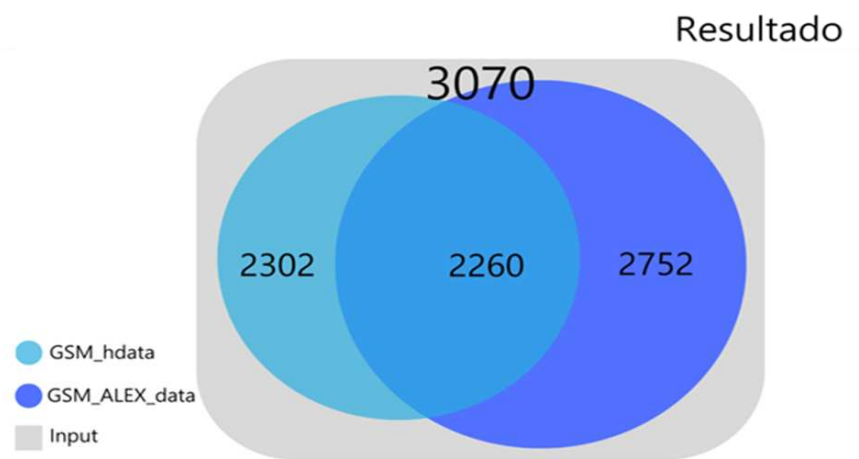
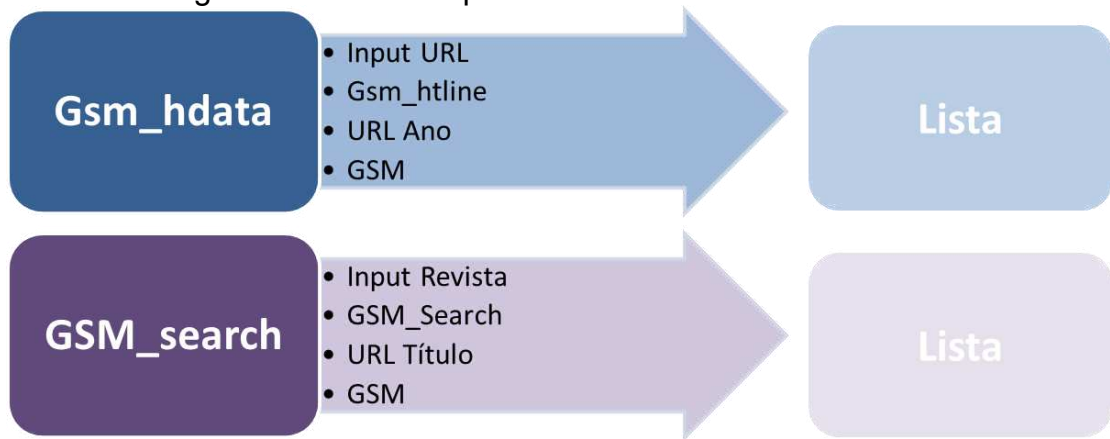
Dessa forma, a lista de 3070 periódicos foi processada pelos dois sistemas: o Gsm_hdata, conforme publicado por Canto *et al.* (2023), e o GSM_search, desenvolvido posteriormente. O Gsm_hdata localizou 2.302 revistas, o que representa cerca de 75% do total listado. Importante mencionar que esse *script* opera a partir das URLs dos periódicos, o que garante maior precisão e evita ambiguidade entre títulos homônimos.

Já com o uso do GSM_search, foram obtidas respostas positivas para 2.752 títulos, correspondendo a 90% da lista. Dos 318 títulos não localizados, 309 não retornaram qualquer resultado na consulta ao GSM, o que indica que provavelmente não estavam mais indexados na base. Comparativamente, o GSM_search apresentou um ganho de 450 títulos a mais em relação ao Gsm_hdata,

representando um aumento de 15% na cobertura direta, mas seu resultado contém duplicações e periódicos que podem divergir da listagem original.

Para avaliar a coincidência entre os registros obtidos por cada procedimento, realizou-se uma comparação direta das URLs recuperadas no GSM. Dos 2.752 títulos identificados pelo GSM_search, 2.260 apresentavam a mesma URL entre os 2.302 identificados pelo Gsm_hdata. Isso indica uma taxa de precisão de 98% no conjunto de dados testado. Ressalta-se ainda que parte do procedimento para compor lista de periódicos, descrito por Canto *et al.* (2021), envolveu verificação manual. No entanto o GSM_search foi totalmente automatizado. A Figura 12 ilustra o fluxo de etapas e os respectivos resultados obtidos.

Figura 12 - Fluxo do procedimento de coleta dos dados



Fonte: Dados da pesquisa.

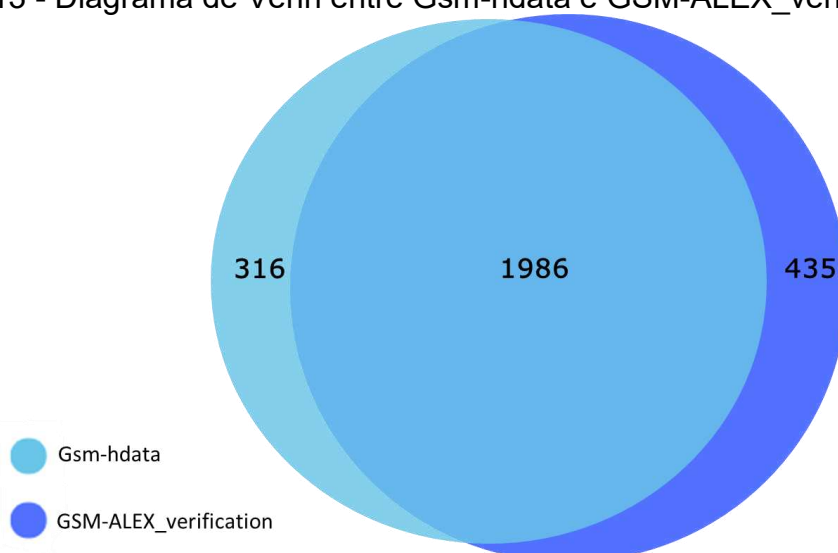
Os 2.752 títulos identificados inicialmente pelo GSM_search foram processados pelo módulo GSM-ALEX_verification, que verifica a equivalência entre os conteúdos do OpenAlex e do GSM com base nos títulos dos artigos. O

procedimento resultou na identificação de 2.421 periódicos que apresentavam um ou mais títulos coincidentes nas duas bases.

Após a validação dos periódicos com base nos títulos dos artigos, etapa essencial para eliminar inconsistências decorrentes da indexação automatizada do GSM, os registros confirmados foram comparados, por meio de suas URLs, com os 2.302 periódicos identificados pelo procedimento Gsm_hdata.

Como resultado, das 2.421 revistas validadas, 1.986 apresentaram a mesma URL das encontradas no Gsm_hdata, o que corresponde a uma taxa de correspondência de aproximadamente 86% em relação ao total identificado por esse procedimento, que pode ser verificado no diagrama da figura 13.

Figura 13 - Diagrama de Venn entre Gsm-hdata e GSM-ALEX_verification



Fonte: Dados da pesquisa.

O procedimento GSM-ALEX_verification identificou 435 periódicos (14% das 3070 revistas) que são exclusivos dessa abordagem em comparação com os 2.302 encontrados pelo Gsm_hdata. Esse resultado representou um avanço modesto na cobertura, com 119 periódicos a mais (aproximadamente 4%), superando o total previamente obtido pelo Gsm_hdata.

Para 316 revistas identificadas pelo Gsm_hdata, não foi possível encontrar um correspondente no OpenAlex. Esse número representa um aumento em relação à diferença observada na comparação entre o Gsm_hdata e o GSM_search, que havia sido de 42 periódicos.

A análise manual das causas dessa divergência revelou diferentes situações: revistas que não estão indexadas no OpenAlex; periódicos com poucos

artigos disponíveis na base, o que inviabilizou a correspondência automática; variação de idioma nos metadados, algumas revistas publicam em mais de um idioma, mas estão indexadas no OpenAlex apenas em um deles; e títulos dos artigos registrados com dois idiomas no mesmo campo textual.

Após a verificação manual, optou-se por manter o resultado gerado pelo processo automatizado e quando identificado que ajustes poderiam aprimorar o sistema, as melhorias foram implementadas e a análise refeita com base na nova versão do *script*.

Nos resultados obtidos pelo GSM-ALEX_verification, foram identificados casos de duplicidades, os quais não foram considerados na comparação com o Gsm_hdata. Foram identificados 45 registros duplicados, que foram excluídos, resultando em uma lista final.

A exclusão das duplicidades foi realizada por meio da análise textual dos títulos de entrada, priorizando a versão mais próxima da forma como os periódicos são apresentados na busca do GSM.

A pesquisa utiliza, para fins de análise bibliométrica, os dados provenientes da etapa GSM-ALEX_verification. Por essa razão, foram examinados detalhadamente os 435 casos que não estavam presentes entre os 2.302 periódicos identificados pelo Gsm_hdata.

A análise bibliométrica teve como objetivo verificar a compatibilidade entre o texto de entrada e o título recuperado no GSM. Como resultado, foram identificadas 13 incompatibilidades, o que representa aproximadamente 3% dos 435 títulos divergentes em relação ao conjunto do Gsm_hdata, mas em relação ao total validado pelo procedimento representa 0,5%.

Os testes realizados demonstraram a efetividade do sistema na recuperação e validação de periódicos indexados no GSM. Como GSM_search, 98% dos títulos estavam contidos na base de comparação. A integração dos módulos permitiu alcançar uma taxa de correspondência com 86% de precisão e um incremento de 4% na cobertura em comparação ao método de base Gsm-hdata. Destaca-se ainda a baixa taxa de incompatibilidade entre o título de entrada e o título recuperado, registrada em 0,5% dos 2.421 periódicos validados.

Os resultados foram considerados satisfatórios, especialmente diante dos desafios enfrentados como ausência de ISSN, inconsistência dos metadados, uniformidade dos títulos, variações linguísticas e ausência de interface programável.

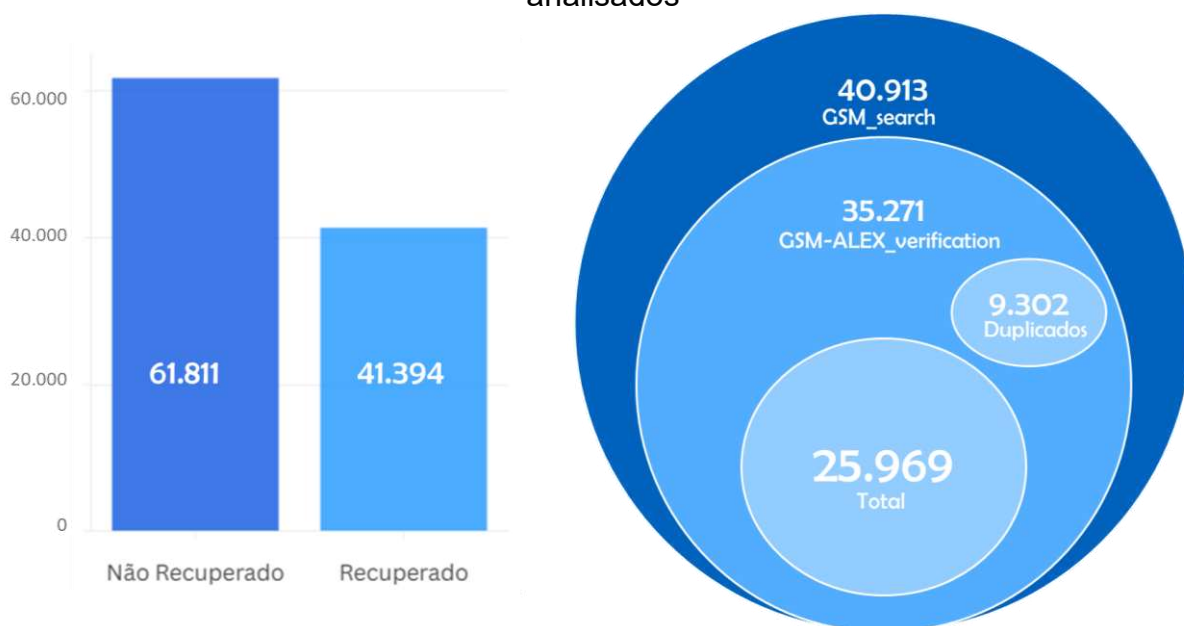
Esses resultados indicam a robustez e a confiabilidade do sistema para aplicações em análises bibliométricas automatizadas.

5 RESULTADOS

Neste capítulo, apresentam-se os resultados da pesquisa em dois momentos distintos. Primeiramente, são descritos os dados referentes ao processo de coleta, que integrou o GSM com a base OpenAlex. Em seguida, é apresentada a análise bibliométrica realizada a partir dos dados obtidos.

A pesquisa teve início com uma base composta por 103.212 títulos, provenientes do OpenAlex, de abrangência global, incluindo periódicos de todos os continentes. A Figura 14 ilustra esse volume inicial, juntamente com o total de registros não recuperados e aqueles validados após os procedimentos metodológicos adotados.

Figura 14 - Síntese do processo de coleta, verificação e integração dos periódicos analisados



Fonte: Dados da pesquisa.

Na figura, a coluna Não Recuperado corresponde aos casos em que não houve retorno de dados do GSM, totalizando 60% do conjunto inicial. Já “Recuperados” referem-se às respostas obtidas positivamente no GSM, representando 40% do total de 103.212 registros.

O GSM_search indica o total de revistas efetivamente recuperadas dentro desses 40%. Nesse procedimento, o sistema conseguiu salvar 99% dos títulos para cada caso analisado, conforme os critérios estabelecidos.

Os casos analisados seguiram quatro critérios de similaridade textual (TM), descritos anteriormente na metodologia. Entre eles, o TM1 respondeu por 28.602

casos, representando aproximadamente 70% do total recuperado. O TM2 somou 12.038 casos (cerca de 29%). Já TM3 e TM4, em conjunto, totalizaram 273 casos, correspondendo a aproximadamente 1% do total.

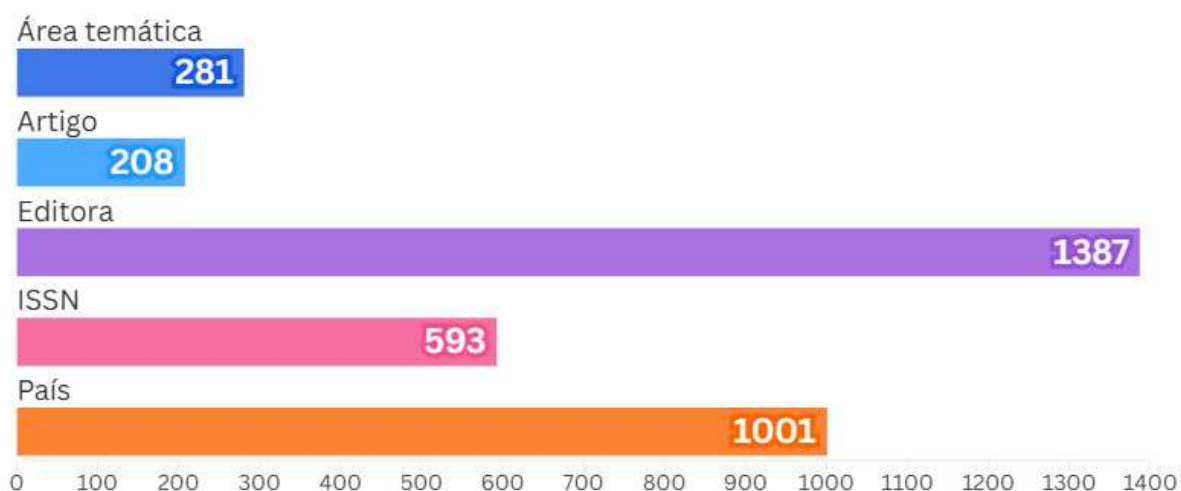
O “GSM-ALEX_Verification” corresponde a um subconjunto do total de revistas recuperadas pelo “GSM_search”, com base na verificação de artigos extraídos do GSM e localizados no OpenAlex. Esse procedimento resultou em um retorno positivo para 86% das revistas verificadas.

No total, foram verificados 122.546 artigos raspados do GSM, comparados a 32.119.995 registros de artigos provenientes do “OpenAlex_data”. Foram identificadas correspondências para 50.314 artigos, validando 35.271 revistas. Para as 5.642 revistas restantes (14%), não foram encontrados artigos correspondentes no OpenAlex.

A partir das 35.271 revistas validadas, verificou-se a ocorrência de duplicidades, uma situação comum nesse tipo de procedimento, em função de variações idiomáticas, nomes homônimos e inconsistências de registro. Constatou-se que 23% das revistas do “GSM_search” (9.302) estavam duplicadas, resultando em 63% (25.969) de revistas únicas a serem analisadas no estudo.

Com base nesse conjunto final de 25.969 revistas, foram identificados os respectivos IDs para verificar sua correspondência no total de 199.833 registros de revistas provenientes do “OpenAlex_data”. Esse procedimento possibilitou agregar informações complementares, como país de publicação, editora e outras variáveis, que são detalhadas a seguir.

Gráfico 2 – Quantidade de revistas com dados incompletos

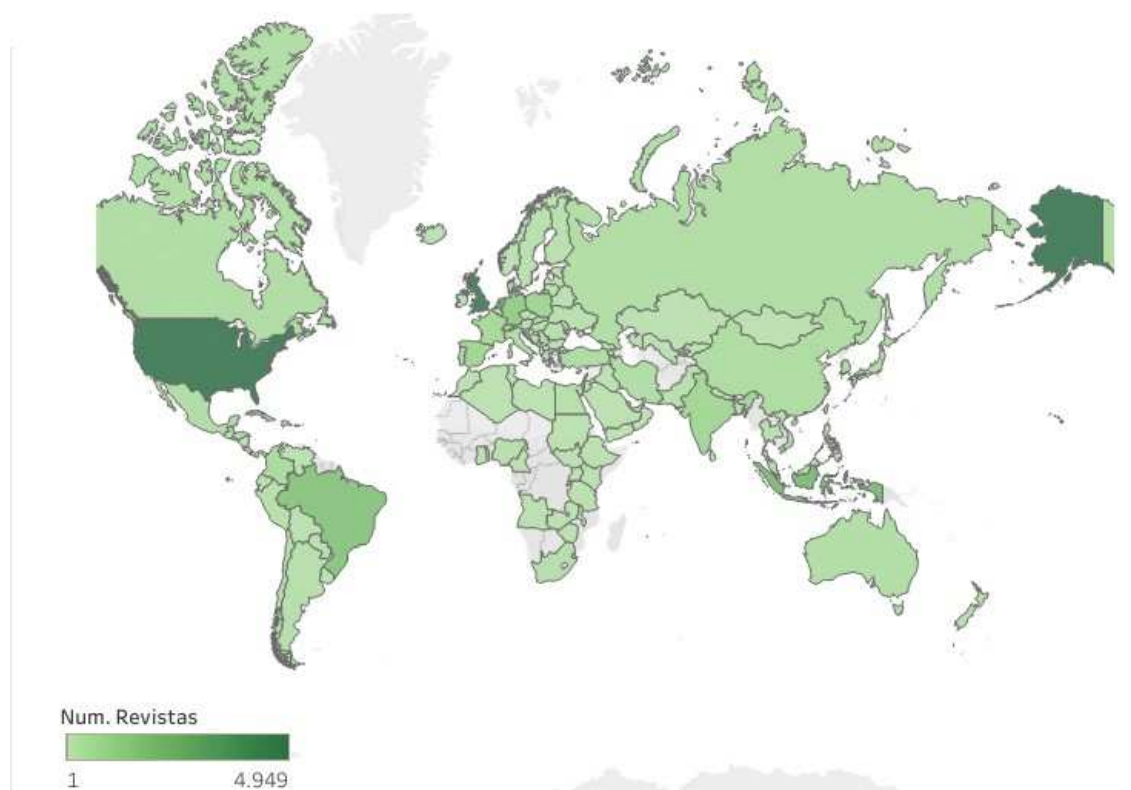


Fonte: Dados da pesquisa.

O Gráfico 2 representa os casos em que os campos esperados na extração de dados via “OpenAlex_data” não estavam preenchidos. A variável “Área temática” corresponde a 1% dos casos, indicando registros sem atribuição de conceito no campo `x_concepts_display_name`. A ausência de informações sobre o número de artigos publicados por revista ocorreu em menos de 1% dos casos. Para o campo ISSN, 2,4% das revistas não apresentaram esse dado. Já os campos “Editora” e “País” foram os que apresentaram o maior número de registros incompletos, com ausência de informações em 5,3% e 3,8% das revistas, respectivamente.

Devido às lacunas identificadas nos dados, a análise dos campos mencionados foi realizada apenas para as revistas que apresentavam todas as informações devidamente preenchidas. Os resultados são apresentados a seguir, iniciando com a Figura 15 que ilustra a distribuição das revistas por país.

Figura 15 - Representação dos periódicos por países



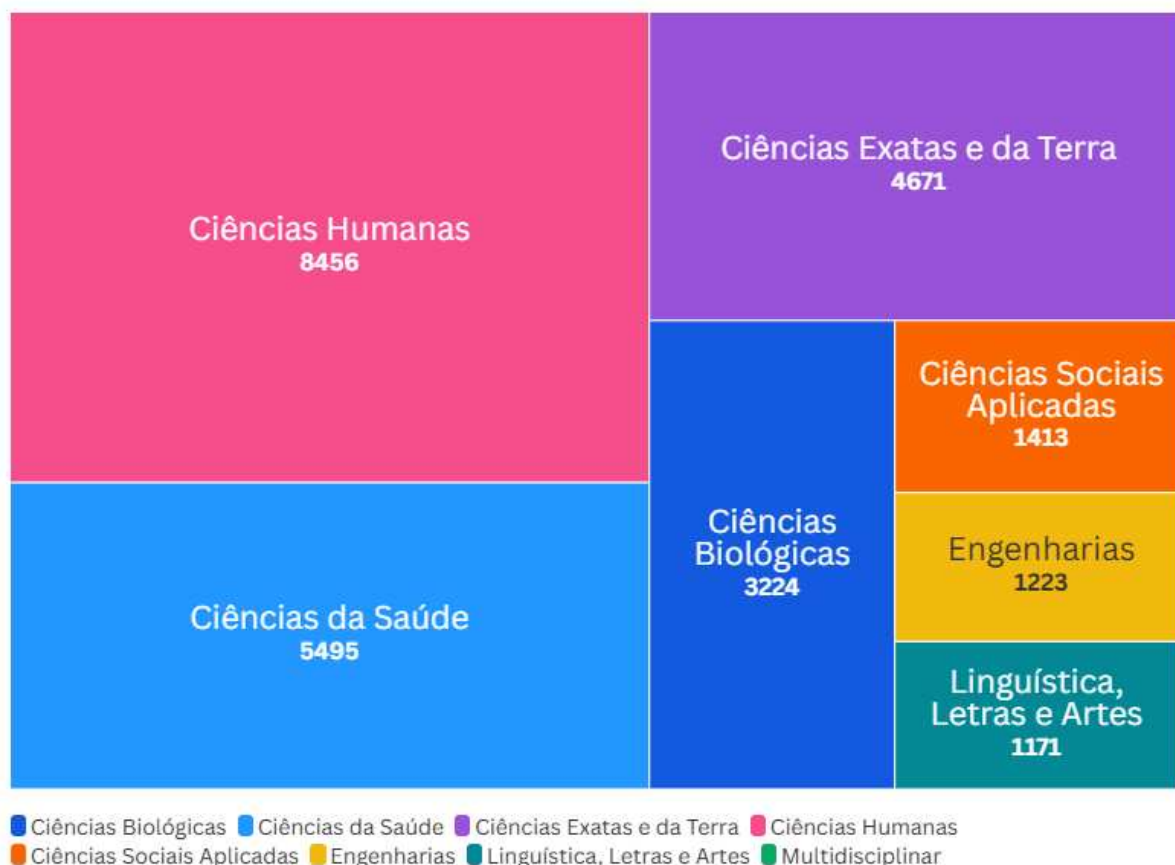
Fonte: Dados da pesquisa.

A quantidade de revistas por país está representada pela variação de cor: quanto mais intensa a cor, maior o número de revistas identificadas para aquele país. Essa representação permite visualizar tanto a concentração de periódicos quanto as regiões melhor cobertas pelo GSM. Observa-se que América e Europa concentram a maior parte das revistas indexadas: a Europa corresponde a 47% do total, com 11.777 revistas, enquanto a América representa 31%, com 7.852 revistas. Em seguida, destaca-se a Ásia, com 19% (4.610 revistas). As regiões com menor representatividade são África e Oceania, com 429 e 300 revistas, respectivamente, totalizando 3% do conjunto analisado.

Em relação à distribuição dos periódicos nos países por continente, foram identificados periódicos em 38 países na Europa, 35 na Ásia, 24 na América, 23 na África e 2 na Oceania. Os dez países com maior número de revistas analisadas respondem por 71% do total recuperado. São eles: Estados Unidos, com 4.949 revistas (19,82%); Reino Unido, com 4.648 (18,62%); Indonésia, com 1.862 (7,46%); Países Baixos, com 1.735 (6,75%); Brasil, com 1.465 (5,87%); Alemanha, com 1.169 (4,68%); Polônia, com 667 (2,67%); Suíça, com 642 (2,57%); Espanha, com 629 (2,52%) e Índia, com 627.

Foi verificada a distribuição temática das revistas incluídas na base analisada. A Figura 16 apresenta esse panorama, agrupando os periódicos conforme as grandes áreas do conhecimento definidas pela CAPES.

Figura 16 - Concentração de revistas por grandes áreas da CAPES

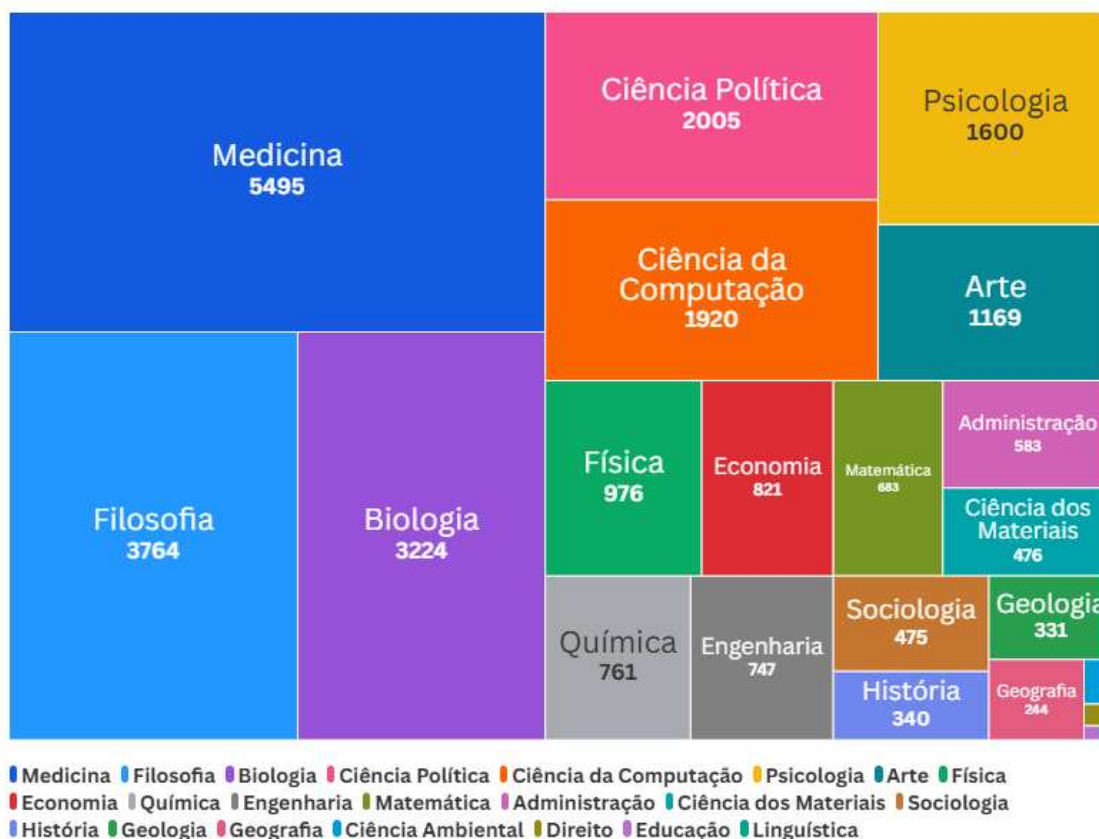


Fonte: Dados da pesquisa

Ciências Humanas concentra o maior volume de periódicos (32,85%), seguidas pelas Ciências da Saúde (21,39%), Ciências Exatas e da Terra (18,18%) e Ciências Biológicas (12,55%). Em contrapartida, grandes áreas como Ciências Sociais Aplicadas (5,54%), Engenharias (4,76%), Linguística, Letras e Artes (4,56%) e Multidisciplinar (41 revistas; 0,16%) apresentam uma quantidade menor de periódicos, sugerindo possíveis limitações na cobertura desses campos.

A distribuição evidencia uma diversidade temática para os dados encontrados e serve como indicativo sobre o alcance e as lacunas na representação de determinadas grandes áreas do conhecimento no GSM. A partir dessa visão geral, as ilustrações seguintes detalham essa composição ao nível de áreas, permitindo identificar nuances dentro de cada grande área, bem como variações na concentração de periódicos em temas mais específicos.

Figura 17 - Concentração de revistas por área da CAPES



Fonte: Dados da pesquisa.

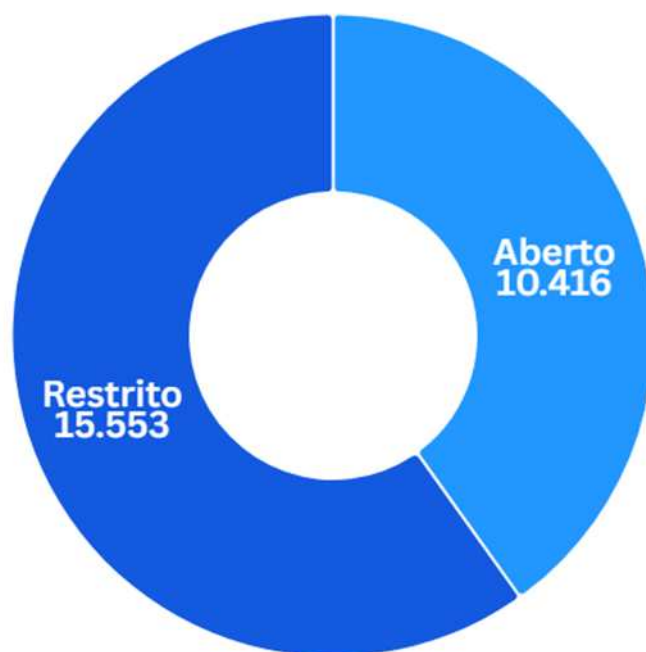
Observa-se que as áreas com maior representatividade foram Medicina (21,4%), Filosofia (14,7%), Biologia (12,6%) e Ciência Política (7,8%), demonstrando uma forte concentração de títulos nesses campos científicos. Destacam-se também áreas como Ciência da Computação (7,5%), Psicologia (6,2%) e Arte (4,6%), que, embora não apresentem o maior volume absoluto, fazem parte do grupo que fica acima da média dessa distribuição.

Por outro lado, algumas áreas apresentaram um número significativamente menor de periódicos, como Linguística (3), Educação (10), Direito (20) e Ciências Ambientais (41). Todas com valores abaixo da média e que, somadas, não correspondem a 1% do total de revistas analisadas. Esse cenário indica possíveis lacunas de cobertura ou uma menor expressão editorial dessas áreas na amostra considerada. Nesse sentido, a figura a seguir apresenta a distribuição das revistas por editoras.

a 100 periódicos; 12% (2.878) concentram-se na faixa de 11 a 50; e 24% (9.086) encontram-se na faixa de 1 a 10.

Essa distribuição evidencia uma forte concentração em editoras de grande porte, majoritariamente comerciais. Na sequência, será apresentada a forma como esses periódicos estão indexados em relação ao tipo de acesso, discriminando a proporção de títulos em acesso aberto e restrito.

Gráfico 3 - Revistas indexadas por modelo de acesso



Fonte: dados da pesquisa.

O Gráfico 3 apresenta a distribuição dos periódicos segundo o tipo de acesso, evidenciando a presença de 15.553 periódicos com acesso restrito e 10.416 com acesso aberto. Observa-se que, embora o acesso aberto represente uma parcela significativa (cerca de 40% do total), os periódicos de acesso restrito ainda predominam na amostra analisada. Esse cenário reflete a convivência de diferentes modelos editoriais, marcados por tensões entre barreiras de acesso e estratégias de democratização do conhecimento científico.

A Tabela 1 apresenta a distribuição das revistas analisadas de acordo com os critérios aplicados na etapa de comparação de similaridade textual entre o título original de entrada e o título recuperado no GSM. Esses critérios, designados como (TM1, TM2, TM3 e TM4), foram definidos com base em diferentes níveis de tolerância para variações textuais, sendo detalhados na subseção de GSM_search.

Tabela 1 - Critérios de comparação entre título de entrada e do GSM

Critério	Revistas	Percentual
TM1	23.749	91,45%
TM2	2.113	8,14%
TM3	90	0,35%
TM4	17	0,07%
Total Geral	25.969	100,00%

Fonte: Dados da pesquisa.

Observa-se que o critério TM1 concentrou a maior parte dos casos validados, totalizando 23.749 revistas, o que representa 91,45% do total de títulos verificados. Esse critério compara diretamente o texto com variação entre 90 a 100% de similaridade. O critério TM2 correspondeu a 8,14% (2.113 revistas), sua base de comparação permite uma variação maior entre os textos.

Como previsto na definição metodológica, os critérios TM3 e TM4 apresentaram baixa incidência na amostra, com apenas 90 (0,35%) e 17 (0,07%) revistas, respectivamente. Essa baixa frequência se explica pelo fato de o TM3 contemplar situações em que parte do título de entrada está contida no título recuperado, ou vice-versa, enquanto o TM4 desconsidera os espaços entre palavras para captar variações mínimas de formatação. Esses critérios são aplicados de forma hierárquica: registros que não atendem aos critérios de similaridade (TM1 e TM2) são reavaliados por TM3 e TM4 como tentativa final de validação, garantindo que possíveis casos marginais não fiquem de fora da base analisada.

Esses valores indicam que a maioria das correspondências entre os títulos de entrada e os registros recuperados no GSM ocorreu com alto grau de similaridade textual, não sendo necessárias permissões mais amplas de tolerância para variações maiores.

Para uma compreensão mais aprofundada dos conjuntos de dados, foram analisadas variáveis adicionais relacionadas à qualidade da correspondência, tais como a similaridade textual, subtítulos e a variação do idioma e casos em que o texto está contido. A variável de similaridade textual, neste caso, difere da etapa anterior por comparar o nome da revista recuperado no GSM com o título do periódico validado no procedimento GSM-ALEX_Verification. A distribuição da variável de similaridade textual encontra-se apresentada na tabela a seguir.

Tabela 2 - Concentração de revistas por faixa de similaridade textual

Similaridade	Contagem	%
90 a 100%	23.589	90,84%
80 a 89%	395	1,52%
70 a 79%	219	0,84%
60 a 69%	292	1,12%
50 a 59%	263	1,01%
Abaixo de 50%	1.211	4,66%
Total Geral	25.969	100,00%

Fonte: Dados da pesquisa.

A Tabela 2 apresenta a distribuição dos periódicos analisados de acordo com a faixa de similaridade textual. Observa-se que a grande maioria dos casos (90,84%) se encontra na faixa de 90% a 100% de similaridade, o que evidencia um elevado grau de correspondência entre os registros. Conforme apontado por Donner (2016), correspondências situadas nesse intervalo apresentam precisão de 100%, o que dispensa verificações adicionais de consistência. Assim, para fins desta pesquisa, esse grupo foi considerado de alta confiabilidade, não sendo submetido a análises complementares mais detalhadas. As demais faixas de similaridade, por apresentarem maior potencial de variação textual, foram analisadas em conjunto com as outras variáveis, idioma e título contido.

Nas faixas intermediárias, compreendidas entre 50% e 89% de similaridade, verifica-se uma presença menor, que somadas as faixas, representam cerca de 4,5% do total. A faixa abaixo de 50% de similaridade concentra 4,66% dos casos, sinalizando registros com maior divergência textual.

Para os casos abaixo de 90% de similaridade, foi percebido que muitas vezes o título original está integralmente contido no texto recuperado ou vice-versa, além de similaridades entre os títulos alternativos e os títulos do GSM. Para melhor entender essas discrepâncias, foi deliberado pormenorizá-las na tabela abaixo.

Tabela 3 - Distribuição dos periódicos por similaridade abaixo 90%

Similaridade	Frequência	TI Alter. 90%<		TI contido		TI não contido	
80 a 89%	395	17	4,3%	271	68,6%	107	27,1%
70 a 79%	219	9	4,1%	164	74,9%	46	21,0%
60 a 69%	292	8	2,7%	193	66,1%	91	31,2%
50 a 59%	263	8	3,0%	183	69,6%	72	27,4%
Abaixo de 50%	1.211	91	7,5%	770	63,6%	350	28,9%
Total Geral	2.380	133	5,6%	1.581	66,4%	666	28,0%

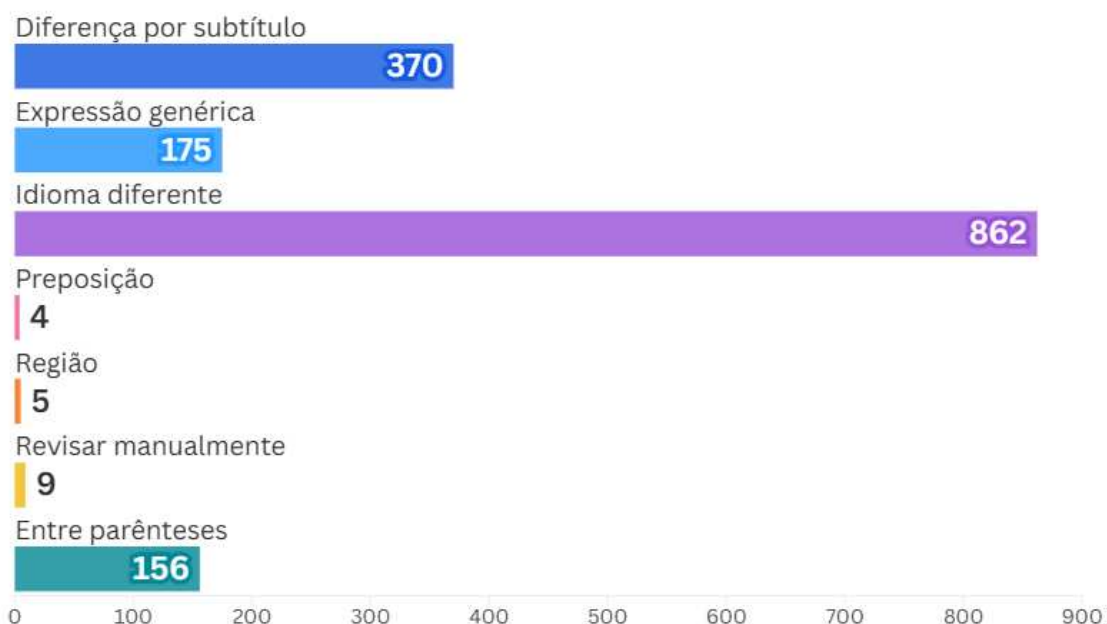
Fonte: Dados da pesquisa.

A Tabela 3 mostra que, do total de revistas analisadas, 2.380 casos (9,16%) apresentaram similaridade considerada de moderada a baixa. Sabe-se que alterações nos títulos, como a presença de palavras genéricas (journal, anais, entre outras) ou abreviações, podem contribuir para essa variação. Por isso, esses casos foram detalhadamente verificados.

Antes de categorizar as razões que levaram a essas discrepâncias, verificou-se a existência de títulos alternativos nos registros. Essa verificação indicou que, para 5,6% dos 2.380 casos, havia títulos alternativos ou abreviados no OpenAlex, que quando comparados aos títulos recuperados no GSM, explicam parte das variações observadas. É importante destacar que a maioria desses casos de títulos alternativos também se concentra na faixa de similaridade abaixo de 50%, o que demonstra uma variação textual substancial, possivelmente relacionada a traduções, formas abreviadas ou alterações editoriais.

Para os casos restantes (2.247 registros), foi inicialmente verificado se o título estava contido integralmente no outro, em ambos os sentidos. Essa situação corresponde à maior parte desses registros: 1.581 casos (66,4% dos 2.380). Esses títulos foram então analisados quanto às possíveis causas da diferença de similaridade, cujos resultados estão ilustrados no gráfico abaixo.

Gráfico 4 - Diferença textual entre os títulos contidos em ambos os sentidos

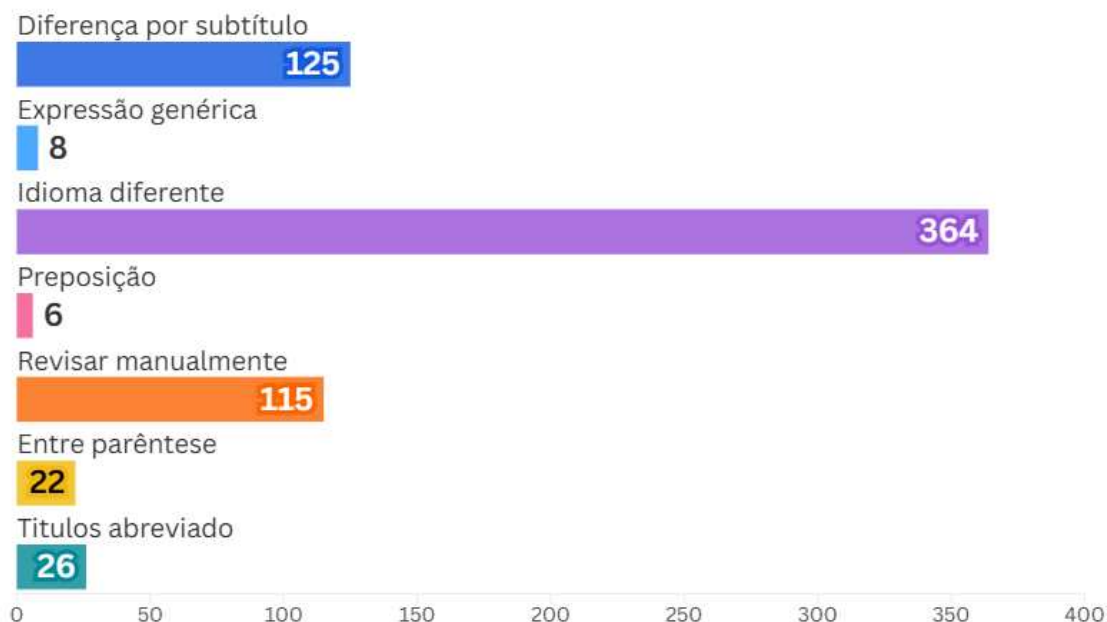


Fonte: Dados da pesquisa.

Para os casos em que os títulos estavam contidos mutuamente (em ambos os sentidos), foram identificadas seis razões principais para a diferença observada. Entre essas, destacam-se quatro categorias com maior concentração: diferença de idioma, responsável por 54,5% dos casos (quando se constatou discrepância linguística entre os títulos); presença de subtítulos, com 23,4% (variações marcadas por pontuação, como “:” ou “—”); uso de expressões genéricas, com 11% (diferenças relacionadas a palavras como Revista, Anais, Journal); e ocorrências de informações entre parênteses, que representaram 10% dos casos.

As razões menos frequentes, que somadas correspondem a pouco mais de 1% dos casos, incluem variações relacionadas a preposições, títulos compostos por uma única palavra ou referências regionais explícitas no nome do periódico. Além disso, para 9 casos (0,6%), não foi possível identificar uma causa específica. O gráfico abaixo apresenta essa categorização para os casos em que o título não estava contido no outro, em ambos os sentidos, destacando como esses fatores podem impactar a análise de similaridade textual.

Gráfico 5 - Diferença textual entre os títulos não contidos em ambos os sentidos



Fonte: Dados da pesquisa.

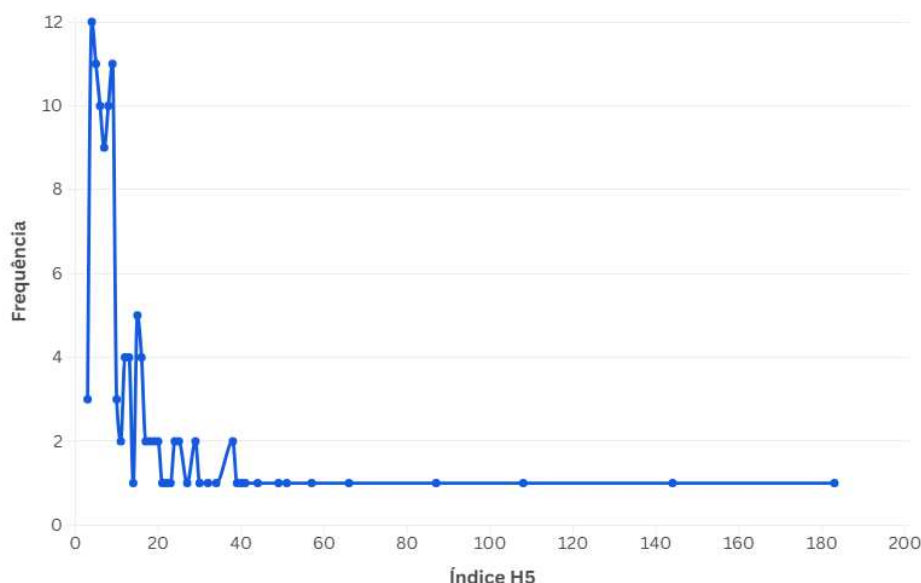
O Gráfico 5 evidencia as principais causas de divergência na similaridade textual, que foram: “Idioma diferente”, com 54,7% dos casos, e “Diferença por subtítulo”, com 18,8%. Outras causas, identificadas com menor frequência, incluem “Preposição”, “Texto entre parênteses”, “Expressão genérica” e “Título abreviado”, que, em conjunto, correspondem a 9,3% dos casos.

Observa-se ainda que, para 17,3% das revistas analisadas, não foi possível identificar uma razão clara que justificasse a diferença na comparação textual entre os títulos.

Os dois Gráficos 4 e 5 ilustram que o procedimento de verificação foi bem-sucedido para 2.123 revistas, que, somadas ao total de registros com similaridade superior a 90%, representam 25.712 casos, atingindo assim uma precisão de aproximadamente 99% na base validada.

Por outro lado, em 124 casos (1% do total de revistas analisadas), não foi possível determinar uma causa aparente para a divergência de similaridade. Para esses casos, procedeu-se à verificação do índice h5, conforme representado no gráfico complementar abaixo, a fim de avaliar se há relação entre baixo impacto bibliométrico e inconsistências na correspondência textual.

Gráfico 6 - Distribuição do índice h5 das revistas que precisaram de revisão manual



Fonte: Dados da pesquisa.

O gráfico acima ilustra a distribuição dos índices h5 das revistas recuperadas no GSM que precisaram de revisão manual (124 revistas), com o objetivo de verificar em quais grupos se concentram a maior parte dos casos. Observa-se que a maior concentração ocorre na faixa de h5 entre 1 e 9, que reúne 66 revistas, correspondendo a 53% do total. Já o grupo restante, que representa 47% (58 revistas), apresenta uma distribuição mais ampla dos índices h5, variando de 10 até 183, embora a frequência seja mais expressiva na faixa inicial, até 40.

Esse padrão de concentração nas faixas mais baixas do índice era esperado, considerando que revistas com menor impacto tendem a apresentar maior instabilidade ou inconsistências na coleta de dados pelo GSM. Por esse motivo, esses registros são apresentados aqui apenas para fins de contextualização e não serão incluídos na análise bibliométrica principal. As possíveis implicações para os casos com h5 mais elevados serão exploradas na seção de Discussão.

Para uma compreensão mais precisa de como foi realizada a categorização dos casos com similaridade abaixo de 90%, destaca-se que o processo envolveu dois momentos complementares. No primeiro momento, de forma automatizada, utilizou-se um *script*¹⁴ que detecta variações de idioma entre os textos comparados. O mesmo *script* também identifica a presença de subtítulos, por meio de

¹⁴ Python, versão 3, com biblioteca langdetect.

separadores comuns, como “:”, “—”, “-” e “|”, o que explica os casos mais frequentes observados.

No segundo momento, procedeu-se a uma verificação manual dos registros remanescentes, revisando título a título para identificar outras causas potenciais para a divergência de similaridade. Assim, foram verificados casos envolvendo abreviações, presença de palavras genéricas típicas de publicações (journal, revista, anais), informações entre parênteses que indicam região ou edições específicas, entre outras situações descritas anteriormente.

Esses procedimentos complementares reforçam a consistência da análise de similaridade textual, permitindo capturar variações legítimas na forma escrita, traduções ou diferenças editoriais. Com isso, estima-se que 99% das revistas coletadas estejam adequadamente habilitadas para compor a análise bibliométrica, que constitui o foco do próximo tópico.

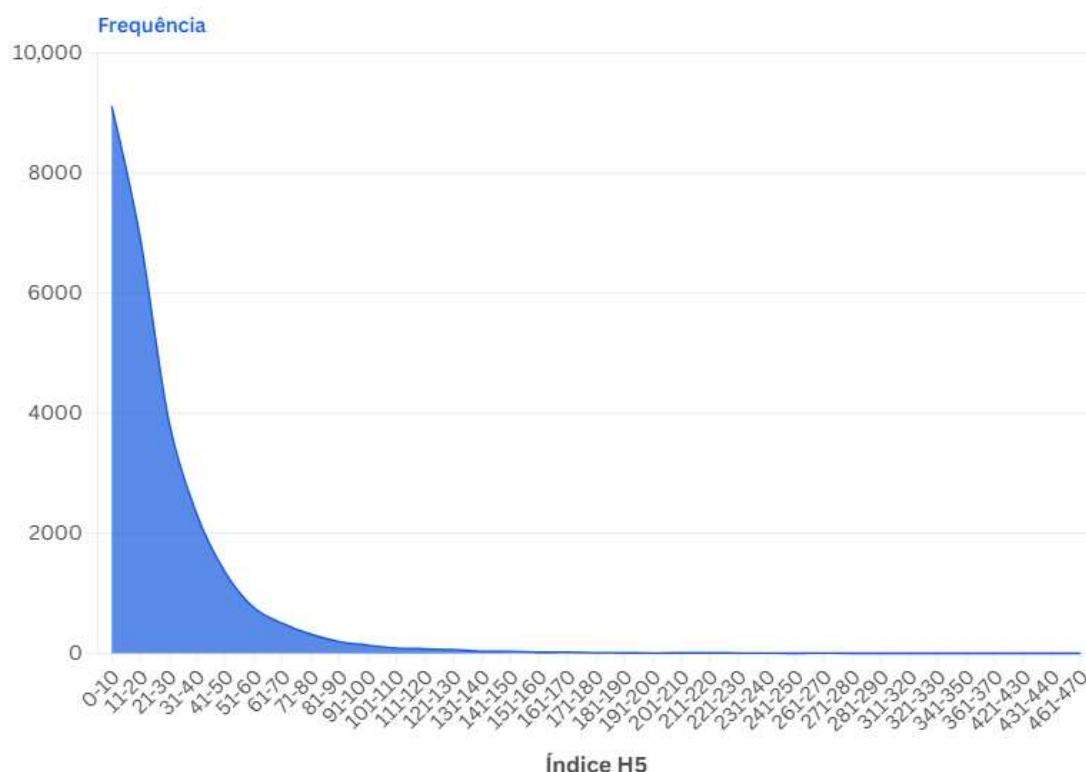
5.1 ANÁLISES BIBLIOMÉTRICAS

Nesse tópico, será apresentada a análise dos dados de citações das 25.845 revistas levantadas pelo processo de coleta e validação dos dados. Como descrito anteriormente, para 124 revistas, não foi possível certificar que eram as revistas esperadas, precisando de uma checagem complementar em outras fontes de informação, além das do presente estudo. Portanto, as mesmas serão desconsideradas da análise.

Inicialmente, serão descritos os dados bibliométricos coletados do GSM, das 25.845 revistas, e posteriormente serão complementados com os dados do OpenAlex e as métricas desenvolvidas para essa pesquisa. Os dados provenientes do GSM são o índice h5 e sua mediana, já do OpenAlex foi coletado o índice h. Com base nos artigos publicados nos cinco anos analisados do OpenAlex, foram calculadas as métricas de índice h5, percentual de não citados, índice de Gini, média de citação dos 5 anos e contabilizado o total de citações e de artigos publicados.

O próximo gráfico irá demonstrar como os dados estão distribuídos com base no Índice h5 do GSM. Esse gráfico representa um panorama geral da frequência do Índice h5 em classes de 10 em 10.

Gráfico 7 - Distribuição de frequência do índice h5 do GSM



Fonte: Dados da pesquisa.

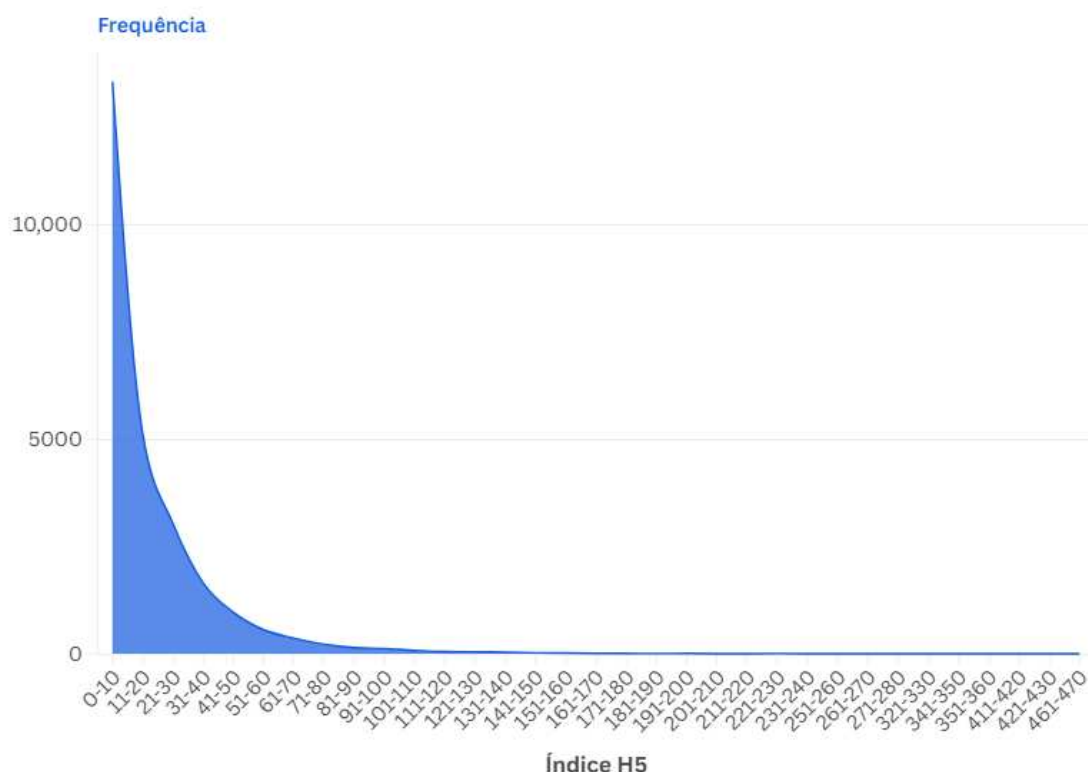
Nota: Métrica coletada do GSM.

Este gráfico evidencia a distribuição assimétrica do índice analisado, com forte concentração nos intervalos iniciais. A maior parte dos periódicos está concentrada na faixa de 0 a 10 (9.119 registros), seguida pelas faixas de 11 a 20 (6.924 registros) e 21 a 30 (3.887 registros). Em conjunto, essas três faixas representam aproximadamente 75% do total de casos, indicando que a maior parte dos periódicos apresenta índice relativamente baixo.

Os indicadores estatísticos reforçam esse padrão: a média geral é de 22,46, a mediana é 15, e a moda é 6, confirmando uma distribuição com assimetria positiva. Esse perfil sugere que, embora existam periódicos de altíssimo impacto, a grande maioria está agrupada em patamares mais modestos, compondo uma distribuição de impacto fortemente concentrada entre os índices h5 de 1 a 30, evidenciado no gráfico.

Na análise seguinte, serão utilizados os dados do índice h5 elaborados nesta pesquisa, com o objetivo de verificar sua distribuição e identificar possíveis semelhanças com os padrões observados anteriormente. O Gráfico 8 apresenta a distribuição desse índice, permitindo uma visualização detalhada das faixas de impacto e sua frequência relativa.

Gráfico 8 - Distribuição de frequência do índice h5 da pesquisa



Fonte: Dados da pesquisa.

Nota: Métrica elaborada na pesquisa.

A distribuição dos periódicos segundo o índice h5 revela um padrão semelhante ao observado para o índice h5 recuperado do GSM, indicando coerência entre esses dois indicadores de impacto. A maior concentração de casos está na faixa de 0 a 10, com 13.333 registros, representando aproximadamente 51,6% do total de periódicos analisados (25.845). Essa concentração se mantém significativa nas faixas de 11 a 20 (5.128 registros; ~19,8%) e 21 a 30 (3.038 registros; ~11,8%), evidenciando que cerca de 83% dos periódicos apresentam h5 igual ou inferior a 30.

Os indicadores estatísticos corroboram esse padrão de concentração em níveis mais baixos de impacto: a média do índice h5 é de 17,48, enquanto a mediana é ainda menor (10), o que confirma a assimetria positiva da distribuição. A moda é igual a 2, o que reforça a alta frequência de periódicos com índice h5 baixo. O valor máximo identificado é 461, muito superior à média, o que confirma a existência de outliers que alongam a cauda direita da distribuição.

Ao comparar o índice h5 com o índice h5 do GSM, ambos mostraram forte concentração em faixas de até 30 pontos, o que sugere uma consistência na medição de impacto: a maioria dos periódicos está agrupada em patamares de baixa visibilidade, enquanto apenas uma parcela atinge valores mais expressivos. Esse

padrão indica uma distribuição com poucos periódicos de alto desempenho e grande maioria com alcance mais restrito.

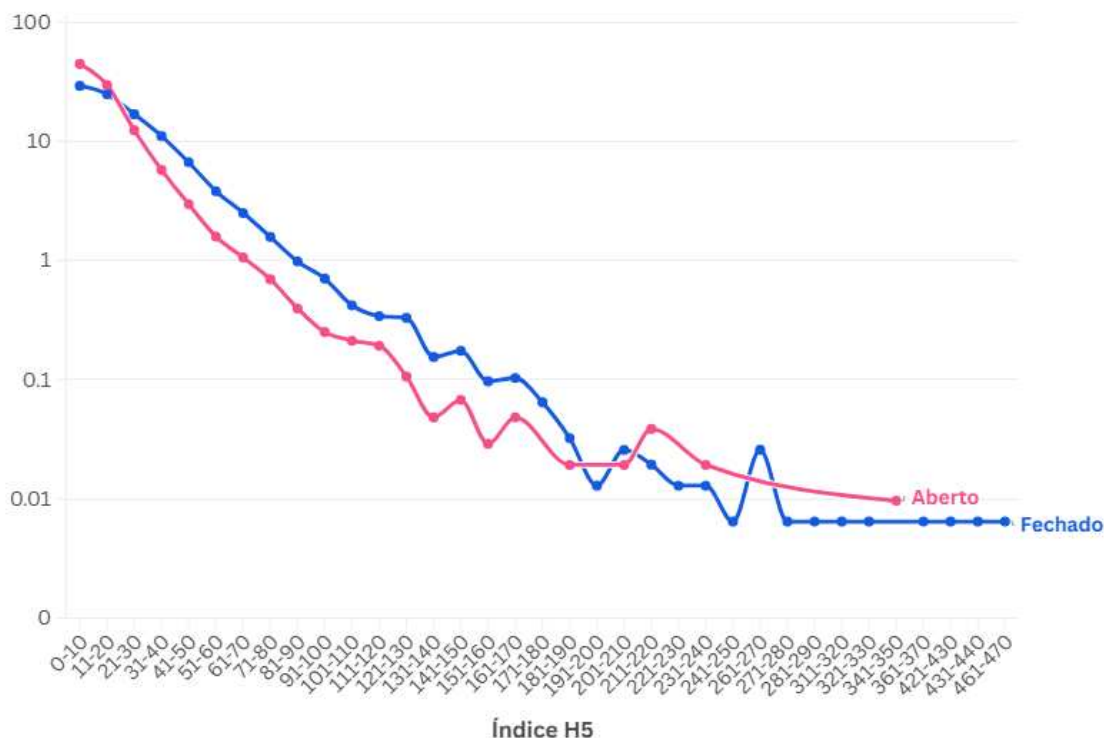
Para verificar a consistência entre os resultados obtidos na presente pesquisa e os dados do GSM, realizou-se uma análise de correlação utilizando o coeficiente de Spearman. O valor obtido ($\rho = 0,91$) indica uma correlação positiva muito forte entre o índice h5 calculado a partir dos dados bibliográficos coletados e o índice h5 disponibilizado pelo GSM, demonstrando que ambos os indicadores apresentam uma relação monotônica elevada.

Assim, os gráficos anteriormente encontram respaldo estatístico, confirmando que os periódicos com maior desempenho em uma base tendem a manter sua posição relativa na outra.

Além disso, foi realizada uma análise de correlação entre o índice h geral obtido a partir dos dados do OpenAlex e o índice h5 disponibilizado pelo GSM, com o intuito de verificar o grau de alinhamento entre essas duas métricas de impacto. O coeficiente de Spearman calculado ($\rho = 0,84$) indica uma correlação positiva forte, demonstrando que, embora as metodologias de apuração dos índices apresentem diferenças em sua estrutura temporal, existe uma tendência consistente de que periódicos mais bem posicionados mantenham desempenho ao longo do tempo.

De forma complementar, foi verificado se as revistas de acesso aberto tendem a apresentar um índice h5 do GSM mais elevado em comparação às revistas de acesso restrito. O gráfico abaixo mostra a distribuição percentual relativa dos periódicos por faixas do índice h5, considerando um total de 15.482 periódicos de acesso restrito e 10.363 de acesso aberto. A visualização em escala logarítmica torna possível identificar diferenças sutis, sobretudo nas faixas superiores de impacto, onde a concentração é menor, mas pode indicar variações relevantes entre os grupos.

Gráfico 9 - Distribuição do percentual relativo dos periódicos por faixa do índice h5



Fonte: Dados da pesquisa.

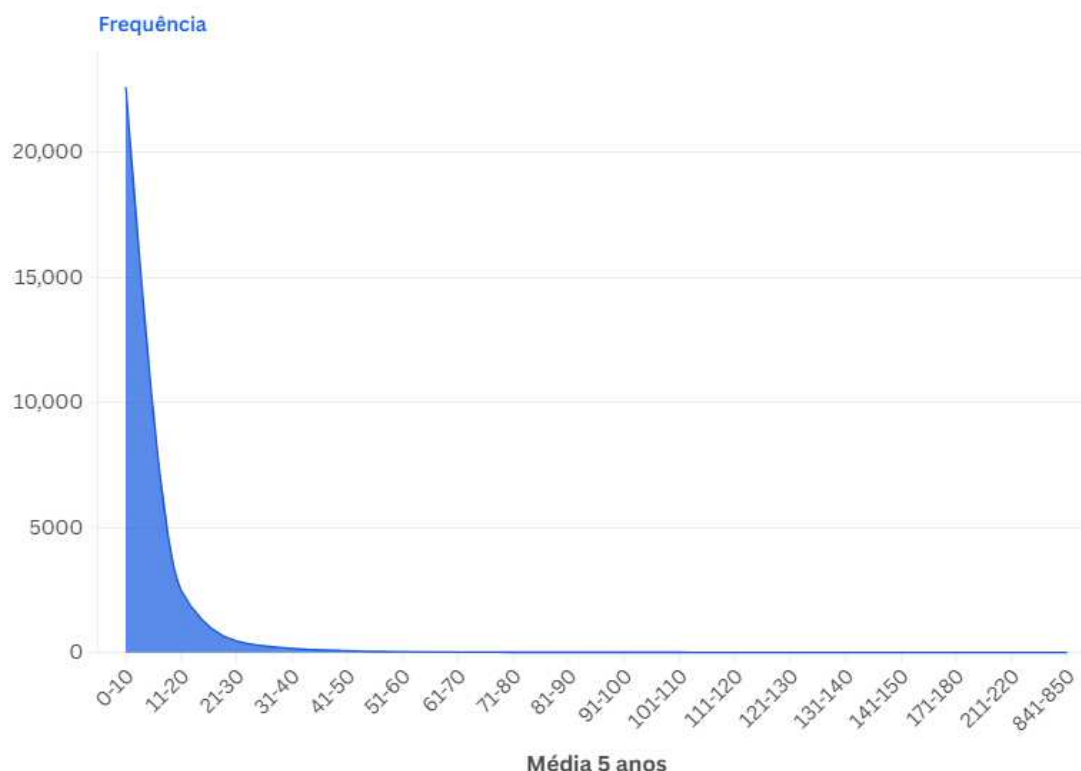
O gráfico revela que, nas classes mais baixas, de índice h5 inferior a 30, as proporções entre revistas de acesso aberto e restrito são bastante próximas, mantendo um padrão semelhante de concentração. No entanto, à medida que os valores do índice aumentam, observa-se um distanciamento progressivo entre as linhas, indicando diferenças no impacto em relação ao tipo de acesso. Esse comportamento evidencia que periódicos de acesso restrito concentram uma proporção maior de casos em faixas de alto impacto, enquanto os periódicos de acesso aberto tendem a se concentrar em níveis mais baixos. A parte final do gráfico, que representa a cauda longa da distribuição, reforça essa constatação, mostrando que as classes superiores chegam a registrar a presença isolada de apenas uma revista em cada faixa.

A análise estatística complementa essa tendência ao mostrar que as revistas de acesso restrito apresentam, em média, um índice h5 mais elevado, 25,81, e mediana superior, 19, em comparação aos periódicos de acesso aberto, cuja média de 17,46 e mediana de 12. A moda coincide em ambos os grupos (6) e evidencia a concentração predominante nos intervalos iniciais, mas o desvio padrão maior no grupo restrito 25,62 contra 18,40 aponta para uma maior heterogeneidade, com

presença de casos extremos de alto impacto, máximo de 467 frente a 349 nos abertos. Esses resultados reforçam a interpretação visual do gráfico, indicando que, em geral, periódicos de acesso restrito, nessa amostra, alcançaram valores mais altos de índice h5, embora a maioria permaneça agrupada em faixas de impacto moderado a baixo.

Para verificar a relação entre o volume de produção científica e o impacto dos periódicos, foi realizada uma análise de correlação entre o número total de artigos publicados e o índice h5 calculado no presente estudo. Optou-se por utilizar o índice h5 gerado a partir dos dados da pesquisa, uma vez que o GSM não disponibiliza o número total de artigos utilizados no cálculo de seus indicadores. A correlação de Spearman obtida ($\rho = 0,64$) indica uma associação positiva moderada entre as variáveis, sugerindo que periódicos com maior volume de publicações tendem a apresentar maior impacto, ainda que essa relação não seja linear. Esse resultado reforça a influência do tamanho da produção na construção de indicadores bibliométricos como no caso do índice h, além disso mostra que o volume de produção influencia o impacto, mas não o determina. Com base nisso, foi verificada a distribuição da média de citações dos artigos nos últimos cinco anos em relação ao total de publicações. O gráfico abaixo ilustra em que pontos há mais concentração de periódicos.

Gráfico 10 - Distribuição da média de 5 anos



Fonte: Dados da pesquisa.

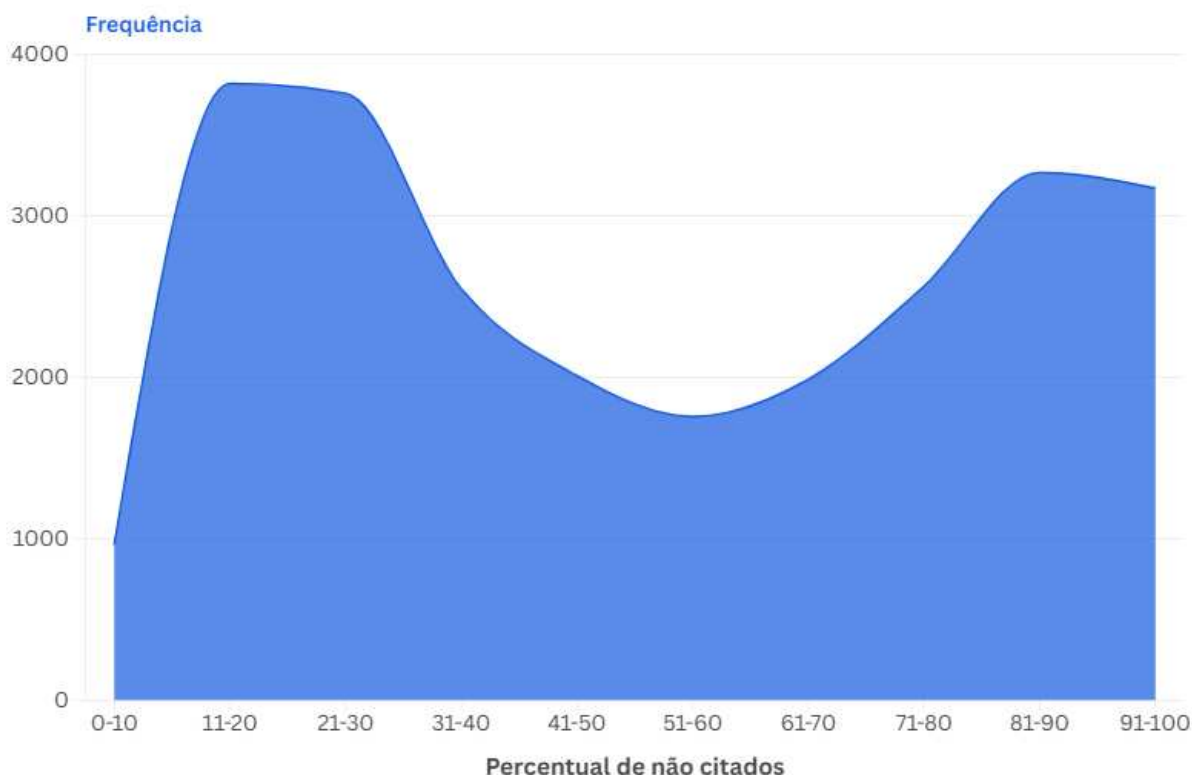
Nota: Métrica elaborada na pesquisa.

A análise da média de citações revela um padrão de concentração semelhante ao identificado nos dados do índice h5. A maioria dos periódicos (87,5%) apresenta médias de citação inferiores a 10, com destaque para a mediana inferior a 2 e a moda igual a 0, o que indica que uma parte significativa dos artigos recebeu poucas ou nenhuma citação nesse período.

Apesar disso, a presença de valores extremos, como o máximo de 847, revelam uma cauda longa na distribuição, caracterizada por poucos periódicos com média de citação muito acima da média geral (4,56). A distribuição por classes reforça esse cenário: apenas 8 periódicos superam a faixa das 100 citações em média.

Com base na análise de impacto, observou-se uma forte concentração de periódicos em faixas com indicadores mais modestos. De forma complementar, foi analisado o percentual de artigos não citados, como forma de entender o desempenho dos periódicos sobre esse aspecto. O gráfico a seguir apresenta essa distribuição.

Gráfico 11 - Distribuição das revistas por percentual de artigos não citados



Fonte: Dados da pesquisa.

Nota: Métrica elaborada na pesquisa.

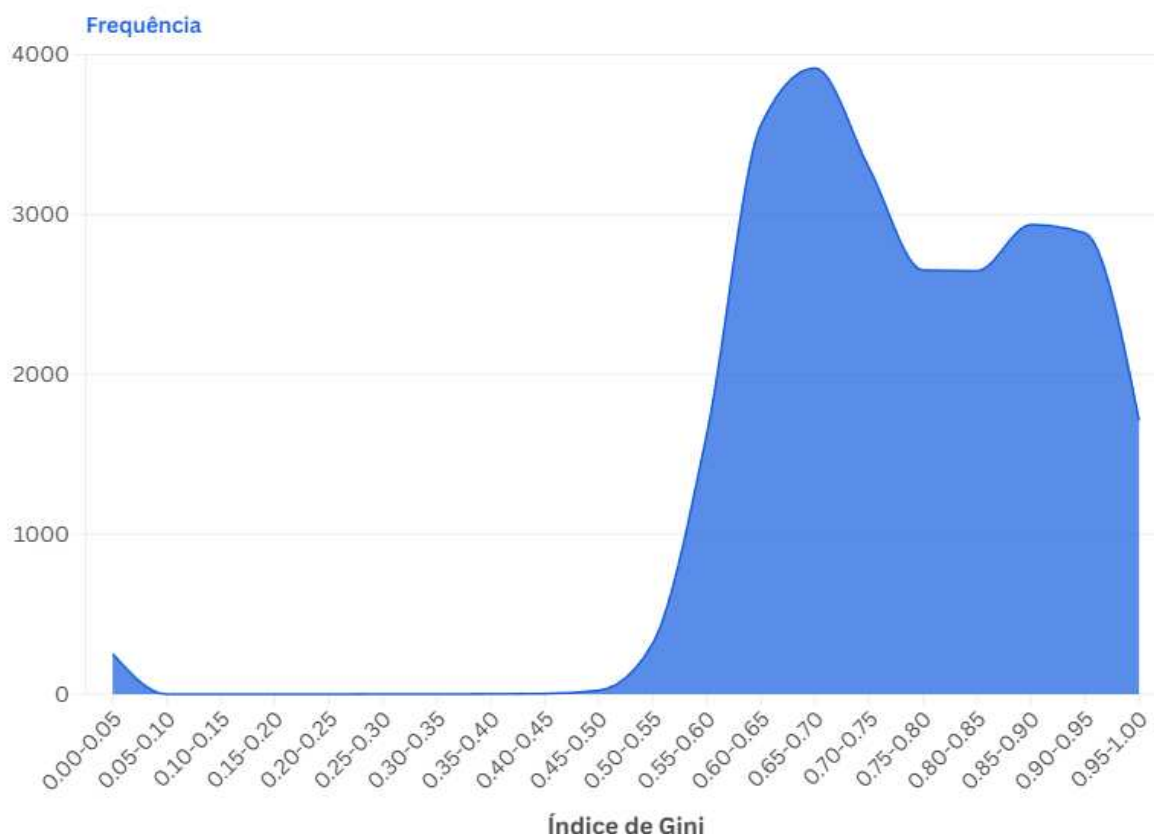
O gráfico evidencia a presença de uma distribuição bimodal, com duas curvas distintas: uma situada à direita, representando periódicos com um elevado percentual de artigos não citados, e outra concentrada na faixa entre 20% e 30% de não citação, seguida por uma queda acentuada. Esse padrão pode ser influenciado por limitações na indexação dos artigos na base OpenAlex, especialmente em periódicos com menor cobertura ou atualização irregular. Tais fatores podem comprometer a contabilização precisa das citações, resultando em percentuais artificialmente elevados de não citação em determinadas revistas.

Sobre a medida de tendência central da análise, em média, 51,73%, dos artigos publicados, indicam que mais da metade das revistas analisadas não receberam citações. A mediana de 49,27% reforça essa tendência, sugerindo que pelo menos metade das revistas possuem um percentual de não citados próximo ou acima de 50%.

Dando continuidade à análise dos casos de artigos não citados, foi realizado o cálculo do índice de Gini, com o objetivo de medir o grau de concentração de

citações entre os artigos. O gráfico a seguir apresenta a distribuição dos valores desse indicador.

Gráfico 12 - Concentração de citações por artigo com base no índice de Gini



Fonte: Dados da pesquisa.

Nota: Métrica elaborada na pesquisa.

A análise do índice de Gini revela um elevado grau de desigualdade na distribuição de citações entre os artigos. A média do índice foi de aproximadamente 0,76, com mediana de 0,75, indicando uma forte tendência de que as citações se concentram em poucos artigos.

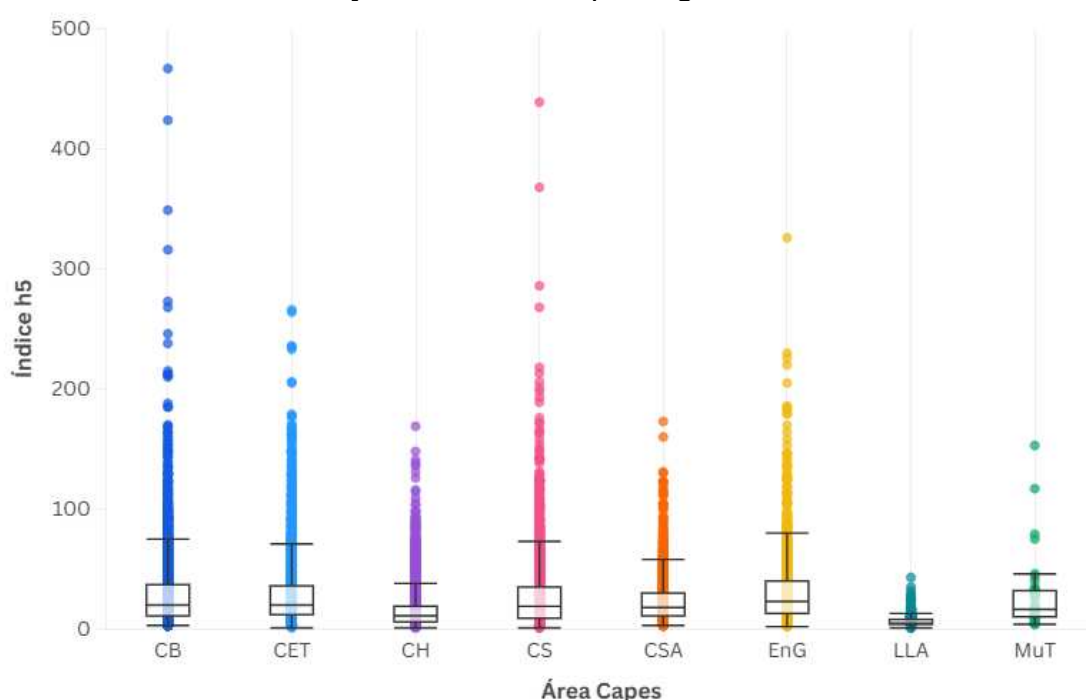
Esse padrão é ainda mais evidente na distribuição por faixas. A maioria dos periódicos apresenta valores de Gini entre 0,60 e 0,95 (85%) com destaque para a faixa 0,65–0,70, que concentra 3.917 casos, seguida por outras faixas superiores a 0,70, evidenciando que poucos artigos acumulam a maior parte das citações.

É importante destacar que 254 periódicos aparecem na extremidade esquerda do gráfico, na faixa de Gini entre 0,00 e 0,05, o que, à primeira vista, poderia sugerir uma boa distribuição de citações entre os artigos. No entanto, esses casos refletem, na realidade, ausência de citações, ou seja, periódicos cujos artigos não foram citados.

Considerando todas as faixas com índice de Gini abaixo de 0,55, que correspondem às revistas com melhor equilíbrio na distribuição de citações, observa-se apenas 352 casos, o que representa aproximadamente 1% do total analisado. Esse dado reforça a constatação de que a concentração de citações em poucos artigos é uma característica predominante entre os periódicos da amostra.

Não obstante, torna-se relevante observar como essa dinâmica se distribui entre as diferentes grandes áreas do conhecimento. Nesse sentido, a seguir apresenta a distribuição dos periódicos conforme as grandes áreas da CAPES.

Gráfico 13 - Distribuição do índice h5 pelas grandes áreas da CAPES



Fonte: Dados da pesquisa.

Notas: CB (Ciências Biológicas), CET (Ciências Exatas e da Terra), CH (Ciências Humanas), CS (Ciências da Saúde), CSA (Ciências Sociais Aplicadas), EnG (Engenharias), LLA (Linguística, Letras e Artes), MuT (Multidisciplinar).

A análise da distribuição do índice h5 por grandes áreas CAPES, representada nos Boxplots, confirma a tendência geral observada anteriormente: há uma alta concentração de periódicos com valores abaixo de 50, com os quartis 2, 3 e 4 situando-se nesse intervalo para praticamente todas as grandes áreas. As medianas oscilam entre 5 e 23, evidenciando que a maior parte das revistas tem impacto moderado quando comparadas aos valores máximos registrados em cada área.

As grandes áreas de Ciências Biológicas (CB) e Ciências da Saúde (CS) concentram os periódicos de maior impacto, com médias de índice h5 de 29,3 e

25,7, respectivamente, e medianas semelhantes (20 e 19). Ambas apresentam valores máximos elevados (CB com 467 e CS com 439), o que se reflete em caixas mais altas e presença de outliers nos boxplots.

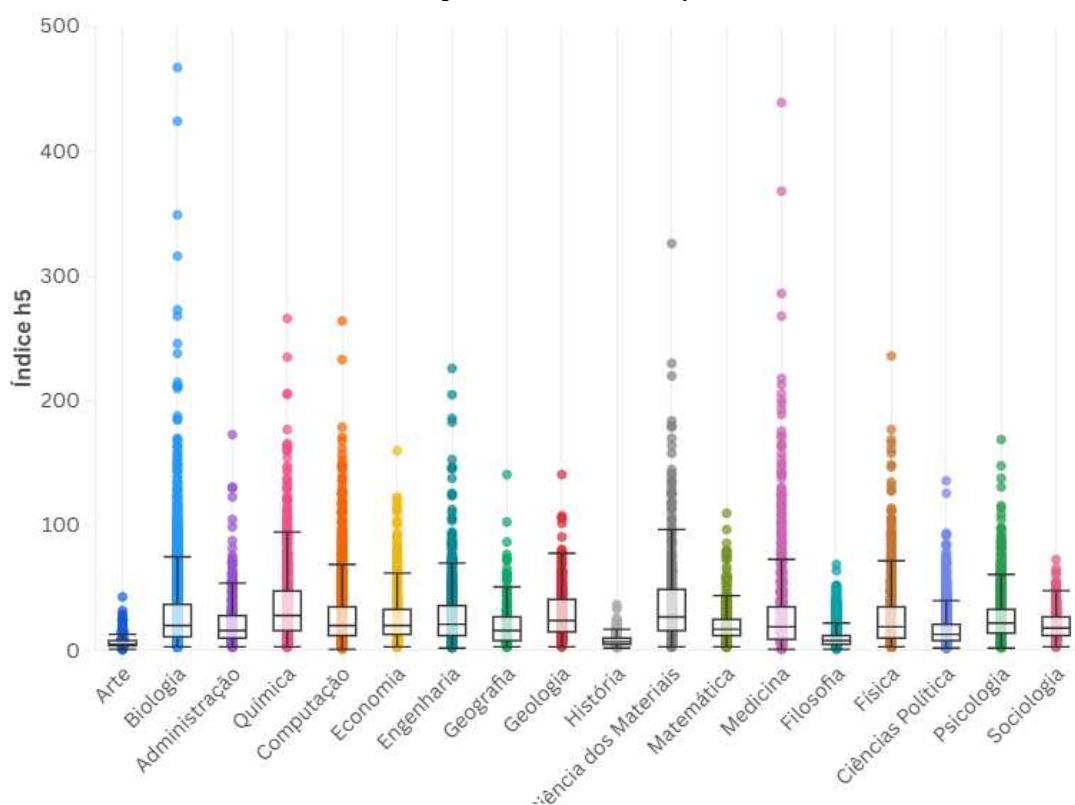
Na sequência, destacam-se Ciências Exatas e da Terra (CET) e Engenharias (EnG), com distribuições amplas e valores de tendência central também elevados, média de 28,5 e mediana 20 para CET; média de 32,3 e mediana 23 para EnG. A grande área de Ciências Sociais Aplicadas (CSA), embora com uma dispersão menor, apresenta média de 24 e mediana de 18, também indicativas de bom desempenho.

Por outro lado, grandes áreas como Ciências Humanas (CH) e Linguística, Letras e Artes (LLA) apresentam indicadores visivelmente inferiores: CH com média de 15 e mediana de 11, enquanto LLA registra média de apenas 6,3 e mediana de 5, com caixas estreitas que indicam baixa variabilidade.

A grande área Multidisciplinar (MuT) apresenta média de 27,5 e mediana de 16,5, sugerindo desempenho relativamente elevado. No entanto, com apenas 41 periódicos na amostra, os resultados devem ser interpretados com cautela, uma vez que o número reduzido pode comprometer a representatividade estatística.

A seguir será apresentada a distribuição do índice h5 por áreas de avaliação. Essa abordagem permite identificar variações internas dentro de cada grande área, destacando quais áreas concentram os periódicos de maior ou menor impacto, e contribuindo para uma compreensão mais refinada da dinâmica de citações e desempenho bibliométrico em diferentes campos do conhecimento.

Gráfico 14 - Distribuição do índice h5 por área da CAPES



Fonte: Dados da pesquisa.

A análise da distribuição do índice h5 por área confirma os padrões já identificados em níveis mais acima descritos. Nota-se uma predominância de valores de índice h5 abaixo de 50, o que é evidenciado pelas medianas que, em sua maioria, variam entre 16 e 28, mesmo em áreas com valores máximos mais elevados, como Biologia e Medicina. A concentração de periódicos em faixas de impacto moderado segue como característica comum, embora algumas áreas apresentem desempenho levemente superior.

As áreas de Geologia, Química e Ciência dos Materiais se destacam com os melhores desempenhos em termos de índice h5 entre as analisadas. Geologia tem média de 30,1 e mediana de 24 e valor máximo de 141, acima vem a Química que apresenta média de 38 e mediana de 28 e valor máximo de 266, refletindo uma distribuição elevada. A Ciência dos Materiais, por sua vez, possui média ainda mais alta de 38,8, mediana de 27 e desvio padrão de 37,3, com valor máximo atingindo 326.

As áreas de Psicologia e Matemática apresentam desempenhos consistentes em termos de índice h5, destacando-se especialmente por concentrarem a moda próxima à mediana, o que sugere uma distribuição mais

equilibrada entre os periódicos. Psicologia registra média de 25,8, mediana de 22 e moda de 18 e valor máximo de 169. De forma semelhante, Matemática apresenta média de 20,6, mediana de 17 e moda de 16, com menor dispersão, desvio padrão de 13,8, e valor máximo de 110. A proximidade entre os valores centrais e a moda em ambas as áreas reforça um padrão de desempenho sólido e bem distribuído entre os títulos analisados.

Algumas áreas apresentam valores médios elevados, puxados por revistas com índice h5 alto, o que se reflete também na alta dispersão dos dados. É o caso da Biologia com média de 29,4, desvio padrão de 30,4 e máximo de 467, Engenharia com média, 28,3, desvio padrão de 26,1 e máximo 226, Física com média de 27,2, desvio padrão de 25,8 e máximo de 236, e Medicina com média de 25,7, desvio padrão de 24,7 e valor máximo de 439. Nesses casos, poucos periódicos com alto impacto elevam a média do grupo, enquanto a maioria permanece concentrada em faixas médias ou baixas. Isso é confirmado pelo desvio padrão e sua mediana entre 19 e 21 para essas áreas.

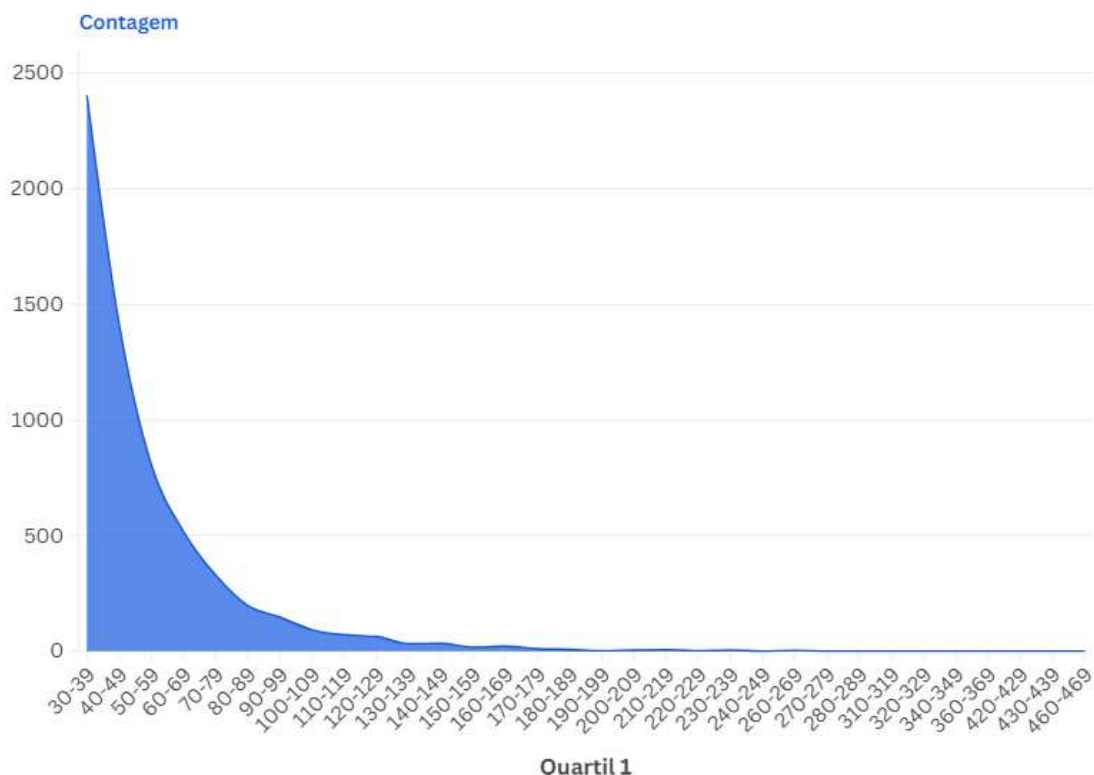
Por outro lado, áreas como Artes (média 6,3; mediana 5), História (média 8,2; mediana 7) e Filosofia (média 9,4; mediana 8) apresentaram os menores valores médios e medianas, com distribuições concentradas em faixas de baixo impacto. Isso pode estar relacionado ao perfil de citação dessas áreas, que tende a ser mais localizado e menos sensível a métricas baseadas em volume de citações.

Áreas como Geografia, Ciência Política, Economia, Sociologia e Administração revelaram desempenhos intermediários, com médias entre 16 e 26 e medianas entre 13 a 20 pontos. Sugerindo um impacto moderado e relativamente homogêneo dentro dos respectivos campos.

É importante destacar que não foram incluídos nos gráficos os dados das áreas de Educação, Ciência Ambiental, Direito e Linguística, devido ao número reduzido de periódicos nesses campos. A análise por áreas, portanto, permite identificar com maior detalhe os focos de concentração de impacto e os campos que apresentam maior heterogeneidade na visibilidade de seus periódicos.

Após a análise por grande área e área, buscou-se aprofundar a compreensão da distribuição dos periódicos ao segmentá-los por quartis. O gráfico abaixo destaca essa visão do quartil com impacto mais alto, permitindo observar a variação do índice h5 neste grupo.

Gráfico 15 – Distribuição do índice h5 por quartil 1



Fonte: Dados da pesquisa.

Nota: Métrica elaborado para a pesquisa.

O Quartil 1 (Q1) reúne os periódicos com maior impacto da amostra, totalizando 6.203 revistas, todas com índice h5 igual ou superior a 30. Esse ponto de corte evidencia que cerca de 76% de todas as revistas analisadas têm índice h5 inferior a 30.

Dentro do Q1, que abrange índices entre 30 e 467, a média é 53,07, refletida pela presença de periódicos altamente citados. Pois à medida que o índice h5 aumenta, observa-se uma queda progressiva na frequência, em que apenas 30 revistas ultrapassam a faixa de h5 igual a 200, o que representa menos de 0,5% da amostra.

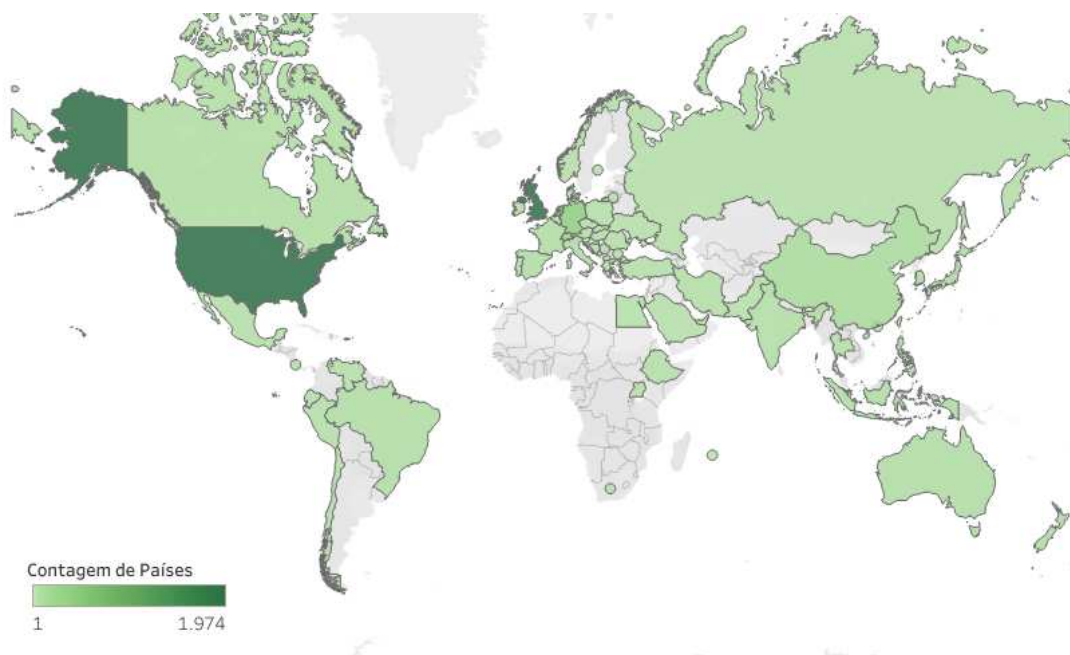
A distribuição dos periódicos revela uma forte concentração nas faixas iniciais da escala do índice h5. Aproximadamente 73% das revistas estão concentradas entre os intervalos de 30 a 59, com destaque para a classe 30–39, que sozinha reúne 2.404 títulos, correspondendo a mais de um terço do total.

Esse padrão evidencia que, ainda dentro do grupo mais qualificado, a maior parte das revistas está concentrada nos intervalos intermediários e poucos títulos

dominando os maiores índices de impacto. A distribuição reforça a ideia de que o alto impacto extremo é alcançado por um grupo seletivo e restrito de periódicos.

Abaixo será apresentado um mapa dos países com mais revistas dentro do Q1. Esse mapa possibilita identificar quais países concentram as revistas mais influentes.

Figura 19 - Mapa dos países com revistas no quartil 1



Fonte: Dados da pesquisa.

Com base no mapa dos países com maior número de periódicos no Q1, é possível observar padrões marcantes de concentração geográfica na produção científica de alto impacto, medida pelo índice h5.

Os Estados Unidos e o Reino Unido se destacam de forma expressiva, com 1.974 e 1.930 revistas no Q1, respectivamente. Juntos, esses dois países concentram mais de 60% dos periódicos de maior impacto analisados. Em seguida, aparece um grupo secundário formado por Holanda com 887 periódicos, Alemanha com 348 e Suíça com 274, reforçando o predomínio europeu nesse estrato superior.

A presença de países asiáticos como China, Coreia do Sul, Japão, Singapura e Índia, embora mais modesta em número absoluto, indica o avanço de centros emergentes na ciência global, com inserção crescente em revistas de maior visibilidade. O Brasil aparece com 41 periódicos no Q1, evidenciando avanço dos países emergentes entre as revistas com maior impacto.

Outros países da América Latina e da África têm representações muito pequenas ou pontuais, como México (6), Chile (4), Colômbia e Peru (3 cada), África do Sul (7) e Uganda (2), revelando a barreira que esses países enfrentam para visibilidade internacional de suas revistas.

A análise bibliométrica realizada com base no índice h5 permitiu identificar padrões consistentes de concentração do impacto científico entre revistas, áreas do conhecimento e países. A seguir, parte-se para a discussão desses achados.

5.2 DISCUSSÃO

O levantamento global por meio de um procedimento de conferência manual de periódicos realizado por Gu e Blackmore (2017) identificou 10.354 revistas de 41.787 periódicos consultados. Isso confere à presente pesquisa um resultado satisfatório, visto que se trata de um procedimento automatizado, e que encontrou um total de 25.969 revistas válidas para análise, de 41.394 de respostas possíveis.

Embora deva-se considerar a diferença temporal, uma vez que, em 2017, o GSM poderia indexar um número menor de revistas, a diferença observada entre os resultados das pesquisas é significativa. Ademais, a técnica utilizada por Gu e Blackmore foi manual, enquanto o presente estudo utilizou uma abordagem automatizada, o que colabora para o bom desempenho do GSM-ALEX_metrics desenvolvido. A eficiência do processo automatizado possibilita não apenas uma cobertura mais ampla, mas também maior reprodutibilidade dos resultados por meio da aplicação do GSM-ALEX_metrics.

Outro estudo de referência, voltado à América Latina e Caribe, foi o de Canto *et al.* (2022), que identificou 3.070 periódicos. A presente pesquisa encontrou 2.612 revistas da mesma região, o que demonstra uma eficiência de recuperação. Apesar da variação de 458 revistas, isso pode ser atribuído tanto à exclusão de periódicos que podem ter deixado de atender aos critérios do GSM durante esse período, quanto à ausência de alguns artigos indexados na base OpenAlex, o que impediu a validação final desses títulos.

A maior parte dos estudos acaba se concentrando em levantamento por determinadas disciplinas, como o caso de Gonzalez-Pardo, Repiso e Arroyave-Cabrera (2020), que analisaram 297 revistas de comunicação ibero-americanas, ou como Ahmad *et al.* (2019) que analisaram apenas 10 revistas da Arábia Saudita.

Ainda, Camacho *et al.* (2019) realizaram levantamento por base e analisou 506 revistas indexadas no Latindex. De forma geral, talvez pela dificuldade de extrair dados em larga escala do GSM, não foram identificados muitos estudos que objetivam uma verificação em múltiplas áreas.

Outro ponto sobre os estudos que envolvem o GSM é o fato de que poucos fazem o uso automatizado da coleta de dados. Alguns identificados na literatura como o estudo de Costa, Canto e Pinto (2020), que usaram Publish or Perish, ou como Costa *et al.* (2023), Pinto *et al.* (2020), Canto *et al.* (2023) que fazem uso de procedimentos próprios automatizados. Estes últimos estão relacionados a esta pesquisa, uma vez que este estudo é um aperfeiçoamento desses sistemas para coleta de dados do GSM.

Singh *et al.* (2021) estudaram a cobertura das grandes áreas do conhecimento e verificaram que a Web of Science tem cerca de 1.800 títulos na grande área de Ciências Humanas, e a Scopus, com aproximadamente 5 mil títulos da mesma grande área. O presente estudo encontrou 8.464 revistas no GSM. As áreas com mais revistas foram de Filosofia, com 3.783, Ciência Política, com 2.009 e Artes, com 1.171. Esse resultado demonstra o potencial do GSM em captar indicadores para áreas que historicamente são sub-representadas em bases tradicionais. Esse nível de cobertura amplia significativamente as possibilidades de avaliação de impacto em campos tradicionalmente marginalizados pelas grandes bases comerciais.

O Google Scholar, por meio do GSM, oferece uma alternativa relevante a bases como Web of Science e Scopus, ao criar rankings de periódicos baseados no índice h e no h5, que compartilham o mesmo fundamento (Delgado López-Cózar; Orduña-Malea; Martín-Martín, 2019). A principal vantagem desses indicadores é a capacidade de sintetizar, em um único número, tanto a quantidade quanto a qualidade da produção científica, estabelecendo uma relação direta entre o volume de artigos publicados e o impacto medido pelas citações recebidas (Barnes, 2017). Essa característica confere ao índice h e ao h5 uma aplicação versátil, permitindo seu uso na avaliação de pesquisadores, departamentos acadêmicos e periódicos, e sendo considerados por muitos uma métrica mais equilibrada e justa em comparação a outros indicadores tradicionais (Bornmann, 2014).

Apesar de sua relevância, o índice h é alvo de críticas por considerar apenas os artigos mais citados, desconsiderando os menos citados e o total de citações

recebidas, o que limita sua capacidade de oferecer uma visão abrangente da produção científica (Bornmann, 2014; Silva; Dobránszki, 2018).

Com relação aos dados analisados, considerando que o índice h5 é uma métrica que combina produção e impacto, é esperado que a maioria dos periódicos se concentre nas faixas mais baixas da distribuição. Para que esse índice se eleve, é necessário que haja uma maior distribuição das citações por artigo. A distribuição por faixa de 10 pontos confirmou essa tendência, com cerca de 76% das revistas com índice h5 inferior a 30. Além disso, a medida do percentual de não citados e o índice de Gini mostram também que há uma concentração de citação em grupos pequenos de artigos.

O índice de Gini, utilizado para mensurar a desigualdade de citação entre os artigos de cada periódico, revelou concentração excessiva de impacto em poucos trabalhos. Apenas 1% dos periódicos analisados apresentaram Gini abaixo de 0,55, o que indica uma distribuição desigual, pois quanto mais próximo de zero, mais igualitária seria a distribuição das citações. A maioria dos periódicos concentra suas citações em poucos artigos.

Asubiaro e Onaolapo (2024) analisaram a regionalização dos periódicos e observaram uma forte probabilidade de revistas provenientes da Europa, América do Norte e Oceania serem indexadas em bases como Web of Science e Scopus, em comparação a outras regiões do mundo. No presente estudo, verificou-se um cenário semelhante, com maior concentração de periódicos oriundos da Europa e América do Norte. A diferença principal foi em relação à Oceania, que aqui apresentou menor representatividade, enquanto a Ásia se destacou por uma participação proporcionalmente maior.

Ao verificar quais regiões apresentam os melhores indicadores h5, observa-se que a Europa e a América, somadas, concentram 78% do total de revistas identificadas, com destaque para Estados Unidos e Reino Unido, que juntos somam mais de 3.900 periódicos no quartil mais alto (Q1). Esse cenário reforça o papel do GSM como ferramenta voltada majoritariamente para países ocidentais, com menor representatividade para regiões do Oriente. Tais resultados apontam para um desequilíbrio de visibilidade internacional e destacam a necessidade de ampliar a diversidade geográfica em bases de dados científicas.

Essa concentração também se manifesta nas métricas de citação. Aproximadamente 50% de todos os artigos não receberam nenhuma citação no

período analisado de cinco anos, o que evidencia uma elevada proporção de não citados. Como destacado por Dorta-González *et al.* (2020), essa taxa depende fortemente da janela de tempo, da disciplina e do tipo de documento. Portanto, uma análise por área se torna fundamental para compreender as nuances e os padrões de citação específicos de cada campo do conhecimento, bem como identificar as razões pelas quais metade dos artigos não foi citada.

Foi realizada uma análise de correlação entre o índice *h* obtido no OpenAlex e o índice *h5* do GSM, que revelou uma relação positiva. Esse resultado indica uma tendência de a mesma revista manter sua posição no ranking ao longo do tempo, corroborando o chamado “efeito Mateus”, descrito por Merton (2013).

Outro aspecto investigado foi se as revistas de acesso aberto apresentam índice *h5* e desempenho de citação superior em relação às revistas de acesso restrito, considerando-se a vantagem de não possuírem barreiras monetárias para leitura. Contudo, os resultados deste estudo não confirmaram essa hipótese, divergindo de parte da literatura sobre o tema. Uma revisão sistemática conduzida por Langham-Putrow, Bakker e Riegelman (2021) identificou que 47,8% dos estudos analisados apontam maior número de citações para revistas de acesso aberto, 27,6% não encontraram diferença significativa e 23,9% observaram esse aumento apenas em subconjuntos específicos. Portanto, os achados desta pesquisa alinham-se ao grupo de estudos que não identificaram aumento significativo de citações decorrente do acesso aberto. Esse resultado pode estar relacionado ao fato de que cerca de 36% das editoras com maior número de revistas analisadas são comerciais e concentram os melhores indicadores. Em razão disso, é provável que muitos autores priorizem esses periódicos, reforçando, em cadeia, o impacto já acumulado por esse grupo.

Durante o processo, foram identificadas inconsistências tanto no GSM quanto no OpenAlex. No GSM, 380 revistas foram excluídas por duplicidade de registros com URLs diferentes, mas IDs iguais. Além disso, em 124 casos não foi possível determinar com precisão a causa das divergências entre os títulos comparados, o que pode ter sido provocado por mudanças de nome da revista ou pela indexação de artigos de outras revistas sob o mesmo título.

No OpenAlex, foram observadas lacunas no preenchimento de metadados. Houve casos de revistas com múltiplos IDs em razão da indexação por título em diferentes idiomas, além de artigos com títulos bilíngues no mesmo registro. Campos

importantes como editora e país também apresentaram ausência de preenchimento, assim como o campo “conceito”, gerado automaticamente. Essas limitações impactam diretamente a capacidade de análise por instituições, países e áreas temáticas.

De forma geral, essas inconsistências não comprometem os resultados do estudo, uma vez que os problemas de indexação representam menos de 1% do total analisado. As ausências nos campos de editora e país também foram pontuais, afetando cerca de 4% do conjunto de dados. Portanto, os resultados apresentados são estatisticamente robustos e representam de maneira fidedigna a composição e o impacto dos periódicos indexados no GSM e OpenAlex.

6 CONSIDERAÇÕES FINAIS

Este estudo é fruto de um grupo de pesquisa que iniciou, de forma informal, em 2020, a exploração do GSM, verificando a viabilidade de implementar um procedimento automatizado para a coleta de seus dados, o que resultou no desenvolvimento do primeiro *script*. Com o avanço das atividades, o grupo consolidou-se e publicou, ainda em 2020, o artigo “Periódicos científicos brasileiros indexados no Google Scholar Metrics” na revista Informação & Sociedade.

Em 2022, com o aprofundamento dos estudos sobre a ferramenta, foi publicado o artigo “Latin American and Caribbean journals indexed in Google Scholar Metrics” na revista Scientometrics, marcando a primeira publicação internacional do grupo. Em 2024, foi lançado o artigo “Gsm_hdata: a bibliometric tool to analyze data from Google Scholar Metrics” na revista Mobile Networks and Applications. Mais recentemente, outro trabalho foi submetido à revista Transinformação o qual se encontra em processo de avaliação e é resultado deste estudo.

A presente pesquisa, fundamentada na tese de Canto (2022), constitui um complemento e uma expansão do trabalho, ao integrar o GSM ao OpenAlex. Ambas as iniciativas fazem parte do histórico do grupo de estudos e, dessa forma, o conhecimento acumulado ao longo dos anos foi determinante para a realização desta investigação. Esse acúmulo permitiu não apenas o domínio sobre o GSM, mas também uma compreensão aprofundada das implicações envolvidas na coleta de seus dados, o que viabilizou a proposição e a execução do estudo atual.

O estudo estabelece cinco objetivos específicos, sendo o primeiro o desenvolvimento de um processo automatizado para coleta, certificação e integração de dados do GSM com base em informações bibliométricas do OpenAlex. Esse objetivo foi alcançado com a construção de um sistema que compreende o módulo de coleta de dados do GSM-ALEX_metrics, estruturado em três procedimentos principais: GSM_search, GSM-Alex_verification e OpenAlex_data.

O GSM_search sistematizou a raspagem dos dados do GSM a partir da consulta de títulos na URL da plataforma, permitindo a extração do índice h5 e de informações sobre artigos. Esse procedimento viabilizou a verificação de correspondência entre as revistas indexadas no GSM e aquelas registradas no

OpenAlex, além de garantir a coleta estruturada dos dados para integração posterior.

O `OpenAlex_data` consistiu no processo de obtenção dos dados disponibilizados nos diretórios públicos do OpenAlex, por meio dos *scripts* `OpenAlex_source` e `OpenAlex_work`. Juntos, eles geraram um conjunto de dados para análise detalhada de revistas indexadas no GSM. A extração sistematizada em etapas superou limitações de armazenamento e processamento, resultando em arquivos CSV com os campos originais do OpenAlex, facilitando a integração e futuras análises. Esses dados complementam os dados do GSM, permitindo análises bibliométricas com maior riqueza de detalhes das revistas indexadas no GSM, como por eixo temático, países e dados bibliométricos disponíveis no OpenAlex.

O `GSM-Alex_verification` validou a correspondência entre as duas bases, conferindo se os artigos coletados no GSM estavam efetivamente presentes no OpenAlex. Essa etapa assegura a identificação precisa das revistas comuns às duas fontes, fortalecendo a confiabilidade do conjunto final de dados utilizado para as análises bibliométricas.

A execução desse objetivo resultou na análise de uma lista de 103.212 revistas consultadas no GSM. Esse processo de integração e limpeza de duplicações resultou em uma lista final de 25.969 revistas. Trata-se de um resultado expressivo quando comparado ao estudo de Gu e Blackmore (2017), que, aplicando um procedimento manual, recuperou aproximadamente 10 mil revistas.

O segundo objetivo consistiu em validar o processo de coleta de dados, verificando o desempenho do sistema diante de um conjunto de periódicos previamente conhecidos por estarem indexados no GSM. Para isso, utilizou-se um conjunto de 2.302 revistas, identificando-se 2.260 correspondências exatas, o que representa 98% de precisão quando comparado com o procedimento `GSM_search` e 86% de precisão no `GSM-Alex_verification`.

A verificação dos procedimentos `GSM_search`, `GSM-ALEX_verification` e `Gsm_hdata` permitiu confirmar a precisão dos *scripts* desenvolvidos e também identificar pontos de aprimoramento relacionados à qualidade dos metadados, à duplicidade de registros e às limitações da indexação por idioma.

O terceiro objetivo consistiu em identificar e analisar as características editoriais dos periódicos recuperados do GSM com base nos dados do OpenAlex,

suprindo a ausência dessas informações para parte das revistas no GSM. Para isso, utilizaram-se os dados provenientes da integração entre as duas bases, o que possibilitou examinar aspectos como região geográfica, área temática, editora e tipo de acesso.

A análise revelou que a Europa concentra a maioria das revistas indexadas no GSM, seguida pela América e Ásia. África e Oceania, juntas, representam apenas 3% do total. Foram identificados periódicos em 38 países europeus, 35 asiáticos, 24 americanos, 23 africanos e 2 da Oceania. Os dez países com maior número de revistas são Estados Unidos, Reino Unido, Indonésia, Países Baixos, Brasil, Alemanha, Polônia, Suíça Espanha e Índia.

Quanto à distribuição por grandes áreas, a maior concentração de periódicos foi para Ciências Humanas, Ciências da Saúde, Ciências Exatas e da Terra e Ciências Biológicas. Já as áreas que se destacam pela representatividade foram Medicina, Filosofia, Biologia e Ciência Política. Em contrapartida, campos como Linguística, Educação, Direito e Ciências Ambientais somam menos de 1% do total, evidenciando lacunas de cobertura e menor expressão editorial na amostra.

A análise das editoras mostra forte concentração em poucos grupos, como Elsevier, Springer, Taylor & Francis, SAGE e Wiley reúnem cerca de 36% de todos os títulos. Quanto ao tipo de acesso, 60% dos periódicos são pagos, evidenciando a predominância do modelo de assinaturas para acesso aos artigos.

O quarto objetivo consistiu em analisar os dados bibliométricos dos periódicos presentes no GSM e no OpenAlex. A integração entre as duas bases, realizada por meio do identificador das revistas, viabilizou a execução dessa análise de forma consistente e integrada.

A distribuição do índice h5 revelou que a maioria dos periódicos está concentrada com indicadores que variam de 1 a 30. Essa característica se mantém com índice h do OpenAlex. Foi realizada uma correlação entre esses índices que indicou um forte alinhamento, sugerindo que os periódicos tendem a manter sua posição no ranking ao longo do tempo.

A comparação por tipo de acesso mostrou que, embora ambos os grupos se concentrem em faixas iniciais, os periódicos de acesso restrito apresentam, em média, índices mais elevados e maior heterogeneidade, com casos extremos de alto impacto, enquanto os de acesso aberto tendem a permanecer em níveis mais baixos.

A análise do índice h5 por grandes áreas da CAPES indicou que Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra e Engenharias apresentaram os maiores desempenhos médios, enquanto Ciências Humanas e Linguística, Letras e Artes registraram valores mais baixos. Entre as áreas, destacaram-se Química, Ciência dos Materiais e Geologia, de forma mais consistente. Por outro lado, Artes, História e Filosofia concentram os menores índices.

A análise por quartis Q1 (com indicadores mais elevados) apresentou que geograficamente, o Q1 é dominado por Estados Unidos e Reino Unido, que concentram mais de 60% dos títulos de maior impacto, seguidos por Holanda, Alemanha e Suíça. Outros países têm participação nesse grupo reduzido, refletindo a desigualdade na visibilidade internacional das revistas, confirmando o efeito Mateus de Merton (2013).

O quinto e último objetivo consistiu em desenvolver indicadores bibliométricos complementares para o GSM-ALEX_metrics. Esse processo utilizou o identificador das revistas para percorrer o arquivo CSV proveniente do diretório *Works* do OpenAlex, o mesmo empregado na validação das revistas com base no título dos artigos. Como resultado, foi possível calcular métricas como percentual de artigos não citados, índice de Gini, média de citações e índice h5 das revistas analisadas.

A análise desse índice h5 evidenciou um padrão de concentração semelhante ao índice h5 do GSM. Foi investigado também se há uma correlação entre o volume de publicações e um índice h elevado. Foi identificado que há uma associação positiva moderada sugerindo que revistas com mais artigos tendem a apresentar maior impacto.

Os demais indicadores bibliométricos reforçam esse cenário, com a maioria das revistas concentrada em médias abaixo de 10, percentual alto de artigos não citados em quase todos os periódicos, e uma concentração de citação em número ainda menor de artigos, calculado pelo índice de Gini.

De forma geral, os resultados demonstram que o GSM-ALEX_metrics desenvolvido é capaz de integrar e analisar de forma eficiente, automatizada, sistemática e escalável, grandes volumes de dados do GSM e do OpenAlex, oferecendo indicadores complementares e detalhados sobre a distribuição geográfica, temática, editorial e de impacto dos periódicos. O estudo confirma

padrões de concentração do impacto científico e desigualdades na visibilidade entre países e áreas, contribuindo para ampliar a compreensão sobre a cobertura e as limitações das bases analisadas. Esses achados reforçam a utilidade do GSM-ALEX_metrics como instrumento de estudo sobre a visibilidade e a distribuição do conhecimento.

Outro mérito do estudo reside na proposição e validação de indicadores bibliométricos complementares, que ampliam a capacidade analítica além dos índices tradicionais. A incorporação de métricas como percentual de não citados e índice de Gini trouxe novas perspectivas para compreender as dinâmicas de citação e a concentração do impacto em poucos periódicos e artigos. Esse aspecto evidencia o potencial do GSM-ALEX_metrics para oferecer análises mais refinadas e multifacetadas, adequadas a diferentes contextos de avaliação científica.

Por fim, este trabalho representa um marco no percurso do grupo de pesquisa, consolidando anos de experiência acumulada sobre o GSM e ampliando-o com a integração ao OpenAlex. O GSM-ALEX_metrics aqui proposto abre caminho para um ambiente de investigação contínua, em que novas bases, indicadores e metodologias poderão ser incorporados. Assim, a pesquisa não apenas alcançou seus objetivos, mas estabeleceu fundamentos para a evolução de estudos bibliométricos em larga escala, reafirmando sua relevância para a compreensão da ciência e de seus modos de circulação no cenário global.

6.1 LIMITAÇÕES

Uma limitação relevante está na dependência direta da estrutura do GSM. O GSM-ALEX_metrics desenvolvido baseia-se no padrão atual da plataforma para realizar a raspagem de dados. Qualquer alteração na interface, no código-fonte ou na forma de disponibilização das informações pode comprometer a execução automatizada do procedimento.

Outra limitação refere-se à cobertura parcial das bases consideradas. Nem todas as revistas listadas no GSM estão indexadas no OpenAlex, e nem todas as revistas do OpenAlex aparecem no GSM. Esse desencontro entre as duas fontes gera lacunas de representatividade e pode impactar diretamente a completude da amostra analisada.

A qualidade e padronização dos metadados também constituem uma limitação do estudo. Foram observadas inconsistências na grafia dos títulos, variações de idioma e uso de abreviações, fatores que podem gerar duplicidades ou falhas de correspondência entre as bases. Apesar do uso de procedimentos de verificação, tais problemas permanecem como um desafio metodológico.

A distribuição dos periódicos por áreas do conhecimento revelou baixa representatividade em determinados campos, como Linguística, Educação, Direito e Ciências Ambientais. Essa limitação disciplinar dificulta a generalização dos resultados, uma vez que as conclusões refletem sobretudo os campos mais bem representados na amostra.

Observa-se também a predominância de periódicos de grandes editoras, em especial aqueles de acesso pago, que tendem a apresentar melhores indicadores bibliométricos. Essa assimetria pode influenciar os resultados, já que reforça a centralidade de revistas comerciais de grande porte em detrimento de periódicos independentes, de acesso aberto ou de regiões periféricas.

Apesar da automação, em alguns casos foi necessária a revisão manual para confirmar a correspondência entre títulos de periódicos. Esse procedimento, embora pontual, evidencia que nem todas as situações podem ser resolvidas exclusivamente por meio de algoritmos de correspondência textual.

Outra limitação enfrentada foi a existência de casos de títulos compostos por uma única palavra ou termos muito genéricos, como Acervo ou Estudos Políticos. Esse tipo de ocorrência aumenta a ambiguidade nas correspondências, podendo gerar falsos positivos ou dificultar a validação automática.

Situações em que um título apresentava subtítulo em uma base, mas não na outra, também representaram uma dificuldade. Nesses casos, a ausência de padronização na representação do nome da revista pode prejudicar o processo de pareamento entre os registros, demandando maior esforço de validação.

Algumas revistas apresentaram versões do título em dois idiomas distintos, o que comprometeu a correspondência direta entre os registros. Esse aspecto ressalta a importância da normalização linguística, que ainda representa um desafio não totalmente superado no estudo.

A presença de títulos com erros de grafia constitui outra limitação. Essas inconsistências, muitas vezes decorrentes da própria indexação nas bases, podem

ter causado a não correspondência de determinados registros ou a geração de duplicatas.

Por fim, mudanças no nome de determinados periódicos ao longo do tempo também afetaram o processo de correspondência. Esse fator dificultou a rastreabilidade de algumas revistas e pode ter resultado na exclusão de títulos válidos do conjunto final analisado.

6.2 TRABALHOS FUTUROS

Como em qualquer investigação científica, os avanços obtidos também revelam novas possibilidades de exploração e aprimoramento. Assim, os trabalhos futuros aqui propostos visam expandir o escopo da ferramenta, superar limitações identificadas e explorar dimensões complementares da análise bibliométrica, tornando-a mais abrangente e acessível para o estudo da produção científica.

Futuros estudos podem investir no aprimoramento dos procedimentos de coleta automatizada, de modo a torná-los mais resilientes a possíveis mudanças estruturais no GSM. Isso inclui a implementação de algoritmos capazes de adaptar-se dinamicamente às alterações da página, bem como o uso de técnicas mais sofisticadas de raspagem e conferência, garantindo a continuidade e a reprodutibilidade das coletas em longo prazo.

Outro caminho promissor é a integração do GSM-ALEX_metrics com bases adicionais, que podem complementar lacunas existentes no GSM e no OpenAlex. Essa ampliação, além de enriquecer as análises, contribuiria para resolver problemas relacionados à gestão de títulos problemáticos, como mudanças de nome, duplicidades ou versões em diferentes idiomas, possibilitando maior precisão na correspondência entre registros.

Outro desenvolvimento futuro relevante seria a aplicação do GSM-ALEX_metrics em análises longitudinais, acompanhando a evolução dos periódicos ao longo dos anos e a ampliação dos indicadores utilizados, incorporando métricas como autocitações. Essa abordagem pode ser fortalecida pela combinação das técnicas desenvolvidas neste estudo com a ferramenta Gsm_hdata, permitindo avaliar não apenas a posição atual das revistas, mas também suas trajetórias históricas de impacto e autocitações.

Sugere-se ainda a realização de investigações específicas sobre as inconsistências dos dados presentes no GSM, utilizando a técnica proposta nesta tese como instrumento de verificação. Esse tipo de estudo pode evidenciar falhas de indexação, duplicidades e lacunas, contribuindo para maior transparência dos dados.

De forma complementar, outro campo de pesquisa refere-se à análise das inconsistências no OpenAlex, especialmente no que diz respeito à qualidade dos metadados fornecidos. A avaliação sistemática desses aspectos permitiria identificar pontos de melhoria e propor ajustes que beneficiem a comunidade científica usuária da base.

Por fim, um trabalho futuro relevante consiste no desenvolvimento de uma interface gráfica mais intuitiva, capaz de integrar todos os procedimentos do GSM-ALEX_metrics. Tal aprimoramento facilitaria o uso por pesquisadores de diferentes perfis, ampliando o alcance da ferramenta e sua aplicabilidade em investigações bibliométricas de maior escala, além de realizar o depósito dos dados e scripts da pesquisa em repositórios abertos em conformidade com os princípios da Ciência Aberta.

REFERÊNCIAS

AHMAD, Shakil *et al.* A scientific influence and reflectivity of Saudi Arabia research journals. **Library Philosophy and Practice (ejournal)**, 2019.

AKSNES, D.W., LANGFELDT, L.. How Citations Relate to Research Quality. *In*: SIVERTSEN, G., LANGFELDT, L. (Eds) **Challenges in Research Policy. SpringerBriefs in Political Science**. [S.l.]: Springer, 2025

ALPERIN, J. P.; ROZEMBLUM, C. La reinterpretación de visibilidad y calidad en las nuevas políticas de evaluación de revistas científicas. **Revista Interamericana de Bibliotecología**, v. 40, n. 3, p. 231–241, 2017.

ARAÚJO, Ronaldo Ferreira *et al.* Ciência aberta na perspectiva de especialistas brasileiros: proposta de taxonomia. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 26, p. 1-27, 2021.

ASUBIARO, Toluwase; ONAOLAPO, Sodiq; MILLS, David. Regional disparities in Web of Science and Scopus journal coverage. **Scientometrics**, v. 129, n. 3, p. 1469-1491, 2024.

BARBETTA, Pedro Alberto. **Estatística aplicada a ciências sociais**. 5. ed. Florianópolis: Ed. Ufsc, 2002.

BARNES, C. The h-index Debate: An Introduction for Librarians. **The Journal of Academic Librarianship**, v. 43, n. 6, p. 487–494, 1 nov. 2017. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0099133316301732>. Acesso em: 10 mar. 2023.

BORKO, Harold. Information science: what is it?. **American documentation**, v. 19, n. 1, p. 3-5, 1968.

BORNMANN, L. *et al.* The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. **Journal of the Association for Information Science and Technology**, v. 67, n. 11, p. 2778–2789, 1 nov. 2016. Disponível em: <http://doi.wiley.com/10.1002/asi.23627>. Acesso em: 10 set. 2020.

BORNMANN, L. h-Index research in scientometrics: A summary. **Journal of Informetrics**, v. 8, n. 3, p. 749–750, jul. 2014. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1751157714000649>. Acesso em: 16 mar. 2023.

BORNMANN, Lutz e DANIEL, Hans-Dieter. What do citation counts measure? A review of studies on citing behavior. **Journal of Documentation**, v. 64, n. 1, p. 45-80, 2008.

CABEZAS-CLAVIJO, Alvaro; DELGADO LÓPEZ-CÓZAR, Emilio. Google Scholar and the h-index in biomedicine: the popularization of bibliometric assessment. **Medicina Intensiva (English Edition)**, v. 37, n. 5, p. 343-354, 2013.

CAMACHO, Oscar *et al.* An approach to measure the impact of Scientific Regional Journals: The Latindex case. **Revista Ibérica de Sistemas e Tecnologias de Informação**, n. E20, p. 13-23, 2019.

CANTO, F. L. *et al.* **Latin American and Caribbean journals indexed in Google Scholar Metrics**. [Conjunto de dados]. Zenodo, 2021. Disponível em: <https://doi.org/10.5281/zenodo.5704895>. Acesso em: 10 out. 2024.

CANTO, Fabio Lorens. **Avaliação de impacto de periódicos Ibero-Americanos com base no índice h5 do Google Scholar Metrics**. 2022. 193 p. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Santa Catarina, Florianópolis, 2022.

CANTO, Fabio Lorensi do *et al.* Journal self-citation on the h5-index of Ibero-American journals. **Transinformação**, v. 35, p. e237318, 2023.

CANTO, Fábio Lorensi *et al.* Gsm_hdata: a bibliometric tool to analyze data from google scholar metrics. **Mobile Networks and Applications**, v. 29, n. 3, p. 754-761, 2024.

CANTO, Fabio Lorensi *et al.* Latin american and caribbean journals indexed in Google Scholar Metrics. **Scientometrics**, v. 127, n. 2, p. 763-783, 2022.

CARVALHO SEGUNDO, W. L. R. de; PINTO, A. L.; CANTO, F. L. do; NEUBERT, P. Projeto Laguna: infraestrutura de um lago de dados científicos em acesso aberto. **BiblioCanto**, [S. l.], v. 9, n. 2, p. 133–138, 2023. DOI: 10.21680/2447-7842.2023v9n2ID33825. Disponível em: <https://periodicos.ufrn.br/bibliocanto/article/view/33825>. Acesso em: 1 jul. 2024.

CHAWLA, Dalmeet Singh. Massive open index of scholarly papers launches. **Nature**, 2022.

CHIZZOTTI, Antonio. A pesquisa qualitativa em ciências humanas e sociais: evolução e desafios. **Revista portuguesa de educação**, v. 16, n. 2, p. 221-236, 2003.

CLARIVATE. **Journal Citation Reports (JCR)**. 2023. Acesso restrito a assinantes. Disponível em: <https://jcr.clarivate.com/>. Acesso em: 30 mar. 2023.

CLARKE, Rachel Ivy. Breaking records: the history of bibliographic records and their influence in conceptualizing bibliographic data. **Cataloging & Classification Quarterly**, v. 53, n. 3-4, p. 286-302, 22 dez. 2014.

COSTA, Heloisa *et al.* Correlation between Latin American And Caribbean Journals Impact Indicators. **Bibliotecas. Anales de investigación**, v. 19, n. 1, p. 3, 2023.

COSTA, Heloisa; CANTO, Fábio Lorensi; PINTO, Adilson Luiz. Google Scholar Metrics e a proposta do novo Qualis: impacto dos periódicos brasileiros de Ciência da Informação. **Informação & Sociedade**, v. 30, n. 1, 2020.

D'ALMONTE, Edson Fernando *et al.* Periódicos científicos na área Comunicação e Informação: consolidação do índice e as questões éticas sobre sua utilização. **Em Questão**, v. 30, p. e-136673, 2024.

DE FILIPPO, D.; ALEIXANDRE-BENAVENT, R.; SANZ-CASADO, E. Toward a classification of Spanish scholarly journals in social sciences and humanities considering their impact and visibility. **Scientometrics**, v. 125, n. 2, p. 1709–1732, 2020.

DELGADO LÓPEZ-CÓZAR, E.; CABEZAS-CLAVIJO, Á. Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank? **Learned Publishing**, v. 26, n. 2, p. 101–113, 1 abr. 2013. Disponível em: <http://doi.wiley.com/10.1087/20130206>. Acesso em: 10 set. 2020.

DELGADO LÓPEZ-CÓZAR, E.; ORDUÑA-MALEA, E.; MARTÍN-MARTÍN, A. **Google Scholar as a Data Source for Research Assessment**. Springer Handbooks. [S.l.: s.n.], 2019. p. 95–127. Disponível em: http://link.springer.com/10.1007/978-3-030-02511-3_4. Acesso em: 10 mar. 2020.

DELGADO LÓPEZ-CÓZAR, Emilio; CABEZAS-CLAVIJO, Álvaro. Google Scholar Metrics: an unreliable tool for assessing scientific journals. **El Profesional de la Información**, v. 21, n. 4, p. 419-427, 2012. Disponível em: <https://recyt.fecyt.es/index.php/EPI/article/view/epi.2012.jul.15>. Acesso em: 10 dez. 2024.

DELGADO LÓPEZ-CÓZAR, Emilio; ROBINSON-GARCÍA, Nicolás. Repositories in Google Scholar Metrics or what is this document type doing in a place as such? **Cybermetrics**, v.16, n. 1, 2012.

DIAS, Thiago Magela Rodrigues *et al.* Brcris: plataforma para integração, análises e visualização de dados técnicos-científicos. **Informação & Informação**, v. 27, n. 3, p. 622-638, 2022.

DIAZ, Alexandre Paim *et al.* Journal Metrics in Psychiatry: What do the rankings tell us?. **Journal of Affective Disorders**, v. 287, p. 354-358, 2021.

DONNER, Paul. Enhanced self citation detection by fuzzy author name matching and complementary error estimates. **Journal of the Association for Information Science and Technology**, v. 67, n. 3, p. 662-670, 2016.

DORTA-GONZÁLEZ, Pablo; SUÁREZ-VEGA, Rafael; DORTA-GONZÁLEZ, María Isabel. Open access effect on uncitedness: a large-scale study controlling by discipline, source type and visibility. **Scientometrics**, v. 124, p. 2619–2644, 2020.

DRUMOND, Larissa Barbara Borges. **Caminhos para consolidação da Ciência Aberta no âmbito dos periódicos científicos brasileiros**: uma proposta de abertura de dados por meio do repositório da Universidade Federal de Goiás. 2023. Dissertação (Mestrado em Comunicação) - Programa de Pós-Graduação em

Comunicação da Faculdade de Informação e Comunicação da Universidade Federal de Goiás, 2023.

EGGHE, L. Expansion of the field of informetrics: Origins and consequences. **Information Processing & Management**, v. 41, n. 6, p. 1311–1316, dez. 2005. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0306457305000245>. Acesso em: 28 jan. 2023.

EGGHE, Leo; ROUSSEAU, Ronald. **Introduction to informetrics**: quantitative methods in library, documentation and information science. Amsterdam: Elsevier, 1990.

ELSEVIER. **CiteScore**. 2025. Disponível em: <https://www.elsevier.com/products/scopus/metrics/citescore>. Acesso em: 10 mar. 2025.

FRANDSEN, Tove Faber; NICOLAISEN, Jeppe. Defining the unscholarly publication: a bibliometric study of uncited and barely cited publications. **Scientometrics**, v. 128, p. 1337–1350, 2023.

GARFIELD, Eugene. Citation analysis as a tool in journal evaluation: journals can be ranked by frequency and impact of citations for science policy studies. **Science**, Washington, v. 178, n. 4060, p. 471–479, nov. 1972. Disponível em: <http://science.sciencemag.org/content/178/4060/471>. Acesso em 20 jan. 2023.

GLÄNZEL, Wolfgang. The multi-dimensionality of journal impact. **Scientometrics**, v. 78, n. 2, p. 355–374, 2009.

GLÄNZEL, Wolfgang; MOED, Henk F. Journal impact measures in bibliometric research. **Scientometrics**, v. 53, n. 2, p. 171–193, 2002.

GONZALEZ-PARDO, Rafael; REPISO, Rafael; ARROYAVE-CABRERA, Jesús. Revistas iberoamericanas de comunicación a través de las bases de datos Latindex, Dialnet, DOAJ, Scopus, AHCI, SSCI, REDIB, MIAR, ESCI y Google Scholar Metrics. **Revista española de documentación científica**, v. 43, n. 4, p. e276–e276, 2020.

GONZÁLEZ-PEREIRA, B.; GUERRERO-BOTE, V. P.; MOYA-ANEGÓN, F. A new approach to the metric of journals' scientific prestige: The SJR indicator. **Journal of Informetrics**, v. 4, n. 3, p. 379–391, jul. 2010. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1751157710000246>. Acesso em: 21 fev. 2023.

GU, X.; BLACKMORE, K. Characterisation of academic journals in the digital age. **Scientometrics**, v. 110, n. 3, p. 1333–1350, 3 set. 2017. Disponível em: <http://link.springer.com/10.1007/s11192-016-2219-4>. Acesso em: 10 fev. 2023.

GUSENBAUER, M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. **Scientometrics**, v. 118, n. 1, p. 177–214, 10 jan. 2019. Disponível em: <http://link.springer.com/10.1007/s11192-018-2958-5>. Acesso em: 15 fev. 2023.

HAGOOD, Alex *et al.* Endorsement of reporting guidelines and clinical trial registration across urological medical journals: a cross-sectional study. **PeerJ**, v. 12, p. 18619, 2024.

HALEVI, Gali; MOED, Henk; BAR-ILAN, Judit. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the literature. **Journal of informetrics**, v. 11, n. 3, p. 823-834, 2017.

HARLIANTO, Netanja I.; HARLIANTO, Zaneta N. Health Care Industry Payments to Editorial Board Members of Major Neurosurgery Journals Between 2017 and 2022. **Neurosurgery**, v. 95, n. 4, p. 816-824, 2024.

HENNING, P.; MOREIRA, J. Ciência aberta, dados abertos e princípios FAIR: uma contribuição dos Países Baixos. In: SHINTAKU, M.; SALES, L. F; COSTA, M. (Org). **Tópicos sobre dados abertos para editores científicos**. Botucatu, SP: ABEC, 2020. p. 145-158. DOI: 10.21452/ 978-85-93910-04-3.cap12.

HENNING, Patricia *et al.* Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 11, n. 1, 2018.

HERNANDEZ SAMPIERI, Roberto; FERNÁNDEZ COLLADO, Carlos; BAPTISTA LUCIO, Pilar. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, v. 102, n. 46, p. 16569–16572, 15 nov. 2005. Disponível em: www.pnas.org/cgi/doi/10.1073/pnas.0507655102. Acesso em: 10 mar. 2023.

HUGHES, Griffin *et al.* Improving transparency in publishing: gaps in standardised reporting across surgical pathology and laboratory medicine journals. **Journal of Clinical Pathology**, 2025.

IBRAHIM, Hazem *et al.* Citation manipulation through citation mills and pre-print servers. **Scientific reports**, v. 15, n. 1, p. 5480, 2025.

Initiative for Open Citations. [2024]. Disponível em: <https://i4oc.org/>. Acesso em: 3 jul. 2024.

JACSÓ, P. Google Scholar duped and deduped: the aura of “robometrics”. **Online Information Review**, v. 35, n. 1, p. 154–160, 22 fev. 2011. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/14684521211241503/full/html>. Acesso em: 28 jan. 2023.

JACSÓ, P. Google Scholar Metrics for Publications. **Online Information Review**, v. 36, n. 4, p. 604–619, 3 ago. 2012. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/14684521211254121/full/html>. Acesso em: 10 mar. 2023.

JACSÓ, Péter. The need for end-user customization of the journal sets of the subject categories in the SCImago journal ranking database for more appropriate league

lists. A case study for the library & information science field. **El profesional de la información**, v. 22, n. 5, p. 459-473, 2013.

JAMES, C. *et al.* CiteScore metrics: Creating journal metrics from the Scopus Citation index. **Learned Publishing**, v. 32, n. 4, p. 367–374, 19 out. 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1246>. Acesso em: 23 mar. 2023.

JESUS, Ananda Fernanda de. **Recomendações teórico-metodológicas para a publicação de dados bibliográficos abertos e conectados**. 2021. Dissertação (Mestrado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos, 2021.

JOHNSON, Austin L. *et al.* Data-sharing across otolaryngology: comparing journal policies and their adherence to the FAIR principles. **Annals of Otology, Rhinology & Laryngology**, v. 133, n. 1, p. 105-110, 2024.

KHURSHID, Zahiruddin. Measuring the quality of contributions of saudi authors to lis journals using Journal Impact Factor (JIF), SCImago Journal Rank (SJR), and Google Scholar Metrics (GSM). **The Serials Librarian**, v. 67, n. 1, p. 81-98, 2014.

KOLI, Paresh G.; KULKARNI, Ankita; SHETTY, Yashashri C. Evaluation of issues affecting time between study completion, manuscript submission, acceptance, and publication in medical journals. **Cureus**, v. 14, n. 3, 2022.

KOZLOWSKI, Diego *et al.* The decrease in uncited articles and its effect on the concentration of citations. **Journal of the Association for Information Science and Technology**, v. 75, n. 2, p. 188-197, 2024.

KRAMPL, A. Journal Citation Reports. **Journal of the Medical Library Association**, v. 107, n. 2, p. 280-283, Apr. 2019. DOI [dx.doi.org/10.5195/jmla.2019.646](https://doi.org/10.5195/jmla.2019.646). Acesso em: <http://jmla.pitt.edu/ojs/jmla/article/view/646>. Acesso em: 28 fev. 2023.

LAENDER, Alberto Henrique Frade *et al.* **Abertura e gestão de dados: desafios para a ciência brasileira**. Rio de Janeiro: Academia Brasileira de Ciências, 2020.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Metodologia científica: ciência e conhecimento científico, métodos científicos, teoria, hipóteses e variáveis**. 5. ed. São Paulo: Atlas, 2007. 312 p.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Metodologia do trabalho científico: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos**. 4. ed. São Paulo: Atlas, 1992. 214p.

LAKATOS, I. **The Methodology of Scientific Research Programmes: Philosophical Papers**. Cambridge: Cambridge University Press, 1978. v. 1.

LANGHAM-PUTROW, Allison; BAKKER, Caitlin; RIEGELMAN, Amy. Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. **PloS one**, v. 16, n. 6, 2021.

LESHEM, S.; TRAFFORD, V. Overlooking the conceptual *framework*. **Innovations in Education and Teaching International**, v. 44, n. 1, p. 93–105, fev. 2007. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/14703290601081407>. Acesso em: 10 maio 2025.

LOTBINIERE-BASSETT, Madeleine P.; RIVA-CAMBRIN, Jay; MCDONALD, Patrick J. Conflict of interest policies and disclosure requirements in neurosurgical journals. **Journal of Neurosurgery**, v. 131, n. 1, p. 264-270, 2018.

MACEDO, Marcelo; SOUZA, Mariane Ramos. Teoria, modelos e *frameworks*: conceitos e diferenças. *In*: Congresso Internacional de Conhecimento e Inovação–ciki, 12. **Anais...** [S.l: s.n], 2022.

MARCONI, Marina de Andrade. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisa, elaboração, análise e interpretação de dados. 8. ed. São Paulo: Atlas, 2018. 312p.

MARTÍN-MARTÍN, A. *et al.* Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. **Journal of Informetrics**, v. 12, n. 4, p. 1160–1177, 1 nov. 2018. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1751157718303249>. Acesso em: 15 mar. 2023.

MARTÍN-MARTÍN, Alberto *et al.* Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. **Scientometrics**, v. 126, n. 1, p. 871-906, 2021.

MARTINS, Gilberto de Andrade; THEÓPHILO, Carlos Renato. **Metodologia da investigação científica para Ciências Sociais Aplicadas**. 2.ed. São Paulo: Atlas, 2009.

MERTON, Robert K. **Ensaio de sociologia da ciência**. Organização de Anne Marcovich, Terry Shinn. São Paulo: Editora 34, 2013.

MICHEL, M. H. **Metodologia e pesquisa científica em ciências sociais**: um guia prático para acompanhamento da disciplina e elaboração de trabalhos monográficos. 2. ed. São Paulo: Atlas, 2009.

MOED, Henk F. Citation analysis of scientific journals and journal impact measures. **Current Science**, p. 1990-1996, 2005.

MONTEIRO, Gabriela; LUCAS, ER de O. Dados científicos abertos: identificando o papel das políticas de gestão e das agências de fomento. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 8, n. 1, p. 13-20, 2019.

MUGNAINI, Rogério; SALES, Denise Peres. Mapeamento do uso de índices de citação e indicadores bibliométricos na avaliação da produção científica brasileira. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. 12, p. 2361-2372, 2011.

MUGNAINI, Rogério; STREHL, Letícia. Recuperação e impacto da produção científica na era Google: uma análise comparativa entre o Google Acadêmico e a

Web of Science. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, n. Esp, p. 92-105, 2008.

NEDERHOF, A. Bibliometric monitoring of research performance in the Social Sciences and the Humanities: a review. **Scientometrics**, v. 66, n. 1, p. 81-100, 2006. Disponível em: <https://link.springer.com/article/10.1007/s11192-006-0007-2>. Acesso em: 15 dez. 2022.

NIFORATOS, Joshua D.; NARANG, Jatin; TRUEGER, N. Seth. Financial conflicts of interest among emergency medicine journals' editorial boards. **Annals of Emergency Medicine**, v. 75, n. 3, p. 418-422, 2020.

OKAGBUE, Hilary I.; SILVA, Jaime A Teixeira da. Correlation between the CiteScore and Journal Impact Factor of top-ranked library and information science journals. **Scientometrics**, v. 124, n. 1, p. 797-801, 2020. Disponível em: <https://doi.org/10.1007/s11192-020-03457-x>. Acesso em: 10 fev. 2023.

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, n. 1, p. 433-450, 24 out. 2019. Disponível em: <http://link.springer.com/10.1007/s11192-019-03208-7>. Acesso em: 10 fev. 2023.

ORDUÑA-MALEA, E.; DELGADO LÓPEZ-CÓZAR, E. Google Scholar Metrics evolution: na analysis according to languages. **Scientometrics**, v. 98, n. 3, p. 2353-2367, 23 set. 2014. Disponível em: <http://link.springer.com/10.1007/s11192-013-1164-8>. Acesso em: 10 mar. 2023.

PABLOS COELLO, Jose Manuel; MATEOS MARTIN, Concha; TUNEZ LOPEZ, Miguel. Google, the change of paradigm of the scientific metrics. **Historia y Comunicacion Social**, v. 18, p. 225-235, 2013.

PACKER, Abel Laerte. Indicators of national focus of the research published by Public Health journals edited in Brazil. **Ciência & Saúde Coletiva**, v. 20, p. 1983-1995, 2015.

PATRA, Nihar K. From electronic resources to electronic resource management. In: PATRA, Nihar K. (ed.). **Digital disruption and electronic resource management in libraries**. Cambridge: Chandos Publishing, 2017. p. 13-25.

PINTO, Adilson Luiz *et al.* Periódicos científicos brasileiros indexados no Google Scholar Metrics. **Informação & Sociedade: Estudos**, v. 30, n. 4, p. 1-18, 2020.

REPISO, Rafael; JIMÉNEZ-CONTRERAS, Evaristo; AGUADED, Ignacio. Revistas Iberoamericanas de Educación en SciELO Citation Index y Emerging Source Citation Index. **Revista española de documentación científica**, v. 40, n. 4, p. e186-e186, 2017.

RODRIGUES, David; LOPES, António Luís; BATISTA, Fernando. Web of Science citation gaps: An automatic approach to detect indexed but missing citations. In: SYMPOSIUM ON LANGUAGES, APPLICATIONS AND TECHNOLOGIES, 12, 2023. **Proceedings...** Schloss Dagstuhl--Leibniz-Zentrum für Informatik, 2023.

ROMANCINI, R. O que é uma citação? A análise de citações na ciência. **Intexto**, Porto Alegre, n. 23, p. 5–17, 2010. Disponível em: <https://seer.ufrgs.br/index.php/intexto/article/view/15885>. Acesso em: 21 out. 2024.

ROUSSEAU, Ronald; EGGHE, Leo; GUNS, Raf. **Becoming metric-wise: A bibliometric guide for researchers**. Cambridge: Elsevier, 2018.

RUDIO, Franz Victor. **Introdução ao projeto de pesquisa científica**. 14. ed. Petrópolis: Vozes, 1978. 120 p.

SALES, Luana Farias; SAYÃO, Luís Fernando. A grande e a pequena ciência: análise das diferenças na gestão de dados de pesquisa. **Inf. & Soc.: Est.**: João Pessoa, v.29, n.3, p. 151-170, 2019.

SANTAELLA, Lucia. **Comunicação e pesquisa: projetos para mestrado e doutorado**. 2. ed. São José do Rio Preto: Hacker, 2010. 144 p.

SANTOS, Paula Xavier dos; GUANAES, Paulo. Ciência Aberta, dados abertos: desafio e oportunidades. **Trabalho, educação e saúde**: Rio de Janeiro. v.16, n.1, editorial, jan./abr. 2018.

SCHEIDSTEGER, T.; HAUNSCHILD, R. Which of the metadata with relevance for bibliometric analyses are covered in Microsoft Academic Graph and OpenAlex? **Profesional de la Información**, v. 32, n. 2, p. 320209, 2023.

SCHWARTZ, Charles A. The Rise and Fall of Uncitedness. **College & Research Libraries**, v. 58, n. 1, p. 19-29, jan. 1997.

SCIMAGO. **SJR**: SCImago Journal & Country Rank. 2023. Disponível em: <http://www.scimagojr.com> Acesso em: 28 mar. 2023.

SEMLER, Alexandre Ribas; PINTO, Adilson Luiz. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, v. 48, n. 1, 2019.

SILVA, Jaime A. T.; DOBRÁNSZKI, Judit. Multiple versions of the h-index: Cautionary use for formal academic purposes. **Scientometrics**, v. 115, n. 2, p. 1107-1113, 2018.

SILVEIRA, Murilo Artur Araújo da. **Produção e distinção de domínio da organização e representação do conhecimento no Brasil**. 2016. 155 p. Tese (Doutorado em Comunicação e Informação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

SILVEIRA, Murilo Artur Araújo; CAREGNATO, Sônia Elisa. Demarcações epistemológicas dos estudos de citação: o fenômeno da citação. **Informação & Sociedade**, v. 27, n. 3, 2017.

SIMS, Matthew T. *et al.* Do emergency medicine journals promote trial registration and adherence to reporting guidelines? A survey of “Instructions for Authors”. **Scandinavian journal of trauma, resuscitation and emergency medicine**, v. 24, p. 1-8, 2016.

SINGH, Vivek Kumar *et al.* The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. **Scientometrics**, v. 126, n. 6, p. 5113-5142, 2021.

SPIROSKI, Mirko. Current biomedical scientific impact (2013) of institutions, academic journals and researchers in the Republic of Macedonia. **Contributions/Prilozi**, v. 35, n. 3, 2015.

TARGINO, Maria das Graças. Comunicação Científica: uma revisão de seus elementos básicos. **Informação & Sociedade**, João Pessoa, v. 10, n. 2, p.67-85, 2000. Disponível em: <<http://www.ies.ufpb.br/ojs/index.php/ies/article/view/326/248>>. Acesso em: 27 jan. 2017.

TOMHAVE, Benjamin L. Alphabet soup: **Making sense of models, frameworks, and methodologies**. Washington: George Washington University, 2005.

TORRES-SALINAS, D.; RUIZ-PÉREZ, R.; DELGADO LÓPEZ-CÓZAR, E. Google Scholar como herramienta para la evaluación científica. **El Profesional de la Informacion**, v. 18, n. 5, p. 501–510, 1 set. 2009. Disponível em: <https://recyt.fecyt.es/index.php/EPI/article/view/epi.2009.sep.03>. Acesso em: 10 mar. 2020.

UNESCO. **Recomendação da UNESCO sobre Ciência Aberta**. [S.l.: s.n.], 2022. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>. Acesso em: 15 set. 2023.

VANTI, Nadia Aurora Peres. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 369-379, 2002.

VISSER, Martijn; VAN ECK, Nees Jan; WALTMAN, Ludo. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. **Quantitative science studies**, v. 2, n. 1, p. 20-41, 2021.

VOLPATO, Gilson Luiz. **Bases teóricas para redação científica**: porque seu artigo foi negado? São Paulo: Cultura Acadêmica, 2007.

WALTERS, William H. Citation-based journal rankings: Key questions, metrics, and data sources. **IEEE Access**, v. 5, p. 22036-22053, 2017.

WALTERS, William H. Evaluating journals in business and related fields: A guide for faculty. **Business Information Review**, v. 39, n. 3, p. 90-97, 2022.

WALTERS, William H. Information sources and indicators for the assessment of journal reputation and impact. **The Reference Librarian**, v. 57, n. 1, p. 13-22, 2016.

WALTMAN, L. A review of the literature on citation impact indicators. **Journal of Informetrics**, v. 10, n. 2, p. 365–391, maio 2016. Disponível em: <http://dx.doi.org/10.1016/j.joi.2016.02.007>. Acesso em: 24 mar. 2023.

WALTMAN, L.; VAN ECK, N. J. The relation between Eigenfactor, audience factor, and influence weight. **Journal of the American Society for Information Science and Technology**, v. 61, n. 7, p. 1476–1486, 15 abr. 2010. Disponível em: <http://doi.wiley.com/10.1002/asi.21354>. Acesso em: 16 mar. 2023.

WILKINSON, M. D. *et al.* A design *framework* and exemplar metrics for FAIRness. **Scientific Data**, v. 5, n. 180118, 2018. Disponível em: <https://www.nature.com/articles/sdata2018118>. Acesso em: 25 nov. 2019.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016.

ZHANG, Lin *et al.* Missing institutions in OpenAlex: Possible reasons, implications, and solutions. **Scientometrics**, v. 129, n. 10, p. 5869-5891, 2024.