DOI: 10.5433/1981-8920.2022v27n3p402

AVALIANDO A REPRESENTATIVIDADE DOS PRINCIPAIS TÓPICOS DE PESQUISA EM PUBLICAÇÕES DE ACESSO ABERTO NO BRASIL

EVALUATING THE REPRESENTATION OF THE MAIN RESEARCH TOPICS IN OPEN ACCESS PUBLICATIONS IN BRAZIL

Patrícia Mascarenhas Dias[a]
Thiago Magela Rodrigues Dias[b]
Gray Farias Moita[c]

RESUMO
Objetivo: neste trabalho, o principal objetivo é identificar e analisar a frequência dos principais tópicos de pesquisa utilizados no conjunto de publicações em periódicos de acesso aberto por pesquisadores do Brasil. **Metodologia:** para se contemplar o objetivo proposta, inicialmente são identificados no conjunto de currículos da Plataforma Lattes todas as publicações realizadas em periódicos de acesso aberto. Posteriormente, com auxílio de técnicas de mineração de textos e com a adoção da Lei de Zipf são identificados os principais tópicos utilizados, tendo como fonte de dados, os títulos das publicações inicialmente identificadas. **Resultados:** como resultados das análises foi possível verificar como alguns tópicos são frequentemente utilizados pelos pesquisadores brasileiros na descrição de seus estudos. **Conclusões:** com o estudo e a análise apresentada neste trabalho, foi possível verificar como alguns tópicos que indicam localidades são frequentemente utilizados, bem como, termos que em geral indicam métodos adotados nas pesquisas.

Descritores: Acesso Aberto. Plataforma Lattes. Bibliometria. Dados Abertos.

[a] Doutora em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET). Docente da Universidade do Estado de Minas Gerais (UEMG), Divinópolis, Brasil. E-mail: patriciamdias@gmail.com

[b] Doutor em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET). Docente do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Divinópolis, Brasil. E-mail: thiagomagela@cefetmg.br

[c] Doutor em Aeronautica pelo Imperial College London. Docente do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, Brasil. E-mail: gray@cefetmg.br

1 INTRODUÇÃO

Compreender a evolução do desenvolvimento científico e tecnológico de um país é de extrema importância, tendo em vista que isso possibilita identificar como o progresso das pesquisas nas diversas áreas do conhecimento tem evoluído historicamente. Além disso, tal compreensão permite identificar os principais tópicos de investigação, o perfil dos pesquisadores e as suas colaborações científicas, o que pode servir como base para diversas políticas de fomento à pesquisa científica.

Meadows (1999) afirma que a competição científica está estreitamente ligada ao grau de interação que os cientistas têm com seus pares nos mais variados canais de comunicação. A comunicação científica é tão vital para ciência quanto a própria pesquisa.

O tradicional formato impresso de comunicação da ciência vem aos poucos dando espaço para os novos formatos eletrônicos, devido à ascensão da tecnologia de informação e comunicação. No contexto das pesquisas e estudos científicos, a comunicação científica surge nos dias atuais como um elemento central em diversos níveis de discussão, com ênfase na divulgação de artigos científicos em periódicos, atualmente um dos principais meios de comunicação para esse fim.

Mueller (1999) afirma que o periódico científico desempenha pelo menos quatro funções essenciais: certificação da ciência com o aval da comunidade cientifica; canal de comunicação entre os cientistas e de divulgação mais ampla da ciência; arquivo ou memória científica e registro da autoria da descoberta científica.

De acordo com vários estudos, os periódicos — principalmente os disponíveis em formato eletrônico — estão em crescimento desde a última década. Pode-se falar que os periódicos, em todas as áreas do conhecimento, têm o papel de ser um filtro para o reconhecimento dos trabalhos aceitos. Para Rodrigues e Oliveira (2012), a publicação em uma revista reconhecida pela área é a forma mais aceita para registrar a originalidade do trabalho e para confirmar que os trabalhos foram confiáveis o suficiente para superar o ceticismo da

comunidade científica.

Neubert, Rodrigues e Goulart (2012) afirmam que o acesso aberto assume um importante papel em todo o contexto da atividade científica, pois permite ao pesquisador ter acesso aos resultados de outros estudos sem as barreiras de custo e as dificuldades de acesso, além de promover a visibilidade e a divulgação dos resultados das atividades cientificas de cada pesquisador e de cada universidade.

A publicação científica em acesso aberto faz parte de um cenário mais amplo em prol da abertura do conhecimento em geral (acesso aberto, dados abertos, recursos educacionais abertos, software livre, licenças abertas) e constitui essencialmente um movimento em direção à concepção da informação e do conhecimento como bens públicos (FURNIVAL; SILVA-JEREZ, 2017).

Tendo em vista que grande parte das pesquisas científicas no país é financiada com recursos públicos, geralmente em instituições de ensino ou centros de pesquisa públicos, é de se esperar que os resultados de tais estudos sejam divulgados sem nenhum tipo de barreira, principalmente financeira. Nesse contexto, aliados às vantagens que as publicações em acesso aberto apresentam, como disponibilidade, visibilidade e acessibilidade, diversos esforços estão sendo empregados para que cada vez mais artigos científicos sejam publicados em periódicos de acesso aberto.

Governantes de vários países reconhecem que o acesso aberto aos dados, à informação e ao conhecimento contribui decisivamente para os avanços da pesquisa científica e inovação, além de maximizar o valor derivado de investimentos públicos, trazer benefícios para a economia e a sociedade e inserir os países em desenvolvimento no sistema global de ciência, contribuindo para seu desenvolvimento econômico e social (OCDE, 2004).

Diante disso, compreender quais os principais tópicos de pesquisa estão sendo investigados nos artigos publicados em periódicos de acesso aberto possibilita identificar um panorama das principais temáticas estudadas. Permite, ainda, verificar a representatividade de determinados tópicos presentes nos artigos analisados.

2 TRABALHOS RELACIONADOS

No trabalho de Silva e Alcará (2008), são analisadas as políticas de acesso aberto à informação científica e as propostas de ação, com ênfase nas iniciativas governamentais em diferentes países. Foi identificado que o movimento de acesso livre à informação científica já era preocupação oficialmente registrada em vários países, embora com diferentes graus de desenvolvimento. Entre tais diferenças estão as próprias determinações das políticas, já que algumas obrigam instituições públicas e pesquisadores a disponibilizar em acesso aberto os resultados de suas pesquisas, enquanto outras apenas sugerem o envolvimento e a participação de pesquisadores e instituições no movimento.

Oliveira e Chalhub (2009) analisaram as revistas científicas ibero-americanas que aderiram ao movimento de acesso aberto, integrantes do Directory of Open Access Journal (DOAJ). Especificamente, esse estudo visou identificar as revistas da região ibero-americana incluídas no repositório, suas instituições editoras, periodicidade e áreas de cobertura, além de verificar sua inserção no movimento de acesso livre, por meio da análise de sua participação em outros espaços virtuais. Os resultados apontaram crescente adesão das revistas científicas da região. Em âmbito internacional, o Brasil e a Espanha ocupam a segunda e a quarta posições, respectivamente. As unidades de ensino e pesquisa representam a maioria das instituições editoras. Destaca-se, ainda, a diversidade das áreas das revistas, que vão desde a Engenharia até a Linguística, com participação expressiva da Medicina. Conclui-se que dados específicos de um diretório parecem fortalecer a legitimação do movimento de acesso livre por parte dos atores envolvidos no sistema de comunicação científica.

Já em Chalhub e Pinheiro (2011), são identificados os principais canais de comunicação científica de acesso aberto utilizados por pesquisadores, e analisados os fatores intervenientes na adesão ao auto arquivamento de sua produção científica. O trabalho tinha como objetivo identificar os principais canais de comunicação científica em acesso aberto utilizados por pesquisadores de

universidades públicas do estado do Rio de Janeiro. Foi utilizada a listagem de 47 Comitês de Assessoramento do CNPq para Bolsas de Produtividade em Pesquisa e efetuada uma amostragem probabilística estratificada por área de conhecimento, seguindo a divisão por Comitê de Assessoramento (Ciências Agrárias, Ciências Biológicas, Ciências Exatas e da Terra, Ciências da Saúde, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias, e Linguística Letras e Artes). A partir da seleção dos pesquisadores contemplados pelo programa de Bolsa de Produtividade em Pesquisa do CNPq no ano de 2010, cuja relação está disponível no site desse órgão federal, foram identificados aqueles vinculados a universidades públicas do estado do Rio de Janeiro com cursos de pós-graduação stricto sensu. Após a identificação dos endereços eletrônicos dos selecionados, foi enviada correspondência contendo em anexo um formulário com questões fechadas e abertas sobre estas categorias: comportamento informacional, publicação de acesso aberto e adesão a repositório institucional.

De maneira geral, os resultados da pesquisa apontam para uma mudança na postura desses pesquisadores com relação à publicação de resultados de pesquisa em canais de acesso aberto. Algumas áreas apresentam publicações em canais formais de comunicação científica, como em periódicos eletrônicos, e auto arquivamento em repositórios institucionais ou temáticos. Outras se inserem mais em iniciativas individuais ou de grupos de pesquisa, muitas vezes antecipando as políticas institucionais. Os pesquisadores foram unânimes com relação às vantagens da publicação em acesso aberto, e a democratização do conhecimento foi apontada pela maioria como a principal vantagem dessa adesão. Além desse aspecto, também aparece nas falas dos pesquisadores o benefício da comunicação entre pares — “trocas”, “parcerias” e “diálogos” — no processo de produção do conhecimento. Ainda, é sinalizada a importância da utilização desse canal aberto de comunicação em dois momentos distintos: para que o pesquisador acesse a informação para suas pesquisas e para que disponibilize seus resultados, possibilitando-lhes maior visibilidade e impacto.

Percebe-se que diversos outros trabalhos também utilizaram questionários junto a conjuntos de pesquisadores para coleta de dados. É o caso

do trabalho de Furnival e Silva-Jerez (2017), que se propôs a explorar de que modo várias dimensões do acesso aberto à literatura científica são percebidas por pesquisadores brasileiros, identificando também seus hábitos de publicação e de uso e citação de fontes em acesso aberto. Os autores levantaram opiniões de uma amostra de pesquisadores da comunidade científica brasileira, buscando identificar os fatores que afetam sua aceitação ou resistência à publicação de suas pesquisas nas revistas de acesso aberto ou ao depósito de cópias dos seus trabalhos em repositórios de acesso aberto. Foi aplicado um questionário a 643 doutores vinculados a universidades brasileiras, oriundos de todas as áreas do conhecimento. O questionário era composto por 29 questões de múltipla escolha. Foram coletadas informações de pesquisadores de 31 institutos de ensino superior e institutos de pesquisa do Brasil. A estratificação por áreas do conhecimento mostrou uma distribuição relativamente equilibrada entre elas, sendo as Ciências da Saúde (21%), seguidas pelas Ciências Exatas e da Terra (20,1%), as áreas mais representativas, e a menor porcentagem, de 1,1%, relativa à área de Linguística, Letras e Artes. Como resultado, foi observado que a maioria dos pesquisadores que responderam ao questionário detém conhecimento sobre o acesso aberto e apoiam seus princípios e algumas de suas ações, principalmente em relação à publicação de revistas nesse formato, o que se reflete também nos seus hábitos de uso e citação dessas fontes. Porém, existe certa indefinição por parte deles com relação ao status do copyright, o que tem limitado a publicação de seus trabalhos em repositórios de acesso aberto como os repositórios institucionais. Por isso, é destacada a necessidade de maiores iniciativas informativas para esclarecer esses aspectos.

Na literatura revisada, verificou-se um grande esforço para compreender a opinião dos pesquisadores sobre a divulgação do resultado de suas pesquisas em formato que possa ser acessado de forma fácil, livre e sem restrições. Dentre os principais trabalhos, destacam-se algumas iniciativas que avaliam, com o auxílio de questionários, o sentimento de pesquisadores com reconhecida relevância em suas áreas de atuação sobre a divulgação de suas pesquisas em formato de acesso aberto.

No entanto, apesar de ser possível identificar certo interesse por parte dos

pesquisadores que responderam a tais questionários, percebe-se que, entre eles, os interessados em realizar publicações em acesso aberto ainda são pouco representativos, e que vários deles ainda possuem dúvidas sobre questões legais relativas à divulgação nesse formato.

Outro elemento a ser destacado diz respeito à quantidade de indivíduos que foram avaliados. Esses estudos trabalharam com conjuntos específicos e, consequentemente, com uma pequena parcela dos autores de publicações em acesso aberto, que não representam de forma significativa a produção científica brasileira nesse formato. Tendo em vista que, em geral, os avaliados são de um mesmo nível de formação ou de uma mesma região geográfica, as conclusões não podem representar de forma efetiva o perfil nacional dos autores.

3 DESENVOLVIMENTO

A bibliometria tem como objetivo desenvolver padrões e modelar matematicamente os processos para as medições e, a partir dos resultados, traçar previsões e tomar as possíveis decisões. Por meio de suas técnicas, a bibliometria procura estudar os aspectos quantitativos da ciência e da produção científica como uma atividade que envolve características sociais, econômicas e políticas. Ela fornece um instrumental para estudos que visam mapear o conhecimento científico e extrair informações, bem como a compreensão de como a produção científica tem sido realizada (HAYASHI, 2012).

Dentre as principais leis bibliométricas, tem se a Lei de Zipf. A Lei de Zipf está relacionada à frequência de ocorrência de palavras em um dado texto. Essa lei desenvolveu e estendeu uma lei empírica observada por Estoup em 1916, a qual estabelece uma relação entre a posição de uma palavra e a frequência de seu aparecimento em um texto longo. A Lei de Zipf é assim formulada: r .f = c, sendo que “r” é a posição da palavra, “f” é a frequência e “c” é uma constante. Zipf extraiu sua lei de um princípio geral do “esforço mínimo”, segundo o qual uma palavra cujo custo de utilização seja pequeno ou cuja transmissão demande um esforço mínimo é frequentemente usada em um texto grande (KLEINUBING, 2010).

No contexto deste trabalho, no intuito de melhor compreender os

principais tópicos de pesquisa, investigados pelos pesquisadores brasileiros em periódicos de acesso aberto, foi utilizado o repositório de dados curriculares da Plataforma Lattes.

Grande parte dos editais de financiamento de projetos de pesquisa, realizados por diversos órgãos de fomento, utiliza dados cadastrados nos currículos dos proponentes como uma das formas de avaliação das propostas. Logo, há um grande incentivo para que os pesquisadores mantenham as informações de seus currículos atualizadas, o que torna esses currículos uma excelente fonte de dados para análises. Por essa mesma razão, vários trabalhos têm utilizado a Plataforma Lattes como fonte de dados para diversos estudos, como redes de colaborações, análise de produção científica, genealogia acadêmica, entre outros.

Como já mencionado, atualmente a Plataforma Lattes conta com mais de 7.8 milhões de currículos cadastrados (segundo dados de dezembro de 2022) e o aumento no número de usuários tem sido constante, impulsionado, principalmente, pelos órgãos governamentais e por agências de fomento que incentivam o cadastro e atualizações dos currículos.

Considerando que a maioria dos trabalhos correlatos analisou apenas grupos específicos de indivíduos, e tendo em vista que a manipulação de grandes quantidades de currículos da Plataforma Lattes não é uma tarefa trivial, já que existem problemas que envolvem recuperação de informação e algoritmos eficientes para manipulação de grandes volumes de dados, o *LattesDataXplorer* (DIAS, 2016), um framework para extração e tratamento dos dados curriculares, foi utilizado.

Conforme já explanado, um currículo cadastrado na Plataforma Lattes pode conter diversas informações capazes de auxiliar na compreensão da evolução da ciência brasileira sob diversas perspectivas. No entanto, para atender aos propósitos desta tese, somente dados de publicações de artigos em periódicos de acesso aberto, bem como de seus autores, foram levados em conta. Diante disso, foi proposta uma extensão do *LattesDataXplorer*, com a inclusão de componentes a priori inexistentes, a qual avaliasse, para cada artigo publicado em periódico (a saber, 7.841.860), de cada um dos indivíduos (a saber,

6.548.210), se o periódico no qual aquele artigo havia sido publicado era de acesso aberto. Logo, com a proposta dessa extensão, somente os autores e as publicações de artigos em periódicos de acesso aberto foram analisados.

A seleção desse conjunto de dados para as análises tem como motivação o fato de que a maioria dos pesquisadores brasileiros possui currículos cadastrados, e mesmo pesquisadores em início de carreira são incentivados a se registrar e manter seus currículos atualizados. Assim, o conjunto de indivíduos analisado neste trabalho compreende grande parte dos pesquisadores em atuação no Brasil.

Para a definição do conjunto de dados a ser analisados neste trabalho, optou-se por extrair as palavras dos títulos dos artigos publicados em periódicos de aceso aberto (a saber 2.090.015 artigos). Para a identificação dos artigos a serem considerados foi realizado um cruzamento com os periódicos de cada artigo com a relação de periódicos de acesso aberto obtida no DOAJ. A escolha da extração das palavras dos títulos dos artigos em detrimento das palavras-chave vinculadas aos artigos, se deve ao fato de que aproximadamente, apenas 17% dos artigos analisados possuíam palavras-chave vinculados a eles. Além disso, diversos trabalhos têm utilizados as palavras dos títulos das publicações como objeto de análise (CUNHA *et al.*, 2013, VINKERS *et al.*, 2015, MRYGLOD *et al.*, 2016; RONDA-PUPO, 2016).

Além disso, tendo em vista que o cadastramento das palavras-chave de um artigo científico em seus currículos é de inteira responsabilidade dos respectivos pesquisadores, e isso é feito livremente por eles, significa que podem ser inseridos quaisquer conjuntos de caracteres como uma palavra-chave. A partir disto, geralmente tem-se uma coleção muito grande de palavras-chave e sem nenhum padrão (GOMES, 2018).

Logo, para as análises aqui realizadas, os títulos de todas as publicações do conjunto identificado foram considerados. Os títulos passaram por um processo de tratamento de dados que visou identificar as palavras que que posteriormente serão objeto de análise. Todas as etapas do processo de tratamento podem ser visualizadas na Tabela 1.

Tabela 1 – Etapas do Processo de Tratamento dos Dados

ETAPA DA DO ALGORITMO	RESULTADO
Recebimento do Título	UMA ESTRATÉGIA PARA IDENTIFICAÇÃO DE ARTIGOS EM PERIÓDICOS DE ACESSO ABERTO NA PLATAFORMA LATTES.
LowerCase	uma estratégia para identificação de artigos em periódicos de acesso aberto na plataforma lattes
StopWords_PT	estratégia identificação artigos periódicos acesso aberto plataforma lattes
StopWords_EN	estratégia identificação artigos periódicos acesso aberto plataforma lattes
Identificação de Termos	estratégia
	identificação
	artigos
	periódicos
	acesso
	aberto
	plataforma
	lattes

Fonte: Os autores (2022)

Como pode ser observado, para cada artigo seu título é recuperado e dessa forma, o processo de tratamento dos dados é inicializado. Na etapa de LowerCase, todas as palavras são convertidas para minúsculo com a proposta de padronizar o conjunto, bem como, evitar que palavras sejam mapeadas em tópicos distintos por algumas possuírem letras em maiúsculo e outras não. Já no processo de remoção de stopWords (StopWords_PT e StopWords_EN), são removidos todos os termos que não possuem valores semânticos significativos para caracterizar um tópico de pesquisa e, com isso, diminuir o volume de palavras a serem processadas e analisadas. Foram removidos os *stopWords* inicialmente em português e posteriormente em inglês, tendo em vista que são os idiomas mais utilizados, conforme já apresentado.

Como o objeto inicial de análise é o título dos artigos, em que, se existe uma preocupação com a descrição geral do estudo a ser apresentado, a quantidade de *stopWords* é significativa, diferentemente das palavras-chave, justificando a remoção para as análises a serem realizadas. Após, na última etapa Identificação de Termos as palavras são separadas em tópicos, que irão compor um dicionário de termos para a contagem das frequências.

4 RESULTADOS

No contexto desta pesquisa, considerando a produção científica, foram analisados somente os artigos publicados em periódicos de acesso aberto. Logo, o principal elemento de análise é o conjunto de artigos publicados em periódicos científicos. Esse conjunto é composto por quase oito milhões de artigos (7.841.860) e, considerando somente os artigos únicos, representa 41,94% do conjunto global.

Utilizando a extensão proposta neste trabalho para o *LattesDataXplorer*, foram identificados todos os autores que publicaram pelo menos um artigo em periódico de Acesso Aberto (370.388). Esses autores — apesar de serem uma pequena quantidade de indivíduos, em relação a todo o conjunto cadastrado na Plataforma Lattes (5,65%) — possuem uma grande representatividade no contexto das publicações em periódicos. Aproximadamente 78% dos artigos globais e 81% dos artigos únicos publicados em periódicos foram realizados pelo conjunto de indivíduos que possuem publicações em periódicos de acesso aberto.

Ao analisar os autores das publicações realizadas em periódicos de acesso aberto pelas suas grandes áreas de atuação, é possível verificar sua distribuição por essas áreas, bem como analisar quais delas têm maior representatividade, considerando a quantidade de indivíduos que já publicaram pelo menos um trabalho nesse meio de publicação.

Ressalta-se que o primeiro registro identificado em cada um dos currículos foi utilizado para determinar a principal grande área de atuação dos autores, tendo em vista que nos currículos da Plataforma Lattes é possível realizar o registro de até três grandes áreas de atuação, segundo a distribuição de áreas do CNPq. Como pode ser observado, a grande área de Ciências da Saúde se destaca com a maior quantidade de indivíduos (24,53%), seguida pelas grandes áreas de Ciências Humanas (13,49%), Ciências Biológicas (12,14%), Ciências Agrárias (10,85%) e Ciências Sociais Aplicadas (9,89%). Já a menor quantidade de autores pertence à grande área de Linguística, Letras e Artes (4,17%), com quantidade próxima à de Engenharias (4,87%).

Destaca-se, ainda, uma significativa quantidade de indivíduos (10,81%) que não informaram grande área de atuação em seus currículos, o que foi categorizado como “Não Informado”. Na análise desse conjunto de indivíduos que não informaram grande área, nota-se que a maioria é composta por autores que ainda estão em processo de formação e que possivelmente ainda não definiram suas áreas de atuação. Além disso, também foi possível identificar um pequeno conjunto de indivíduos (1,03%) que informaram a grande área “Outros”.

Com o intuito de verificar as áreas do conhecimento mais representativas do conjunto analisado, destaca-se Medicina (32.966), compondo a grande área de Ciências da Saúde, como a mais representativa, conforme já mostrado. A representatividade da área de Medicina é tão considerável que somente ela possui praticamente a mesma quantidade de indivíduos que a soma total das grandes áreas de Linguística, Letras e Artes e Engenharias juntas. Na sequência, destacam-se as áreas de Educação (16.840), Agronomia (15.719), Enfermagem (12.710), Administração (11.717), Química (10.501) e Psicologia (10.483). Somente essas sete áreas são responsáveis por abrigarem aproximadamente 30% do conjunto total de indivíduos analisados.

Em análise aos títulos das publicações, foram identificados 28.636.958 tópicos, considerando todas as palavras dos títulos dos artigos. Após a remoção das duplicatas o conjunto foi reduzido a 423.364 palavras únicas. Posteriormente, com a remoção das *stopWords*, o conjunto passou a ter um total de 393.896 palavras que se tornaram objeto de análise.

Considerando a característica dos títulos das publicações que em geral necessitam da utilização de *stopWords* na sua composição, todos os primeiros 15 termos identificados são *stopWords* em português ou inglês.

Aproximadamente 64% das publicações em periódicos de acesso aberto analisadas neste trabalho são em português, logo se justifica uma quantidade considerável de termos neste idioma. A Tabela 2 apresenta o resultado da extração e ordenação pela frequência das palavras dos títulos de cada artigo analisado, após todo o tratamento dos dados.

Tabela 2 – Distribuição das palavras por posição (x) e suas frequências (y)

Posição (x)	Frequência(y)	Palavras
1	88.485	brazil
2	85.470	brasil
3	72.100	estudo
4	71.618	avaliação
5	69.314	análise
6	58.977	saúde
7	46.863	educação
8	44.131	study
9	43.411	brazilian
10	42.365	rio
11	41.411	patients
12	38.754	diferentes
13	37.788	produção
14	37.586	ensino
15	37.395	estado
⋮	⋮	⋮
393.894	1	zzaa
393.895	1	zzgam
393.896	1	zzgamma

Fonte: Os autores (2022)

Como pode ser observado, mesmo após a remoção das *stopWords* é possível verificar que dentre as palavras mais frequentes, a maioria está em português, com algumas destas palavras em sua versão em inglês, como por exemplo, as duas palavras mais frequentes. Já nas últimas posições, se encontram palavras com uma frequência muito baixa. Percebe-se que tais palavras não possuem conteúdo semântico, sendo uma hipótese para a existência de tais palavras, erros de digitação no momento de cadastro do título da publicação em um determinado currículo. Percebe-se ainda, que dentre as palavras mais frequentes, se encontram tópicos que geralmente fazem parte dos títulos das publicações, já que são importantes para indicar métodos, técnicas, objetos ou localidades.

No intuito de avaliar o conjunto de palavras que estão vinculadas as publicações de artigos em periódicos de acesso aberto, utilizou-se a Lei de Zipf. No trabalho de QUONIAM (1992), o autor descreve a curva de Zipf, em que a mesma é dividida em três zonas de distribuição:

• **Zona I** - Informação trivial ou básica: define os temas centrais da análise

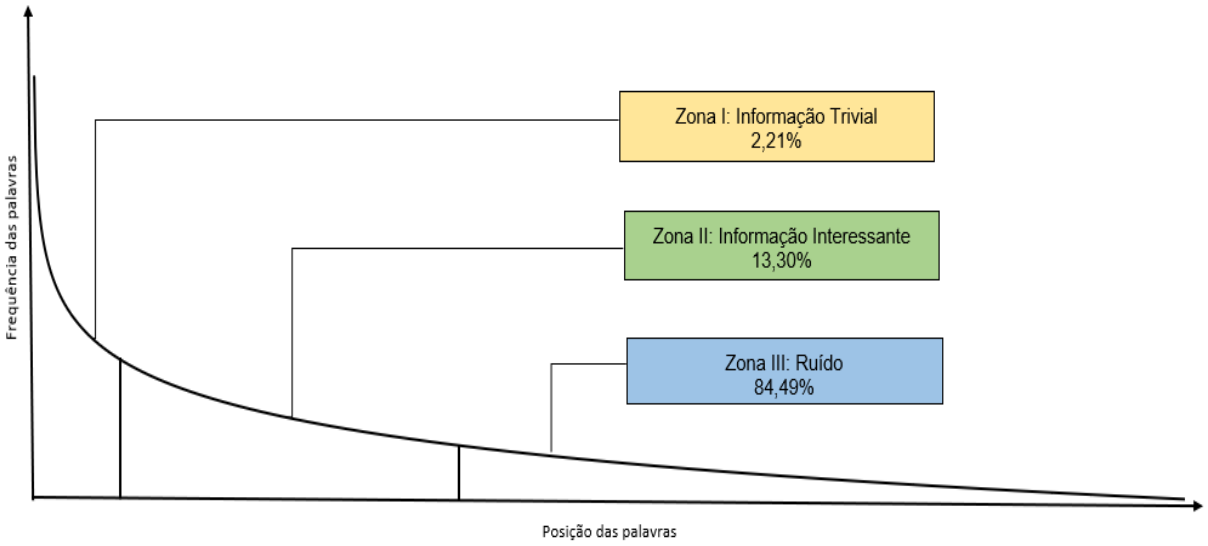
bibliométrica;

• **Zona II** - Informação interessante: localiza-se entre as Zonas I e III e mostra os temas periféricos, uma informação potencialmente inovadora. É aí que as transferências de tecnologia relacionadas aos novos temas devem ser consideradas;

• **Zona III** - Ruído: tem como característica possuir conceitos ainda não emergentes, onde é impossível afirmar se eles serão emergentes ou se são apenas ruído estatístico.

Neste contexto, o conjunto de palavras identificadas nesta tese, após todo o tratamento de dados já apresentado, foi dividido em três zonas textuais de distribuição (Figura 1).

Figura 1 – Divisão das Palavras Identificadas nas Três Zonas Textuais



Fonte: Os autores (2022)

A primeira zona identificada (Zona I), possui 2,21% das palavras analisadas, tais palavras que são as mais frequentes, descrevem quais são os temas centrais do conjunto analisado. Apesar de contemplar um baixo percentual de palavras, a frequência delas, corresponde a 47,93% de todo o conjunto, comprovando a sua representatividade. Já na Zona II, que possui 13,3% das palavras, engloba um conjunto de tópicos que ocorrem em menor frequência que os da Zona I, e por não serem palavras utilizadas com tanta frequência são caracterizadas como temas emergentes, já que se caracterizam

como informação potencialmente inovadora. Por fim, na Zona III que possui a grande maioria das palavras (84,49%), se caracteriza por agregar tópicos com baixa frequência, considerados ruídos. Aqui, cabe destacar como já apontado, os problemas originados do livre cadastramento dos dados das publicações por parte dos indivíduos em seus currículos, em que é real a inserção de dados incorretos, seja por erros de digitação ou até mesmo de codificação no momento de copiar um texto de documentos em diversos formatos. Um total de 149.891 palavras possuem apenas uma ocorrência.

Como pode ser observado, existe uma grande disparidade entre os conjuntos de palavras que compõem cada uma das Zonas identificadas. No intuito de melhor compreender cada uma destas Zonas, diversas técnicas de análise e visualização de dados podem ser aplicadas. A Figura 2, apresenta uma nuvem de palavras da Zona 1.

Figura 2 – Zona I: Informação Trivial ou Básica

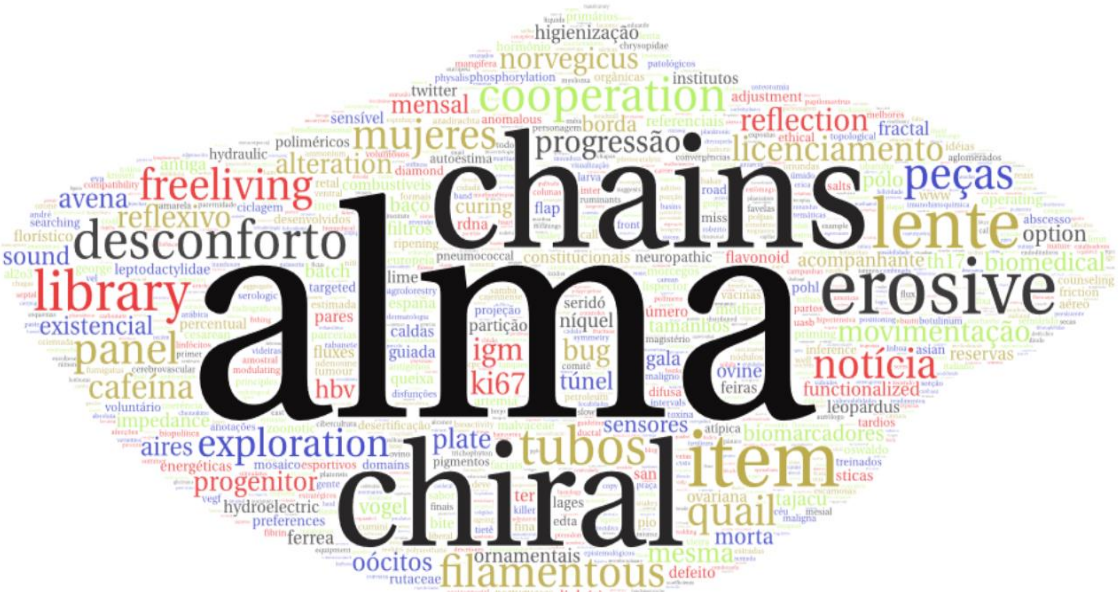
Fonte: Os autores (2022)

Observa-se que diante das palavras mais frequentes no conjunto analisado, as palavras “brazil” e “brasil” se destacam significamente. Como uma hipótese, pode-se inferir que tais palavras são amplamente utilizadas por indicar localidades de aplicação das pesquisas, principalmente por considerar praticamente em sua totalidade artigos publicados por brasileiros, nos idiomas inglês e português. Além disso, destaca-se ainda as palavras “estudo”, “avaliação”, “análise”, “study”, e “analysis”, que geralmente indicam métodos utilizados para a realização das pesquisas. Importante também, destacar as

palavras “saúde” e “educação” com 58.977 e 46.863 ocorrências respectivamente, apresentando-se também como tópicos muito representativos nas pesquisas realizadas. Ressalta-se aqui, a ocorrência de outras palavras como “rio”, “paulo”, “caso” e “meio” que também possuem frequência significativa, mas que podem ter sofrido influência do método utilizado para identificar as palavras dos títulos, tendo em vista que são palavras que também podem ter sido derivadas de palavras compostas.

No intuito de se obter uma visão geral dos principais tópicos que compõem a Zona II, uma nuvem de palavras pode ser observada na Figura 3.

Figura 3 – Zona II: Informação Interessante



Fonte: Os autores (2022)

Já a segunda Zona, denominada de Informação Interessante, possui como destaque a palavra “alma”, que está inserida em títulos de autores de diversas áreas do conhecimento, com 296 ocorrências. Apesar de “alma” surgir como destaque na nuvem de palavras por causa da ordem alfabética na classificação, outras 24 palavras também possuem essa mesma frequência. Destaca-se nesse conjunto alguns tópicos relacionados diretamente a algumas áreas do conhecimento, como: “chiral”, “erosive”, “chains”, “filamentous” e “tubos”, podendo indicar elementos de estudo em destaque nessas áreas. Além dessas, também se destacam alguns outros tópicos, como “cooperation”, “mujeres”, “library”, “exploration”, “licenciamento”, “progressão” e “notícia”, tendo

sido esses tópicos interdisciplinares objeto de estudo em diversas pesquisas.

Por fim, na Zona III, que possui a maior quantidade de palavras, aquelas com as maiores frequências (2.426 palavras) possuem 16 ocorrências, e se caracterizam por possuir, em geral, tópicos pouco utilizados, caracterizados como ruídos (Figura 4).

Figura 4 – Zona III: Ruído

Fonte: Os autores (2022)

Tendo em vista a grande quantidade de palavras com a mesma frequência, na nuvem de palavras estão destacadas aquelas que iniciam com a letra “a”, devido à ordem alfabética considerada pelos algoritmos de geração das nuvens. Ressalta-se uma quantidade significativa de palavras para todas as frequências até a frequência mínima igual a um (149.891 palavras). Em geral, as palavras com apenas uma ocorrência são resultado de termos novos propostos, novos acrônimos, fórmulas, ou termos técnicos, como os nomes de proteínas inseridos nos títulos. No entanto, em sua maioria são ruídos originados de erros ortográficos, provavelmente de digitação, como “wonen”, “brsil”, “wwww”, “qualitatva” e “vrasileiro”.

5 CONSIDERAÇÕES FINAIS

Ressalta-se que no estudo aqui apresentado, foi adotada a Lei de Zipf no intuito de identificar os principais tópicos de pesquisa dos pesquisadores brasileiros em publicações em periódicos de acesso aberto. Para tanto, foi inicialmente no repositório de dados curriculares da Plataforma Lattes, todos os artigos publicados neste meio de divulgação. Considerando que “Brazil e Brasil” indicam localidades, e que os termos “Estudo, Avaliação e Análise” indicam métodos utilizados, destacam-se neste contexto, os termos “Saúde e Educação” como os mais representativos no conjunto analisado. Além disso, também é importante destacar que não foi possível realizar a unificação das palavras no singular e plural, bem como, a utilização de palavras compostas, tendo em vista que seriam necessárias a adoção de técnicas como por exemplo de Processamento de Linguagem Natural, como radicalização e n-gramas que vão além do escopo deste trabalho.

REFERÊNCIAS

CHALHUB, T.; PINHEIRO, L. V. R. Publicações de acesso livre: tendências entre pesquisadores de universidades públicas do estado do Rio de Janeiro. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília. **Anais** [...]. Brasília: Ancib, Unb, 2011. p. 2225-2241. Disponível em: http://ridi.ibict.br/handle/123456789/84. Acesso em: 27 mar. 2023.

CUNHA, M. V.; ROSA, M. G.; FADIGAS, I. de S.; MIRANDA, J. G. V.; PEREIRA, H. B. de B. Redes de títulos de artigos científicos variáveis no tempo. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2., 2013, Maceió. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2013. p. 194-205. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/6842. Acesso em: 27 mar. 2023.

DIAS, T. M. R. **Um estudo sobre a produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. 181 f. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2033874&key=d8d1d2008e1ebe20f0f136527af3a222. Acesso em: 27 mar. 2023.

FURNIVAL, A. C. M.; SILVA-JEREZ, N. S. Percepções de pesquisadores brasileiros sobre o acesso aberto à literatura científica. **Informação & Sociedade:** Estudos, João Pessoa, v. 27, n. 2, 2017. Disponível em: https://biblat.unam.mx/es/revista/informacao-sociedade/articulo/percepcoes-de-pesquisadores-brasileiros-sobre-o-acesso-aberto-a-literatura-cientifica. Acesso em: 27 mar. 2023.

GOMES, J. O. **Uma análise temporal dos principais tópicos de pesquisa da ciência brasileira a partir das palavras-chave de publicações científicas**. 2018, 127 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2018. Disponível em: https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2276931&key=a4f163edd015be78d38300e0a55933db. Acesso em: 27 mar. 2023.

HAYASHI, M. C. P. I. Sociologia da Ciência, Bibliometria e Cientometria: Contribuições para a Análise da Produção Científica. *In*: SEMINÁRIO DE EPISTEMOLOGIA E TEORIAS DA EDUCAÇÃO (EPISTED), 4., 2012, São Paulo. **Anais** [...]. São Paulo: Episted, 2012.

KLEINUBING, L. S. Análise bibliométrica da produção científica em gestão da informação na base de dados LISA. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 8, n. 2, p. 1-11. 2010. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/1943. Acesso em: 27 mar. 2023.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999. 268 p.

MRYGLOD, O.; HOLOVATCH, Y.; KENNA, R.; BERCHE, B. Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster. **Scientometrics**, [*S.l.*], v. 106, n. 3, p. 1151-1166, 2016. Disponível em: https://link.springer.com/article/10.1007/s11192-015-1820-2. Acesso em: 27 mar. 2023.

MUELLER, S. P. M. O círculo vícios o que prende os periódicos nacionais. **Datagramazero**, Brasília, v. 0, n. 4, p.1-8, dez. 1999. Disponível em: http://eprints.rclis.org/11196/. Acesso em: 27 mar. 2023.

NEUBERT, P. S.; RODRIGUES, R. S.; GOULART, L. H. Periódicos da Ciência da Informação em acesso aberto: uma análise dos títulos listados no DOAJ e indexados na Scopus │ Open access journals in information Science. **Liinc em Revista**, [*S.l.*], v. 8, n. 2, p. 389-401, dez. 2012. Disponível em: http://dx.doi.org/10.18617/liinc.v8i2.497. Acesso em: 27 mar. 2023.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO (OCDE). Declaration on Access to Research Data from Public Funding. Anexo 1 do Science, Technology and Innovation for the 21st Century.

Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 Jan. 2004. Disponível em: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html. Acesso em: 19 out. 2018.

OLIVEIRA, E. C. P.; CHALHUB, T. O movimento de acesso livre à informação e repercussões nas revistas científicas Ibero-americanas. *In:* FORO IBERO-AMERICANO DE COMUNICAÇÃO E DIVULGAÇÃO CIENTÍFICA, 1., 2009, Campinas. **Anais** [...]. Campinas: Ibict, 2009. p. 1-7. Disponível em: http://repositorio.ibict.br/handle/123456789/400. Acesso em: 27 mar. 2023.

QUONIAM, L. Bibliométrie sur des référence bibliographiques: methodologie. *In:* DESVALS H.; DOU, H. (org.). **La veille technologique**: l'information scientifique, technique et industrielle. Paris: Dunod, 1992. p. 243-262.

RODRIGUES, R. S.; OLIVEIRA, A. B. Periódicos Científicos na America Latina: títulos em Acesso Aberto indexados no ISI e SCOPUS. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 17, n. 4, p. 76-99, dez. 2012. Disponível em: https://www.scielo.br/j/pci/a/P4GSxYP4sL4XHdchpNmzXLL/abstract/?lang=pt. Acesso em: 27 mar. 2023.

RONDA-PUPO, G. A. Knowledge map of Latin American research on management: Trends and future advancement. **Social Science Information**, [*S.l.*], v. 55, n. 1, p. 3-27, 2016. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/0539018415610225?casa_token=Ucs3xXavP14AAAAA:4vK-EUJQja9wS8rQW-LvnaORA56aO9gdciT5BX_B5OzE04ov59ezu8PqqhmKg-7j-n9RBMdsCLLg2g. Acesso em: 27 mar. 2023.

SILVA, T. E.; ALCARÁ, A. R. Políticas de acesso aberto à informação científica: iniciativas governamentais. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais** [...]. São Paulo: USP, 2008. Disponível em: https://brapci.inf.br/index.php/res/v/180390. Acesso em: 27 mar. 2023.

VINKERS C. H.; TIJDINK J. K.; OTTE, W. M. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. **Thebmj**, [*S.l.*], v. 351, 2015. Disponível em: https://www.bmj.com/content/351/bmj.h6467/. Acesso em: 27 mar. 2023.

EVALUATING THE REPRESENTATION OF THE MAIN RESEARCH TOPICS IN OPEN ACCESS PUBLICATIONS IN BRAZIL

ABSTRACT
Objective: in this work, the main objective is to identify and analyze the frequency of the

main research topics used in the set of publications in open access journals by researchers in Brazil. **Methodology**: in order to achieve the proposed objective, all publications in open access journals are initially identified in the set of curricula of the Lattes Platform. Subsequently, with the help of text mining techniques and with the adoption of Zipf's Law, the main topics used are identified, having as a data source, the titles of the initially identified publications. **Results**: as a result of the analysis, it was possible to verify how some topics are frequently used by Brazilian researchers in the description of their studies. **Conclusions**: with the study and analysis presented in this work, it was possible to verify how some topics that indicate locations are frequently used, as well as terms that generally indicate methods adopted in research.

Descriptors: Open Access. Lattes Platform. Bibliometrics. Open Data.

EVALUACIÓN DE LA REPRESENTACIÓN DE LOS PRINCIPALES TEMAS DE INVESTIGACIÓN EN LAS PUBLICACIONES DE ACCESO ABIERTO EN BRASIL

RESUMEN
Objetivo: en este trabajo, el objetivo principal es identificar y analizar la frecuencia de los principales temas de investigación utilizados en el conjunto de publicaciones en revistas de acceso abierto por investigadores en Brasil. **Metodología**: para lograr el objetivo propuesto, todas las publicaciones en revistas de acceso abierto se identifican inicialmente en el conjunto de planes de estudio de la Plataforma Lattes. Posteriormente, con la ayuda de técnicas de minería de textos y con la adopción de la Ley de Zipf, se identifican los principales temas utilizados, teniendo como fuente de datos, los títulos de las publicaciones inicialmente identificadas. **Resultados**: como resultado del análisis, fue posible verificar cómo algunos temas son utilizados frecuentemente por los investigadores brasileños en la descripción de sus estudios. **Conclusiones**: con el estudio y análisis presentado en este trabajo, fue posible verificar cómo algunos temas que indican lugares son frecuentemente utilizados, así como términos que generalmente indican métodos adoptados en la investigación.

Descriptores: Acceso Abierto. Plataforma de Lattes. Bibliometria. Información abierta.

Recebido em: 21.12.2022
Aceito em: 24.03.2023