Tendências da Pesquisa Brasileira em Ciência da Informação

ANÁLISE BIBLIOMÉTRICA DE USO E REUSO DE DADOS DE PESQUISA[1]

BIBLIOMETRIC ANALYSIS OF USE AND REUSE OF RESEARCH DATA

Skrol Salustiano[2]
Fábio Castro Gouveia[3]

Resumo: O uso de dados de pesquisa tem ganhado destaque como elemento essencial na produção científica, mas sua citação formal ainda é pouco adotada. Este estudo tem como objetivo realizar uma análise bibliométrica quantitativa sobre o uso e reuso de *datasets* depositados no repositório *Figshare*, com base em publicações indexadas na base *Scopus*. A pesquisa utiliza bibliometria para quantificar e descrever as citações de *datasets* entre 2016 e 2023, considerando áreas do conhecimento, países e autores mais produtivos. Os procedimentos envolveram a coleta de 17.324 documentos que mencionam o uso de dados de *Figshare* e a análise dos padrões de crescimento dessas citações ao longo do tempo. Os resultados indicam um aumento contínuo no volume de citações, com a área de "Computer Science" sendo a que mais utiliza esses dados. No entanto, o crescimento percentual dessas citações desacelerou nos últimos anos, sugerindo uma menor dependência de dados de repositórios externos. As considerações finais indicam que, embora a citação de *datasets* esteja crescendo, ainda há desafios quanto à sua formalização e à disseminação dessa prática em outras áreas do conhecimento.

Palavras-chave: citação de dados; dados de pesquisa; bibliometria; repositório de dados; reuso de dados.

Abstract: *The use of research data has gained prominence as an essential element in scientific production, but its formal citation is still rarely adopted. This study aims to conduct a quantitative bibliometric analysis of the use and reuse of datasets deposited in the Figshare repository, based on publications indexed in the Scopus database. The research employs bibliometrics to quantify and describe the citation of datasets between 2016 and 2023, considering fields of knowledge, countries, and the most productive authors. The procedures involved the collection of 17,324 documents mentioning the use of Figshare data and the analysis of citation growth patterns over time. The results indicate a continuous increase in citation volume, with the "Computer Science" field being the most frequent user of these datasets. However, the citation growth rate has slowed in recent years, suggesting less reliance on external repository data. The final considerations indicate that*

[1] Este artigo foi submetido, avaliado, aprovado, apresentado e premiado no GT7 do XXII ENANCIB.

[2] Mestre em Ciência da Informação. IBICT/UFRJ. E-mail: sallustiano@gmail.com. ORCID: https://orcid.org/0000-0002-1396-1199.
[3] Doutor em Química Biológica. IBICT/UFRJ - Fiocruz. E-mail: fgouveia@gmail.com. ORCID: https://orcid.org/0000-0002-0082-2392.

while dataset citation is growing, challenges remain regarding its formalization and dissemination across other fields of knowledge.

Keywords: *data citation; research data; bibliometrics; data repository; dataset reuse.*

1 INTRODUÇÃO

Nas últimas cinco décadas, os dados têm ganhado visibilidade crescente em diversos setores, fenômeno que foi intensificado com a digitalização e a popularização da internet. Com a ampla adoção das redes sociais e o surgimento de novas oportunidades para monitorar a prática científica, como previsto por Bossy (1995), a internet tornou-se um campo fértil para observar a 'ciência em ação'. Nesse contexto, os dados de pesquisa, ou *datasets*, quando seu uso é finalizado, têm sido tratados, muitas vezes, como meros instrumentos auxiliares para validar hipóteses ou contextualizar estudos. No entanto, em muitos casos, os dados também são reconhecidos como fontes valiosas de conhecimento que merecem destaque equivalente às produções científicas que os utilizam.

Apesar dessa crescente valorização, os estudos sobre o ciclo de vida desses *datasets* ainda são escassos, muitas vezes prejudicados pela forma como os dados são referenciados. Mayernik (2013, p. 1) observou que os conjuntos de dados utilizados em artigos científicos frequentemente são mencionados apenas nas seções de metodologia ou agradecimentos, e não como citações formais na bibliografia. Isso limita a visibilidade e o reconhecimento dos *datasets* como parte integral da pesquisa científica. Robison-Garcia, Mongeon, Jeng e Costas (2017, p. 2) reforçam essa observação, destacando a necessidade de uma mudança no comportamento dos pesquisadores, ao citar as fontes de dados utilizadas, de forma que essas referências recebam o devido reconhecimento.

Com base nessa premissa, de que os dados de pesquisa precisam ser mais bem referenciados e valorizados, esta pesquisa tem como objetivo realizar uma análise

bibliométrica quantitativa do uso e reuso de *datasets* depositados no repositório *Figshare*[4], com base em artigos indexados na base *Scopus*, da *Elsevier*. A pesquisa é caracterizada como exploratória, pois busca abrir novos caminhos para investigações futuras, e descritiva, ao identificar as características das citações de dados.

O contexto atual, em que as revistas científicas estão tornando mais rigorosas suas políticas de compartilhamento de dados, torna essa pesquisa oportuna. Ao abordar o tema da citação de dados e a rastreabilidade do uso desses *datasets*, o estudo contribui para o debate crescente sobre o reconhecimento dos dados como elementos essenciais da produção científica. Além disso, espera-se que este trabalho sirva de base para a ampliação de estudos na área e o aprimoramento de metodologias para identificar e compreender a dispersão e o reuso dos dados de pesquisa em diversas áreas do conhecimento.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa teve como objeto de estudo os dados da base *Scopus*, da *Elsevier*, selecionada por disponibilizar metadados de publicações científicas, os quais permitem realizar diversos estudos sobre a produção (como vínculos institucionais, agências de fomento, tipo de documento, entre outros) e o consumo de tais publicações, por meio de métricas de citação.

Com base nesses parâmetros, a pesquisa adotou a bibliometria como metodologia, conforme destacado por Okubo (1997) ao observar que os estudos bibliométricos evoluíram e passaram a ser utilizados em vários campos do saber com técnicas que são combinadas para fornecer medições mais detalhadas e mais eficazes.

Nesta pesquisa, a bibliometria foi empregada para realizar uma análise quantitativa das citações de dados de pesquisa depositados no repositório *Figshare*,

[4] Disponível em: https://figshare.com/. Um repositório de dados da Digital Science.

que foi escolhido por demonstrar maior capilaridade, atração de público e por estar vinculado a periódicos de diversas áreas, utilizando o ambiente de programação Python, com a IDE Jupyter (Kluyver *et al.*, 2016).

Dessa forma, a pesquisa caracteriza-se como quantitativa, sendo descritiva e exploratória, ao buscar ampliar o debate sobre a importância da correta citação de dados de pesquisa, bem como identificar como esses dados estão sendo utilizados em diferentes áreas do conhecimento. A pesquisa também possui uma natureza aplicada, ao visar gerar conhecimentos práticos sobre o uso de *datasets* em publicações científicas.

O ponto de partida para a exploração dos dados foi a definição do termo de pesquisa e da janela temporal. Os parâmetros utilizados, descritos no Quadro 1, resultaram na seguinte cláusula de pesquisa: (REF(WEBSITE(figshare)) AND PUBYEAR > 2015 AND PUBYEAR < 2024).

Quadro 1 - Parâmetros de Busca

Parâmetros	
Base	*Scopus* (*Elsevier*)
Janela Temporal	Anos de 2016 a 2023
Termo de pesquisa	*Figshare*
Delimitação do campo	Somente as referências dos documentos
Critérios de Seleção	Ter nas referências link para documentos depositados no *Figshare*
Critérios de Exclusão	Não aplicável
Operadores utilizados	REF, AND, LIMIT TO, EXCLUD TO

Fonte: Elaborado pelos autores (2024).

A busca recuperou 17.324 documentos científicos[5] (*Article, Conference Paper, Data Paper, Review, Book Chapter, Note, Book, Letter, Editorial, Short Survey, Erratum, Retracted*), que foram analisados para observar a curva de crescimento no uso/reuso de dados do *Figshare*, além de identificar as áreas que mais citam esses dados, os

[5] O dataset pode ser acessado no link: https://doi.org/10.6084/m9.figshare.23750679.

pesquisadores que mais demonstram esse uso e os países com maior prevalência na citação de dados de pesquisa.

Contudo, a base *Scopus* apresentou uma limitação importante: a ausência de informações detalhadas sobre os países e áreas de conhecimento de cada autor de forma individualizada, uma vez que esses dados são fornecidos de maneira consolidada. Por essa razão, não foi possível identificar com precisão os países e as áreas de atuação dos autores mais produtivos.

3 DADOS DE PESQUISA NO CONTEXTO CIENTÍFICO

A pesquisa adotou como marco inicial dos estudos sobre as potencialidades dos *datasets*, ou dados de pesquisa, o artigo de Bisco (1964). Nele, o autor descrevia o grande volume de pesquisas políticas realizadas desde 1940 e destacava a necessidade de pensar em formas de arquivar e disponibilizar essas informações para trabalhos futuros ou acompanhamento de tendências políticas e socioeconômicas nos Estados Unidos. Bisco (1964) sugeriu o uso de sistemas computacionais emergentes e marcações para melhorar a recuperação da informação e, simultaneamente, utilizar os dados gerados pelas consultas para aperfeiçoar o sistema. Este trabalho influenciou Dodd (1979), que conduziu uma pesquisa sobre a citação de dados de pesquisa.

Na pesquisa, Dodd (1979) observou a falta de parâmetros adequados para a citação de dados, ou a ausência de identificação correta ao descrever dados aplicados em diferentes mídias. Ele propôs um modelo de arquivamento com uso consistente de título, autor e edição, incluindo a data. Contudo, somente em 1997 foi publicado o primeiro documento que buscava padronizar a citação de dados, incorporando o tratamento de bancos de dados com a exigência de título, autor e versão (ISBD(ER): International..., 1997).

Embora a padronização das citações ainda estivesse em discussão, Fienberg, Martin e Straf (1985) destacaram a importância do compartilhamento de dados de pesquisa. Segundo os autores, se dados diferentes, coletados de forma independente, forem usados para estudar o mesmo problema, a reanálise é chamada de replicação. No entanto, quando se compartilham os mesmos dados, a reanálise passa a ser verificação. Em uma análise secundária, os dados coletados para um conjunto de problemas podem ser usados para estudar outro problema. Entre os principais objetivos estava a possibilidade de agregar dados com outros e reexaminá-los sob uma perspectiva transtemporal.

Pollak (2006) discutiu a importância da correta e completa citação de dados de pesquisa. Ele observou que, em muitos casos, as informações eram citadas como notas de rodapé, e identificou que muitas publicações omitiram esse tipo de citação para reduzir custos, seguindo manuais de estilo e tipografia tradicionais.

Diversos trabalhos subsequentes apresentaram questões essenciais para o avanço no compartilhamento e citação de dados, como os de Arzberger *et al*. (2004); Paton (2008); Tenopir *et al*. (2011); Yakel; Faniel e Robert (2024), abordaram a publicação e compartilhamento de formas específicas de dados em certas áreas do saber, enquanto Parsons, Duerr e Minster (2010) e Lawrence *et al.* (2011) ressaltaram a necessidade de revisão por pares para os *datasets*, como forma de conferir credibilidade e validade antes de seu compartilhamento.

Mayernik (2013) destacou, após o workshop "*Bridging Data Lifecycles: Tracking Data Use via Data Citations*", que as citações de dados ainda eram consideradas fora do padrão em diversas áreas do conhecimento. Além disso, muitos pesquisadores tinham o hábito de citar *datasets* nas seções de metodologia ou agradecimentos, em vez de citá-los formalmente na bibliografia. A falta de recompensas institucionais,

como promoções baseadas em citações de conjuntos de dados, foi um dos fatores apontados para o baixo uso formal dessas citações.

Seguindo o mesmo conceito da importância da publicação dos dados de pesquisa, mas com abordagem diferente, Quarati e Raffaghelli (2022) discutiram sobre a subutilização dos dados e a qualidade dos metadados que acompanham os datasets públicos, e identificaram que estes podem estar sendo influenciados pela falta da "obrigatoriedade" da divulgação de documentos de suporte juntamente com a publicação do resultado final da pesquisa.

Konkiel (2013) observou que os dados de pesquisa são uma oportunidade para ampliar a mensuração dos impactos de documentos científicos. No entanto, destacou a incipiente padronização de dados e a falta de revisão por pares como obstáculos para o melhor aproveitamento desses documentos.

Altman e Crosas (2014) revisou a evolução dos padrões e práticas de citação de dados e enfatizou a importância de vincular artigos científicos aos dados utilizados na pesquisa, como forma de aumentar a visibilidade desses documentos.

Torres-Salinas, Jiménez-Contreras e Robinson-García (2014) realizaram as primeiras análises de cobertura e citação de dados de pesquisa utilizando o *Data Citation Index* (DCI), constatando que 88% dos *datasets* permaneciam praticamente não citados. Peters *et al.* (2016) confirmaram esses achados, identificando uma tendência crescente na citação de *datasets* a partir de 2008.

Khan, Thelwall e Kousha (2021), em um estudo sobre biodiversidade, encontraram inconsistências na citação de *datasets*, com 27% das menções ocorrendo nas referências e 13% nas seções de metodologia ou declarações de acesso aos dados. O estudo também indicou que blogs eram uma das fontes mais informativas, apesar de raras, enquanto a maioria dos *tweets* e postagens no *Facebook* eram de natureza promocional.

Quarati e Raffaghelli (2022) analisaram a subutilização de dados e a qualidade dos metadados que os acompanham, sugerindo que a falta de obrigatoriedade na divulgação de documentos de suporte pode ser um dos gargalos para a citação de *datasets*.

Esses trabalhos revelam um crescente interesse na citação e compartilhamento de dados de pesquisa, destacando a urgência de políticas e diretrizes mais rígidas para garantir a correta citação desses dados nas referências de documentos científicos. Apesar dos avanços, ainda há desafios relacionados à contextualização e incentivos ao pesquisador para incluir os dados utilizados em seus artigos.

3.1 DADOS DE PESQUISA COMO DOCUMENTOS CIENTÍFICOS

O conceito de 'documento científico' tradicionalmente está relacionado a artigos de periódicos, livros, relatórios técnicos e outras formas de produção científica em formato textual que registram os resultados de pesquisas. Essa definição, no entanto, está evoluindo, influenciada por novos estudos e abordagens como a de Guinchat e Menou (1994, p. 53), ao afirmarem que "todo conjunto de suporte de informação e dos dados nele registrados, que possam servir para consulta, estudo ou prova" pode ser considerado um documento.

Com o crescimento do debate sobre compartilhamento de dados e sua correta citação, a definição de 'documento científico' tem se expandido para incluir os *datasets*, que podem representar um recorte temporal, uma comprovação científica, entre outros.

Embora, em um primeiro momento, possa parecer desconexo o debate entre a tipologia documental dos dados de pesquisa e sua possível citabilidade, ambos se entrelaçam porque os dados de pesquisa são registros que contêm informações detalhadas sobre as observações, métodos, protocolos, instrumentação e contextos

em que foram coletados. Esses registros podem incluir metadados que descrevem a proveniência dos dados, os critérios de seleção, as unidades de medida e as informações de licenciamento. Com esse conjunto de informações, é possível entender o contexto em que os dados foram gerados, possibilitando a validação, interpretação e reprodução dos resultados.

Anteriormente, os *datasets* eram muitas vezes tratados apenas como apêndices suplementares aos artigos científicos, onde os dados brutos eram disponibilizados, mas nem sempre com a devida contextualização e descrição. No entanto, como observa Altman e Crosas (2014, p. 2), a citação de dados de pesquisa "está finalmente emergindo como uma norma fundamental para promover a acessibilidade e a responsabilidade dos dados", proporcionando uma nova forma de atribuição de crédito acadêmico.

Atualmente, espera-se que os *datasets* sejam disponibilizados de forma independente, seja em repositórios específicos, em plataformas de compartilhamento ou em anexos aos artigos. Essa mudança de paradigma tem levado ao reconhecimento dos *datasets* como documentos científicos completos, capazes de fornecer informações essenciais para validar e reproduzir a pesquisa original. À medida que os *datasets* ganham relevância como documentos científicos, principalmente por terem identificadores persistentes como o *Digital Object Identifier* (DOI), passam a ter uma identidade única, o que facilita as citações em publicações acadêmicas.

Esse novo cenário acaba por estimular a citação dos *datasets*, reconhecendo o trabalho do autor na coleta, organização e disponibilização dos dados. Isso também influencia na cultura de compartilhamento responsável, que encoraja pesquisadores a tornarem seus dados acessíveis para a comunidade científica e o público em geral.

No entanto, o reconhecimento dos *datasets* como documentos científicos não exclui sua relação com documentos tradicionais, como artigos e relatórios. Pelo

contrário, a interseção entre essas duas formas de produção de conhecimento é enriquecedora para a ciência. Enquanto os artigos fornecem a interpretação e análise dos dados, os *datasets* fornecem o suporte empírico e a base para a investigação. Juntos, esses documentos científicos complementares garantem uma abordagem holística e transparente para a pesquisa, permitindo uma visão e compreensão mais abrangente e detalhada das descobertas científicas.

Dessa forma, ao destacarmos os dados de pesquisa ou *datasets* como documentos científicos, também apoiamos a afirmativa de que o "conceito de 'documento' está em constante mutação, uma vez que acompanha todo o desenvolvimento das tecnologias de informação" (Ribeiro; Mesquita; Miranda, 2014, p. 32).

3.2 OPORTUNIDADES E DESAFIOS COM DADOS DE PESQUISA

Na era digital, a coleta e o compartilhamento de dados de pesquisa têm se tornado atividades corriqueiras entre os pesquisadores. Esse fenômeno ocorre pela crescente importância dos dados de pesquisa, uma vez que eles se referem a informações primárias ou secundárias coletadas ou geradas durante uma investigação científica. Esses dados podem incluir resultados experimentais, observações, entrevistas, respostas de questionários, medições, simulações, análises estatísticas, imagens, áudios, vídeos e outros tipos de registros. Em essência, os dados de pesquisa constituem a base empírica de qualquer estudo científico e permitem a reprodução e verificação das conclusões e resultados obtidos.

Os dados de pesquisa se tornam uma das principais vertentes para a construção de novos conhecimentos e para a evolução do conhecimento científico como um todo. Altman e Crosas (2014) reforçam essa visão ao argumentar que o compartilhamento de dados de pesquisa tem "o potencial de permitir novas formas de publicação

acadêmica, promover pesquisas interdisciplinares, fortalecer o vínculo entre política e ciência e reduzir os custos de replicação" (Altman; Crosas, 2014, p. 1, tradução nossa). Essa opinião é complementada por Piwowar e Vision (2013), ao destacar que artigos com conjuntos de dados publicamente disponíveis recebem um número maior de citações do que estudos semelhantes sem dados disponíveis.

Além disso, os dados de pesquisa são diversificados e podem variar amplamente de acordo com o campo de estudo e a natureza da pesquisa. Eles podem ser estruturados (por exemplo, bancos de dados, planilhas) ou não estruturados (como arquivos de texto, áudios, vídeos). Também podem ser dados brutos, processados ou derivados, dependendo do estágio de análise e tratamento.

A disponibilização de dados de pesquisa possibilita que outros pesquisadores tenham a oportunidade de examinar, verificar e reproduzir os resultados. Esse acesso confere transparência, aumenta a confiabilidade e reforça a credibilidade da pesquisa, ao permitir a confirmação ou contestação das conclusões. Ao mesmo tempo, facilita a detecção de possíveis fraudes, como apontado por Altman e Crosas (2014) ao afirmarem que o acesso aos dados e sua documentação pode identificar erros ou fraudes antes e depois da publicação.

A transparência e a abertura facilitam a colaboração entre pesquisadores, promovem o compartilhamento de conhecimento, reduzem custos e ampliam experiências. Essa cultura de colaboração e compartilhamento também permite avanços mais rápidos e significativos em diferentes campos do conhecimento. Altman e Crosas (2014) observam que o compartilhamento de dados tem o potencial de promover novas formas de publicação acadêmica, pesquisas interdisciplinares e o fortalecimento do vínculo entre política e ciência.

Por outro lado, a visão de abertura dos dados representa desafios para pesquisadores, instituições e agências de fomento. Borgman (2012) destaca obstáculos

como a interoperabilidade, padronização e curadoria dos dados. Para Prost e Schöpfel (2019), os pesquisadores, que são os grandes produtores de dados, podem ter receios quanto ao uso dos resultados de suas pesquisas por terceiros, além de se preocupar com a criação de pesquisadores especializados apenas em gerar dados.

Além disso, todo o processo de coleta, seleção, preservação e disponibilização de quantidades crescentes de dados científicos requer uma infraestrutura adequada, como observa Wittenburg *et al*. (2010). A necessidade de curadoria e a construção de uma infraestrutura confiável e sustentável é um desafio contínuo. Os autores também destacam a questão da metanálise:

> Como transmitiremos o contexto e a proveniência dos dados? [...] Como os usuários de uma ampla variedade de origens entenderão e consultarão os dados que estão acessando e reconhecerão as circunstâncias especiais sob as quais foram coletados? (Wittenburg *et al*., 2010, p. 16-17, tradução nossa).

Ao levar o foco para Gorgolewski, Margulies e Milham (2013), identificaram o temor de outros pesquisadores descobrirem erros ou discrepâncias nos conjuntos de dados ou em sua interpretação e afirmam que “os documentos de dados não podem ser considerados definitivos no momento da publicação, uma vez que erros serão, sem dúvida, descobertos ao longo do tempo” (Gorgolewski; Margulies; Milham, 2013, p. 4, tradução nossa). Tenopir *et al.* (2011) também observam gargalos como financiamento insuficiente e falta de tempo para preparar os dados para reutilização. Ainda assim, Wittenburg *et al*. (2010) ponderam que esses desafios podem ser mitigados com infraestrutura adequada e curadoria de dados que gerem confiança na interpretação correta dos dados.

3.2 OPORTUNIDADES E DESAFIOS COM DADOS DE PESQUISA

A ciência moderna é cada vez mais caracterizada pela interdisciplinaridade e pela abordagem de problemas complexos e multifacetados. Pesquisas de ponta

frequentemente envolvem equipes multidisciplinares, que trazem expertises diversas para a solução desses problemas.

Métricas tradicionais, como o Fator de Impacto, o Índice h e o número de citações, embora amplamente utilizadas, não refletem adequadamente a complexidade e profundidade das contribuições científicas em contextos interdisciplinares. No âmbito da dinâmica de citação de dados de pesquisa, essas métricas podem fornecer uma avaliação superficial do valor científico, deixando de lado aspectos como inovação, aplicabilidade e impacto real na solução de desafios globais.

Uma das principais limitações dessas métricas tradicionais é sua incapacidade de capturar a complexidade das contribuições científicas. Embora o Fator de Impacto e o número de citações indiquem a popularidade ou visibilidade de uma revista ou artigo, eles não discernem a qualidade, originalidade ou relevância de uma pesquisa para o avanço científico. Isso é especialmente problemático em áreas emergentes, onde a inovação pode não ser refletida imediatamente por índices de citação elevados.

Com o aumento da diversidade nos formatos de publicação científica, como conferências, relatórios técnicos, *preprints* e compartilhamento de dados de pesquisa, as métricas tradicionais tornam-se ainda menos representativas do impacto real de uma pesquisa. Esses formatos não são totalmente capturados por métricas baseadas apenas em revistas e citações, criando desafios para pesquisadores e instituições que buscam avaliar de maneira mais ampla o impacto do trabalho científico.

Além disso, a crescente ênfase na ciência aberta e no compartilhamento de dados traz novos desafios à avaliação da produtividade científica. Enquanto as métricas tradicionais focam na visibilidade das publicações, a disponibilidade e reutilização de dados compartilhados em repositórios como o *Figshare* estão sendo cada vez mais valorizadas como indicadores de relevância e impacto. Contudo, medir o impacto

desses dados de maneira justa e padronizada continua sendo um desafio em constante evolução. O próprio termo "impacto" muitas vezes é confundido com "repercussão", o que pode distorcer a verdadeira influência do trabalho científico.

Outro ponto crucial é a ausência de consideração pelo impacto social da pesquisa nas métricas tradicionais. Focadas principalmente no impacto acadêmico (citações e publicações), essas métricas frequentemente negligenciam o efeito que a pesquisa tem sobre a sociedade, políticas públicas, indústria e outras áreas. Isso resulta numa avaliação incompleta do valor real que a ciência oferece à sociedade como um todo. Percepção que segundo Gonçalves (2022, p. 107) deveria levar em consideração outros parâmetros, como o contexto regional, como forma de buscar equiparar as discrepâncias existentes entre os países e periódicos.

Outro ponto que torna esse debate urgente é o crescente volume de dados de pesquisa que trazem oportunidades de serem utilizados em contextos não tradicionais, como análises de políticas públicas, empreendedorismo, inovação e em aplicações nas ciências humanas e sociais. As métricas tradicionais, porém, ainda não estão preparadas para medir ou entender plenamente a relevância dessas novas aplicações, destacando uma lacuna significativa na avaliação da produtividade científica.

Esses desafios evidenciam a necessidade de abordagens mais abrangentes e inovadoras para a avaliação da produtividade acadêmica. A crescente disponibilidade de dados e a evolução das práticas acadêmicas exigem o desenvolvimento de novas métricas e indicadores que possam refletir, de forma mais precisa e justa, o valor das contribuições científicas para a comunidade acadêmica e para a sociedade. Métricas complementares, que considerem inovação, impacto social e o uso de dados, podem fornecer uma visão mais holística da produtividade científica, contribuindo para uma avaliação mais justa e eficaz no cenário contemporâneo.

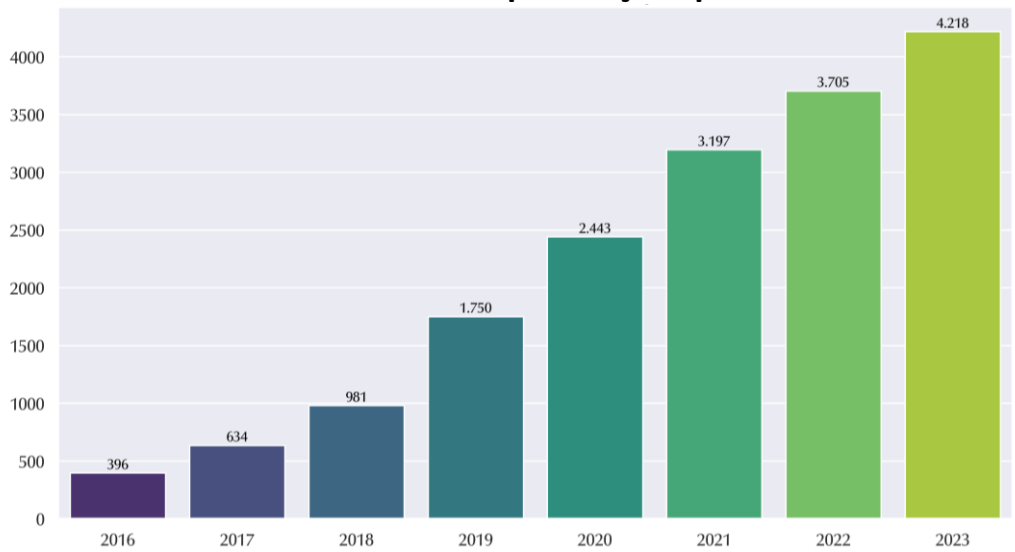
Embora exista o debate sobre a utilização das métricas tradicionais para mensurar e/ou avaliar a produção científica, esta pesquisa foi estruturada com a percepção de que mesmo sendo considerada uma das métricas mais estabelecidas na avaliação científica, como apontam Chueke e Amatucci (2015), que observam os estudos bibliométricos com a possibilidade de auxiliar no processo de sistematização das pesquisas nas mais variadas áreas do saber, além de auxiliar na geração de inteligência para a resolução de problemas.

Com essa percepção de que a Bibliometria, como as mais variadas métricas que buscam avaliar a produção científica, tem suas limitações ou 'ponto cego', que a pesquisa foi desenvolvida, com a certeza que seria possível extrair conhecimento, e poderá fornecer uma visão sobre o comportamento de citação e uso de *datasets*, não pretende ser uma métrica definitiva de impacto ou relevância. Dessa forma, para uma avaliação mais completa, é necessário o desenvolvimento e uso de métricas complementares que possam capturar tanto a inovação quanto o impacto social.

4 RESULTADOS

A pesquisa analisou dados de 17.324 artigos publicados entre 2016 e 2023, todos contendo citações a *datasets* depositados no repositório *Figshare*. No Gráfico 1, observa-se um crescimento constante no volume de documentos que citam a utilização de *datasets* ao longo dos anos, embora a taxa de crescimento percentual apresente uma desaceleração nos últimos períodos.

Gráfico 1 - Total de publicações por ano

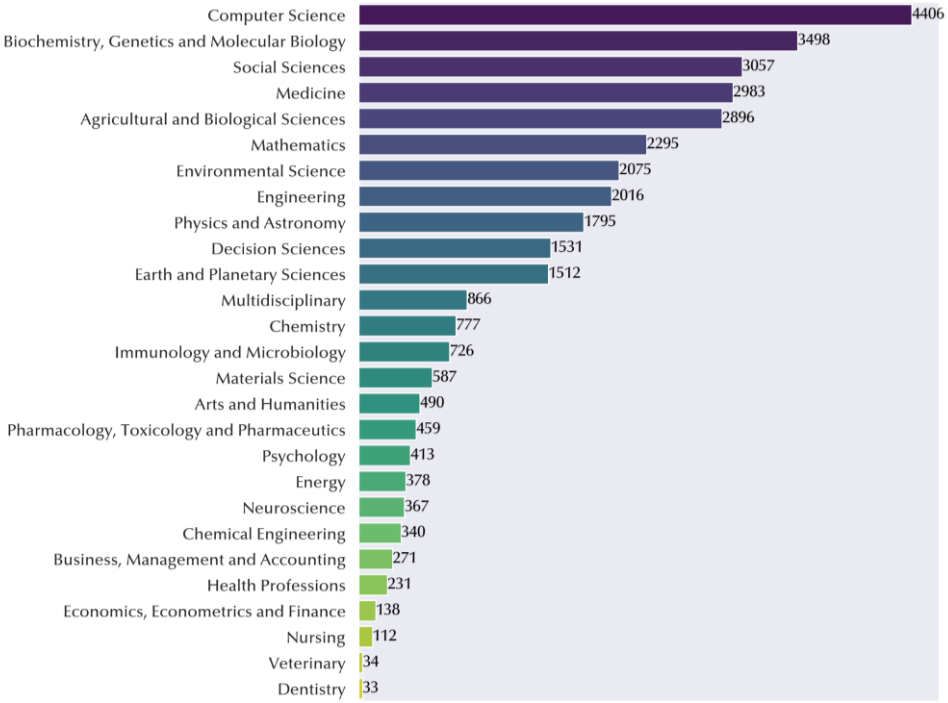


Fonte: Dados da pesquisa (2024).

Entre 2016 e 2019, o crescimento anual foi expressivo, com variações percentuais superiores a 50%, chegando a 78% em 2019. Contudo, a partir de 2020, a taxa de crescimento começou a diminuir, com variações de 39,6% em 2020, 30,86% em 2021, 15,89% em 2022 e 13,85% em 2023. Essa desaceleração sugere que, apesar do crescimento absoluto contínuo, o ritmo de aumento de citações de *datasets* tem se tornado mais modesto. Uma possível explicação para essa mudança é o uso crescente de *datasets* próprios pelos pesquisadores, diminuindo a dependência de dados de terceiros.

Considerando que o número de documentos indexados na base *Scopus* não cresceu mais do que 7% ao ano neste período, com valores que vão de 1,63% de 2021 para 2022 até 6,98% de 2020 para 2021, podemos dizer que houve aumento efetivo no número de citações à *datasets* ao longo do período analisado. Ao mesmo tempo, não foi encontrada uma relação direta entre a quantidade de citações e o aumento no volume de documentos indexados.

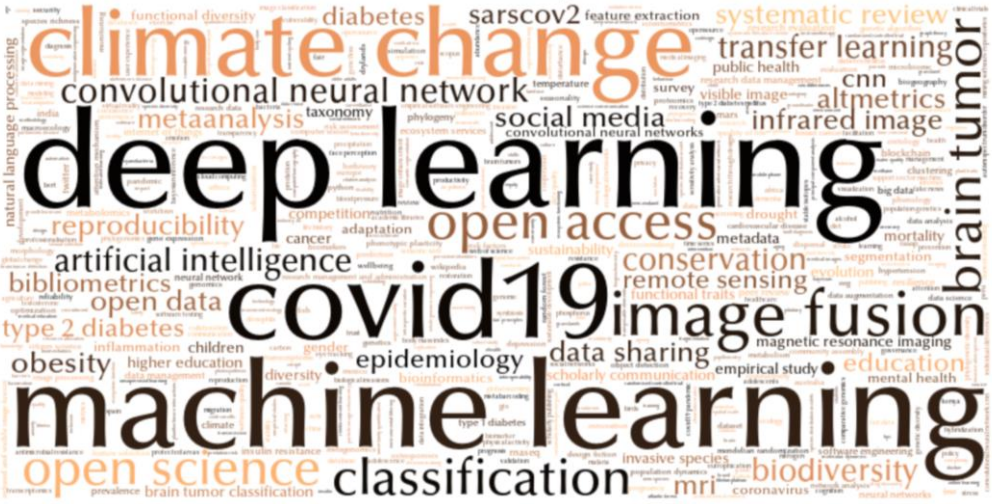
Gráfico 2 – Número de publicações por área



Fonte: Dados de pesquisa (2024).

Em relação às áreas do conhecimento, o Gráfico 2 demonstra que "Computer Science" é o campo com maior incidência de citações de *datasets* sendo esta área a terceira em documentos indexados na base, atrás de medicina e engenharia que na nossa coleta ficam em quarto e oitavo lugar respectivamente. Para entender melhor esse domínio, foi gerada uma nuvem de palavras (Figura 1) com base nas palavras-chave fornecidas pelos autores. Os termos foram normalizados para minúsculas, o que minimizou problemas causados por variações ortográficas.

Gráfico 3 – Nuvem de palavras



Fonte: Dados da pesquisa (2024).

Os principais termos observados na nuvem de palavras refletem as temáticas predominantes nas pesquisas, com destaque para *Machine Learning*, *Deep Learning*, *Climate Change*, e o impacto da Covid-19. Esses temas indicam que áreas tecnológicas e científicas contemporâneas são as principais responsáveis pelo uso de dados de pesquisa no *Figshare* citados.

Ao expandir a análise para a produtividade por países, verificou-se que três nações concentram quase 30% de toda a produção de artigos que citam *datasets*.

Tabela 1 - Países mais produtivos

Países	Produção	%
United States	2470	11,57%
United Kingdom	1811	8,91%
China	1558	8,77%
Germany	1145	6,13%
Australia	1030	6,10%
Canada	986	5,58%
France	859	5,11%
Spain	781	4,66%
Italy	769	4,13%
Netherlands	680	2,08%
India	626	1,29%
Switzerland	576	1,19%
Sweden	413	0,35%

Japan	352	0,19%
Brazil	305	0,11%

Fonte: Elaborada pelos autores (2024).

Ao buscar regionalizar os dados para identificar como os pesquisadores brasileiros se posicionam em relação aos demais países, foi identificada baixa participação. No ranking, o país ocupa a 15ª posição na lista de países com o maior volume de citações de dados, o que reflete uma participação relativamente baixa em comparação com sua 13ª posição no volume de registros indexados na base *Scopus*.

Em relação à produtividade individual dos autores, a Tabela 2 destaca os pesquisadores mais produtivos.

Tabela 2 - Pesquisadores mais produtivos

Autores	Produção	Total Citações
Wang Y.	241	4.819
Zhang Y.	201	17.459
Wang X.	146	2.937
Li X.	140	3.444
Wang Z.	140	2.835
Li Y.	138	2.333
Zhang X.	136	2.238
Wang J.	134	2.782
Chen Y.	125	3.809
Liu Y.	117	3.849

Fonte: Elaborado pelos autores (2024).

Os dados sugerem uma adoção crescente da prática de citação de *datasets* do *Figshare*, especialmente entre autores com nomes comuns na China. Esse fato pode refletir uma maior produção científica dessa região em áreas que utilizam *datasets*. A verificação de autocitações indicou que entre os dez autores mais produtivos, houve apenas uma ocorrência de autocitação. No total, identificaram-se 21 autocitações, representando apenas 0,12% da base de dados.

Esses números indicam que, embora a autocitação esteja presente, ela não é uma prática comum no contexto da citação de *datasets*. O maior percentual de autocitações ocorreu em 2017 (0,31%), seguido por 2021 (0,19%) e 2023 (0,16%).

5 CONSIDERAÇÕES FINAIS

Nesta pesquisa o objetivo principal foi o de mapear e analisar quantitativamente o uso de *datasets* do *Figshare* ao longo dos anos, por área de pesquisa e país de afiliação dos autores. O uso da Bibliometria foi crucial para fornecer uma visão abrangente sobre o volume de citações, e revelou um aumento contínuo nessas citações, mesmo quando comparado ao número de registros completos indexados na base *Scopus*.

O crescimento observado é significativo, mas ainda é cedo para afirmar se esse aumento se deve a uma conscientização voluntária sobre o reuso de dados ou se é simplesmente impulsionado pelo crescimento na quantidade de *datasets* compartilhados em repositórios como o *Figshare*.

Uma limitação importante deste estudo foi a utilização de uma única base de dados (*Scopus*) e a análise centrada exclusivamente em citações de dados de pesquisa depositados no repositório *Figshare*. Embora a *Scopus* ofereça uma cobertura abrangente, o escopo restrito pode deixar de captar nuances presentes em outros repositórios e bases de dados. Além disso, a análise focada no *Figshare* pode não refletir a totalidade das práticas de citação e reutilização de dados em outros contextos. Dessa forma, para uma assertividade dos resultados apontados nesta pesquisa uma ampliação para outras bases e repositórios de dados, como forma de obter uma visão mais robusta das práticas de citação de *datasets*, pode ser positiva.

Por fim, essa pesquisa pretende abrir caminhos para que com outras variáveis inseridas no debate, e a ampliação do escopo de repositórios analisados, se possa

buscar uma visão aprofundada de como os dados estão sendo compartilhados, e quais são as políticas editoriais que favorecem e/ou podem engessar a reutilização de dados. Outros estudos poderão levar a desdobramentos para gerar inteligência em relação ao reuso de dados de pesquisa.

REFERÊNCIAS

ALTMAN, Micah; CROSAS, Mercè. The evolution of data citation: from principles to implementation. ***IASSIST* Quarterly**, [*s. l.*], v. 37, n. 1-4, p. 62-70, 2013. Disponível em: https://iassistquarterly.com/index.php/iassist/article/view/504. Acesso em: 11 dez. 2023.

ARZBERGER, Peter; *et al*. Promoting Access to Public Research Data for Scientific, Economic, and Social Development. **Data Science Journal**, [*s. l.*], v. 3, p. 135-152, 2004. Disponível em: http://datascience.codata.org/articles/abstract/10.2481/dsj.3.135/. Acesso em: 1 jul. 2023.

BISCO, Ralph L. Information Retrieval from Data Archives: The ICPR System. **American Behavioral Scientist**, [*s. l.*], v. 7, n. 10, p. 45-48, jun. 1964. Disponível em: https://journals.sagepub.com/doi/10.1177/000276426400701013. Acesso em: 2 set. 2023.

BORGMAN, Christine L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, [*s. l.*], v. 63, n. 6, p. 1059-1078, jun. 2012. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.22634. Acesso em: 22 set. 2024.

BOSSY, Márcia Jurkiewicz. The last of the litter: Netometrics. **Solaris Information Communication**, v.2, p.245-250, 1995. Disponível em: http://gabriel.gallezot.free.fr/Solaris/d02/2bossy.html. Acesso em: 25 ago. 2023.

CHUEKE, Gabriel Vouga; AMATUCCI, Marcos. O que é bibliometria? Uma introdução ao Fórum. **Internext**, [*s. l.*], v. 10, n. 2, p. 1-5, 9 set. 2015. Disponível em: https://internext.espm.br/internext/article/view/330. Acesso em: 2 mai. 2023.

DODD, Sue A. Bibliographic references for numeric social science data files: Suggested guidelines. **Journal of the American Society for Information Science**, [*s. l.*], v. 30, n. 2, p. 77-82, mar. 1979. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.4630300203. Acesso em: 1 jul. 2023.

FIENBERG, Stephen E.; MARTIN, Margaret E.; STRAF, Miron L. **Sharing research data**. Washington, D.C.: National Academies Press, 1985. *E-book.* Disponível em: http://www.nap.edu/catalog/2033. Acesso em: 2 jun. 2024.

GONÇALVES, Andréa Ferreira. Métricas para avaliação da produção científica. *In*: RODE, Sigmar De Mello; GALLETI, Silvia Regina; MORAIS, Ana Marlene Freitas de. (org.). **Desafios e perspectivas da editoria científica**. São Paulo: ABEC Brasil, 2022. p. 109-118. *E-book.* Disponível em: https://www.abecbrasil.org.br/arquivos/Desafios_e_perspectivas_da_editoria_cientifica_2021.pdf#cap10. Acesso em: 1 ago. 2023.

GORGOLEWSKI, Krzysztof J.; MARGULIES, Daniel S.; MILHAM, Michael P. Making data sharing count: a publication-based solution. **Frontiers in Neuroscience**, [*s. l.*], v. 7, art. 9, p. 1-7, feb. 2013. Disponível em: http://journal.frontiersin.org/article/10.3389/fnins.2013.00009/abstract. Acesso em: 21 abr. 2024.

GUINCHAT, Claire; MENOU, Michael. **Introdução geral às ciências e técnicas da informação e documentação**. Brasília: IBICT, 1994.

ISBD(ER): INTERNATIONAL Standard Bibliographic Description for Electronic Resources. [S. l.]: IFLA, 1997. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/isbd/isbder.pdf. Acesso em: 15 jun. 2023.

KHAN, Nushrat; THELWALL, Mike; KOUSHA, Kayvan. Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics. **Scientometrics**, [*s. l.*], v. 126, n. 4, p. 3621-3639, abr. 2021. Disponível em: https://link.springer.com/10.1007/s11192-021-03890-6. Acesso em: 14 out. 2023.

KLUYVER, Thomas; *et al*. Jupyter Development Team. Jupyter Notebooks: a publishing format for reproducible computational workflows. *In:* LOIZIDES, Fernando; SCHMIDT, Birgit. (ed.). **Positioning and Power in Academic Publishing**:

Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing. Amsterdam: IOS Press, 2016. p. 87-90. *E-book*. Disponível em: https://ebooks.iospress.nl/book/positioning-and-power-in-academic-publishing-players-agents-and-agendas-proceedings-of-the-20th-international-conference-on-electronic-publishing. Acesso em: 24 set. 2024.

KONKIEL, Stacy. Tracking citations and altmetrics for research data: challenges and opportunities. **Bulletin of the American Society for Information Science and Technology**, [*s. l.*], v. 39, n. 6, p. 27-32, aug. 2013. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/bult.2013.1720390610. Acesso em: 16 nov. 2022.

LAWRENCE, Bryan; *et al*. Citation and peer review of data: moving towards formal data publication. **International Journal of Digital Curation**, [*s. l.*], v. 6, n. 2, p. 4-37, 26 jul. 2011. Disponível em: http://ijdc.net/article/view/181. Acesso em: 12 maio. 2024.

MAYERNIK, Matthew. Bridging data lifecycles: **Tracking data use via data citations workshop report.** [*S. l.*]: NCARLIB, 2013. Nota técnica. Disponível em: http://opensky.ucar.edu/islandora/object/technotes:505. Acesso em: 2 maio 2023.

OKUBO, Yoshiko. **Bibliometric Indicators and analysis of research systems:** methods and examples. Paris: OECD, 1997. Disponível em: https://www.oecd-ilibrary.org/science-and-technology/bibliometric-indicators-and-analysis-of-research-systems_208277770603. Acesso em: 5 mai. 2024.

PARSONS, Mark A.; DUERR, Ruth; MINSTER, Jean-Bernard. Data Citation and Peer Review. **Eos, Transactions American Geophysical Union**, [*s. l.*], v. 91, n. 34, p. 297-298, 24 ago. 2010. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2010EO340001. Acesso em: 2 jul. 2024.

PATON, Norman W. Managing and sharing experimental data: standards, tools and pitfalls. **Biochemical Society Transactions**, [*s. l.*], v. 36, n. 1, p. 33-36, 1 fev. 2008. Disponível em: https://portlandpress.com/biochemsoctrans/article/36/1/33/83705/Managing-and-sharing-experimental-data-standards. Acesso em: 21 nov. 2023.

PETERS, Isabella; *et al*. Research data explored: an extended analysis of citations and altmetrics. **Scientometrics**, [*s. l.*], v. 107, n. 2, p. 723i744, may 2016. Disponível em: http://link.springer.com/10.1007/s11192-016-1887-4. Acesso em: 25 set. 2024.

PIWOWAR, Heather A.; VISION, Todd J. Data reuse and the open data citation advantage. **PeerJ**, [*s. l.*], v. 1, p. e175, 1 out. 2013. Disponível em: https://peerj.com/articles/175. Acesso em: 3 mar. 2023.

POLLAK, Oliver B. The Decline and Fall of Bottom Notes, *op. cit., loc. cit*., and a Century of the Chicago Manual of Style. **Journal of Scholarly Publishing**, [*s. l.*], v. 38, n. 1, p. 14-30, out 2006. Disponível em: https://utpjournals.press/doi/10.3138/jsp.38.1.14. Acesso em: 19 fev 2023.

PROST, Hélène; SCHÖPFEL, Joachim. Data repositories in information and communication sciences: an empirical study. **Etudes de communication**, [*s. l.*], v. 52, n. 1, p. 71-98, 2019. Disponível em: https://shs.cairn.info/journal-etudes-de-communication-2019-1-page-71?lang=en. Acesso em: 26 set. 2024.

QUARATI, Alfonso; RAFFAGHELLI, Juliana E. Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case. **Journal of Information Science**, [*s. l.*], v. 48, n. 4, p. 423-448, ago 2022. Disponível em: http://journals.sagepub.com/doi/10.1177/0165551520961048. Acesso em: 16 nov 2023.

RIBEIRO, Maria Cristina; MESQUITA, Walma; MIRANDA, Marcos Luiz Cavalcanti. A tese otletiana para gestão, organização e disseminação do conhecimento. **Revista RACIn**, João Pessoa, v. 3, n. 2, p. 2-22, jul./dez., 2014. Disponível em: http://arquivologiauepb.com.br/racin/edicoes/v2_n2/racin_v2_n2_artigo01.pdf. Acesso em: 11 maio 2023.

ROBINSON-GARCIA, Nicolas; MONGEON, Philippe; JENG, Wei; COSTAS, Rodrigo. DataCite as a novel bibliometric source: coverage, strengths and limitations. **Journal of Informetrics**, [*s. l.*], v. 11, n. 3, p. 841-854, ago 2017. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1751157717300834. Acesso em: 19 maio 2023.

TENOPIR, Carol; *et al*. Data sharing by scientists: practices and perceptions. **PLoS ONE**, [*s. l.*], v. 6, n. 6, p. e21101, 29 jun. 2011. Disponível em: https://dx.plos.org/10.1371/journal.pone.0021101. Acesso em: 19 mai 2023.

TORRES-SALINAS, Daniel; JIMÉNEZ-CONTRERAS, Evaristo; ROBINSON-GARCÍA, Nicolas. How many citations are there in the Data Citation Index? **arXiv preprint**, [*s. l.*], 2014. Disponível em: https://arxiv.org/abs/1409.0753. Acesso em: 7 ago. 2023.

WITTENBURG, Peter; *et al*. **Riding the wave: how Europe can gain from the rising tide of scientific data**. [*S. l.*]: High-Level Group on Scientific Data, 2010. *E-book.* Disponível em: https://cds.cern.ch/record/1298248/files/HLEG-report.pdf. Acesso em: 11 ago. 2023.

YAKEL, Elizabeth; FANIEL, Ixchel M.; ROBERT, Lionel P. An empirical examination of data reuser trust in a digital repository. **Journal of the Association for Information Science and Technology**, [*s. l.*], v. 75, n. 8, p. 898-915, ago. 2024. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24933. Acesso em: 29 jul 2024.

AGRADECIMENTOS

A pesquisa foi realizada com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e Conselho Nacional de Desenvolvimento Científico e Tecnológico, Processos: 430982/2018-6, 315521/2020-1 e 315689/2023-4.

tpbci@ancib.org @anciboficial @ancib_brasil