23° enancib
ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO
06 a 10 de novembro de 2023
ARACAJU-SE

GT 7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

ISSN 2177-3688

CITAÇÕES A DADOS DE PESQUISA DO FIGSHARE: UMA ANÁLISE BIBLIOMÉTRICA

CITATIONS TO FIGSHARE RESEARCH DATA: A BIBLIOMETRIC ANALYSIS

Skrol Salustiano - Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Fabio Castro Gouveia - Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) / Fundação Oswaldo Cruz (FIOCRUZ)

Resumo: Os dados de pesquisa têm ganhado protagonismo na produção científica e igualmente nos debates sobre boas práticas de citação e na necessidade de referenciar corretamente esses documentos informacionais. **Objetivo.** O estudo tem como objetivo realizar uma análise bibliométrica quantitativa da referenciação de dados de pesquisa em um repositório utilizado por revistas científicas internacionais. **Metodologia.** A pesquisa tem como base metodológica a Bibliometria, por ter o objetivo de identificar a disseminação de datasets. Para a exploração dos dados foi escolhida a base da Scopus, da Elsevier, por ter maior alcance e disponibilizar o download de dados em diversos formatos. O termo base utilizado nas buscas foi “figshare”, que foi definido, por ser um repositório utilizado por grandes periódicos para o depósito de dados. O período analisado foi de 2016 a 2022 e foram recuperados 13.012 documentos. **Resultados.** A pesquisa identificou um aumento na referenciação aos datasets em publicações científicas. Os resultados também apresentam um crescimento da China, como um dos principais atores em referenciar o uso/reuso desse tipo de documento. **Considerações finais.** Foi observado que existe uma movimentação, tanto de autoria como de citação de datasets. Isso demonstra a necessidade de estudos ampliando o escopo da base estudada para identificar e aprofundar as discussões abordadas.

Palavras-chave: Figshare; dados de Pesquisa; Bibliometria.

Abstract: Research data has gained prominence in scientific production as well as in discussions on good citation practices and the need to properly reference these informational documents. **Objective:** The study aims to conduct a quantitative bibliometric analysis of the referencing of research data in a repository used by international scientific journals. **Methodology:** The research is based on Bibliometrics, as it aims to identify the dissemination of datasets. The Scopus database from Elsevier was chosen for data exploration due to its wider reach and availability of data downloads in various formats. The search term used was "figshare," which was selected because it is a repository used by major journals for data deposition. The analyzed period was from 2016 to 2022, and a total of 13,012 documents were retrieved. **Results:** The research identified an increase in the referencing of datasets in scientific publications. The results also indicate the growing involvement of China as one of the main actors in referencing the use/reuse of this type of document. **Final considerations:** It was observed that there is movement both in authorship and citation of datasets. This demonstrates the need for studies to expand the scope of the analyzed database to identify and further explore the discussed issues.

Keywords: Figshare; research data; Bibliometrics.

1 INTRODUÇÃO

A Revolução Digital, iniciada nos anos de 1950, atualmente se caracteriza pelo volume de dados produzidos que modelam vários aspectos de nossas vidas. Na academia essa realidade não é diferente. Influenciados pelo movimento Open Data, e pela demanda de periódicos em publicitar os registros de pesquisas, observa-se o aumento vertiginoso do número de dados compartilhados em repositórios digitais, como Figshare e Zenodo.

No entanto, os estudos sobre o ciclo de vida destes documentos ainda são poucos, em muitos casos influenciados pela forma como eles são referenciados. Mayernik (2012, p.1) observou que as referências aos conjuntos de dados utilizados no desenvolvimento de um artigo científico, normalmente, são identificadas nas seções de métodos de pesquisa ou agradecimentos de seus artigos, e não como citações formais na bibliografia de um artigo.

A opinião é complementada por Robison-Garcia *et al*. (2017, p. 2) que observou a necessidade de uma mudança no comportamento de pesquisadores ao citar fontes de dados utilizadas no desenvolvimento do trabalho científico.

Com base nessa premissa, de que os dados precisam ser mais bem referenciados, esta pesquisa tem o objetivo de realizar uma análise bibliométrica quali-quantitativa do uso/reuso de datasets depositados no repositório Figshare, com base em artigos indexados na base Scopus, da Elsevier. O estudo pode ser caracterizado como exploratório, pois não objetiva se encerrar em si próprio e descritivo por buscar identificar características intrínsecas nas referências aos dados.

No contexto atual, em que revistas e periódicos científicos tornam rigorosas suas políticas de compartilhamento do conjunto de dados, essa pesquisa se torna oportuna por se aliar a temática que tem ganhado espaço nos debates sobre citação de dados e a rastreabilidade dos usos destas informações científicas. Ao mesmo tempo, pretende-se que ela possa contribuir para a ampliação dos estudos e aprimoramento de metodologias para pesquisas que tenham o objetivo de identificar a dispersão de dados de pesquisa.

2 OS DATASETS E A PRODUÇÃO CIENTÍFICA

Os estudos sobre a citação de dados, ou seja, citações formais incluídas nas listas de referências de artigos publicados, não são recentes. Dodd (1979) identificou problemas no modelo utilizado para a descrição de dados, quando aplicados em formatos diferentes de

mídias e propôs um modelo de arquivamento e uso consistente de Título, Autor e Edição, com a inclusão da data.

Porém, somente em 1997, é publicado o primeiro documento que buscava padronizar a citação (ISBD 1990, ISO 690-2 1997) e incorporou o tratamento de bancos de dados com a exigência de Título, Autor e Versão (ISBD CF REVIEW GROUP *et al*., 1997).

Trabalhos posteriores apresentaram diferentes questões essenciais para a o processo de evolução no compartilhamento e citação dos dados. Pollak (2006), discutiu a necessidade e importância da correta e completa citação dos dados de pesquisa. Já Mayernik (2013), editou o compilado do material do workshop "*Bridging Data Lifecycles: Tracking Data Use via Data Citations*", quando houve a busca por parâmetros e melhores práticas para a citação de dados. Crosas (2014), fez uma revisão sobre a evolução dos padrões e práticas de citação de dados e afirmou ser imperioso a vinculação das publicações de artigos com os dados que serviram de base para a publicação.

Embora esses trabalhos tenham como o foco a necessidade e debates sobre os padrões de compartilhamento de dados de pesquisa, todos têm em comum o debate sobre a urgência de políticas ou diretrizes mais rígidas para o compartilhamento e correta citação, na lista de referências, dos dados utilizados na produção de documentos científicos.

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 os dados de pesquisa como uma oportunidade para ampliar e/ou aperfeiçoar a mensuração dos impactos de documentos científicos. No entanto, destacou que ainda é incipiente a padronização de dados e a revisão por pares de dados de pesquisa e afirma que esses dois fatores possam ser obstáculos para o melhor aproveitamento destes documentos científicos.

Em comum a todos estes trabalhos está o interesse da comunidade científica sobre esta temática, principalmente, se levarmos em conta a sua importância para o desenvolvimento de pesquisas. Porém, ainda existem alguns gargalos como a dependência "de uma forte contextualização para serem interpretados e transmitirem informação e conhecimento ao longo do tempo" (SAYÃO; SALES, 2019, p. 81). Além disso, existe a falta de

estímulo ao pesquisador em integrar na publicação do seu artigo os dados utilizados para sua realização.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa tem como base metodológica a Bibliometria, por ter o objetivo de identificar a disseminação de datasets. Para a exploração dos dados foi escolhida a base Scopus, da Elsevier, por ter maior alcance e disponibilizar o download de dados em diversos formatos e permitir a busca por links em citações. O termo base utilizado nas buscas foi "WEBSITE (figshare)", que foi definido, por ser um repositório utilizado por grandes periódicos para o depósito de dados. O período analisado foi de 2016 a 2022. Assim, foi possível a extração de dados para a realização da análise da utilização de dados hospedados na plataforma Figshare.

Tabela 1 – Elaborada pelos autores

Paramêtros	Variáveis
Base	Scopus (Elsevier)
Janela Temporal	Anos de 2016 a 2022
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
Delimitação de documentos	Não aplicável
Delimitação de idiomas	Não aplicável
Indicadores extraídos	Periódico; Ano de Publicação; Autoria; Co-autoria; Pais.

Fonte: Elaborado pelos autores (2023).

A Scopus possibilita a realização de pesquisa por vários parâmetros, mas download de apenas dois mil resultados. Por esse motivo e para conseguir baixar todos os dados foi necessário fatiar o download dos arquivos.

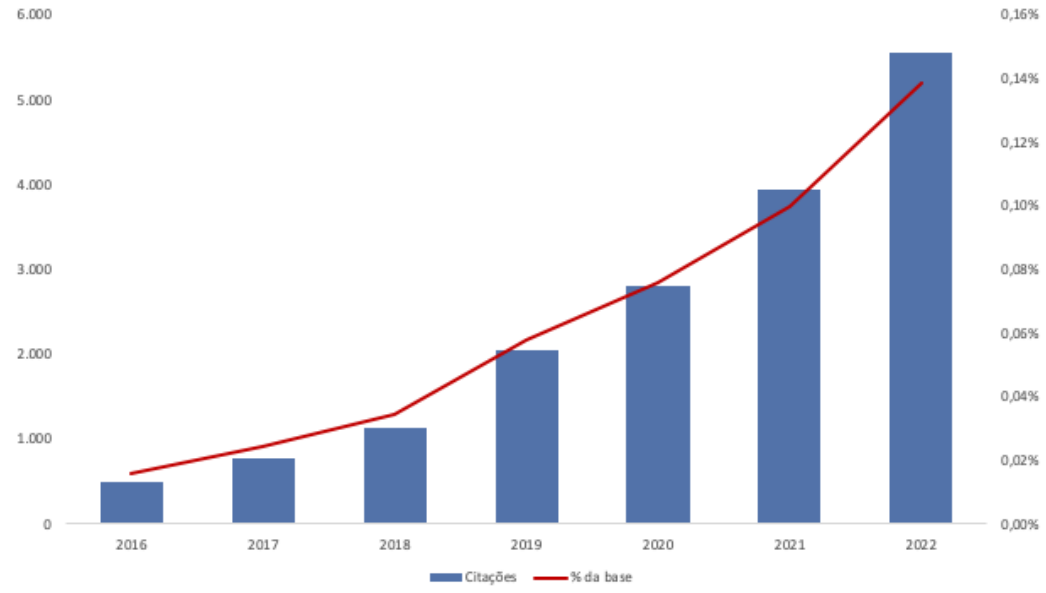
Após o download dos dados, a modelagem foi realizada no software Rstudio. Nesse ponto foi identificado que existiam coautorias de autores de países diferentes e como consequência a duplicação dos dados. Para evitar erro de vieses foi realizada a limpeza da base utilizando como parâmetro a coluna DOI das publicações.

Foi efetuada a análise dos dados coletados e a contextualização dos resultados, que nos permitiu observar a curva de crescimento no uso/reuso de fontes de dados no Figshare, bem como as categorias da Scopus onde elas são mais utilizadas.

4 RESULTADOS

O gráfico 1 apresenta o volume de citações por datasets ao longo do período analisado. Nele é possível observar que existe um crescimento no volume de citações a conteúdos no Figshare, e que mesmo quando se efetua uma fatoração pelo total da base Scopus esta curva é claramente ascendente.

Gráfico 1 – Evolução das citações

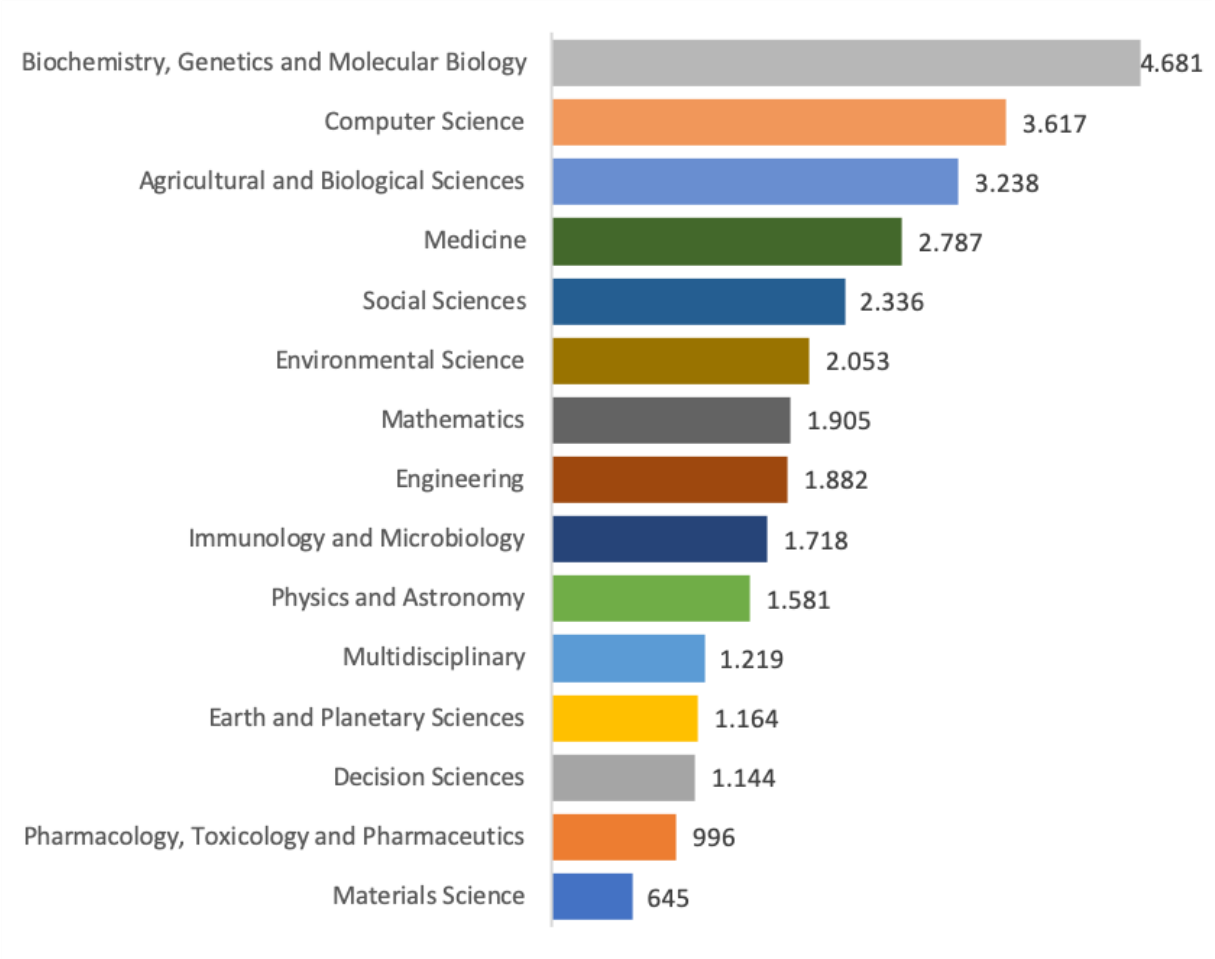


Fonte: Dados da pesquisa (2023).

No Gráfico 2 é possível observar as áreas com o maior volume de uso/reuso de dados de pesquisa. As áreas de Bioquímica, Genética e Biologia Molecular, e Ciência da Computação,

mesmo não sendo as mais presentes na base Scopus (as áreas de maior volume de registros indexados são Medicina e Engenharia), demonstram ter maior facilidade em trabalhar com dados publicados.

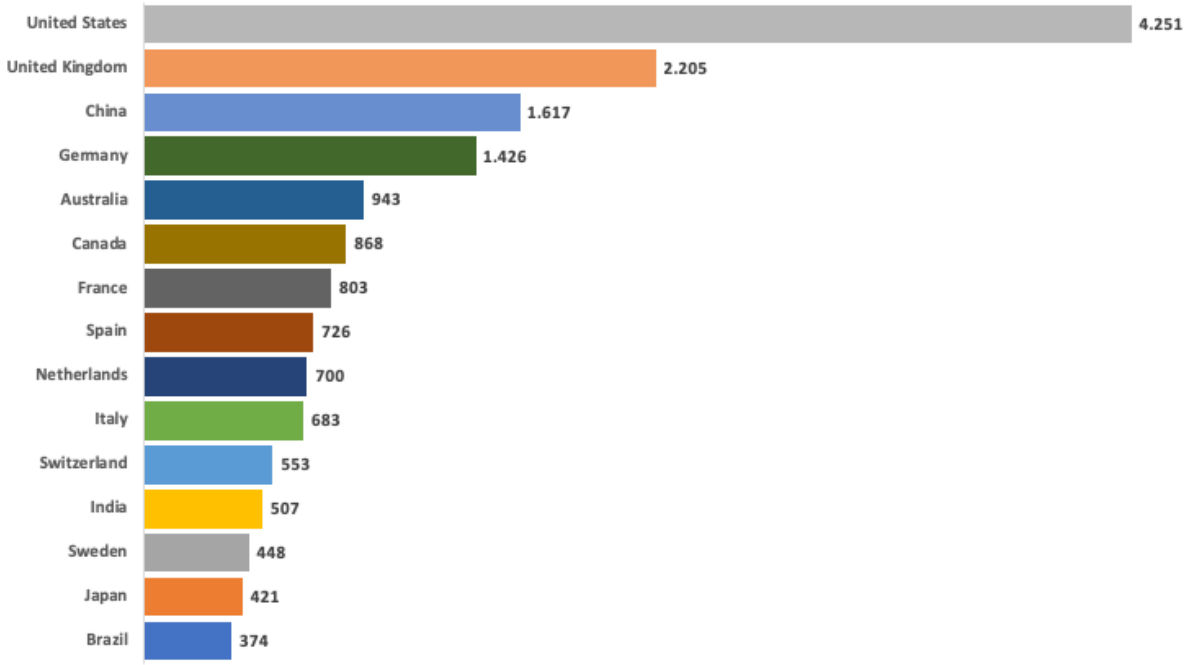
Gráfico 2 – Áreas com maior volume de citação



Fonte: Dados da pesquisa (2023).

No gráfico 3 os dados consolidados mostram os Estados Unidos com o maior volume de pesquisadores que citaram datasets. Porém, um olhar detalhado sobre os dados mostrou que nos dois últimos anos o país registrou uma ligeira retração e no mesmo período a China se consolidou em terceiro.

Gráfico 3 – Países com maior volume de citação



Fonte: Dados da pesquisa (2023).

Ao buscar regionalizar os dados para identificar como os pesquisadores brasileiros, por meio de citação, se posicionam em relação aos demais países identificou-se baixa participação. No ranking das Instituições, das 160 que figuraram na lista, somente a Universidade de São Paulo (USP) apareceu na 41ª posição, o que proporcionou ao país ocupar a 15ª posição na lista de países com o maior volume de citações de dados, um pouco abaixo da 13ª posição que o país ocupa no volume de registros indexados na base Scopus.

5 CONSIDERAÇÕES FINAIS

Com base nos dados coletados, o objetivo da pesquisa foi cumprido ao conseguir mapear e analisar de forma quantitativa a utilização dos datasets ao longo dos anos, por área de pesquisa e por país de afiliação dos autores. O uso da Bibliometria para traçar esse panorama foi fundamental, pois conseguiu demonstrar um aumento no volume de citações, quando comparado com o volume de registros completos indexado na base Scopus, ainda é possível identificar que há claro crescimento no volume de citações. No entanto, ainda é prematuro afirmar que essa dinâmica no processo de citações e reuso dos dados de pesquisa seja voluntário ou apenas influenciado pelo crescimento no volume dos datasets compartilhados em repositórios como o Figshare.

Como limitação desde estudo temos que considerar que mesmo com a capilaridade da base Scopus, a pesquisa se limita a uma base de dados e o foco centrado em apenas um repositório de dados. Dessa forma, para uma assertividade dos resultados preliminares apontados nesta pesquisa é necessário a ampliação da base e do repositório de dados.

Por fim, essa pesquisa não pretende encerrar em si própria, existem outras variáveis que podem ser inseridas no debate, como a ampliação do escopo de repositórios analisados, buscar uma amostragem de como os dados estão sendo compartilhados, e quais são as políticas editoriais que favorecem e/ou podem engessar a reutilização de dados. Enfim, essa pesquisa se encerra, abrindo vários possíveis desdobramentos que podem ser utilizados para gerar inteligência em relação ao reuso de dados de pesquisa.

AGRADECIMENTOS

O presente trabalho foi realizado 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, Processo 430982/2018-6 e 315521/2020-1.

REFERÊNCIAS

CROSAS, Mercè. The Evolution of Data Citation: From Principles to Implementation. *IASSIST* **Quarterly**, [*s.l.*], v. 37, n. 1–4, p. 62, 2014. Disponível em: https://iassistquarterly.com/index.php/iassist/article/view/504. Acesso em: 11 dez 2022.

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, 1979. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.4630300203. Acesso em: 1 jul 2023.

ISBD CF REVIEW GROUP ET AL. (ER): International Standard Bibliographic Description for Electronic Resources: Revised from the **ISBD (CF) International Standard Bibliographic Description for Computer Files**. [*s.l.*]: De Gruyter, 1997. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/isbd/isbder.pdf>. Acesso em: 15 jun 2023.

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

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

MAYERNIK, Matthew S. Data citation initiatives and issues. **Bulletin of the American Society for Information Science and Technology**, [s. l.], v. 38, n. 5, p. 23–28, 2012. Disponível em: http://doi.wiley.com/10.1002/bult.2012.1720380508. Acesso em: 24 jun. 2019.

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, 2006. Disponível em: https://utpjournals.press/doi/10.3138/jsp.38.1.14. Acesso em: 19 fev 2023.

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 2022.

ROBINSON-GARCIA, Nicolas e colab. DataCite as a novel bibliometric source: Coverage, strengths and limitations. **Journal of Informetrics**, [*s.l.*], v. 11, n. 3, p. 841–854, 2017. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1751157717300834. Acesso em: 19 maio 2023.

SAYÃO, Luis Fernando; SALES, Luana Farias. Subsídios para a contrução de um modelo de avaliação de sistemas de gestão de dados de pesquisa. **PontodeAcesso**, [*s.l.*], v. 12, n. 3, p. 80, 2019. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/28965. Acesso em: 2 jul 2023.